Journal of Political Institutions and Political Economy, 2020, 1: 341–378

Toward a General Causal Framework for the Study of Racial Bias in Policing

Dean Knox[1] and Jonathan Mummolo[2]*

[1] *Department of Operations, Information and Decisions, The Wharton School, University of Pennsylvania, Philadelphia, PA, USA; dcknox@upenn.edu*
[2] *Department of Politics, School of Public and International Affairs, Princeton University, Princeton, NJ, USA; jmummolo@princeton.edu*

ABSTRACT

A series of controversial police-involved killings and nationwide protests have recently reinvigorated the study of racial bias in policing. But a fractured interdisciplinary literature presents contradictory claims, and scholars have struggled to reconcile a dizzying array of seemingly incompatible analytic approaches that often rely on implausible and/or unstated assumptions. This confusion arose in part because data constraints have prompted researchers to examine only isolated aspects of the police–civilian encounters they seek to understand — focusing only on traffic stops in one study, or fatal shootings in another — while neglecting the complex, multi-stage nature of these interactions. The result is a conflicting and at times misleading body of evidence. To move toward a scientific consensus, scholars should converge on a common empirical framework that unites these disparate approaches under a shared conceptual umbrella, acknowledges the causal nature of the study of racial bias, accounts for the fundamental limitations of policing data, and yields substantively interpretable results that are useful to policymakers. We present such a framework and demonstrate its capacity to adjudicate conflicting claims, accumulate knowledge, and characterize the severity of one of the most pressing problems of institutional performance of our time.

Keywords: Policing; racial bias; causal inference; research design; partial identification; meta-analysis

*We thank Phillip Atiba Goff, Justin Grimmer, Erik Peterson, and Shea Streeter for their helpful feedback.

Online Appendix available from:
http://dx.doi.org/10.1561/113.00000018_app
ISSN 2689-4823; DOI 10.1561/113.00000018

Introduction

The empirical study of racial inequity in law enforcement has moved well beyond typical spheres of inquiry like criminology and law, and is now a fixture in political science (Baumgartner *et al.*, 2018; Knox *et al.*, 2020; Lerman and Weaver, 2014), psychology (Eberhardt *et al.*, 2004; Johnson *et al.*, 2019), economics (Fryer, 2019; West, 2018), and sociology (Legewie, 2015). As prominent criminal justice scholars have noted, the intensified focus of social scientists on how and why the coercive arm of the state dispenses violence and protection is long overdue (Soss and Weaver, 2017). With this enhanced scholarly attention, we might reasonably expect knowledge in this area to more rapidly accumulate and for social science to begin coalescing on answers to central questions, including the focus of this paper: whether and to what degree police behavior is racially biased.

Unfortunately, this progress has not materialized nearly as quickly as many had hoped. While recent years have seen several innovative papers that approach the substantial challenge of inferring racial bias in policing with the care and innovation the task deserves, they have also seen several prominent and widely publicized studies — research that has survived peer review to appear in some of the most prestigious and widely-cited social scientific journals in the world — that were either inattentive to fundamental sources of bias in police administrative data, or trumpeted conclusions that rest on fallacies. The extent of the confusion in this literature was underscored in 2019, when the same leading general science journal in the same month published two studies that analyzed similar data on fatal police encounters, but came to conflicting conclusions (Edwards *et al.*, 2019; Johnson *et al.*, 2019).

This incoherence is an outgrowth of longstanding data constraints. The roughly 18,000 local police agencies in the United States are not subject to many of the mandated data collection practices that exist in other public institutions, like courts and legislatures (Goff and Kahn, 2012). As a result, scholars have for decades mostly gathered records of police behavior through personal relationships with single law enforcement agencies or targeted freedom-of-information requests, though often the necessary records did not exist at all. This constraint has limited the scope of many analyses not only to single places or times, but also to isolated aspects of policing within those settings, such as traffic-stop reports (Knowles *et al.*, 2001), arrest records (Ousey and Lee, 2008), or use-of-force incidents (Kahn *et al.*, 2016). As a result, scholars have employed a fragmented collection of statistical tests — each relying on partial information about the complex, multi-stage process of police–civilian interactions — with no clear relation or ultimate common goal. Besides producing a difficult-to-synthesize body of evidence, the result is an incomplete and at times inaccurate portrait of the role of race in law enforcement.

If this literature is to cohere and progress, scholars must recognize the complexity of the data-generating process — the sequence of decisions and events that ultimately lead observations to appear in police administrative records — and adopt a common analytic framework that accounts for these features and targets substantively meaningful statistical quantities. As research on other pressing policy matters such as climate change has shown (Rogelj *et al.*, 2019), holistic models of complex processes make it possible to synthesize decades of disparate results — a crucial first step for both scholarly progress and evidence-based policy reform. Such frameworks encourage transparency and facilitate consensus on the shared building blocks of quantitative analyses, opening the door to rigorous evaluation and meta-analysis.

To this end, in "Research Design in the Study of Racial Bias in Policing," we use established tools of causal inference to outline a unifying framework that accounts for the role of race at each step in this process. Though it is not always acknowledged as such, assertions about racial bias in police–civilian interactions are inherently causal claims: they are statements that police would or would not have behaved differently during a police encounter had it involved a civilian of a different race, *counterfactually*, holding all other relevant factors constant. This is not to suggest that descriptive and qualitative research has not made substantial contributions to the study of racial bias writ large. Work such as Alexander (2010) examines structural racism, which can exert powerful effects even in the absence of any bias during police–civilian encounters. On this point we should be clear — the approach we outline does not illuminate macro-level disparities in how society allocates resources, which can contribute to racial inequality in policing outcomes — from the geographic deployment of police officers, to patterns in educational spending, to lending and housing practices. (Such macro-institutional factors could also be incorporated into a causal framework, but that task is beyond the scope of the current analysis.) However, our framework does not seek to merely describe non-causal disparities at the micro level, though these disparities are important to document even if they do not result from biased police behavior.

Rather, we focus on tools for estimating precisely defined causal quantities that capture police bias in micro-level encounters: the average difference in the way officers would behave when encountering minority civilians, relative to white civilians, *all else equal.* When it comes to quantifying racial bias in police–civilian interactions, the strongest evidence is a rigorous causal analysis. In its absence, disparate outcomes that stem from other sources will be misdiagnosed — precluding the most effective policy solutions — and actual racial bias will be too easily dismissed by alternative explanations.

Yet, despite the indisputable value of causal inference in this setting, quantitative research in the policing literature almost never clarifies the crucial ingredients of a rigorous causal analysis: the unit of analysis, the hypothetical manipulation being studied, the estimand about which claims are made, and the identifying assumptions needed to arrive at the stated conclusions. The resulting patchwork of approaches has left the foundations of this research enterprise shaky or entirely undefined, leaving readers unclear even about elementary questions like whether the unit under study is the civilian, the officer, or something else entirely. And even when studies do carefully enumerate these steps, they often invoke heroic assumptions — for example, with parametric models that place strict and unverifiable structure on what civilians know about officers and vice versa (Anwar and Fang, 2006; Knowles *et al.*, 2001; Simoiu *et al.*, 2017).

We propose an alternative framework centered on *police–civilian encounters* — every instance of police contact with civilians, e.g., in a sighting on the street — as the primary unit of analysis. This conceptual approach emphasizes the fact that police records contain only a minuscule fraction of the events of interest. In this framework, civilian race and police bias can play a role at every stage of the causal process, from the initial act of approaching an encountered civilian to the interlocking decisions to question, search, arrest, injure, or even kill — and every step in between. Crucially, our approach is nonparametric, and it focuses on estimating bounds which contain the full range of possibilities consistent with observed data, rather than filling in the gap between data and results using heavy-handed and implausible functional-form assumptions. In the absence of consensus on correct specification of formal models and regression analyses, nonparametric approaches emphasize necessary conditions for principled inference and offer a path forward through unresolvable disagreements on these points. However, our approach does not preclude analysts from using parametric models when they are justified.

After outlining this general causal framework, we present in "Reinterpreting Seemingly Disconnected Approaches to Studying Racial Bias" a structured overview of the most common statistical tests currently employed in the literature on racial bias in policing: (1) simplistic counts with racial encounters in police data in which police take some action, e.g., assessing whether shootings of minority civilians are more numerous than those of white civilians; (2) "benchmark tests," which compare these racial counts with some reference distribution; (3) analyses of post-detainment police action that assumes away bias in detainment, e.g., comparing rates of police violence between stopped minority and stopped white civilians; (4) "outcome tests" that compare civilian racial groups and examine how often police officers are retrospectively "justified" in their behavior, e.g., by discovering contraband; and (5) analyses which compare officers of different racial groups. The difficulty in reconciling these seemingly disparate approaches is perhaps the central obstacle to progress in this literature. Without understanding how each of these analytic strategies relates to one another and why they sometimes appear to produce contradictory results, knowledge aggregation is virtually impossible.

The goal of this paper is to elucidate these connections by expressing each analytic strategy in a common mathematical dialect under the proposed causal framework.

By clarifying the underpinnings of each strategy, our analysis reveals implicit assumptions in some tests and surprising new insights about others, including how some disparate approaches relate to the same underlying causal estimand, or how other approaches can enhance one another. For example, we show that outcome tests, previously thought to only indicate the presence of bias, in fact imply a lower bound on its magnitude under relatively modest assumptions. We also show the so-called "veil of darkness" strategy for uncovering bias in traffic stops (Grogger and Ridgeway, 2006) relates closely to a more established approach, the benchmark test, for which we introduce a sensitivity analysis to probe the vulnerability of results to often-imperfect benchmarks. And we demonstrate our improved interpretation of outcome tests can help analysts correct for sample selection bias in studies of detainment data (e.g., stop or arrest records), suggesting a path toward meaningful meta-analysis. Perhaps most importantly, our approach revolves around the estimation of substantively meaningful quantities that can more easily facilitate policy reform: the number of policing events (e.g., stops, arrests, or uses of force) imposed on minority civilians that would not have occurred had the same police encounters involved white civilians.

Our analysis also points to avenues for enriching the study of racial bias moving forward. In "Moving beyond Data on Detainments," we discuss strategies for future improvements in data collection that are suggested by our causal framework. In particular, we stress the advantages of recording data on the number of police–civilian encounters across racial groups — i.e., the number of times civilians are sighted by police in various settings, regardless of whether police engage further — a crucial quantity which presently is largely unknown. Such data will be difficult to collect, but as our analysis reveals, would greatly improve scholars' ability to produce credible estimates while invoking minimal assumptions.

We note that the study of racial bias presents daunting challenges for the empirical social scientist in any domain. These obstacles are, however, dramatically exacerbated by data scarcity in the study of policing. As a result, a scattershot set of approaches has emerged, some of which, we demonstrate, have likely done more to impede rather than advance the state of knowledge on this important question. The study of racial discrimination in the coercive arm of the state demands a higher standard. By adopting a common and rigorous analytic framework — one that clarifies the conditions necessary for principled inference while accounting for the inherent deficiencies of police administrative data — the study of racial discrimination in law enforcement can progress more rapidly and contribute to meaningful and effective policy reform.

Research Design in the Study of Racial Bias in Policing

Racial bias in policing — that is, disparate police treatment of civilians *because* of their race — is widely discussed but rarely defined in precise terms. What does it mean for civilian race to *cause* police behavior, and how would we know such a discriminatory process was occurring?

As is often the case in causal inference, outlining the "ideal experiment" is a useful way to clarify these issues. The first step in this process is to clearly define the unit of analysis. For example, analysts studying police traffic records must decide whether to analyze police departments, officers, drivers, stops, or something else entirely. A key second step is to carefully specify the counterfactual claim, or the desired estimand, a task that is often glossed over in studies of racial bias. Depending on these choices, all subsequent components of the causal analysis — including identifying assumptions, estimators, and the feasibility of obtaining credible evidence — can differ immensely. We examine each of these components of research design in turn below.

Units of Analysis and Potential Outcomes

As a running example, we consider the task of estimating racial bias in the use of force during an encounter between a police officer and pedestrian on the street. (Here, an officer's use of force can stand in for any other police action, such as a search or arrest.) Building on Knox *et al.* (2020), we take as the unit of analysis the *police–civilian encounter*, which begins at the moment of initial contact — e.g., when a civilian is first sighted by police. Over the course of any officer's shift, hundreds or even thousands of such encounters occur, but because officers are not required to record data on civilians that they observe but do not detain, most encounters leave no administrative trace.

This hyper-granular perspective is often revealing. Many questions relating to racial bias toward civilians are inherently based on aggregated statistics on encounters. For example, a benchmark test may ask whether arrest counts involving a particular group are disproportionate to that group's share of the population. However, the necessary conditions for principled inference are often obscured by aggregate-level views because they do not allow for any interrogation — whether empirical or conceptual — of whether the encounters contributing to those aggregate totals are otherwise similar.

By focusing on police–civilian encounters, we can formalize statements about racial bias in policing. In this context, "civilian race causing differential police behavior" means that counterfactually, substituting an *individual of differing race who is otherwise observationally equivalent to police* into an encounter would have produced a different sequence of police behaviors. For illustration, consider a single encounter, i. As all other characteristics are held fixed, the minority status of the encountered civilian, $D_i \in \{0, 1\}$, determines whether they are stopped by police. We denote this with the potential-outcome notation $M_i(d) \in \{0, 1\}$, which helps clarify the differing counterfactual police behaviors that *would have arisen* if a minority civilian ($d = 1$) had been substituted into the encounter, as opposed to an otherwise identical white civilian ($d = 0$), *holding everything else about the encounter fixed* (Rubin, 1974).[1] For some encounters, $M_i(0) = M_i(1)$, meaning that civilian race has no causal effect on police stopping decisions; in others, $M_i(0) \neq M_i(1)$, indicating racial bias in stopping.

Because it comes at the end of the multi-stage police–civilian encounter, an officer's decision to use force is more complex — it depends on not only civilian race but also whether the civilian was detained (i.e., stopped, $m = 1$) or not ($m = 0$). We denote this joint dependence with $Y_i(d, m)$. Because the decision to stop may itself be a product of civilian race, this is an instance of causal mediation (Imai *et al.*, 2011; Pearl, 2001). When the actual civilian in an encounter is of race D_i, the observed outcome is $Y_i(D_i, M_i(D_i))$; by substituting different values for d and m in the potential-outcome function, analysts can interrogate various counterfactual scenarios.[2]

Having defined the key elements of this causal process, we can now consider the ideal experiment, or the optimal study that could be designed, given unlimited resources and control. Here, the benefit of analyzing the encounter (rather than, say, the civilian) manifests most clearly — it permits the causal question to be probed with a well-defined experimental procedure that closely parallels the counterfactual scenario described above: randomly assigning white and minority civilians to enter pre-existing police beats and act in some prescribed way, then observing the consequent patterns of officer behavior. On average, both sets of civilians would behave identically and appear identical on observable features, *except for their race.* Under these conditions, the analyst could observe the rate at which civilians of each racial group are (1) detained or (2) subject to force, then straightforwardly obtain valid estimates of the average treatment effect (ATE, defined below) on each quantity among the population of individuals that come into the presence of police.

In contrast, attempting to formulate the ideal experiment at the civilian level reveals that the very notion of a "causal effect" of an individual's race is

[1] 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 $D_i = 1$. (This can easily be relaxed by allowing D_{ii} to take on additional values and redefining the potential outcome function accordingly. It is straightforward to extend our analyses to the categorical treatment case.) 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.

[2] 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.

conceptually fraught. Scholars disagree vigorously about whether it is even meaningful to speak of counterfactually manipulating an individual's race while leaving pre-existing aspects of their persona and circumstances untouched (Greiner and Rubin, 2011; Hernán, 2016; Holland, 1986; Pearl, 2018). At a minimum, there is consensus that any conceivable real-world intervention on an individual's race would lead to an inevitably tangled mess of downstream implications in access to education (Orfield *et al.*, 2005), credit/housing (Pager and Shepherd, 2008), medical care (Williams and Wyatt, 2015), and other public goods that do not directly bear on the policy of interest but may distort inferences. This underscores a useful rule of thumb in causal inference: if an ideal experiment is difficult to imagine, even given infinite resources, the causal question may not be well defined.

By analyzing encounters rather than individuals, we make counterfactual claims of the form "if a similar civilian of a different race had been in these circumstances...," rather than the far-fetched "if *this* civilian had been of a different race... ." However, even in the context of encounters, analysts must take into account the complex and multifaceted nature of racial identity. In the words of Sen and Wasow (2016), race is a "bundle of sticks," that affects outcomes through myriad channels. Therefore, when studying racial bias, researchers must be precise about which specific facets of race are under consideration — e.g. which combination of features lead an officer to perceive a civilian as belonging to one racial group rather than another — and which fall under "all else equal" (Greiner and Rubin, 2011). We leave the task of defining "race" — be it conceived as skin tone, cultural features, or some combination thereof — to the analyst, and take up the task of estimating the effect of that factor, however defined.[3]

A word of caution is warranted. While the strategy we outline here offers analytic traction, it also limits the scope of analysis. Specifically, such analyses will necessarily — and intentionally — fail to capture any influence of racial bias that occurs prior to the start of an encounter. For example, if patrol officers behave even-handedly in each encounter, there exists no racial bias *in officer behavior during encounters*. Yet, this does not mean that the system itself is fair: racial bias may still lead policymakers and police commanders to over-allocate officer patrols to minority neighborhoods, resulting in excess uses of force against minorities. As we demonstrate below and elsewhere, this narrowing of scope allows for credible estimation of defined causal quantities of great social and policy importance, but remains an important limitation for readers to keep in mind.

In addition, our approach explicitly makes no attempt at parsing "taste-based discrimination" (racial animus) from so-called "statistical discrimination"

[3] However, given the nature of policing records, which typically include only coarse indicators for the race of civilians as perceived and documented by officers, analysts often have little choice over how to operationalize race in practice.

(Arrow, 1972, 1998; Becker, 1957; Eberhardt *et al.*, 2004; Phelps, 1972) as mechanisms for racially biased policing. While some may view this as a drawback, perhaps due to the widely held view that statistical discrimination is more innocuous, we view these semantic differences as a second-order concern. Statistical discrimination, while not the product of animosity toward minorities, is nonetheless an illegal act of racial profiling in which officers detain civilians not due to their own observed actions but rather the actions of their racial group. Quantifying the causal effect of civilian race on police behavior is imperative, regardless of the motive for discrimination.

Causal Quantities of Interest

A broad literature on race and policing shares the common high-level goal of using data to test whether police behave in a racially biased manner toward civilians of color. However, the precise statistical goals in most of these studies — that is, the specific quantity of interest, and the assumptions necessary to credibly estimate it — are rarely made explicit. Rather, researchers typically express analytic goals informally, stating that they aim to gain insight on an "implicit bias effect" (Nix *et al.*, 2017, p. 317), or estimate "anti-Black disparities" in the outcomes of police–civilian encounters (Johnson *et al.*, 2019, p. 15878). This lack of specificity is a serious hindrance for the accumulation of knowledge, which requires a precise understanding of the claims made in each study. Stating that the goal of a study is to estimate "the effect of race" simply does not provide enough information to judge whether a given statistical test will achieve its goal. As we outline below, there are many types of causal effects relevant to the study of racial bias, each with particular identifying assumptions depending on the analytic strategy and data environment. When imprecise claims are made without reference to a clearly defined target quantity, it becomes exceedingly difficult to judge the validity of any analysis, let alone synthesize results across studies. Here, we present a non-exhaustive set of target quantities that analysts may wish to estimate, along with similarly non-exhaustive assumptions that may be invoked. As we demonstrate in "Reinterpreting Seemingly Disconnected Approaches to Studying Racial Bias", this clarity and transparency is crucial in reconciling diverse methodological approaches and divergent results.

The counterfactual manipulation we describe — substituting an otherwise identical civilian of differing race into an encounter — leads to a widely used, easily interpretable, and substantively meaningful causal quantity of interest, the *average treatment effect* on the outcome of interest, Y_i. We write this estimand as $\text{ATE}^Y \equiv \mathbb{E}[Y_i(1, M_i(1)) - Y_i(0, M_i(0))]$, indicating the hypothetical average change in police behavior that would result if minority civilians were inserted into every police–civilian encounter, instead of inserting white civilians. This estimand accounts for the contribution of all intermediating officer behaviors (Pearl, 2009) — in particular, allowing officers to make stops as they normally would, possibly depending on civilian race. (Analysts interested only in the first stage of the encounter, the decision to stop, M_i, may also consider the related quantity, $\text{ATE}^M = \mathbb{E}[M_i(1) - M_i(0)]$, which represents the average treatment effect of civilian race on this intermediating behavior.) Importantly, this quantity does not presume that the treatment, civilian race, D_i, exerts the same effect across encounters, a feature that existing approaches targeting a parametric quantity often lack.

As we discuss in more detail below, a practical challenge is that the vast majority of these encounters are unobserved; scholars of policing currently have no data on the number of such unobserved encounters, or even its order of magnitude. For this reason, it is somewhat more tractable to focus on the *average treatment effect among stops*, $\text{ATE}^Y_{M=1} \equiv \mathbb{E}[Y_i(1, M_i(1)) - Y_i(0, M_i(0))|M_i = 1]$. This estimand asks the following question: among the subset of encounters that led to detainment ($M_i = 1$), and thus have some record of occurring, what is the hypothetical average difference in police behavior if minority civilians had been present, as opposed to white civilians? This estimand is not only highly policy relevant, as it concerns encounters in which police take some action toward civilians, but also inherently more straightforward to estimate than the ATE^Y for the pragmatic reason that in this subset of observed events, scholars working with police administrative data have information about the circumstances of the encounter. For the same reason, researchers may also focus on the *average treatment effect among stopped minorities*, $\text{ATT}^Y_{M=1} \equiv \mathbb{E}[Y_i(1, M_i(1)) - Y_i(0, M_i(0))|D_i = 1, M_i = 1]$. This causal quantity is closely related to the excess number of police actions toward minorities — the number of incidents that would not have occurred had civilians been white — which may be of even greater policy interest.

Though rarely explicitly stated, some scholars appear to target the *controlled direct effect* of race on the outcome, an alternative quantity denoted $\text{CDE}^Y \equiv \mathbb{E}[Y_i(1, 1) - Y_i(0, 1)]$. The CDE^Y and its subset counterpart, $\text{CDE}^Y_{M=1} \equiv \mathbb{E}[Y_i(1, 1) - Y_i(0, 1)|M_i = 1]$, ask a different question: what would happen if, in addition to manipulating the race of the encountered civilian, a researcher also *forced officers to make a stop* regardless of civilian race. However, this estimand is notoriously difficult to work with, even when studying actual police detainment records. The challenge arises because of racial bias in police stops — in many circumstances, police will stop minorities in situations where white civilians would be allowed to pass. Practically speaking, white civilians in these circumstances never appear in police records, so there is simply no data that allows an analyst to interrogate this question.[4] And

[4] As we show in Appendix A.3 of Knox *et al.* (2020), analysts claiming to estimate this quantity must implicitly rely on highly implausible assumptions to fill the gap between data and claims.

more conceptually, the CDE and CDE$_{M=1}$ are based in part on "cross-world" scenarios that never occur naturally. For example, if police use force against a Black civilian stopped for jaywalking — a situation in which a white civilian might not have been stopped — the CDE$_{M=1}$ would ask whether police would have used force against a white civilian detained for the same reason, even though the stop *would never have happened.* Because this manipulation is impractical, these quantities are extremely difficult to estimate and are of limited use from a policy perspective.

Identifying Assumptions in Policing Research

Regardless of which quantity of interest the analyst targets, the next step in the analysis is to assess whether it is causally identified — in other words, if it can be estimated well. Because practical, financial, or ethical concerns typically preclude the ideal experiment, achieving the conditions necessary to credibly estimate a causal effect is extremely challenging, especially when studying policing. It has been said that the "fundamental problem of causal inference" (Holland, 1986) is a missing data problem: we seek to compare the outcomes for a given unit (e.g., a police–civilian encounter) under two states of the world (involving a white and nonwhite civilian), but we can only observe one of the two potential outcomes. However, because most police encounters are never recorded, the missing data problem here is far worse than usual. For example, police are not required to record encounters in which they do not engage a civilian (e.g., a civilian walks by an officer on the street without any further action taken by police). Because of this feature, rather than simply being unable to observe one counterfactual outcome for a particular encounter, we typically do not observe the encounter *at all*, making causal inference extremely difficult.

This missing data problem is visualized in the upper panel of Figure 1, which depicts four types of police–civilian encounters: those with minority civilians and those with white civilians, each divided into encounters that resulted in a detainment (i.e., a stop) and those that did not. We use the terms "stop" and "detainment" broadly here, and stress that these can refer to many types of police actions so long as they trigger a reporting requirement that causes an encounter to appear in police administrative data. Depending on the setting, this may refer to stops of drivers or pedestrians, responses to 911 calls, arrests, drawing a weapon against a civilian, or any other type of intermediate behavior by police officers which leaves an administrative trace but is temporally prior to the outcome being studied (e.g., use of force during an encounter). Note that this perspective applies regardless of whether an interaction is initiated by officers or civilians. Because researchers analyzing policing data only observe encounters that involve a detainment, police administrative records paint a misleading portrait of the broader universe of encounters. As we show later,

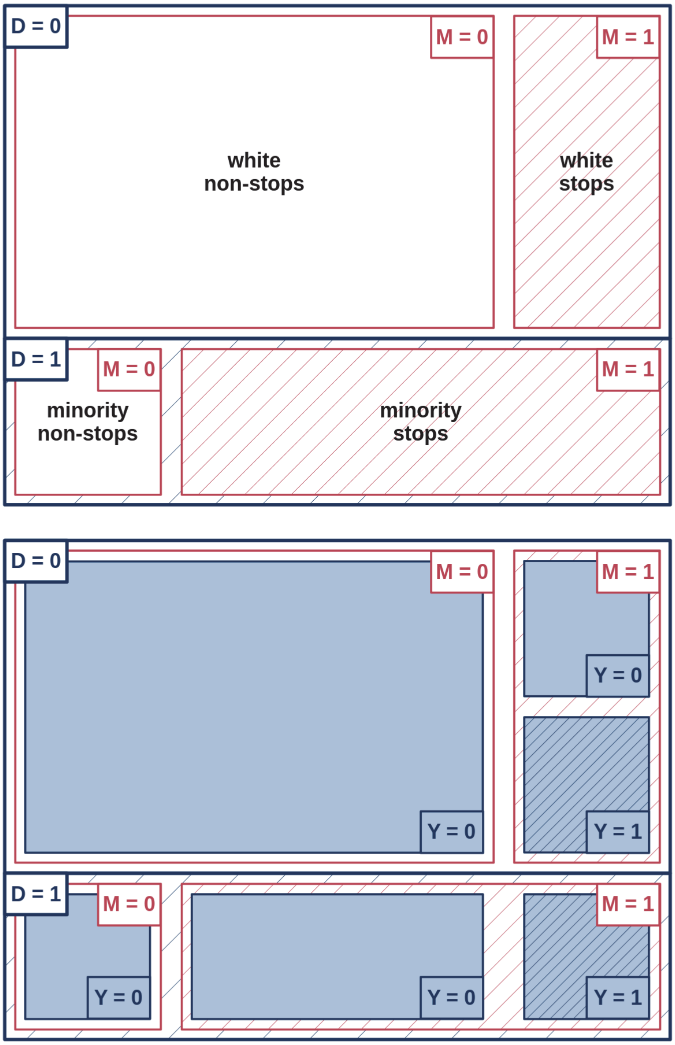

Figure 1: **Universe of police–civilian encounters**: The top panel of the diagram partitions the universe of police–civilian encounters into those that are treated (those involving minority civilians, $D_i = 1$) and control (those involving white civilians, $D_i = 0$), as well as the decision by officers to detain civilians ($M_i = 1$) or let them pass ($M_i = 0$). The bottom panel further divides the encounter space based on the action taken by the police officer, e.g., the decision to use force ($Y_i = 1$) or not ($Y_i = 0$). In our discussion of popular strategies to estimate racial bias, we refer back to this diagram to illustrate the types of encounters utilized in each approach.

given this challenging data environment, scholars of policing must proceed with caution to avoid drawing mistaken conclusions. The lower panel subdivides these cells into encounters that do or do not result in some police behavior, like use of force.

Conceptually, encounters can be subdivided further into *principal strata* based on how they would respond to different treatment scenarios (Frangakis and Rubin, 2002). Briefly, these principal strata represent four basic types of encounters that analysts can only partially distinguish: *always-stop* encounters in which officers would stop any civilian regardless of race, such as violent crimes observed in progress; *anti-minority (anti-white) racial stops* in which only minority (white) civilians would be stopped; and *never-stop* encounters in which neither group would be detained, e.g., when the civilian behaves inconspicuously. Because these categories of encounters cannot be readily distinguished in data, making "apples-to-apples" comparisons across encounters becomes extremely challenging. However, acknowledging the existence of these latent types, we show below, is extremely useful for identifying or bounding causal effects in policing data.[5]

The inherent deficiencies of policing data highlighted above require identifying assumptions to estimate causal effects. Carefully enumerating all such assumptions is a crucial step in any causal analysis that allows scholars and critics alike to answer the following question with precision: given the nature of policing data, what would have to be true about the world to interpret the result of a given empirical test as valid evidence of racial bias in policing? As we discuss in "Reinterpreting Seemingly Disconnected Approaches to Studying Racial Bias," identifying assumptions in the existing literature often fall into two extremes: (1) unstated, implicit assumptions that obscure the conditions required for the stated conclusions to hold, or (2) unnecessarily strong and restrictive parametric assumptions that almost certainly are not satisfied in real-world encounters. Below, we outline four minimal, nonparametric, and substantively motivated assumptions that allow for the study of racial bias using police administrative data. In the analysis that follows in "Reinterpreting Seemingly Disconnected Approaches to Studying Racial Bias," we appeal to some or all of these assumptions as needed in order to clarify the basis of each approach to estimating racial bias.

Assumption 1, ***Mandatory Reporting***, states that a record is made when police take some action of interest. In other words, analysts assume that if a record does not exist, the police behavior of interest did not occur. This could be violated if, for example, police fail to record instances of force. Though there exists wide variability in data-recording practices across jurisdictions, this assumption is plausible in many major police departments. For example,

[5] For an in-depth discussion of principal strata in this context, we refer readers to Knox *et al.* (2020).

New York Police Department (NYPD) officers are required to report a number of variables, including the specific type of force used, following each “stop, question, and frisk” encounter. Based on these and other reports, the NYPD releases detailed annual use-of-force reports (NYPD, 2017). The completeness of these reports with respect to fatalities is informally enforced by standard journalistic 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 scrutiny of police brutality (Fisher and Hermann, 2015) makes unobserved uses of force increasingly unlikely. We note that this assumption is implicit in all analyses of police use of force that rely on administrative data.

Assumption 2, ***Mediator Monotonicity***, holds that no anti-white bias exists in detainment. That is, the assumption states that there are no circumstances in which a white civilian would be detained by police, $M_i(0) = 1$, but an identically situated non-white civilian would be allowed to pass, $M_i(1) = 0$. This is clearly a stylized representation of a complex reality, and it would be violated if minority officers discriminate against white civilians. However, to the extent anti-white bias exists in the decision to stop civilians, the staggering differences in the volume of stops involving white and nonwhite civilians (Gelman *et al.*, 2007; Mummolo, 2018) suggest that their prevalence is minimal.

Assumption 3, ***Relative Non-severity of Racial Stops***, holds that among encounters with a particular civilian race, the severity of the police behavior applied in always-stop encounters is greater than or equal to the severity applied in racial-stop encounters, on average. We regard this assumption, which compares violence rates within encounters that hold civilian race fixed, as highly plausible. As one hypothetical example, this assumption would imply that police are as or more likely to use force against a minority civilian observed committing assault (where a white civilian would also be detained) than a minority civilian observed jaywalking (where a white civilian might be allowed to pass).

The strongest assumption we discuss is **Assumption 4**, ***Treatment Ignorability*** with respect to both M_i and Y_i. In order to clarify what this assumption requires, it is useful to visualize the causal process that generates actual police records, as in the directed acyclic graph (DAG) in Figure 2. Potential confounding factors, X_i, play a role in police behavior, while also making officers more likely to encounter civilians of a particular racial group. For example, when an officer is deployed to a minority neighborhood, he or she will naturally encounter more minority civilians and, perhaps due to department quotas or crime rates, also be more willing to initiate stops. In other words, neighborhood racial composition is a potential confounder since

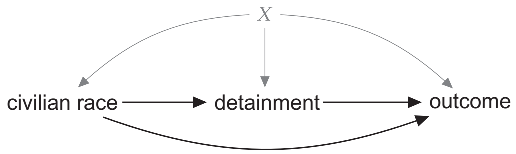

Figure 2: **Causal process in police–civilian encounters.** Encounters result in (i) detainment, M_i, and (ii) subsequent police behavior, such as the use of force, Y_i. The key quantity of interest is the extent to which the race of the encountered civilian (D_i) influences any of these police decisions, either directly or indirectly. Unobserved confounding factors, X, may produce spurious correlations between any of these variables. For example, changing precincts will result in officers encountering civilians of different races. These complicate the inferences that can be drawn from observed data, as discussed in the main text. All analytic results assume faithfulness to this directed acyclic graph, following standard practice in causal inference. While knife-edge cases can in theory arise (e.g., when two forms of statistical bias exactly cancel each other), the implausibility of such scenarios render them irrelevant for applied research.

it affects both D_i and M_i. The analyst will need to adjust for some or all of the confounders contained in X_i, depending on the causal quantity of interest, before comparisons of white and minority encounters will yield unbiased estimates of the desired causal effect. The required adjustment strategy can be determined using the procedure of Pearl (1993); for a primer, we refer the reader to Elwert and Winship (2014).

This assumption, while strong, is necessary to apply any of the approaches we describe below. Fortunately, Assumption 4 has become more plausible in recent years as administrative data sets have come to include a host of encounter attributes observable to police. Many of these attributes capture factors that correlate with suspect race and the potential for force. 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. However, Assumption 4 is difficult to test, even indirectly, without data on non-stopped individuals. For this reason, we elaborate in "Moving beyond Data on Detainments" on the need to collect data on police–civilian encounters in which detainments did not occur.

Though these assumptions are much less stringent than many currently invoked in the literature on racial bias, we readily acknowledge that some skeptical readers may still find them implausible. We nonetheless stress that without invoking some or all of these assumptions, (or others like them), quantitatively estimating racial bias in policing is virtually impossible. It is therefore imperative to state these assumptions explicitly so that scholars can attempt to validate them in practice by collecting additional data and debate their plausibility when reviewing research.

With the building blocks of a causal analysis in hand — namely, the unit of analysis, causal estimand(s), and identifying assumptions — we turn in the next section to re-evaluating established approaches to estimating racial bias in police behavior. As our analysis shows, situating each strategy in this causal framework clarifies both the validity and substantive interpretation of each approach.

Reinterpreting Seemingly Disconnected Approaches to Studying Racial Bias

Because analysts almost never have information about all stages of police–civilian interactions, statistical tests of racial bias are tailored to whatever limited data is available in a given place and time. While no essay can feasibly describe every empirical procedure in this wide-ranging literature, "Review of Prominent Approaches" analyzes four broad approaches that together account for the vast majority of applied research. In "Approaches Incorporating Officer Race," we show how these basic approaches can be expanded with additional information on officer race.

In what follows, we discuss published examples of each approach, then reinterpret them in our general causal framework. This exercise illuminates the substantive meaning of each method as well as the connections between seemingly unrelated methods. Along the way, we show how this framework reveals a number of hidden assumptions and limitations in widely used approaches. We note that the published examples we highlight are not exhaustive; our goal is to present one or two illustrative examples of each leading analytic strategy to illustrate pros, cons, and interconnections. We refer readers to Goff and Kahn (2012) and Ridgeway and MacDonald (2010) for more extensive literature reviews on racial bias in policing.

Review of Prominent Approaches

We begin with an analysis of the four most common approaches used in analyzing police data. The first, *retrospective "predictions" of civilian race*, works within a particular convenience sample of encounters — the relatively easy-to-identify instances when a police behavior occurs, such as the use of force — and examine how often each civilian group appears in this subset. *Benchmarking* analyses similarly work with this convenience sample (or the broader convenience sample of all civilian stops), but incorporate an external reference such as census data to compute a proxy for the unobserved proportion of encounters with each civilian group. A third prominent approach *assumes away racial bias in detainment.* This strategy analyzes police stop records, ignoring race-based selection into this data set, and compares use-of-force rates in stops involving minority and white civilians. Finally, *outcome tests* introduce the notion of a "successful" or "justified" stop, e.g., a stop based on suspicion of weapon possession that in fact recovers the suspected weapon. Below, we address each strategy in turn, clarifying their meaning as well their ability to provide credible information about racial bias.

"Predicting" Civilian Race, after the Fact

As we note above, data on non-stops — encounters in which civilians are not detained — are almost never collected. Even data on stops of civilians, though available, are difficult to analyze comprehensively. To study these records requires corralling a patchwork of police departments and state bureaucracies, which often employ different reporting thresholds, record different variables, and apply different definitions. However, organizations such as *The Washington Post* and Fatal Encounters (Burghart, 2020) have recently expended substantial effort to collect all cases *in which some particular outcome appears*, such as fatal officer-involved shootings, in a uniform format. As a result, many scholars have sought to sidestep the challenges of cross-jurisdiction data by analyzing these national datasets (e.g., Menifield *et al.*, 2019; Nix *et al.*, 2017; Ross, 2018).

The central drawback of this approach is that such data sets contain no variation in the outcome of interest. Despite this *selection on the outcome*, scholars have still attempted to use these sources to test whether civilian race affects police use of lethal force. Because the outcome of interest does not vary, these studies often substitute a different variable in its place — the race of the civilian — and proceed by computing either the proportion of fatally shot civilians belonging to different racial groups, or testing whether other features of shooting incidents predict civilian race. In other words, this approach substitutes the treatment for the outcome during estimation. However, such tests, when properly understood, have virtually no chance of illuminating whether police are racially biased in their decisions to use force (or engage in any other behavior toward civilians).

A prominent recent example is Johnson *et al.* (2019), which analyzes data on one year of fatal officer-involved shootings across the United States. This paper received widespread media coverage and was cited in a 2019 Congressional oversight hearing on policing practices (Mac Donald, 2019). Using only data on fatal shootings, Johnson *et al.* (2019) state several strong conclusions, including: "a person fatally shot by police was…less likely to be Black than white and…less likely to be Hispanic than white. Thus, in the typical shooting, we did not find evidence of anti-Black or anti-Hispanic disparity…and, if anything, found anti-white disparities" (p. 15880).

We summarize this approach graphically in Figure 3. Formally, approaches like the above involve comparing $\Pr(D_i = 1|Y_i = 1)$ and $\Pr(D_i = 0|Y_i = 1)$.

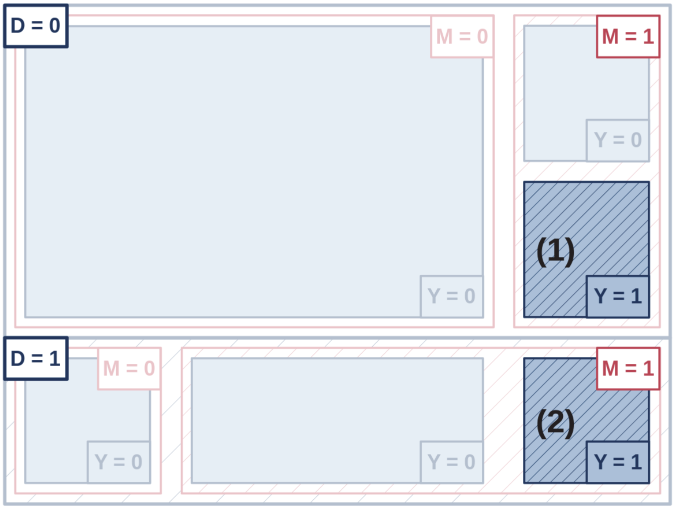

Figure 3: **Selection on the outcome.** Encounters are grouped into cells according to the race of the encountered civilian, whether a stop occurred, and whether the stop resulted in a police behavior such as use of force. The outcome-selection approach focuses on cells labeled (1) and (2), ignoring the remaining faded regions. Analysts argue that anti-white bias exists when cell (1) is larger than cell (2), or anti-minority bias when the reverse is true. However, this approach fails to account for the rate of minority and white encounters — the size of the $D_i = 0$ and $D_i = 1$ boxes — and is uninformative about any causal quantity of interest.

In other words, these studies examine shooting incidents, then compare the proportion that contain each civilian group. This process is logically backward compared to the usual comparison, which is based on $\Pr(Y_i = 1|D_i = 1) - \Pr(Y_i = 1|D_i = 0)$: in words, examine encounters with each civilian group, then compare the proportion that results in a shooting.

Despite the claims of scholars employing this approach, selecting on the outcome is a fatal flaw for studies that attempt to draw causal inferences (Elwert and Winship, 2014). A simple thought experiment illuminates the challenge. Suppose that officers encounter 1,000 minority civilians and 2,000 white civilians in identical circumstances, then fatally shoot 250 of each group. The outcome-selection approach would only analyze the 500 encounters in which shootings occurred, then conclude that there is no "disparity" because equal numbers of each group appear. But a careful inspection of the quantity of interest — "the degree to which Black civilians are more likely to be fatally shot than white civilians" (Johnson *et al.*, 2019, p. 15877), or the ATE^Y — makes the flaw in this reasoning clear. As we show in Knox and Mummolo (2020), Bayes' rule implies that analysts must account for the (unobserved) size of each group of encounters (including non-shootings) when attempting to draw conclusions about the ATE^{Y}. The procedure described above will produce misleading results if officers do not encounter minority and majority civilians in equal number. If officers encounter far more white civilians due to their majority status, as in the thought experiment above, then equal shootings of each group would be evidence of anti-minority bias. Conversely, if they encounter more minority civilians due to police deployment patterns, then equal shooting counts would imply anti-white bias. Without this information, data on shooting incidents alone are uninformative if the goal is to quantify racial bias: any result is consistent with an $\text{ATE}^{Y}_{M=1}$ spanning nearly the entire possible range, from -1 to 1, as we show in Appendix A. We return to this point in "Benchmark Tests."

Some studies extend the outcome-selection approach while retaining much of its basic structure. For example, Streeter (2019) similarly examines fatal encounters, but with the addition of encounter covariates like whether the civilian was engaged in criminal behavior or posed a threat to the officer. The study then evaluates whether each attribute is "predictive of [civilian] race," (p. 1124) conditional on a fatal shooting occurring. In other words, this approach pivots away from merely asking if minority shootings are more or less numerous than white shootings *in general*. Instead, it examines whether minority civilians are more or less common, for example, *in threatening encounters that resulted in shooting* (denoted here as $X_i = 1$) *compared to non-threatening encounters* ($X_i = 0$). Formally, rather than testing if $\Pr(D_i = 1|Y_i = 1) \neq 0.5$, this approach tests if $\Pr(D_i = 1|X_i = 1, Y_i = 1) \neq \Pr(D_i = 1|X_i = 0, Y_i = 1)$. Based on this comparison, the study concludes: "...the racial disparity in the rate of lethal force is most likely driven by higher rates of police contact among African Americans rather than racial differences in the circumstances of the interaction and officer bias in the application of lethal force," (p. 1124).

Though slightly more complex, there is a close parallel between the implicit assumption here and in the simpler case. As we show in Appendix A, the ability of these analyses to inform the study of racial bias hinges entirely on the assumption that minority and white civilians are equally likely to be threatening toward police, or that $\Pr(X_i = 1|D_i = 1) = \Pr(X_i = 1|D_i = 0)$, in other words, that civilian race is as-if random *even before conditioning on covariates*. This is even stronger than Assumption 4, which states that civilian race is as-if random only after conditioning on covariates. Violations of this assumption can lead the analyst astray. To see this, suppose that white civilians are more willing to attack officers. Further suppose that officers always shoot in every threatening encounter, without regard for civilian race, and they shoot at some lower but similarly unbiased rate in non-threatening encounters. If the analyst fails to account for differential threat levels across encounters, the relevant causal quantities $\mathbb{E}[Y_i(1, M_i(1)) - Y_i(0, M_i(0))|X_i = 1]$ (racial effect in threatening encounters) and $\mathbb{E}[Y_i(1, M_i(1)) - Y_i(0, M_i(0))|X_i = 0]$ (racial effect in non-threatening encounters) will both be zero, but analysts would erroneously conclude that racial bias exists because threat level is predictive of race. By analyzing only encounters in which fatal shootings occurred, the analyst has no purchase on whether Assumption 4 (treatment ignorability) is likely to hold, and adjusting for features of fatal encounters, as in Johnson *et al.* (2019), cannot resolve this underlying issue. This approach therefore cannot distinguish whether any observed disparities (or lack thereof) are due to differential rates of contact with civilian groups, differential circumstances across encounters, or racial bias on the part of officers.

Indeed, of all approaches examined here, selecting on the outcome is undoubtedly the most problematic — the identifying assumptions are implausibly strong, rarely made explicit, and virtually impossible to verify in this setting because the data encapsulate such a minuscule proportion of police encounters. Because police behavior does not vary in this approach (e.g., every observation involves a fatal encounter), these tests cannot shed light on whether the race of individuals involved affects the probability that officers shoot civilians. While this approach may have some uses in purely descriptive exercises, we recommend it never be adopted if the goal is to study the causal question of racial bias in police–civilian encounters.

Benchmark Tests

If analysts have additional information about how often each group is encountered, the above approach can be improved dramatically. Even if detailed data are only available for encounters that result in fatal shootings, it may suffice to know the *number* of total encounters from each racial group. This information is sufficient to estimate $\Pr(Y_i = 1|D_i = 1)$ and $\Pr(Y_i = 1|D_i = 0)$ — the proportion of each group's encounters that result in a shooting — and when there is no confounding (i.e., treatment ignorability is satisfied), then the difference in these quantities will yield the ATE^Y. Similarly, if analysts are interested in estimating the effect of race on stops, they can use the number of encounters and stops (instead of shootings) in each group to estimate $\text{ATE}^M = \Pr(M_i = 1|D_i = 1) - \Pr(M_i = 1|D_i = 0)$. In addition, in the presence of confounding, this procedure can only be carried out after conditioning on a sufficient set of encounter circumstances to render white and minority encounters otherwise equivalent.

The chief complication in benchmark tests is that most encounters are unobserved, meaning that total counts are unavailable. A common approach for dealing with this is to "benchmark" against population demographics. For example, researchers have found that racial minorities are killed by police in numbers that are disproportionate to their share of the jurisdiction population (Edwards *et al.*, 2019). As Figure 4 makes clear, benchmarks such as racial demographics, which we refer to as $Q(D_i = 1)$ and $Q(D_i = 0)$, are merely a proxy for the desired missing information, $\Pr(D = 1)$ and $\Pr(D = 0)$, if the

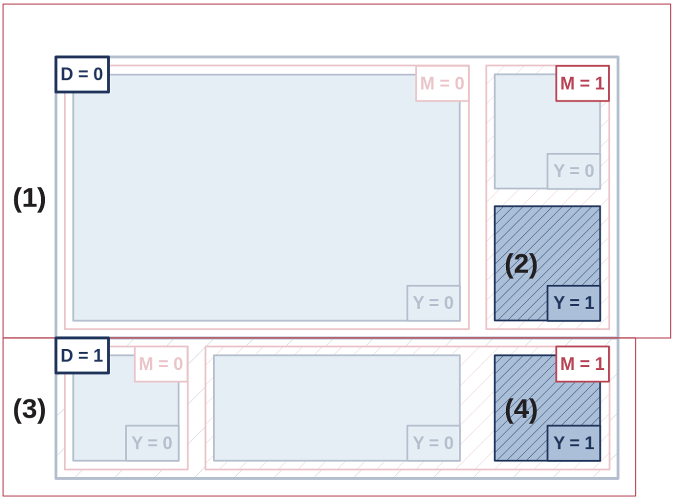

Figure 4: **Benchmark tests.** Encounters are grouped into cells according to the race of the encountered civilian, whether a stop occurred, and whether the stop resulted in a police behavior such as use of force. Outcome-based benchmark tests focus on cells labeled (2) and (4), ignoring the remaining faded regions (stop-based benchmark tests consider the sizes of the $M_i = 1$ boxes instead). Analysts compare these with (1) and (3), proxies for the unobserved racial encounter rate (the sizes of the $D_i = 0$ and $D_i = 1$ boxes). Anti-minority bias is argued to exist when $\frac{(4)}{(3)}$ is larger than $\frac{(2)}{(1)}$. Under assumptions given in the main text, differences in these quantities can reveal the sign of the ATE^Y (outcome-based benchmark tests) or ATE^M (stop-based benchmark tests).

goal is to assess the causal question of racial bias. Formally, the *benchmark test statistic* is

$$\frac{\sum_i \mathbf{1}(D_i = 1, Y_i = 1)}{\sum_j \mathbf{1}(D_j = 1)} - \frac{\sum_i \mathbf{1}(D_i = 0, Y_i = 1)}{\sum_j \mathbf{1}(D_j = 0)}$$
$$\propto \frac{\Pr(D_i = 1 | Y_i = 1)}{Q(D_i = 1)} - \frac{\Pr(D_i = 0 | Y_i = 1)}{Q(D_i = 0)},$$

where i indexes encounters in the observed data set and j indexes observations in the benchmark group, such as the local population. (Again, it is also possible to conduct benchmark tests of police stops, in which case M_i is used instead of Y_i in the above expression.) When the proportion of shootings that involve minority civilians is greater than the proportion of minorities in the benchmark distribution, the first term will exceed one and the benchmark statistic will be positive; when racial proportions are equal in shootings and the benchmark distribution, the statistic will be zero.

While a positive benchmark statistic is consistent with racial bias, it is not direct evidence of it. If the goal is to estimate racial bias in shootings, the crucial assumption when using this approach is that there are no unobserved circumstances that relate to both civilian race and shootings, i.e., treatment ignorability. To see this, note that in some settings officers encounter minorities more often due to deployment patterns. If officers open fire in some fixed percentage of encounters, regardless of civilian race, then the ATE^Y is in truth zero, but a benchmark test will erroneously suggest racial bias during encounters. In recognition of these limitations, scholars have developed more advanced benchmark tests that condition on *observable* confounders, such as residence location or race-specific crime rates. These all share the common goal of seeing whether police violence toward a given group is higher than these factors might predict in the absence of bias (Gelman *et al.*, 2007). (Here, one important caveat is that when conditioning on race-specific crime rates based on historical police data, researchers risk inadvertently introducing past police bias into their analyses. For example, if analysts use historical race-specific arrest counts instead of racial census counts — both common approaches — and if officers have historically over-arrested minorities due to racial bias, then the use of this skewed benchmark will paint a misleading portrait by artificially inflating the "typical" level of criminal activity in this group.)

However, by adopting the framework we propose, and relating the benchmark test to a specific causal quantity, analysts can go a step further by using domain expertise to assess the quality of their benchmark. Specifically, if scholars can determine the maximum possible error between the proxy racial proportions and the true encounter racial proportions, $\delta \equiv \max_d \Pr(D_i = d) - Q(D_i = d)$, then a benchmark test can then a benchmark test can be partially informative about the *sign* of the ATE^Y. Using Bayes' rule, it can be shown that

$$\Pr(Y_i = 1)\left(\frac{\Pr(D_i = 1|Y_i = 1)}{Q(D_i = 1) + \delta} - \frac{\Pr(D_i = 0|Y_i = 1)}{Q(D_i = 0) - \delta}\right)$$
$$\leq \ \text{ATE}^Y \ \leq$$
$$\Pr(Y_i = 1)\left(\frac{\Pr(D_i = 1|Y_i = 1)}{Q(D_i = 1) - \delta} - \frac{\Pr(D_i = 0|Y_i = 1)}{Q(D_i = 0) + \delta}\right).$$

The shooting rate, $\Pr(Y_i = 1)$, is typically not available because the total number of encounters is unknown. Thus, analysts are generally only able to estimate a quantity that is *proportional* to the ATE^Y. However, if a plausible range of shooting rates and proxy errors can be identified, the above inequality can be used as a form of sensitivity analysis, revealing how poor the proxy would have to be to lead analysts to a faulty conclusion.

We stress that this approach breaks down entirely in the presence of confounding. As in the outcome-selection approach, if, for example, white civilians are more willing to engage in threatening behavior toward police, then even a perfect proxy for racial encounter rates will not lead to correct inferences. Instead, analysts must obtain separate proxies for (1) the number of threatening minority and white encounters and (2) the number of non-threatening minority and white encounters, then conduct benchmark tests separately within each group and compute a weighted average of the two to recover the ATE^Y. If additional confounders exist, they would need to be incorporated via the same process. In practice, it is nearly impossible to perform this procedure correctly, so analysts often implicitly assume that no such factors exist.

An additional obstacle to inference in the study of discriminatory policing is sample selection bias: if officers racially discriminate in choosing which civilians to engage, then records of such events will suffer from another source of confounding (Heckman, 1977; Rosenbaum, 1984). We discuss this problem at length in the following section, but note it here because a prominent method of combating sample selection bias — the so-called "veil of darkness" strategy (Grogger and Ridgeway, 2006) — addresses this issue using an approach that our framework reveals to be a special case of the benchmark test. In this approach analysts compare daytime traffic stops ($X_i = 1$) to nighttime stops ($X_i = 0$), drawing on the idea that officers will be unable to determine the race of a driver in evening hours prior to making a stop. If officers are not racially biased in their stopping decisions, then the racial composition of stopped drivers should not differ between daylight and evening hours, as long as the racial composition of drivers also does not change over time.

Our framework reveals a surprising connection to benchmark tests. Formally, the above assumptions are $\Pr(D_i = d|X_i = 0) = \Pr(D_i = d|X_i = 1)$, or that racial composition of drivers is constant between day and night; and $M_i(0) = M_i(1)|X_i = 0$, that officers stop both groups without bias in nighttime encounters, where civilian race is "invisible." When these assumptions are satisfied, it can be seen that $\Pr(D_i = d|M_i = 1, X_i = 0)$ is identical to $\Pr(D_i = d)$, so the racial composition of nighttime stops serves as a perfect benchmark — i.e., $\delta = 0$. (Typically, the first assumption is made more plausible by restricting analysis to the time around sunset or by exploiting the onset of daylight savings time; the latter assumption can be weakened, so that civilian race is merely *less* of a factor at night.) Thus, veil-of-darkness tests identify a quantity that is proportional to the ATE^Y or ATE^M.

Assuming away Racial Bias in Detainment

In response to freedom-of-information requests, transparency laws, court mandates, and in some cases voluntarily, police departments have increasingly released data on events in which civilians are detained, such as traffic stops or pedestrian stops. A number of studies have used these data to study the role

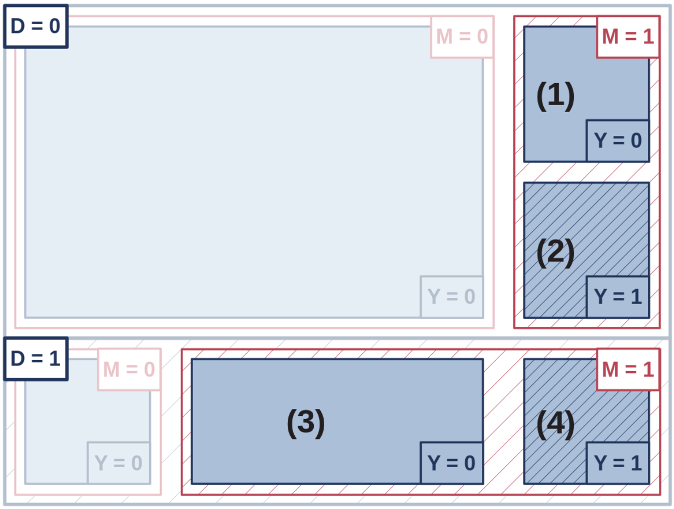

Figure 5: **Stop selection studies and outcome tests.** Encounters are grouped into cells according to the race of the encountered civilian, whether a stop occurred, and whether the stop resulted in a police behavior such as use of force. The stop-selection approach and outcome tests both focus on cells labeled (1)–(4), ignoring the remaining faded regions. In this stop-selection approach, which implicitly assumes away racial bias in the police stops, a positive outcome ($Y_i = 1$) represents some subsequent police behavior, such as the use of force. Analysts argue that anti-minority bias in force exists when $\frac{(4)}{(3)+(4)}$ is *larger* than $\frac{(2)}{(1)+(2)}$, or when force is observed in a larger proportion of minority stops. However, this approach fails to account for the rate of minority and white encounters — the different size of the $M_i = 1$ box within $D_i = 0$, compared to the $M_i = 1$ box within $D_i = 1$ — and is only partially informative about the quantity of interest. In outcome tests, $Y_i = 1$ represents a retrospectively "justified" stop, e.g., due to the discovery of contraband such as a weapon or drugs. Here, interpretation differs enormously. Analysts argue that anti-minority bias in *stops* exists when $\frac{(4)}{(3)+(4)}$ is *smaller* than $\frac{(2)}{(1)+(2)}$, or when contraband is found in a smaller proportion of minority stops. Under assumptions given in the main text, disparities in evidence rates imply a lower bound on $\text{ATE}^M_{D=1,M=1}$, and on the proportion of minority stops that counterfactually would not have occurred if white civilians were substituted into the same circumstances.

of civilian race in police behavior. We refer to this broad analytic strategy as *stop-selection*, though it is important to note that the below also holds for analyses involving other types of detainment, such as arrest records or situations in which officers draw their weapons (Worrall *et al.*, 2018). As we show in Figure 5, stop-selection analyses work with a convenience sample of police–civilian encounters, much as outcome-selection analyses work with readily available nationwide databases of shootings. Though they include a larger set of encounters (all encounters resulting in a stop, whether or not force was used), they are still missing many more.

Unlike approaches that select on the outcome, these data include cases in which police do not use force, but every observation in the data involves a police encounter in which a civilian was detained, i.e., stopped on the street, arrested, or otherwise engaged by officers (Ridgeway, 2016; Wheeler *et al.*, 2017). The analytic strategy here is to examine rates of post-stop police behavior, such as the use of force, and compare these rates across different civilian races. Specifically, analysts estimate $\Pr(Y_i = 1|D_i = 1, M_i = 1)$ and compare with $\Pr(Y_i = 1|D_i = 0, M_i = 1)$. (As with outcome-selection studies, analysts typically make this comparison after attempting to hold fixed or adjust for confounding factors, meaning they attempt to satisfy treatment ignorability.)

One pitfall in stop-selection analyses is to interpret this racial difference in police behavior (e.g., use-of-force rates) as an average causal effect, i.e., the change in police behavior that would result if stopped minority civilians were substituted into every observed encounter, as opposed to white civilians. One prominent example is Fryer (2019), which examines the use of different levels of force, including sub-lethal force like the use of a baton, as well as lethal shootings. Fryer (2019) describes the observed difference in force rates as the effect of civilian race "conditional on an interaction" (Fryer, 2019, Table 2) and concludes that while there is some anti-minority bias in the use of sub-lethal force, there is no evidence of bias in lethal force. As a result, the study received substantial media coverage (Bui and Cox, 2016).

While far superior to studies which select on the outcome, stop-selection studies still rely on implicit assumptions that are highly implausible. Knox *et al.* (2020) shows that in fact, results cannot be interpreted as the study claims. The issue is that if officers do not stop white and minority civilians according to the same criteria, analysts lack a valid comparison set, regardless of the causal quantity of interest. Because there are inevitably unobserved factors that jointly influence the decisions to stop and use force, analyzing only stops introduces selection bias. Therefore, for the conclusions in Fryer (2019) to hold, there must be no racial bias in the decision to detain civilians — in other words, to deem these estimates credible, *analysts must assume away racial bias in a study of racial bias.*

For intuition, recall our discussion of principal strata in police stops in "Identifying Assumptions in Policing Research." Given anti-minority bias in police detainment, minority stops in these data sets would be a mixture of always-stop situations (like violent crimes in progress) and anti-minority stops (e.g., in circumstances like jaywalking), where in the latter case, a stop would not have occurred had the civilian been white. In contrast, there will be no comparable scenarios (i.e., scenarios in which only minorities would be detained) among the recorded white encounters. These different classes of stops cannot be disentangled by the analyst. Importantly, this distinction persists even when stop records contain detailed information on the circumstances of particular stops — we can never see the counterfactual stopping decision, so we cannot determine whether a stop would have occurred had civilian race differed. So long as racial discrimination in stopping decisions occurs, comparisons of average outcomes across stops involving different racial groups of civilians are not "apples to apples," and do not return the causal effect of race "conditional on an interaction" as prior work has claimed.

However, the analyst can still obtain partial information on racial bias in the use of force (or other post-stop outcomes) using only data on stops. Proposition 1 of Knox *et al.* (2020) shows that under the assumptions of "Identifying Assumptions in Policing Research," we can still obtain nonparametric sharp bounds on the $ATE_{M=1}$ and $ATT_{M=1}$ — the tightest possible range of conclusions that are consistent with the data, given issues of sample selection described above. These bounds depend on the severity of racial discrimination in police stops (specifically, the share of encounters involving detained minorities who would not have been detained had they been white), denoted ρ. The following expression defines these bounds:

$$\begin{aligned}
&\mathbb{E}[\hat{\Delta}] + \rho\ \mathbb{E}[Y_i|D_i = 0, M_i = 1]\ (1 - \Pr(D_i = 0|M_i = 1)) \\
&\qquad\qquad\qquad\qquad \leq\ \text{ATE}^Y_{M=1}\ \leq \\
&\mathbb{E}[\hat{\Delta}] + \frac{\rho}{1-\rho}\Bigg(\mathbb{E}[Y_i|D_i = 1, M_i = 1] \\
&\quad - \max\left\{0, 1 + \frac{1}{\rho}\mathbb{E}[Y_i|D_i = 1, M_i = 1] - \frac{1}{\rho}\right\}\Bigg) \Pr(D_i = 0|M_i = 1) \\
&\quad + \rho\ \mathbb{E}[Y_i|D_i = 0, M_i = 1]\ (1 - \Pr(D_i = 0|M_i = 1)),
\end{aligned}$$

where $\hat{\Delta} = \overline{Y_i|D_i = 1, M_i = 1} - \overline{Y_i|D_i = 0, M_i = 1}$, and

$$\text{ATT}^Y_{M=1} = \mathbb{E}[\hat{\Delta}] + \rho\ \mathbb{E}[Y_i|D_i = 0, M_i = 1].$$

Importantly, every term in these expressions is observable, meaning it can be readily estimated from data on detainments, except for ρ. Fortunately, as Knox *et al.* (2020) show, plausible estimates of ρ can be obtained with other techniques, including "outcome tests" for discrimination, discussed further below.

We note that some scholars have identified innovative ways to avoid this form of sample selection bias. For example, West (2018) analyzes police enforcement at traffic accidents on the premise that, conditional on location, the decision by officers to respond to accident scenes does not vary systematically with civilian race. However, such approaches remain rare, and necessarily limit the scope of analyses to very particular aspects of policing.

Outcome Tests

In recognition of the problems that arise from severe selection issues in these convenience samples, an influential line of research in economics has developed the analytic strategy of outcome tests. The basic logic is to compare how often officers find evidence of a crime ($Y_i = 1$) when stopping civilians ($M_i = 1$) of each racial group, a statistic known as the "hit rate." Per Becker (1957), if officers discriminate against minorities by making some stops due to unjustified racial suspicion or for harassment, this will result in lower proportion of stops that turn up evidence. In other words, officers will find weapons or drugs in a smaller proportion of minority stops, $\overline{Y_i|D_i = 1, M_i = 1}$, than in white stops, $\overline{Y_i|D_i = 0, M_i = 1}$. This insight allows analysts to detect the observable implications of discrimination, even when only data on stops are available.

However, a number of issues complicate the interpretation of outcome tests, which are typically framed as hypothesis tests that either reject or fail to reject the null hypothesis of no discrimination. Because we have no information on the underlying features of police encounters at large, and we do not know the precise process that accounts for stopping decisions, the results of outcome tests do not necessarily indicate the presence of discrimination in detainment. For example, police may be unbiased in their decision to stop civilians, choosing to stop any civilian regardless of race if there is at least a 10% chance they are carrying contraband. However, if one group of civilians carries contraband more often, the test would still wrongly indicate discrimination (Simoiu *et al.*, 2017). This problem is often referred to as "infra-marginality" (Ayres, 2002). In essence, this problem relates to unobserved heterogeneity in encounters and requires that analysts using outcome tests invoke Assumption 4, treatment ignorability, in order to make the apples-to-apples comparisons necessary to infer racial discrimination.[6]

Rather than appeal explicitly to a treatment ignorability assumption, prior work has invoked highly restrictive assumptions on officer and human behavior to sidestep this issue. For example, in a study of traffic stops, Knowles *et al.* (2001) assume that "all motorists of a given race, if they are ever searched, will carry contraband with equal probability regardless of their other characteristics that may be observed by the police," and that all officers "have the same racial prejudice against minority motorists," (Anwar and Fang, 2006, p. 129). Hernández-Murillo and Knowles (2004) extend this model, similarly assuming that civilians are perfectly rational and know precisely the probability that they will be searched by officers; they develop a typology of felons and present bounds based on the idea that (under these implausible assumptions) "type-1

[6] As in other tests, if there exists *observed* heterogeneity or confounding, researchers can account for it by appropriate adjustment, either with parametric assumptions as in Simoiu *et al.* (2017) or by conducting analyses within subgroups.

felons do not carry, whereas type-2 felons carry contraband for sure" as a result of the felon's cost-benefit analyses (p. 964). Anwar and Fang (2006) relax the assumptions in Knowles *et al.* (2001) with a formal model that allows officers to use information they gather during traffic stops when determining the likelihood a motorist is carrying contraband, and they draw on richer data that includes officer race. However, such models still rely on particular information structures that dictate how officers and troopers simultaneously assess each other's behavior. For example, Anwar and Fang (2006) assume that drivers of a given race provide officers with the same information during traffic stops regardless of the officer's race (p. 15), which would be violated if drivers were more likely to cooperate with co-racial officers. This model also assumes that "the pools of motorists faced by troopers of different races are the same," (p. 16) which is unlikely to hold given the fact that officers are often assigned to patrol areas with high shares of co-racial, residents (Ba *et al.*, 2020). More recent updates to the outcome test literature similarly impose parametric models that place artificial structure on police–civilian interactions, such as the assumption that officers' perception of a civilian's guilt can be represented by a random draw from the arbitrarily chosen beta distribution (Simoiu *et al.*, 2017). To varying extents, all of these approaches fill in missing information about the process of police–civilian encounters by untestable assertion.

Furthermore, even stipulating to the assumptions in these approaches, analysts using these frameworks can only conclude from a positive outcome test that the officer "has a taste for discrimination," or that his or her utility function "exhibits a preference for searching motorists of one race" (Anwar and Fang, 2006, p. 134) but sheds no light on the magnitude of the problem. Practically speaking, it is highly unclear what this means about the severity of racial bias or the number of civilians impacted.

We show that using outcome tests in the context of a careful causal analysis can reveal far *more* information about discrimination than previously thought, while invoking *less* stringent assumptions. Knox *et al.* (2020) prove that under the nonparametric (and somewhat weaker) assumptions 1, 2 and 4 in "Identifying Assumptions in Policing Research" — mandatory reporting, mediator monotonicity, and treatment ignorability — the outcome test reveals partial information about the proportion of minority stops (those for which $D_i = 1$ and $M_i = M_i(1) = 1$) that are racially motivated ($M_i(0) = 0$, indicating that a white civilian would not have been stopped in the same circumstances). Specifically, we show that

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

where the numerator in the right-hand side represents the observed difference in "hit rates," and the denominator is the hit rate in white stops. Importantly, this result bears on a concrete, policy-relevant causal quantity: the decrease in detainment that would counterfactually result if white civilians were substituted into the circumstances where minority stops occurred, which constitutes strong evidence of racial discrimination. In addition, the potential outcomes framework allows treatment effects to vary across units (Rubin, 1974), negating the need to invoke assumptions about homogeneous behavior or effects.

In "Approaches Incorporating Officer Race," we next turn to extensions of the outcome test and other approaches that relax some of these assumptions by using additional information on officer race.

Approaches Incorporating Officer Race

The approaches described in "Review of Prominent Approaches" average over any heterogeneity in officers. This is not necessarily a problem if the goal is to estimate average causal effects. However, police records sometimes contain information on individual officers, including their race, which can be used to construct richer tests of racial discrimination. The basic logic of these tests is that if white and minority officers are both unbiased in their treatment of civilians, then many aspects of their observed behavior should be identical. To the extent their behavior differs under similar circumstances, analysts have evidence of bias in at least one officer group, if not both. Many methods based on this insight can be thought of as enriched variants of the four basic approaches we outline above; these extensions allow analysts to draw inferences in difficult data environments, but often entail substantial tradeoffs. For example, Anwar and Fang (2006) develop an extension of the outcome test that examines whether white and minority officers have similar hit rates in their searches; their variant is robust to the infra-marginality problem discussed in "Outcome Tests." This approach, and others like it, is valuable because they can be applied in settings where alternatives break down. However, as we discuss below, analysts must proceed with caution because officer-race-based methods require their own untestable assumptions, are sensitive to unobserved data, and must be interpreted carefully.

Here, we present one particular method of incorporating information into officer race, the *proportion test*, to shed light on this broad class of research designs. The proportion test is closely related to the benchmark approach, but rather than comparing the civilian-race proportions in stops with a proxy population, it compares civilian-race proportions in minority-officer stops ($X_i = 1$) with the corresponding proportions in white-officer stops ($X_i = 0$).

Specifically, the test statistic is

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

In Appendix B, we show that the proportion test is closely related to the *difference in differences*, or $\text{ATE}^M_{X=1} - \text{ATE}^M_{X=0}$. However, without additional data or highly implausible assumptions, analysts cannot directly interpret the results of the proportion test in terms of this quantity.[7] Instead, the proportion test can at best offer a limited test of the null hypothesis of no racial discrimination. To see why, suppose that one group of officers discriminates by stopping half of minority civilians, but only one quarter of white civilians in otherwise identical encounters. If another group of officers also discriminates, but does so in a way that is exactly proportional (making stops in all and half of these encounters, respectively), then this comparison will find no difference in the composition of their stops. Thus, the proportion test is blind to the possibility that some officers are more aggressive in their stopping decisions *across the board*, but it can detect whether this severity is unfairly allocated at least some of the time. Intuitively, this is because the only way for both groups of officers to be simultaneously unbiased is if they are identical in their stopping behavior. Even here, the proportion test cannot tell which group of officers is biased — analysts cannot distinguish between the competing possibilities that (i) white officers discriminate against one group of civilians. or (ii) that minority officers discriminate against the other. The proportion test can show only that there is bias somewhere in the system.

The validity of this test hinges critically on the assumption that the two groups of officer face common circumstances (i.e. a "common pool" of civilian encounters). However, unlike every other test we examine, this test does not require that the analyst assumes treatment ignorability (Assumption 4), that white and nonwhite civilians are otherwise identical after accounting for observed characteristics. While the common pool assumption is less stringent than the treatment ignorability assumption, the common pool assumption cannot be sidestepped by merely controlling for circumstances of observed stops, as Johnson *et al.* (2019) incorrectly asserts. This is because the common pool assumption applies to *all* encounters an officer faces — whether or not they result in detainment and thus appear in administrative data. Moreover, even comparing white and minority officers within the same department is insufficient to ensure a common pool. It is well known that minority and white officers are deployed very differently:

[7] Specifically, we would need to assume that (1) white and minority officers stop civilians at equal rates, and (2) both groups of officers stop white civilians at equal rates. Even with these restrictive assumptions, the proportion test would only recover a quantity *proportional* to the difference-in-differences.

in Chicago, for example, officers are often assigned to patrol precincts with large co-racial populations (Ba *et al.*, 2020), and minority officers may receive less desirable assignment shifts due to lack of seniority or internal discrimination. Thus, the common pool assumption is almost never plausible except when conditioning on fine-grained assignment records or when studying plausibly as-if random assignment of officers to encounters (e.g., West, 2018).

Checklist for the Study of Racial Bias

To bring this frayed literature under a common analytic umbrella — a necessary step for the verification of assertions and accumulation of knowledge — we offer a set of guidelines for scholars designing future empirical studies of racial bias. These guidelines are widely applicable to all causal analyses, but given how infrequently they are heeded in the study of racially biased policing, they are worth enumerating here. The steps we list help ensure that empirical claims are on firm footing, and they clarify for both analysts and readers important concepts like the target quantity and the chances that a given study's approach can recover it. They can be distilled to a single recommendation: imagine the ideal experiment.

1. **Define the unit of analysis**

This crucial first step is often overlooked or performed without proper care, perhaps because analysts assume this component is obvious. For example, if one has data on police stops of pedestrians, then it may be assumed that the appropriate unit of analysis is the police stop. But this is only one of several choices — others might imagine stopped civilians, or police–civilian encounters (including non-stops), as alternative units of analysis. The unit of analysis is not a foregone conclusion determined by the data at hand. It is a conceptual choice that emanates from the desired hypothetical experiment. And depending on which unit the analysts chooses, the counterfactual under consideration may not be estimable; it may not even be well defined. This leads to the second item on the checklist:

2. **Define the counterfactual and causal estimand**

Based on the unit of analysis, the analyst must clearly define the counterfactual of interest. Again, the contrast between the study of civilians and the study of police–civilian encounters is instructive. When studying the civilian, the counterfactual of interest may be the manipulation of a given individual's race while holding all else constant. But given the presence of institutional racism, changing a person's race may necessitate changing other characteristics, like their level of income and education, complicating the causal exercise (Holland, 1986). On the other hand, if the unit of analysis is the encounter, we can imagine conducting an experiment in which we randomly assign individuals with otherwise similar observable traits to encounters. As a rule of thumb, the easier it is to imagine conducting the experiment, the less likely it is that the counterfactual manipulation suffers from conceptual issues.

Precisely defining the counterfactual of interest then suggests a number of possible causal estimands, including the average treatment effect in the population, the same quantity in some subset of encounters, or some other quantity. Defining this quantity formally allows the analyst and interested readers to more readily assess whether the analytic strategy is credible.

3. Visualize the causal process of interest

With the unit of analysis and counterfactual in hand, we recommend visualizing the causal process of interest in a directed acyclic graph (DAG), as in Figure 2. Even if done at a very high level, e.g., without naming every possible confounding variable, visualizing the causal process is a useful check. This graph implicitly encodes a number of assumptions about the data-generating process, which can be stated formally (as described in the next step) and helps the analyst to identify which sets of relationships should and should not be conditioned on in order to recover causal quantities of interest. For example, if the analyst is targeting the average treatment effect in the population, and had a representative sample of all police–civilian encounters (whether they resulted in detainment or not), Figure 2 makes clear that they need only condition on confounding factors that jointly cause treatment and outcome to remove selection bias. In contrast, the figure also shows that if the analyst conditions on detainment (M_i), he or she will also allow all common causes of detainment and the outcome to contaminate his or her comparisons, due to collider bias. Ultimately, to determine which of these sets of potential confounders must be adjusted for the analyst must take the next step, and state identifying assumptions.

4. State identifying assumptions

The final step prior to estimation is the explicit statement of the assumptions necessary to identify the estimand in data without bias, similar to our enumeration of Assumptions 1–4 above. This step answers the following question: given the nature of the data, what would have to be true about the world to interpret an empirical result as evidence of discrimination? This step can be difficult if the analyst is not using an established estimation approach, as it requires deriving the statistical bias of the chosen estimator. However, by making identifying assumptions explicit, other scholars engaging with the work will benefit from the clarity of exposition and have a common framework for scrutinizing a study's plausibility. With all of the building blocks of a causal analysis transparently stated, not obscured to the reader, deficiencies can be more easily identified and corrected. Under these conditions, knowledge aggregation is likely to accelerate.

Moving beyond Data on Detainments

In the absence of comprehensive data on police behavior, previous scholars devised a series of inventive but seemingly incompatible approaches to study racially biased policing with the information at hand. By nesting this multidisciplinary inquiry in a common statistical framework and taking advantage of previously unavailable avenues for observing police–civilian interactions, social science has a rare opportunity to meaningfully inform policy decisions that have implications for public safety, trust in government, and the democratic ideal of equal protection under the law.

An additional advantage of the proposed causal framework is that it rigorously clarifies what additional data we would need to make progress on the study of racial bias in policing. As our analysis makes plain, a central limitation of policing data is that it only includes information on encounters in which police take some action, such as a stop, arrest, or use of force. As we show above, if officers are racially biased in decisions to detain civilians, it is difficult to use such data to obtain precise and valid estimates of racial bias in police behavior. To obtain improved estimates of racial bias, scholars must devise ways to collect data on all types of police encounters, including ones in which civilians find themselves in the presence of police, but do not interact with them. This additional data collection would obviate the need to only analyze detainment data, negating the sample selection bias present in so many analyses today. The benefits of these additional data are not limited to analyses which assume away bias in detainment, but spill over to studies employing alternative approaches as well. While detailed data on the features of such encounters would be ideal, even basic information on the frequency of police–civilian encounters across racial groups would facilitate a range of sensitivity analyses, allowing scholars to judge the quality of extant evidence in this vast literature.

In some settings, such as traffic enforcement, such data are already being collected, though it has remained largely untouched by researchers. Passive highway cameras regularly collect images of passing cars regardless of whether drivers are stopped by police. In other situations, public-access cameras record all pedestrians in specific locations. The adoption of police body-worn cameras in many local agencies presents another potential stream of data.[8] Scholars

[8] This approach is less ideal, since procedures governing when officers are required to record vary widely across jurisdictions.

must begin developing ways to harness these data, ideally in collaboration with police agencies, in order to gain purchase on the at-present unknown volume of police encounters across racial and ethic groups. Armed with a random sample of all police encounters in a given setting, researchers could greatly simplify their inferential task. However, in situations in which such extensive data collection presents ethical concerns, bounding approaches like the one outlined above present a tractable alternative.

Regardless of the strategy employed, it is clear that scholars in this literature must adopt a common mathematical dialect and share goals for the study of racial bias in police–civilian encounters. Given the scattershot availability of police administrative data, failure to coalesce on a common analytic framework is unsurprising. But the result is a litany of seemingly disconnected tests that have hindered knowledge accumulation and in some cases produced dangerously misleading results. At present, it is too often ambiguous whether a given analysis satisfies the conditions necessary for rigorous causal inference — often, even the specific objective of the analysis is left undefined. In the few situations where assumptions are laid bare, they are often implausible, and the associated tests often cannot speak to the magnitude of the problem of racial bias. In short, it is virtually impossible at present to reconcile conflicting results across this wide range of approaches, a situation we hope this essay will alleviate.

Scholars across the social sciences have focused their gaze on police–civilian interactions to an unprecedented degree. But for this effort to do the most good, researchers must adopt a common and rigorous analytic approach. By nesting analyses in a general causal framework, evaluating and synthesizing research on this topic will finally be feasible, and scholars can begin to amass credible estimates of inequity in the behavior of the coercive arm of the state.

References

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

Anwar, S. and H. Fang. 2006. "An Alternative Test of Racial Prejudice in Motor Vehicle Searches: Theory and Evidence". *The Review of Economic Studies* 96(1): 127–51.

Arrow, K. J. 1972. "Models of Job Discrimination". In: *Racial Discrimination in Economic Life.* Ed. A. Pascal. D.C. Heath, 83–102.

Arrow, K. J. 1998. "What Has Economics to Say about Racial Discrimination?" *Journal of Economic Perspectives* 12(2): 91–100.

Ayres, I. 2002. "Outcome Tests of Racial Disparities in Police Practices". *Justice Research and Policy* 4(1–2): 131–42.

Ba, B., D. Knox, J. Mummolo, and R. Rivera. 2020. “Diversity in Policing: The Role of Officer Race and Gender in Police-Civilian Interactions in Chicago”. *Working Paper*. https://scholar.princeton.edu/sites/default/files/jmummolo/files/bkmr_diversitypolicingchicago.pdf.

Baumgartner, F. R., D. A. Epp, and K. Shoub. 2018. *Suspect Citizens: What 20 Million Traffic Stops Tell us about Policing and Race*. Cambridge University Press.

Becker, G. 1957. *The Economics of Discrimination*. University of Chicago Press.

Bui, Q. and A. Cox. 2016. “Surprising New Evidence Shows Bias in Police Use of Force but Not in Shootings”. *The New York Times*. https://www.nytimes.com/2016/07/12/upshot/surprising-new-evidence-shows-bias-in-police-use-of-force-but-not-in-shootings.html.

Burghart, D. 2020. “Fatal Encounters”. Database.

Eberhardt, J., P. A. Goff, V. J. Purdie, and P. G. Davies. 2004. “Seeing Black: Race, Crime, and Visual Processing”. *Journal of Personality and Social Psychology* 87(6): 876–93.

Edwards, F., H. Lee, and M. Esposito. 2019. “Risk of Being Killed by Police Use of Force in the United States by Age, Race–Ethnicity, and Sex”. *Proceedings of the National Academy of Sciences* 116(34): 16793–8. https://www.pnas.org/content/early/2019/07/30/1821204116.

Elwert, F. and C. Winship. 2014. “Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable”. *The Annual Review of Sociology* 40: 31–53.

Fisher, M. and P. Hermann. 2015. “Did the McKinney, Texas, Police Officer Know He Was Being Recorded?” *The Washington Post*.

Frangakis, C. E. and D. B. Rubin. 2002. “Principal Stratification in Causal Inference”. *Biometrics* 58(1): 21–9.

Fryer, R. G. 2019. “An Empirical Analysis of Racial Differences in Police Use of Force”. *Journal of Political Economy* 127(3): 1210–61.

Gelman, A., J. Fagan, and A. 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): 813–23.

Goff, P. A. and K. B. Kahn. 2012. “Racial Bias in Policing: Why We Know Less Than We Should”. *Social Issues and Policy Review* 6(1): 177–210.

Greiner, J. D. and D. B. Rubin. 2011. “Causal Effects of Perceived Immutable Characteristics”. *Review of Economics and Statistics* 93(4): 775–85.

Grogger, J. and G. Ridgeway. 2006. “Testing for Racial Profiling in Traffic Stops from behind a Veil of Darkness”. *Journal of the American Statistical Association* 101(475): 878–87.

Heckman, J. 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, M. A. 2016. “Does Water Kill? A Call for Less Casual Inferences”. *Annals of Epidemiology* 26(10): 674–80.

Hernández-Murillo, R. and J. Knowles. 2004. “Racial Profiling or Racist Policing? Bounds Tests in Aggregate Data”. *International Economic Review* 45(3): 959–89.

Holland, P. W. 1986. “Statistics and Causal Inference”. *Journal of the American Statistical Association* 81(396): 945–60.

Imai, K., L. Keele, D. Tingley, and T. Yamamoto. 2011. “Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies”. *American Political Science Review* 105(4): 765–89.

Iyengar, S. 1994. *Is Anyone Responsible?: How Television Frames Political Issues*. University of Chicago Press.

Johnson, D. J., T. Tress, N. Burkel, C. Taylor, and J. Cesario. 2019. “Officer Characteristics and Racial Disparities in Fatal Officer-Involved Shootings”. *Proceedings of the National Academy of Sciences* 116(32): 15877–82. Published online ahead of print July 22, 2019. https://www.pnas.org/content/early/2019/07/16/1903856116.

Kahn, K. B., P. A. Goff, J. K. Lee, and D. Motamed. 2016. “Protecting Whiteness: White Phenotypic Racial Stereotypicality Reduces Police Use of Force”. *Social Psychological and Personality Science* 7(5): 403–11.

Knowles, J., N. Perisco, and P. Todd. 2001. “Racial Bias in Motor Vehicle Searches: Theory and Evidence”. *Journal of Political Economy* 109(1): 203–29.

Knox, D., W. Lowe, and J. Mummolo. 2020. “Administrative Records Mask Racially Biased Policing”. *American Political Science Review*. https://www.cambridge.org/core/journals/american-political-science-review/article/administrative-records-mask-racially-biased-policing/66BC0F9998543868BB20F241796B79B8.

Knox, D. and J. Mummolo. 2020. “Making Inferences about Racial Disparities in Police Violence”. *Proceedings of the National Academy of Sciences* 117: 1261–2.

Legewie, J. 2015. “Racial Profiling and Use of Force in Police Stops: How Local Events Trigger Periods of Increased Discrimination”. *American Journal of Sociology*. Conditionally accepted, http://jlegewie.com/files/Legewie-2016-Racial-Profiling-and-Use-of-Force-in-Police-Stops.pdf.

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

Mac Donald, H. 2019. “Testimony of Heather Mac Donald, Fellow, Manhattan Institute, before the Committee on the Judiciary of the United States House of Representatives”. Oversight Hearing on Policing Practices.

Menifield, C. E., G. Shin, and L. Strother. 2019. “Do White Law Enforcement Officers Target Minority Suspects?” *Public Administration Review* 79(1): 56–68. https://onlinelibrary.wiley.com/doi/abs/10.1111/puar.12956.

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

Nix, J., B. A. Campbell, E. H. Byers, and G. 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): 309–40.

NYPD. 2017. “Use of Force Report, 2017”. *Tech. rep.* https://www1.nyc.gov/assets/nypd/downloads/pdf/use-of-force/use-of-force-2017.pdf.

Orfield, G., P. Marin, and C. L. Horn. 2005. *Higher Education and the Color Line: College Access, Racial Equity, and Social Change*. Harvard Education Press.

Ousey, G. C. and M. R. Lee. 2008. “Racial Disparity in Formal Social Control: An Investigation of Alternative Explanations of Arrest Rate Inequality”. *Journal of Research in Crime and Delinquency* 45(3): 322–55.

Pager, D. and H. Shepherd. 2008. “The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets”. *Annual Review Sociology* 34(6): 181–209. https://www.nytimes.com/2016/07/12/upshot/surprising-new-evidence-shows-bias-in-police-use-of-force-but-not-in-shootings.html.

Pearl, J. 1993. “Graphical Models, Causality and Intervention”. *Statistical Science* 8(3): 266–9.

Pearl, J. 2001. “Direct and Indirect Effects”. *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*: 411–20.

Pearl, J. 2009. *Causality*. Cambridge University Press.

Pearl, J. 2018. “Does Obesity Shorten Life? Or Is It the Soda? On Non-manipulable Causes”. *Journal of Causal Inference* 6(2): 1–7.

Phelps, E. S. 1972. “The Statistical Theory of Racism and Sexism”. *American Economic Review* 62(1): 659–61.

Ridgeway, G. 2016. “Officer Risk Factors Associated with Police Shootings: A Matched Case-Control Study”. *Statistics and Public Policy* 3(1): 1–6.

Ridgeway, G. and J. MacDonald. 2010. “Race, Ethnicity, and Policing: New and Essential Readings”. In: ed. S. K. Rice and M. D. White. NYU Press Chapter Methods for Assessing Racially Biased Policing.

Rogelj, J., P. M. Forster, E. Kriegler, C. J. Smith, and R. Séférian. 2019. “Estimating and Tracking the Remaining Carbon Budget for Stringent Climate Targets”. *Nature* 571(7765): 335–42. https://doi.org/10.1038/s41586-019-1368-z.

Rosenbaum, P. 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): 656–66.

Ross, C. T. 2018. “A Multi-level Bayesian Analysis of Racial Bias in Police Shootings at the County-Level in the United States, 2011–2014”. *PLoS One* 10(11): e0141854.

Rubin, D. B. 1974. “Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies”. *Journal of Educational Psychology* 66(5): 688–701.

Rubin, D. B. 1990. “Formal Mode of Statistical Inference for Causal Effects”. *Journal of Statistical Planning and Inference* 25(3): 279–92.

Sen, M. and O. Wasow. 2016. “Race as a Bundle of Sticks: Designs That Estimate Effects of Seemingly Immutable Characteristics”. *Annual Review of Political Science* 19: 499–522.

Simoiu, C., S. Corbett-Davies, and S. Goel. 2017. “The Problem of Inframarginality in Outcome Tests for Discrimination”. *The Annals of Applied Statistics* 11(3): 1193–216. https://arxiv.org/abs/1706.05678.

Soss, J. and V. Weaver. 2017. “Police Are Our Government: Politics, Political Science, and the Policing of Race–Class Subjugated Communities”. *Annual Review of Political Science* 20: 565–91.

Streeter, S. 2019. “Lethal Force in Black and White: Assessing Racial Disparities in the Circumstances of Police Killings”. *The Journal of Politics* 81(3): 1124–32.

West, J. 2018. “Racial Bias in Police Investigations”. Working Paper https://people.ucsc.edu/~jwest1/articles/West_RacialBiasPolice.pdf.

Wheeler, A. P., S. W. Phillips, J. L. Worrall, and S. A. Bishopp. 2017. “What Factors Influence an Officer’s Decision to Shoot? The Promise and Limitations of Using Public Data”. *Justice Research and Policy* 18(1): 48–76. https://doi.org/10.1177/1525107118759900.

Williams, D. R. and R. Wyatt. 2015. “Racial Bias in Health Care and Health: Challenges and Opportunities”. *Journal of the American Medical Association* 314(6): 555–6. https://www.nytimes.com/2016/07/12/upshot/surprising-new-evidence-shows-bias-in-police-use-of-force-but-not-in-shootings.html.

Worrall, J. L., S. A. Bishopp, S. C. Zinser, A. P. Wheeler, and S. W. Phillips. 2018. “Exploring Bias in Police Shooting Decisions with Real Shoot/Don’t Shoot Cases”. *Crime & Delinquency* 64(9): 1171–92. https://doi.org/10.1177/0011128718756038.

Toward a General Causal Framework for the Study of Racial Bias in Policing

Online Appendix

Contents

A Outcome Selection

Assuming ignorability of civilian race,

$$
\begin{align}
\text{ATE} &= \mathbb{E}[Y_i(1, M_i(1)) - Y_i(0, M_i(0))] \nonumber \\
&= \Pr(Y_i = 1|D_i = 1) - \Pr(Y_i = 1|D_i = 0) \nonumber \\
&= \frac{\Pr(D_i = 1|Y_i = 1)\Pr(Y_i = 1)}{\Pr(D_i = 1)} - \frac{\Pr(D_i = 0|Y_i = 1)\Pr(Y_i = 1)}{\Pr(D_i = 0)} \nonumber \\
&= 1 - \frac{\Pr(D_i = 1|Y_i = 0)\Pr(Y_i = 0)}{\Pr(D_i = 1)} - 1 + \frac{\Pr(D_i = 0|Y_i = 0)\Pr(Y_i = 0)}{\Pr(D_i = 0)} \nonumber \\
&= -\frac{\Pr(D_i = 1|Y_i = 0)\Pr(Y_i = 0)}{\Pr(D_i = 1)} + \frac{\Pr(D_i = 0|Y_i = 0)\Pr(Y_i = 0)}{\Pr(D_i = 0)} \tag{1} \\
&= -\frac{\Pr(D_i = 1|Y_i = 0)\Pr(Y_i = 0)}{\Pr(D_i = 1|Y_i = 0)\Pr(Y_i = 0) + \Pr(D_i = 1|Y_i = 1)\Pr(Y_i = 1)} \nonumber \\
&\quad + \frac{\Pr(D_i = 0|Y_i = 0)\Pr(Y_i = 0)}{\Pr(D_i = 0|Y_i = 0)\Pr(Y_i = 0) + \Pr(D_i = 0|Y_i = 1)\Pr(Y_i = 1)} \tag{2}
\end{align}
$$

In studies that select on the outcome, analysts typically have no information about $\Pr(Y_i = 1)$, how frequently officers engage in the behavior of interest (e.g., what proportion of encounters result in a shooting). Rather, analysts only have data to estimate $\Pr(D_i = 1|Y_i = 1)$ and $\Pr(D_i = 0|Y_i = 1)$. In this case, bounds on the ATE follow by substituting extreme values for the missing information, $\Pr(Y_i = y)$ and $\Pr(D_i = d|Y_i = 0)$.

For example, one extreme possibility is as follows: almost all encounters are unobserved non-shootings ($\Pr(Y_i = 0)$ approaching one and $\Pr(Y_i = 1)$ approaching zero), and all of these non-shooting encounters are with white civilians ($\Pr(D_i = 1|Y_i = 0) = 0$, meaning $\Pr(D_i = 0|Y_i = 0) = 1$). In this scenario—which analysts cannot rule out using the available data—the racial bias in shootings approaches ATE = +1, the highest possible value. Similarly, analysts cannot rule out the reverse, that all of the unobserved non-shooting encounters are with *minority* civilians, so that $\Pr(D_i = 1|Y_i = 0) = 0$ and $\Pr(D_i = 0|Y_i = 0) = 1$, which would imply the ATE approaches −1. Thus, despite having data on all shootings, researchers know nothing more about the quantity of interest than they did before beginning the study—and any conclusions to the contrary are based entirely on assumptions that the data cannot support.

If the shooting rate, $\Pr(Y_i = 1)$, is known, then these bounds can be narrowed somewhat. In this case, plugging in extreme values for $\Pr(D_i = 0|Y_i = 0)$ and $\Pr(D_i = 1 = 0)$ in Equation 2 reveals that the range of possible bias is

$$
-\frac{\Pr(Y_i = 0)}{\Pr(Y_i = 0) + \Pr(D_i = 1|Y_i = 1)\Pr(Y_i = 1)} \leq \text{ATE} \leq \frac{\Pr(Y_i = 0)}{\Pr(Y_i = 0) + \Pr(D_i = 0|Y_i = 1)\Pr(Y_i = 1)}
$$

B Proportion Test

In the proportion test, analysts compare the stops made by minority officers to the stops made by white officers. In particular, analysts compare the proportion of each officer group's stops that are of minority civilians, as opposed to white civilians. The basic logic of this approach is to assess whether both officer groups take the same actions (e.g., stopping civilians) when facing identical pools of civilian behavior. This is a necessary, but not sufficient, condition for both groups to be unbiased: if the two groups of officers behave differently, then at least one must be biased in some direction. However, the converse is not true: if both groups behave identically, it could be that both are equally biased. Thus, the proportion test offers an asymmetric test of officer bias.

This test proceeds by estimating

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

which can be rewritten as

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

invoking the ignorability of civilian race,

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

and finally, the common pool assumption gives $\Pr(D_i = d|X_i = 0) = \Pr(D_i = d|X_i = 1) = \Pr(D_i = d)$, yielding

$$= \frac{\mathbb{E}[M_i(1) - M_i(0)|X_i = 1]\Pr(D_i = 1) - \mathbb{E}[M_i(0)|X_i = 1]\big(\Pr(D_i = 0) - \Pr(D_i = 1)\big)}{\Pr(M_i = 1|X_i = 1)} - \frac{\mathbb{E}[M_i(1) - M_i(0)|X_i = 0]\Pr(D_i = 1) - \mathbb{E}[M_i(0)|X_i = 0]\big(\Pr(D_i = 0) - \Pr(D_i = 1)\big)}{\Pr(M_i = 1|X_i = 0)}.$$

This shows that the desired quantity of interest, the difference in differences ($\mathbb{E}[M_i(1) - M_i(0)|X_i = 1] - \mathbb{E}[M_i(1) - M_i(0)|X_i = 0]$), is not identified by the proportion test: a number of additional assumptions are required to connect the two. Specifically, to draw inferences about the difference in differences, analysts must first assume that overall stopping rates are equal across officer groups, or that $\Pr(M_i = 1|X_i = 0, D_i = d) = \Pr(M_i = 1|X_i = 1, D_i = d)$. In other words, one officer group cannot patrol more actively or be more stringent in enforcement; among other things, this ensures that the denominators are comparable. Then, analysts would need to further assume that both groups treat white civilians equally, so that the second and fourth terms cancel, and the remaining terms contain the desired difference in differences (multiplied by a scaling factor). To be clear, we do not advocate these highly implausible assumptions. Rather, we enumerate them to convey the difficulty in directly interpreting the results of the proportion test in terms of a substantively useful quantity of interest. However, as we discuss in the main text, the proportion test remains a useful test that can reject the null hypothesis that no officer group is biased, and it offers analysts the ability to examine this question when no other statistical test is applicable.