Can Racial Bias in Policing Be Credibly Estimated Using Data Contaminated by Post-Treatment Selection?

Dean Knox, Will Lowe and Jonathan Mummolo*

September 11, 2020

Abstract

Studies of racial bias in policing often rely on data contaminated by selection issues, e.g. using records of stops or arrests—which may themselves be a product of racial bias—to estimate discrimination in subsequent actions like use of force. This feature raises the threat of post-treatment-selection bias, which recent work shows can lead to severe underestimates of discrimination. However, prominent studies continue to ignore this issue, employing standard regression techniques with contaminated data. In this paper, we formally analyze the key identifying assumption undergirding these studies, "subset ignorability," and show it corresponds to the measure-zero set of knife-edge conditions in which differing biases happen to sum to zero. Because there is no substantive reason to believe such accidental cancellation would occur, we conclude this approach is not reliable in applied research, and we emphasize the need for continued caution and increased rigor in high-stakes analyses of discriminatory policing with contaminated data.

Word Count: 7,945

*Dean Knox is an Assistant Professor of Operations, Information, and Decisions at the Wharton School of the University of Pennsylvania, dcknox@upenn.edu. Will Lowe is Senior Research Scientist at the Hertie School of Governance, lowe@hertie-school.org. Jonathan Mummolo is an Assistant Professor of Politics and Public Affairs at Princeton University, jmummolo@princeton.edu.

1 Introduction

Since Heckman's (1977) Nobel-winning work, over four decades of causal-inference research has grappled with the challenge of drawing rigorous conclusions from data contaminated by non-random selection (Rosenbaum, 1984; Greenland, 2014; Elwert and Winship, 2014). Recently, Knox, Lowe and Mummolo (2020) shows how selection bias also contaminates estimates of racial discrimination by police when analyzing records of detainments (e.g. stops, arrests). These police administrative datasets select on officers' post-treatment decisions to detain civilians—decisions that are potentially also discriminatory—thus omitting all data on encounters not resulting in detainments and potentially severely understating the extent of racial bias in policing. Heckman and Durlauf (forthcoming) further note that analyzing only encounters involving detainments is "a classic route to selection bias" (p. 2), and Fryer's (2019) "failure to model interactions between police and civilians as a process," including discrimination in detainment, means that "differences in conditional probabilities for black and white outcomes are not dispositive of discrimination" (pp. 3, 4–5).

Despite these selection issues, decades of research has traditionally used standard regression approaches when estimating racial bias with contaminated records of police-civilian interactions, comparing police behavior across recorded detainments of minority and white civilians (e.g. Smith et al., 1984; Lundman, 1994, 1996; Engel, Sobol and Worden, 2000; Spano, 2003; Novak and Frank, 2005; Schafer et al., 2006; Tillyer and Engel, 2013; Fryer, 2019). While most studies leave the identifying assumptions undergirding these techniques unstated—making it difficult to judge the validity of analyses—a recent paper, Gaebler, Cai, Basse, Shroff, Goel, and Hill (2020), formalizes this identification strategy, allowing for a rigorous evaluation of a workhorse technique. The paper develops new statistical theory aimed at "clarifying the statistical foundations of discrimination analysis" (p. 4)—e.g. by

"estimat[ing] discrimination... based only on data describing those who were arrested" (p. 7). Specifically, Gaebler et al. (2020) formalizes the "often unstated assumptions in studies of discrimination," by stating a theoretical condition, "subset ignorability," that, if credible, would justify standard regression approaches in this setting—even when treatment is not as-if randomly assigned.[1] These arguments, and the vast body of applied work using traditional regression techniques, take a very different stand from methodological work "emphasiz[ing] the difficulties in achieving identification of bias in the presence of differences in the race-specific distributions of unobserved variables" (Heckman and Durlauf, p. 4; referring to Heckman and Siegelman, 1993 and Heckman, 1998). In contrast to this pessimistic view, Gaebler et al. (2020) offers statistical theory arguing that under subset ignorability "a primary quantity of interest in discrimination studies is nonparametrically identifiable" (abstract) and as a result, "in observational studies of of discrimination, concerns about post-treatment bias may be misplaced" (p. 23). In other words, the paper suggests that analysts employing common regression approaches can recover unbiased estimates despite two complicating factors: (i) unobserved baseline differences in the minority and white encounters observed by police, or omitted variable bias; and (ii) the fact that officers may apply different standards for detaining minority and white civilians, or post-treatment selection.

Empirical studies of police violence thus routinely reject the methodological points of Knox, Lowe and Mummolo (2020) and Heckman and Durlauf (forthcoming). However, the recent formal statement of the subset ignorability assumption offers an opportunity to definitively adjudicate these disagreements. If credible, subset ignorability would simultaneously undermine decades of methodological research on the challenges of analyzing post-treatment-selected data, while salvaging decades of empirical research on discrimination that employs standard regression techniques. It therefore merits close investigation. What does this

[1]Specifically, Gaebler et al. (2020) observe that treatment ignorability is difficult to defend, because "there is little reason to think that arrest potential outcomes... would be independent of an individual's race"' (p. 20).

proposal entail? What arguments must be weighed and found compelling if readers of discrimination research—not only researchers, but also civil rights organizations and federal judges—are to be informed consumers?

On close examination, we find that subset ignorability is satisfied *if and only if* the real-world data-generating process happens—even with imperfect controls—to be in the measure-zero set of knife-edge scenarios in which disparate sources of statistical bias happen to sum to precisely zero. More specifically, the approach requires researchers to assume omitted variable bias and post-treatment selection bias perfectly offset one another. In discussing such knife-edge scenarios, Robins et al. (2003) states, "Intuitively, it seems 'unlikely' ... [to have] parameters cancelling each other" (p. 496). Indeed, causal inference textbooks like Spirtes, Glymour and Scheines (1993) often dismiss such "accidents of parameter values," as "rarely occur[ing] in contemporary practice" (p. 53). In *Causality: Models, Reasoning and Inference*, Pearl (2000) says these cases are effectively the same as "see[ing] a picture of a chair" and arguing that it may actually be "two chairs positioned such that one hides the other" (pp. 81–82). In Propositions 1–3, we show formally that subset ignorability implicitly relies on such accidental cancellation, then provide examples of the hyper-specific assumptions that analysts would need to articulate to defend the use of this assumption.

These results reveal that the implicit assumptions of many discrimination studies are far less plausible than standard ignorability assumptions about groups being comparable *given as-if-random assignment of treatment*. In contrast to as-if-random assumptions, subset ignorability only holds if groups are comparable *despite responding differently to treatment*. In the context of police-civilian encounters, even if one could somehow ethically randomly assign civilians of different races to encounter police, subset ignorability amounts to assuming that race is forgotten and then *re*-randomized *after* officers decide to stop civilians because of their race.[2] Critically, subset ignorability is an assertion about the world that cannot be

[2] As Gaebler et al. (2020) states, "we imagine that the perception of race is counterfactually determined after the first-stage decision but before the second-stage decision" (p. 7).

guaranteed *even by gold-standard experimental designs* that randomize actors into police-civilian encounters. In Section 3, we show that even in such ideal settings, the traditional approach essentially assumes away the core problem of post-treatment selection: that if officers are racially biased in their decisions to stop civilians, then minority and white observations in stop data will be fundamentally incomparable. Specifically, given racial bias in stopping, observed encounters will consist of three different groupings (principal strata, Frangakis and Rubin, 2002): circumstances in which officers would stop (i) only minority civilians, e.g. jaywalking; (ii) all civilians, e.g. assault; and (iii) only white civilians, if such cases even exist. (These groups are akin to "compliers," "always takers," and "defiers" in instrumental variables analysis.) Stops of minority civilians will therefore consist of a "jaywalking-assault" mixture, while white civilians will consist of a mixture of "assault" and anti-white stops (whatever these may be). Nevertheless, subset ignorability requires potential outcomes across these groups to exactly balance. And if there are no anti-white stops—a wholly plausible scenario—then subset ignorability is guaranteed to be false *unless officers are equally violent in "jaywalking" and "assault" encounters* (i.e., have identical average potential outcomes across strata). Put another way, the subset ignorability assumption is analogous to assuming that the *complier* average treatment effect, the quantity identified by instrumental variable estimators, is identical to the *full sample* average treatment effect—a position that has been widely rejected by causal inference scholars since Angrist, Imbens and Rubin (1996).

In sum, using detainment records to estimate racial bias in police violence is fraught, because racial bias can affect the decision to detain civilians. Specifically, if there is bias in detainment (e.g. stops and arrests), the white detainments will differ from the nonwhite detainments in unobserved ways, even if they were perfectly comparable at the start of encounters. For example, minority civilians may be arrested for less serious offenses than white civilians, though often these differences—which affect whether police use force—are not indicated in police records. Because of this, the assumption underlying the traditional

approach, subset ignorability, amounts to assuming that various sources of statistical bias happen to exactly offset one another. But since there is no reason to think that accidental cancellation would occur, analysts should instead use techniques to describe the range of possible discrimination (Knox, Lowe and Mummolo, 2020). Ignoring these selection issues risks severely understating the degree of racial bias in policing. Careful research designs, using quasi-experimental scenarios that *justify* assumptions and mitigate sources of statistical bias using expert knowledge and case selection (e.g. West, 2018) offer a second alternative for making reliable inferences. We caution that consumers of high-stakes discrimination research must carefully probe the reliability of work that relies on accidental-cancellation claims. The prioritization of expediency over rigor threatens to damage the credibility of discrimination research at a time when scientific evidence is critically important for reform.

In the remainder of this paper, we first formally define notation and outline concepts for the study of racial bias in Section 2. Section 3 then presents a detailed analysis of subset ignorability, deriving its logical implications and clarifying its applicability to applied research. We conclude by reiterating the need for caution and increased rigor in the study of racial bias using police administrative records.

2 The Causal Problem

We consider the data-generating process in the directed acyclic graph (DAG) in Figure 1. This causal model is general, and applicable to a range of previous studies which use administrative data on police detainments to estimate bias in a subsequent decision, such as the decision to issue a citation, search a vehicle, or the use force—the case we examine in this paper.[3] The units of analysis, indexed by i, are i.i.d. police-civilian encounters (e.g. sightings of a civilian by an officer). Analysts may seek to estimate various average effects

[3]In a stylized example, Gaebler et al. (2020) considers a "two decider" setting in which distinct actors (an arresting officer and a prosecutor) engage in potentially racially biased behavior at different points in time. However, the distinction from single-decider settings,

of the presence of minority civilians in encounters (relative to white civilians), denoted as $D_i = 1$ ($D_i = 0$), on the use or non-use of force, $Y_i \in \{0, 1\}$. Specifically, analysts may estimate the difference in the probability of force that would result from the counterfactual substitution of a different individual with a different racial identity into the encounter, while holding objective context—e.g. location, time of day, criminal activity—fixed (Knox, Lowe and Mummolo, 2020).

This counterfactual is critical to conceptualizing a feasible causal exercise. The choice of police-civilian encounters as the unit of analysis avoids well-known issues regarding nonmanipulable, characteristics; thus, the "ideal experiment" does not entail the difficult-to-imagine manipulation of an individual's race, but rather the substitution of comparable actors into pre-existing scenes.[4] This approach does not seek to estimate the influence of larger systemic factors that contribute to biased outcomes, such as housing discrimination; rather, it seeks to comprehensively evaluate racial bias *during the entire police-civilian encounter*.

As Figure 1 shows, race may affect force through two broad channels: (i) indirectly, via racially biased detainment, $M_i \in \{0, 1\}$; or (ii) directly, via racial bias in post-stop events.[5] Crucially, there almost surely exist unobserved confounders, U_i, such as an officer's level of suspicion or mood, that jointly cause stopping and force decisions, but do not appear in police administrative data. Conditioning on detainment, M_i, results in confounding from U_i by opening a back-door path (Pearl, 1993), creating collider bias (Elwert and Winship, 2014).

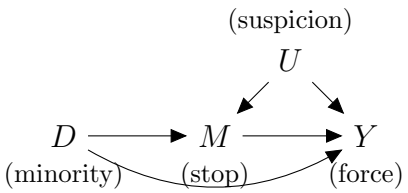
such as the multi-stage process of police stops, is described as being of little importance: "regardless of whether one imagines there are two deciders or a single one, our formal statistical results hold unaltered" (p. 6).

[4]Of course, observational analyses will fail to approximate this ideal experiment if minority- and white-civilian encounters are not comparable on unobserved, pre-treatment characteristics.

[5]For clarity, we sometimes denote observations with treatment status $D_i = 1$ as "minority-civilian encounter" or simply "minority," and those with $D_i = 0$ as "white." Similarly, we refer to $M_i = 1$ observations with "stop" and $M_i = 0$ as "non-stop."

Analyzing only encounters involving detainment is therefore "a classic route to selection bias" (Heckman and Durlauf, forthcoming, p. 2).

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



In the potential outcomes framework (Rubin, 1974), there exist counterfactual states, given a civilian's race, d, of both detainment, $M_i(d)$, and force, $Y_i(d, M_i(d))$. Further, given dichotomous mediator and treatment, encounters each belong to one of four "principal strata" (Frangakis and Rubin, 2002): latent classifications of units based on their counterfactual profiles. These groups, outlined in Figure 2, are: (i) "always stop" encounters with $M_i(1) = M_i(0) = 1$, e.g. encounters with civilians committing assault, an instance where police might theoretically detain civilians regardless of race; (ii) "anti-minority" racial stops, $M_i(1) = 1$ and $M_i(0) = 0$, e.g. encounters with jaywalkers, where minority civilians would be detained but not otherwise similar white civilians; (iii) the somewhat implausible "anti-white" racial stops, $M_i(1) = 0$ but $M_i(0) = 1$; and (iv) "never stop" encounters, $M_i(1) = M_i(0) = 0$, inconspicuous events that never result in detainment. Importantly, these conceptual groups exist even after conditioning on observed pre-treatment features of encounters, X_i. Because the severity of civilian behavior differs dramatically across strata, it strains credulity to say that officers will use violence in the same way across, e.g. "jaywalking" and "assault" type encounters.

Acknowledging the existence of these principal strata illustrates the core challenge with making inferences from post-treatment-selected data: given racial bias in stopping ($D \rightarrow M$), minority detainment records will contain some unknown mix of always stops and anti-*minority* stops, whereas white records will be a non-comparable unknown mix of always stops

Figure 2: **Principal Strata in Police-Civilian Encounters.** The figure displays the four principal strata that comprise police-civilian encounters based on how potential detainment decisions, $M_i(d)$, depend on whether the civilian is a racial minority, D_i.

		Stop if white? $(D_i = 0)$	
		Yes, $M_i(0) = 1$	No, $M_i(0) = 0$
Stop if minority? $(D_i = 1)$	Yes, $M_i(1) = 1$	always stop (e.g. assault)	anti-minority stop (e.g. jaywalking)
	No, $M_i(1) = 0$	anti-white stop (?)	never stop (inconspicuous)

and, to the extent they exist, anti-*white* stops. In practice, this means that even if analysts achieved perfect *pre-detainment* covariate balance, comparisons of post-stop encounters will still be distorted by *post-detainment* non-comparability, absent further assumptions.

Because police administrative data often only capture events that occur post-detainment, analysts may seek to estimate the controlled direct effect among stops,

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

which captures the influence of civilian race *assuming that detainments occur.*[6] Gaebler et al. (2020) asserts that under subset ignorability, the $\text{CDE}_{M=1}$ is nonparametrically identifiable—an advance that, if credible, would justify decades of empirical policing research. Below, we clarify the conditions necessary for this assumption to hold.

[6]We note that this quantity considers an impossible counterfactual for some unknown portion of police encounters—how often force would be used against civilians if officers were forced to stop them—even though, given their principal stratum and hypothetical treatment status, *they would never actually be detained.*

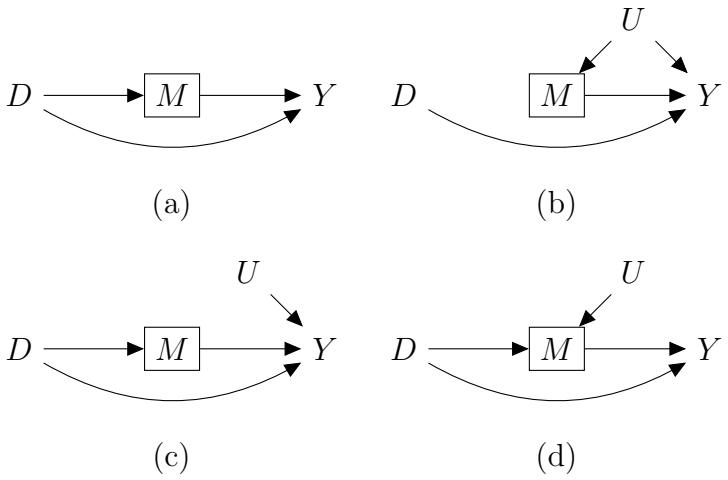
2.1 Additional Assumptions that Guarantee Subset Ignorability

The causal process outlined in Figure 1 represents a general and plausible description of police stops. It allows civilian race to affect the use of force both directly and indirectly, and it allows for unobserved common causes of detainment and force. However, analysts may invoke several assumptions that, if credible, would greatly simplify the task of causal identification. Before analyzing the validity of subset ignorability in the more general case, we first discuss these alternative assumptions, visualized in Figure 3.

Figure 3 depicts scenarios in which subset ignorability is automatically satisfied: if there is either no racial discrimination in detainment (i.e. no $D \rightarrow M$) or no unobserved common causes of detainment and force (i.e. no $U \rightarrow M$, no $U \rightarrow Y$, or both). However, these scenarios are facially implausible in this setting. Panel (a) assumes that there are no unrecorded subjective factors (U, e.g. mood or suspicion) in any officer decisions (either M or Y). Panel (b) assumes no discrimination in detainment—a difficult-to-justify assumption in studies that aim to assess discrimination. And panels (c) and (d) require assuming that while unobserved factors U exist, they do not *jointly* affect mediator (M) and outcome (Y). Because both M and Y are decisions by the same officer, analysts rarely have a substantive basis for this assumption.

Despite their facial implausibility in the police-violence context, these hypothetical data generating processes (DGPs) are helpful for clarifying specific statements about the world that analysts might embrace when invoking subset ignorability. These theoretical justifications may also be useful when considering discrimination in different contexts. However, given our focus on violence during police-civilian encounters, we focus in the remainder of the paper on the conditions under which subset ignorability can be satisfied given the more general DGP displayed in Figure 1.

Figure 3: **Policing causal structures satisfying subset ignorability.** Panels (a–d) present additional assumptions that analysts might make about the directed acyclic graph in Figure 1, describing the relationship between civilian race (D), selection into police data (M, e.g. stops or arrests), subsequent police behaviors (Y, e.g. use of force), and certain unobserved factors (U). Panel (a) assumes away unobserved factors, like officer suspicion, that influence both detainment decisions and use-of-force decisions. This is highly implausible because police records typically do not report such subjective aspects of encounters. Panel (b) allows for unobserved officer suspicion, but assumes that officers are not racially biased in their stopping decisions. However, we find it inadvisable to assume away some aspects of police discrimination in studies intended to analyze other aspects of police discrimination. The remaining panels allow for unreported subjective factors, but assume that such factors will not jointly influence selection into the data and the outcome of interest—an assumption that will often be impossible to justify when analyzing observational data on police-civilian encounters, because both events are decisions made by the same officer.



3 A Formal Analysis of Subset Ignorability

We begin by formally analyzing the argument that the $\text{CDE}_{M=1}$ can be estimated without bias, after selecting on detainment, as long as subset ignorability (Definition 1, below) holds. This claim was formalized in Gaebler et al. (2020) but is implicit in numerous prior studies of racially biased policing that rely on detainment records (e.g. Smith et al., 1984; Lundman, 1994, 1996; Engel, Sobol and Worden, 2000; Spano, 2003; Novak and Frank, 2005; Schafer et al., 2006; Tillyer and Engel, 2013; Fryer, 2019). To evaluate it, we examine the implied relationships that analysts must believe about the world—and justify to readers and policymakers—before invoking this assumption in applied discrimination research.

We begin our formal analysis by first considering a best-case scenario: when treatment ignorability holds at the start of police encounters. This would be satisfied in an experimental setting, where otherwise comparable white and nonwhite civilians were randomly assigned to police encounters, or if observed covariates were sufficiently rich to render treatment as-if random. Even here, discrimination in detainment will still contaminate data received by analysts, but concerns over baseline differences in encounters, at least, can be ruled out. However, even in this ideal case, we find that the subset ignorability assumption is effectively acknowledging *selection*, but assuming away selection *bias*. Specifically, Proposition 1 shows that subset ignorability can be satisfied *if and only if* an extraordinarily difficult knife-edge balancing condition holds: that while officers may stop minority and white civilians in different circumstances due to discrimination (e.g. stopping one group for as little as jaywalking, but another only for crimes as serious as assault), minority and white stops are nonetheless *exactly comparable* in terms of the potential for officer violence.

Because treatment ignorability may well be violated, we then turn to the general case: when analysts must also grapple with baseline differences in encounters due to omitted variables. Gaebler et al. (2020) argue that despite this methodological challenge, the subset ignorability assumption offers a path forward. Using this proposed approach, "a primary quantity of interest in discrimination studies is nonparametrically identifiable" (abstract)

and as a result, "in observational studies of of discrimination, concerns about post-treatment bias may be misplaced" (p. 23). These theoretical arguments are provocative: in contrast, past work has "emphasize[d] the difficulties in achieving identification of [racial] bias in the presence of differences in the race-specific distributions of unobserved variables" (Heckman and Durlauf, p. 4; referring to Heckman and Siegelman, 1993 and Heckman, 1998).

How can subset ignorability solve these well-known issues? In Proposition 2, we formally analyze the proposed method in full generality. We show that under confounding, subset ignorability will hold *if and only if* an even more specific and difficult-to-satisfy knife-edge assumption is true. In Proposition 3, we go a step further, proving that unless post-treatment bias is precisely equal in magnitude and opposite in sign to omitted variable bias, subset ignorability is guaranteed to be false. As a long line of causal inference scholars have noted (see Section 3.2), such knife-edge accidental cancellation cannot be credibly assumed to hold in applied research using real-world data.

3.1 In Ideal Experiments, Subset Ignorability Holds *iff* Cross-principal-strata Knife-edge Balancing Holds

We now state the subset ignorability assumption (Gaebler et al., 2020). The remainder of this section examines it in an idealized experimental setting. For brevity, we implicitly condition on pre-treatment covariates, X_i, here and throughout.

Definition. *Subset ignorability assumption.*

$$Y_i(d, 1) \perp\!\!\!\perp D_i \mid M_i = 1$$

In ideal experimental conditions, invoking subset ignorability means assuming that despite the fact that analysts *selected* on detainments ($M_i = 1$), this selection does not induce selection *bias*. We make one conceptual observation and one formal observation about this "no-selection-bias" assumption. Conceptually, analysts often fail to distinguish between (i) *assuming* a condition holds, which is easy; and (ii) *satisfying* a condition and carefully

justifying it, which is hard. And formally, despite appearing to be a simple statement about the ignorability of civilian race, this no-selection-bias assumption is in fact an extraordinarily strong requirement about the relationship between potential police force across principal strata—in "assault" type always stops, "jaywalking" type anti-minority stops, and (if these exist) anti-white stops—*groups which cannot be fully distinguished by the analyst.* This relationship is given in Proposition 1.

Proposition 1. *With treatment ignorability, the subset ignorability assumption is satisfied if and only if the following knife-edge equality holds:*

$$\begin{aligned}
&\mathbb{E}[Y_i(d,\text{stop}) \mid \text{always stop}] \, \frac{\Pr(\text{always stop})}{\Pr(\text{always stop}) + \Pr(\text{anti-min. stop})} \\
&\quad + \mathbb{E}[Y_i(d,\text{stop}) \mid \text{anti-min. stop}] \, \frac{\Pr(\text{anti-min. stop})}{\Pr(\text{always stop}) + \Pr(\text{anti-min. stop})} \qquad (1) \\
&\qquad = \\
&\mathbb{E}[Y_i(d,\text{stop}) \mid \text{always stop}] \, \frac{\Pr(\text{always stop})}{\Pr(\text{always stop}) + \Pr(\text{anti-white stop})} \\
&\quad + \mathbb{E}[Y_i(d,\text{stop}) \mid \text{anti-white stop}] \, \frac{\Pr(\text{anti-white stop})}{\Pr(\text{always stop}) + \Pr(\text{anti-white stop})} \qquad (2)
\end{aligned}$$

Discussion. The left-hand side of Proposition 1, expression (1), corresponds to the unknown composition of observed minority stops, a "jaywalking-assault" mixture in unknown proportions. The right-hand side, (2), refers to the composition of observed white stops, an unknown mixture of "assault" and anti-white stops (whatever those may be). This shows that, at its core, the no-selection-bias assumption requires perfect balancing (in the frequency-weighted average of potential outcomes) of three fundamentally different types of encounters: "assault," "jaywalking," and (if they exist) anti-white stops. Perturbations in either (i) potential force rates or (ii) strata proportions would cause the assumption to fail.

Figure 4 displays three hypothetical scenarios where both sets of numeric values are precisely tailored to satisfy the Proposition 1 knife-edge balancing condition. For example, panel (c) considers the plausible case where there are no anti-white stops. In this setting,

rearranging terms in Proposition 1 reveals that subset ignorability requires that officers be *equally violent in "assaults" and "jaywalking" encounters* (i.e., have the same average potential outcomes).

To convey concepts with a more specific illustration, panel (b) depicts a world in which $\frac{2}{7}$ of potential detainments are always-stop "assaults," $\frac{4}{7}$ are anti-minority "jaywalking" encounters in which only minority civilians would be detained, and $\frac{1}{7}$ are anti-white encounters (whatever those may be). Thus, the probability fractions in the left-hand side of Proposition 1 (minority stops) are $\frac{2/7}{2/7+4/7} = \frac{1}{3}$ (non-discriminatory) and $\frac{4/7}{2/7+4/7} = \frac{2}{3}$ (discriminatory), respectively; the right-hand-side fractions (white stops) are $\frac{2/7}{2/7+1/7} = \frac{2}{3}$ (non-discriminatory) and $\frac{1/7}{2/7+1/7} = \frac{1}{3}$ (discriminatory). In this case, Proposition 1 holds *if and only if* the "leniency" of officer force in anti-minority stops, defined as $\text{leniency}_{\text{minority}} = \mathbb{E}[Y_i(d,1)|\text{always stop}] - \mathbb{E}[Y_i(d,1)|\text{anti-min. stop}]$, is exactly one half of $\text{leniency}_{\text{white}} = \mathbb{E}[Y_i(d,1)|\text{always stop}] - \mathbb{E}[Y_i(d,1)|\text{anti-white stop}]$.[7]

In Figure 4, to find cases where subset ignorability was not violated, we carefully hand-tuned principal strata sizes and potential force rates until the just-so condition of Proposition 1 was satisfied. Thus, in these unlikely scenarios, selection bias happens to sum to zero. But recall that the analyst has no direct knowledge of, much less control over, precise values for any of these parameters. Critically, even gold-standard experimental designs that randomize treatment at the start of police encounters cannot ensure this knife-edge relationship will hold: standard ignorability assumptions merely require groups to be comparable given as-if random treatment assignment, whereas here, groups must remain comparable *despite responding to treatment differently.* Moreover, because the frequencies of occurrence and the average potential force are almost always different across principal strata, this condition is

[7] Plugging in the above probability fractions, Proposition 1 reduces to $\mathbb{E}[Y_i(d,1)|\text{always stop}] \cdot \frac{1}{3} + \mathbb{E}[Y_i(d,1)|\text{anti-min. stop}] \cdot \frac{2}{3} = \mathbb{E}[Y_i(d,1)|\text{always stop}] \cdot \frac{2}{3} + \mathbb{E}[Y_i(d,1)|\text{anti-white stop}] \cdot \frac{1}{3}$. Subtracting $\mathbb{E}[Y_i(d,1)|\text{always stop}]$ from both sides yields $\text{leniency}_{\text{minority}} \cdot \frac{2}{3} = \text{leniency}_{\text{white}} \cdot \frac{1}{3}$.

almost never satisfied, in a measure-theoretic sense. Thus, knife-edge balancing is essentially a blind hope the analyst expresses about the world.

To examine the impact of post-treatment selection bias in a more general way, Figure 5 examines a fuller range of conditions. These scenarios extend the case of Figure 4(b) while varying key parameters to illustrate the delicacy of the subset ignorability assumption.[8] The two panels display the infinitesimally narrow surface on which subset ignorability holds and shows how post-treatment selection bias varies as a function of two parameters: the ratio of anti-minority to anti-white stops, and the ratio of force across the same two principal strata. The top panel shows that as we depart from the measure-zero set of conditions in which subset ignorability holds—the white curve—analysts employing standard regression approaches will either over- or underestimate racial bias in policing, dubiously inferring anti-white bias in some cases. The steep slopes on the surface displayed in the bottom panel show that the magnitude of this statistical bias grows rapidly as we depart from the conditions necessary to satisfy subset ignorability.

Proof. A detailed derivation is given in Appendix A. Using the definition $M_i = M_i(D_i)$ and treatment ignorability, it is easy to see that the no-selection-bias assumption implies ($\iff$)

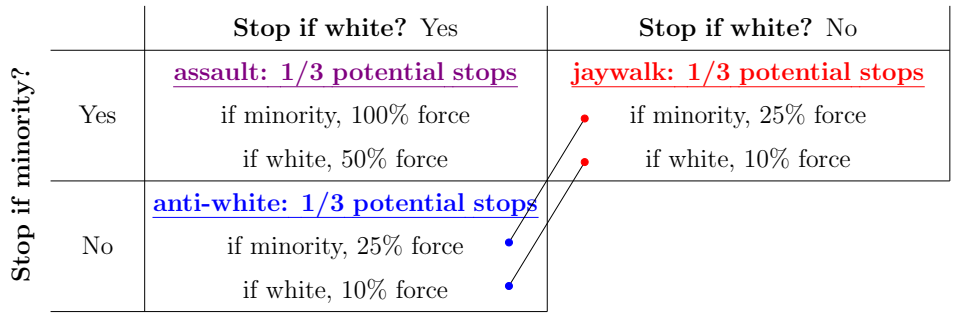
[8] In these hypothetical scenarios, following Figure 4(b), always-stop encounters represent $\frac{2}{7}$ of all potential detainments—i.e. encounters in which either $M_i(0) = 1$ or $M_i(1) = 1$. We also set $\mathbb{E}[Y(1,1)|\text{always stop}] = 1$, $\mathbb{E}[Y(0,1)|\text{always stop}] = 0.7$, $\mathbb{E}[Y(1,1)|\text{anti-white stop}] = 0.5$, and $\mathbb{E}[Y(0,1)|\text{anti-white stop}] = 0.35$, again following Figure 4(b). These key parameters are (i) relative sizes of the anti-minority stop and anti-white stop strata, $\frac{\Pr(\text{anti-min. stop})}{\Pr(\text{anti-white stop})}$; and (ii) relative force rates in these strata, $\frac{\mathbb{E}[Y_i(d,1)|\text{anti-min. stop}]}{\mathbb{E}[Y_i(d,1)|\text{anti-white stop}]}$. Thus, Figure 5 contains the unbiased Figure 4(b) scenario as a special case, but generalizes it to show how bias rapidly increases in magnitude—even in this idealized setting—as either parameter is varied. The patterns depicted in Figure 5 are not specific to this setting. To demonstrate this, we provide code allowing analysts to input any combination of parameters that they find reasonable, then examine the bias that results from any deviation from subset ignorability.

$$
\begin{aligned}
& Y_i(d,1) \perp\!\!\!\perp D_i \mid M_i(D_i) = 1 \\
\iff & \mathbb{E}[Y_i(d,1) \mid \ M_i(D_i) = 1] \ = \ \mathbb{E}[Y_i(d,1) \mid D_i = 0, \ M_i(D_i) = 1] \\
\iff & \mathbb{E}[Y_i(d,1) \mid \ M_i(1) = 1] \ = \ \mathbb{E}[Y_i(d,1) \mid D_i = 0, \ M_i(0) = 1] \\
\iff & \mathbb{E}[Y_i(d,1) \mid M_i(1) = 1] \ = \ \mathbb{E}[Y_i(d,1) \mid M_i(0) = 1] \\
\iff & \mathbb{E}\left[Y_i(d,1) \mid (M_i(0) = 1 \wedge M_i(1) = 1) \vee (M_i(0) = 0 \wedge M_i(1) = 1)\right] \\
& \qquad = \ \mathbb{E}\left[Y_i(d,1) \mid (M_i(0) = 1 \wedge M_i(1) = 1) \vee (M_i(0) = 1 \wedge M_i(1) = 0)\right] \\
\iff & \mathbb{E}[Y_i(d,\text{stop}) \mid \text{always stop OR anti-min. stop}] \\
& \qquad = \ \mathbb{E}[Y_i(d,\text{stop}) \mid \text{always stop OR anti-white stop}],
\end{aligned}
$$

where $\wedge$ ($\vee$) denotes "and" ("or"), and the equivalence between independence and equal expectations is due to binary Y_i. Proposition 1 follows immediately. □

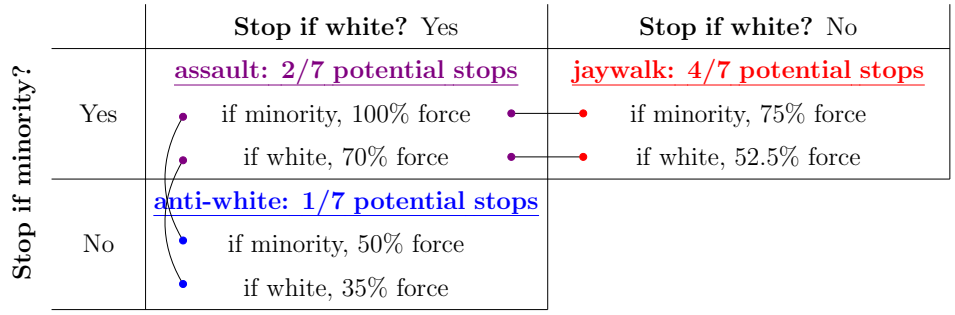
Figure 4: **What would it take for subset ignorability to hold in experiments? Three hypothetical scenarios.** Each panel presents a hypothetical composition of police stops. Subset ignorability is true *if and only if* the described knife-edge condition holds between all cells connected by lines. The first line in each cell gives Pr(strata | $M_i(0) = 1$ or $M_i(1) = 1$); the second and third give $\mathbb{E}[\, Y(d, 1) \mid \text{strata} \,]$.

(a)



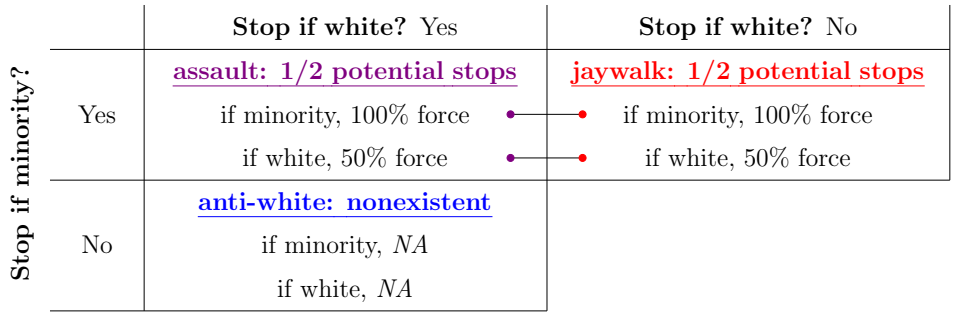
Scenario: All potential detainments are $\frac{1}{3}$ **assaults**, $\frac{1}{3}$ **jaywalking**, $\frac{1}{3}$ **anti-white**
$\Rightarrow$ minority stops are $\frac{1}{2}$ **assaults**, $\frac{1}{2}$ **jaywalking**; white are $\frac{1}{2}$ **assaults**, $\frac{1}{2}$ **anti-white**
To satisfy subset ignorability: requires *exact equality between jaywalking and anti-white encounters* (whatever those may be) in terms of potential officer force.

(b)



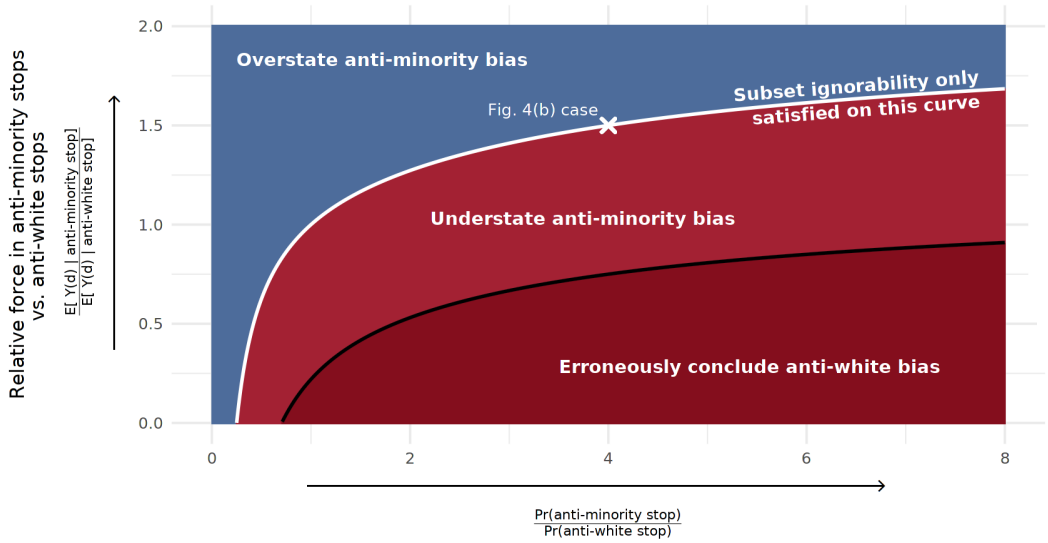
Scenario: All potential detainments are $\frac{2}{7}$ **assaults**, $\frac{4}{7}$ **jaywalking**, $\frac{1}{7}$ **anti-white**
$\Rightarrow$ minority stops are $\frac{1}{3}$ **assaults**, $\frac{2}{3}$ **jaywalking**; white are $\frac{2}{3}$ **assaults**, $\frac{1}{3}$ **anti-white**
To satisfy subset ignorability: requires the difference between assaults and anti-white encounters (whatever those may be), in terms of potential force, to be *exactly double the difference between assault and jaywalking.*

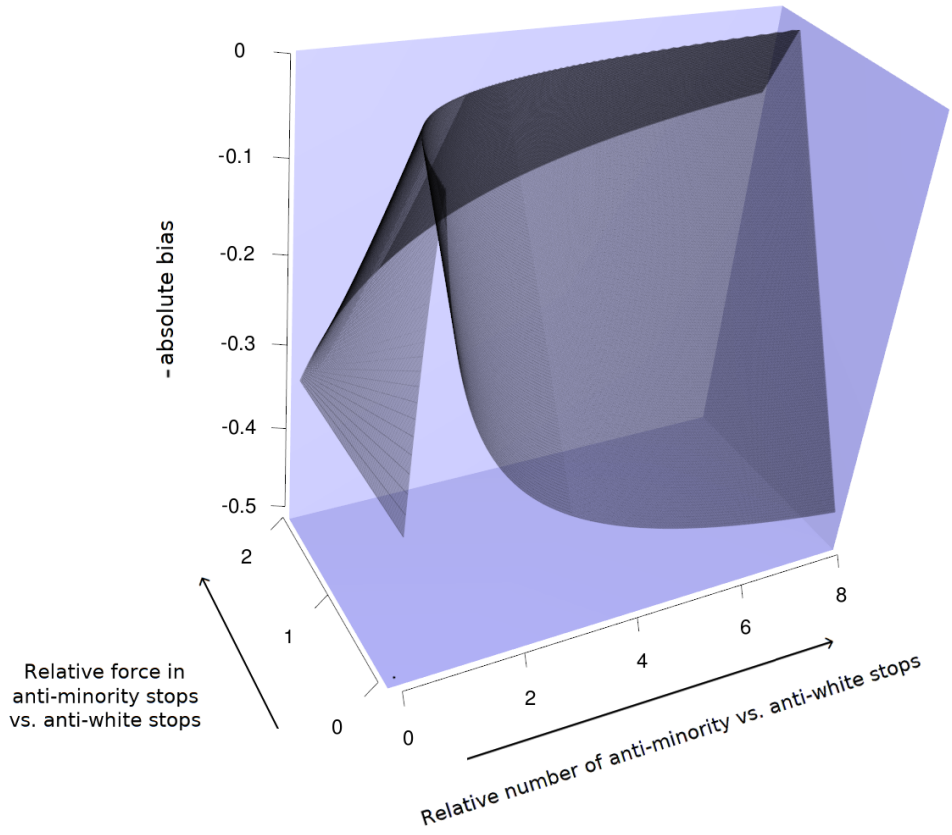
(c)



Scenario: All potential detainments are $\frac{1}{2}$ **assaults**, $\frac{1}{2}$ **jaywalking**
$\Rightarrow$ minority stops are $\frac{1}{2}$ **assaults**, $\frac{1}{2}$ **jaywalking**; white are **all assaults**
To satisfy subset ignorability: requires that there is *absolutely no difference between assaults and jaywalking* in terms of potential officer force.

Figure 5: **Precise balancing is required to achieve accidental bias cancellation.** The figures display hypothetical scenarios illustrating the implicit assumption that analysts make when using naïve approaches to estimate the $\text{CDE}_{M=1}$—that the relative sizes and relative force rates of anti-minority and anti-white stops will exactly balance, per Proposition 1. In both panels, the x axes represent relative size: the ratio of the numbers of anti-minority stops, versus anti-white stops. The y axes represent relative force: the ratio of potential force rates in anti-minority stops, versus anti-white stops. In the top panel, the white curve indicates the narrow slice of scenarios in which subset ignorability would be satisfied and the naïve approach is unbiased. In the lower red regions, the naïve approach is negatively biased, underestimating anti-minority discrimination. The dark red region indicates when this bias is so strong that it leads to a sign error (i.e., on average, analysts erroneously conclude that police are biased against *white* civilians). In the lower panel, the z axis depicts how post-treatment selection bias rapidly grows in magnitude as conditions depart from the subset ignorability knife-edge requirement.





3.2 A Note on Accidental Cancellation in Nonparametric Causal Inference

The knife-edge condition of Proposition 1 (and the condition of Proposition 2, below) is a particularly egregious case of what causal inference scholars refer to as "unfaithfulness"—the notion that in *any* model space, there will always exist an infinitesimally small sliver of just-so data-generating processes that happen to possess "extra independence relationships" (Robins et al., 2003, p. 493) above and beyond those conveyed by the DAG. In their causal inference textbook, Spirtes, Glymour and Scheines (1993) note, "...the Faithfulness Condition can be thought of as the assumption that conditional independence relations are due to causal structure rather [than] to accidents of parameter values" (p. 9). It is typically taken for granted that general nonparametric statements about ranges (e.g. about possible omitted variable bias in the example below) refer to the broad behavior of faithful distributions, with the clear understanding that degenerate unfaithful distributions (often, edge cases and boundaries) can take on specific values within that range.[9]

To understand the nature of accidental cancellation in a more familiar setting, consider the following illustration, extending an example by Robins et al. (2003). Suppose that a true data-generating process has two unobserved confounders, $Z_i^{(1)} = \varepsilon_i^{(Z1)}$ and $Z_i^{(2)} = \varepsilon_i^{(Z2)}$; a treatment $X_i = \alpha_{(1)}Z_i^{(1)} + \alpha_{(2)}Z_i^{(2)} + \varepsilon_i^{(X)}$; an outcome $Y_i = \beta X_i + \gamma_{(1)}Z_i^{(1)} + \gamma_{(2)}Z_i^{(2)} + \varepsilon_i^{(Y)}$; and all i.i.d. errors $\varepsilon_i^{(*)} \sim \mathcal{N}(0,1)$. In these circumstances, a typical causal inference

[9]For example, Knox, Lowe and Mummolo (2020) state at one point that "bias is weakly negative" (Appendix p. 6) for the $\text{CDE}_{M=1}$ under some assumptions. In this case, the statement refers to a broad region in the model subspace defined by the relevant assumptions. "Weakly negative" (i.e., nonpositive) is a statement about the range of the estimator's bias for all data-generating processes in that range, and the term "weakly" is a technical caveat meaning that for specific unfaithful edge cases in this subspace, the bias may in fact be exactly zero.

scholar might first assert that to eliminate omitted variable bias, it is necessary to rule out unobserved confounders $Z_i^{(1)}$ and $Z_i^{(2)}$. The scholar would then state the exact form of the omitted variable bias that would result if these confounders were not addressed either through design or statistical adjustment: $\frac{\gamma_{(1)}\alpha_{(1)}}{\alpha_{(1)}^2+\alpha_{(2)}^2+1} + \frac{\gamma_{(2)}\alpha_{(2)}}{\alpha_{(1)}^2+\alpha_{(2)}^2+1}$. However, mapped to this setting, subset ignorability would hold that the analyst need *not* control for these omitted variables, but instead can assume that the bias induced by one perfectly offsets the bias induced by the other, i.e. that $\gamma_{(1)}\alpha_{(1)} = -\gamma_{(2)}\alpha_{(2)}$. In Appendix A, we demonstrate a step-by-step equivalence between this line of argumentation and the argument for subset ignorability.

Such contrived scenarios, in which statistical bias exists but happens to conveniently cancel itself out, have been dismissed by leading causal inference scholars for decades because they are of little practical use. As Robins et al. (2003) states, "Intuitively, it seems 'unlikely'... [to have] parameters cancelling each other" (p. 496); the premise that analysts will not generally be so fortunate "is implicit in a variety of statistical practices" (p. 494). The reason it seems unlikely is because it is well known that these "accidents," or "cancelling" data-generating processes, have Lebesgue measure zero in the model space (Spirtes, Glymour and Scheines, 1993; Meek, 1995). In other words, the probability that nature draws such a convenient data-generating process from any smooth distribution over possible models is *zero*.

Other scholars have noted that unfaithful edge cases for broader nonparametric results (i) require little effort to produce and (ii) are not particularly helpful in an applied sense. For example, Spirtes, Glymour and Scheines (1993) remarks, "While it is easy enough to construct models that violate... Faithfulness, such models rarely occur in contemporary practice, and when they do, the fact that they have properties that are consequences of unfaithfulness is taken as an objection to them" (p. 53); "Faithfulness... turns out to be the 'normal' relation between probability distributions and causal structures" (p. 56). This is why, in "An Introduction to Causal Inference," Scheines (1997) observes that "assuming faithfulness... is widely embraced by practicing scientists," though "nevertheless, critics continue to create

unfaithful cases and display them" (p. 10).

3.3 In Confounded Settings, Subset Ignorability Holds *only if* Selection Bias Exactly Cancels Omitted Variable Bias

We now turn to the more general case, when treatment ignorability is violated and there exist baseline, pre-detainment differences between minority and white encounters. This issue, as Gaebler et al. (2020) notes, is commonplace: "there is little reason to think that arrest potential outcomes... would be independent of an individual's race" (p. 20) and thus, "treatment ignorability... is unlikely to hold in our setting for the same reason" (p. 21).

Estimating causal effects in this confounded setting is widely seen as more challenging. For example, the entirety of Heckman (1998) revolves around the difficulties posed by "unobserved characteristics for each race" for detecting discrimination. Knox, Lowe and Mummolo (2020) warns "Our aim... is not to assert the plausibility of treatment ignorability, but rather to clarify that deep problems remain even if this well-known issue is somehow solved" (p. 626). Yet, Gaebler et al. (2020) nonetheless asserts that in spite of confounding *and* post-treatment selection, the proposed approach allows analysts to estimate causal effects without bias. They write, "critically, such information about the first stage," discrimination in detainment, "is not necessary to estimate the [$\text{CDE}_{M=1}$], which only quantifies discrimination in the second-stage decision" (p. 21). Rather, "subset ignorability is sufficient to ensure the [$\text{CDE}_{M=1}$] can be identified from data on the second-stage decisions" (p. 22).

This bold assertion, which stands in direct contradiction to a vast body of work by causally oriented discrimination scholars (e.g. Heckman and Siegelman, 1993; Heckman, 1998; Heckman and Durlauf, forthcoming), merits close investigation. What, precisely, does subset ignorability require the analyst to believe? In Proposition 2, we analyze the proposed method formally, and find that under confounding, subset ignorability is logically equivalent to an even more challenging knife-edge assumption than that of Proposition 1. To aid in the interpretation of this knife-edge condition, we introduce Proposition 3, proving that

subset ignorability will hold *only if* omitted variable bias (induced by confounding) is exactly cancelled out by selection bias (induced by post-treatment conditioning).

Proposition 2. *Without treatment ignorability, the subset ignorability assumption is satisfied if and only if the following knife-edge equality holds:*

$$\Big\{ \overbrace{\mathbb{E}[Y_i(d,\text{stop}) \mid \text{always stop},\ \text{white}]\ \tfrac{\Pr(\text{always stop}|\text{minority})}{\Pr(\text{always stop}|\text{minority})+\Pr(\text{anti-min. stop}|\text{minority})} + \mathbb{E}[Y_i(d,\text{stop}) \mid \text{anti-min. stop},\ \text{minority}]\ \tfrac{\Pr(\text{anti-min. stop}|\text{minority})}{\Pr(\text{always stop}|\text{minority})+\Pr(\text{anti-min. stop}|\text{minority})}}^{\text{LHS of Prop. 1, after extracting omitted variable bias (and dropping treatment ignorability)}} \Big\}$$

$$- \Big\{ \overbrace{\mathbb{E}[Y_i(d,\text{stop}) \mid \text{always stop}, \text{white}]\ \tfrac{\Pr(\text{always stop}|\text{white})}{\Pr(\text{always stop}|\text{white})+\Pr(\text{anti-white stop}|\text{white})} + \mathbb{E}[Y_i(d,\text{stop}) \mid \text{anti-white stop}, \text{white}]\ \tfrac{\Pr(\text{anti-white stop}|\text{white})}{\Pr(\text{always stop}|\text{white})+\Pr(\text{anti-white stop}|\text{white})}}^{\text{RHS of Prop. 1 (dropping treatment ignorability)}} \Big\}$$

$$= -\Big\{ \overbrace{\Big(\mathbb{E}[Y_i(d,\text{stop}) \mid \text{always stop},\ \text{minority}] - \mathbb{E}[Y_i(d,\text{stop}) \mid \text{always stop},\ \text{white}]\Big) \times \tfrac{\Pr(\text{always stop}|\text{minority})}{\Pr(\text{always stop}|\text{minority})+\Pr(\text{anti-min. stop}|\text{minority})}}^{\text{newly introduced omitted variable bias: minority and white always-stops are now non-comparable}} \Big\}$$

(a) Previously non-comparable expectations in Prop. 1 because they refer to differing principal strata, now further confounded by unobserved differences in minority & white encounter characteristics

(b) Previously comparable expectations in Prop. 1 that are now only comparable (i.e., have the same conditioning set) after first extracting omitted variable bias (unobserved gaps in potential force between minority & white always-stops, moved to the right-hand side)

(c) Previously non-comparable proportions in Prop. 1 due to differing post-treatment selection criteria for white & minority encounters, now additionally confounded by unobserved differences in minority & white encounter-type frequencies

Proposition 3. *The subset ignorability assumption is falsified unless post-treatment bias is precisely equal in magnitude and opposite in sign to omitted variable bias.*

Discussion. Proposition 2 requires the difference between the first two terms (closely resembling the terms in Proposition 1, relating to post-treatment selection) to be *exactly equal in magnitude and opposite in sign* to the third term (relating to differences in the nature of minority and white always stops). The key difference between Proposition 1 and Proposition 2 is that in the former, because treatment is as-if random, minority always-stop encounters ("assaults") are directly comparable to white "assaults." As a result, the third term is zero, and so the Proposition 1 condition requires the first two terms to be identical to ensure that their difference is zero.

To see the roots of omitted variable bias more clearly, examine the following equality, which is logically equivalent to (merely an algebraic manipulation of) the following restatement of subset ignorability: $\mathbb{E}[Y_i(d,1) \mid D_i = 1,\ M_i(1) = 1] \ = \ \mathbb{E}[Y_i(d,1) \mid D_i = 0,\ M_i(0) = 1]$.

$$
\begin{aligned}
&\overbrace{\mathbb{E}[Y_i(d,\text{stop}) \mid \text{always stop},\ \text{minority}]}\ \tfrac{\Pr(\text{always stop}|\text{minority})}{\Pr(\text{always stop}|\text{minority})+\Pr(\text{anti-min. stop}|\text{minority})} \\
&\quad + \mathbb{E}[Y_i(d,\text{stop}) \mid \text{anti-min. stop},\ \text{minority}]\ \tfrac{\Pr(\text{anti-min. stop}|\text{minority})}{\Pr(\text{always stop}|\text{minority})+\Pr(\text{anti-min. stop}|\text{minority})} \\
&= \overbrace{\mathbb{E}[Y_i(d,\text{stop}) \mid \text{always stop}, \text{white}]}\ \tfrac{\Pr(\text{always stop}|\text{white})}{\Pr(\text{always stop}|\text{white})+\Pr(\text{anti-white stop}|\text{white})} \\
&\quad + \mathbb{E}[Y_i(d,\text{stop}) \mid \text{anti-white stop}, \text{white}]\ \tfrac{\Pr(\text{anti-white stop}|\text{white})}{\Pr(\text{always stop}|\text{white})+\Pr(\text{anti-white stop}|\text{white})}
\end{aligned}
$$

This statement is equivalent to Proposition 1 after dropping treatment ignorability. Above, the two terms marked with braces are non-comparable due to confounding: omitted variables mean that white and minority "assault" (always-stop) encounters have different average potential outcomes. To render them comparable, we must first account for the difference in baselines, $\mathbb{E}[Y_i(d,\text{stop}) \mid \text{always stop}, \underline{\text{minority}}] - \mathbb{E}[Y_i(d,\text{stop}) \mid \text{always stop}, \underline{\text{white}}]$. Only after extracting this term (forming the right-hand side of Proposition 2, the source of the omitted variable bias characterized in Proposition 3) will the resulting terms, marked (b) in the proposition, refer to comparable groups as before. The remaining left-hand side closely

resembles Proposition 1, but with two additional complications. First, in as-if-experimental conditions, analysts could at least deduce that "jaywalking" type encounters were equally common in minority and white encounters, if not in the selected dataset observed by analysts. Without treatment ignorability, however, white encounters may involve differing amounts of "jaywalking," "assault," etc. (i.e., different allocations to principal strata). The affected terms in Proposition 2 are marked (c). And second, in as-if-experimental conditions, analysts using selected data are comparing generic "jaywalking" encounters to, e.g. generic "assault" encounters. These groups were *already* non-comparable due to potentially vast differences between principal strata. Without treatment ignorability, however, analysts must now defend an even more specific knife-edge assumption about the peculiar *white* "assault" encounters and how these relate to peculiar *minority* "jaywalking" encounters. These terms are marked (a).

The proof of Proposition 2, which consists of two algebraic manipulations, is straightforward; interested readers are referred to footnote 10 of Appendix A. Proposition 3 clarifies the interpretation of Proposition 2 further, providing a decomposition of total bias into post-treatment bias (PTB) and a remainder that we show is easily interpretable as omitted variable bias (OVB). We then show that the subset ignorability assumption implicitly requires analysts to also assume PTB $= -$OVB; if PTB $\neq -$OVB, then subset ignorability is guaranteed to be false. However, we caution that when treatment is confounded, the subset ignorability assumption is even stronger than this "accidental cancellation of bias" assumption. Even if analysts could somehow identify cases where omitted-variable bias happened to perfectly cancel out post-treatment bias, this would not be sufficient to guarantee the subset ignorability assumption holds. The proof of Proposition 3 is given in Appendix B.

4 Discussion

The use of traditional regression-based approaches to study discrimination in policing remains widespread. For decades, researchers have applied these workhorse techniques to police administrative datasets, drawing conclusions about patterns of police behavior that in turn form the basis of real-world reform recommendations—which are increasingly relied upon as policy-makers seek opportunities for meaningful change. The recent formalization of the key identifying assumption undergirding this common approach, subset ignorability, offers an opportunity to rigorously assess the reliability of this literature. In examining the statistical underpinnings of this work, our formal analysis reveals that much of this literature rests on implicit, difficult-to-defend assumptions about exact balancing between the disparate types of minority and white encounters that appear in police data. We show that serious issues arise when analysts ignore the process by which police-civilian encounters result in detainment and subsequent police actions, like officers' use of force. The core problem is that discrimination in detainment can lead to minority stops in scenarios where white civilians would be allowed to pass without comment—and vice versa, to the extent that anti-white discrimination exists in detainment. This non-comparability of minority and white detainment records produces potentially large statistical bias, except in highly implausible just-so scenarios that analysts cannot verify.

The study of racial bias in policing faces severe challenges even beyond those examined here. In addition to the inherently selective nature of detainment records, the nature of police reports also means that analysts also only see a *temporally* limited slice of police-civilian encounters: the portion beginning with actions triggering a reporting requirement. Because racial bias may well influence officers' decisions in both dimensions, as well as the accuracy of their reporting, analysts must not only contend with the formidable obstacle of omitted variable bias, but also with vast additional obstacles presented by various forms of post-treatment selection, mismeasurement, and purposeful misrepresentation or fabrication (Lee et al., 2017; Friberg et al., 2019; Gay, 2020).

Despite the familiarity of these issues to methodologists and causal inference scholars, applied discrimination researchers have only recently begun to tackle them in earnest. In addition to the bounding approach offered in Knox, Lowe and Mummolo (2020), recent work by Zhao et al. (2020) thoroughly examines a range of discrimination estimands and shows it may be difficult to extrapolate from the $\text{ATE}_{M=1}$ and $\text{ATT}_{M=1}$ to the ATE. To address this issue, it develops an approach to estimate causal risk ratios that sidestep problems relating to the unknown magnitude of $\Pr(M_i = 1)$. But much work still opts to ignore these challenges. If researchers are to uncover an honest portrait of racial bias in policing, the implausible assumptions underlying vast swaths of the literature must be abandoned. Policing data is generated via a complex, multi-stage process that raises unusual threats to causal inference. Given this indisputable property, science is better served by cautious bounding approaches that acknowledge the limitations of police data, or develop careful research designs to avoid these sources of statistical bias from the start. This will require continued innovation in statistical analysis and data collection. While daunting, these challenges are not insurmountable. But simply ignoring them for the sake of expediency will only serve to distort estimates of the severity of this pressing social problem.

References

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

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

Engel, Robin Shepard, James J. Sobol and Robert E. Worden. 2000. "Further exploration of the demeanor hypothesis: The interaction effects of suspects' characteristics and demeanor on police behavior." *Justice Quarterly* 17(2):235–258.

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

Friberg, Ben, David Barer, Rachel Garza, Josh Hinkle, Robert Sims, Calily Bien, Patrick Tolbert and Chad Cross. 2019. "Texas troopers ticketing Hispanic drivers as white." *kxan* . https://www.kxan.com/investigations/texas-troopers-ticketing-hispanic-drivers-as-white/.

Fryer, Roland G. 2019. "An Empirical Analysis of Racial Differences in Police Use of Force." *Journal of Political Economy* 127(3):1210–1261.

Gaebler, Johann, William Cai, Guillaume Basse, Ravi Shroff, Sharad Goel and Jennifer Hill. 2020. "Deconstructing Claims of Post-Treatment Bias in Observational Studies of Discrimination." https://arxiv.org/abs/2006.12460.

Gay, Mara. 2020. "Why Was a Grim Report on Police-Involved Deaths Never Released? A review shows that the number of people killed by police activity in New York is more than twice what has been reported." *The New York Times* . https://www.nytimes.com/2020/06/19/opinion/police-involved-deaths-new-york-city.html.

Greenland, Sander. 2014. "Quantifying biases in causal models: classical confounding vs collider-stratification bias." *Epidemiology* 14(3):300–306.

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

Heckman, James J. 1998. "Detecting Discrimination." *Journal of Economic Perspectives* 12(2):101–116.

Heckman, James J. and Peter Siegelman. 1993. *Clear and Convincing Evidence*. Washington, DC: Urban Institute Press chapter The Urban Institute Audit Studies: Their Methods and Findings.

Heckman, James J. and Steven N. Durlauf. forthcoming. "Comment on "An Empirical Analysis of Racial Differences in Police Use of Force" by Roland G. Fryer Jr." *Journal of Political Economy* .

Knox, Dean, Will Lowe and Jonathan 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.

Lee, Christopher T., Mary Huynh, Paulina Zheng, Alejandro Castro III, Francia Noel, Darlene Kelley, Jennifer Norton, Catherine Stayton and Gretchen Van Wye. 2017. Enumeration and classification of law enforcement-related deaths — New York City, 2010–2015. Technical report New York City Dept. of Health Maryland: . https://www1.nyc.gov/assets/doh/downloads/pdf/about/law-enforcement-deaths.pdf.

Lundman, Richard J. 1994. "Demeanor or crime? The Midwest City police-citizen encounters study." *Criminology* 32(4):631–656.

Lundman, Richard J. 1996. "Demeanor and arrest: Additional evidence from previously unpublished data." *Journal of Research in Crime and Delinquency* 33(3):306–323.

Meek, Christopher. 1995. Strong completeness and faithfulness in Bayesian networks. In *Proceedings of the Eleventh Conference on Uncertainty in AI*, ed. P. Besnard and S. Hanks. Morgan Kaufmann pp. 411–418.

Novak, Kenneth J., Robert A. Brown and James Frank. 2005. "Police-citizen encounters and field citations." *Policing: An International Journal of Police Strategies & Management* pp. 234–249.

Pearl, Judea. 1993. "Graphical Models, Causality and Intervention." *Statistical Science* 8(3):266–269.

Pearl, Judea. 2000. *Causality.* Cambridge University Press.

Robins, James M., Richard Scheines, Peter Spirtes and Larry Wasserman. 2003. "Uniform Consistency in Causal Inference." *Biometrika* 90(3):491–515.
URL: *http://www.jstor.org/stable/30042062*

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

Rubin, Donald B. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology* 66(5):688–701.

Schafer, Joseph A., David L. Carter, Andra J. Katz-Bannister, and William M. Wells. 2006. "Decision making in traffic stop encounters: A multivariate analysis of police behavior." *Police Quarterly* 9(2):184–209.

Scheines, Richard. 1997. *Causality in Crisis?* University of Notre Dame Press chapter An Introduction to Causal Inference.

Smith, Douglas A., Christy A. Visher, and Laura A. Davidson. 1984. "Equity and discretionary justice: The influence of race on police arrest decisions." *J. Crim. L. & Criminology* 75(1).

Spano, Richard. 2003. "Concerns about safety, observer sex, and the decision to arrest: evidence of reactivity in a large-scale observational study of police." *Criminology* 41(3):909–932.

Spirtes, Peter, Clark Glymour and Richard Scheines. 1993. *Causation, Prediction and Search.* Springer-Verlag.

Tillyer, R. and R.S. Engel. 2013. "The impact of drivers' race, gender, and age during

traffic stops: Assessing interaction terms and the social conditioning model." *Crime & Delinquency* 59(3):369–395.

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

Zhao, Qingyuan, Luke J Keele, Dylan S Small and Marshall M Joffe. 2020. "A note on post-treatment selection in studying racially biased policing." https://arxiv.org/abs/2009.04832.

Supplementary Information

Table of Contents

A Detailed Proofs of Propositions 1 & 2

Our proof of the naïve estimator's bias for the $\text{CDE}_{M=1}$ builds on Appendix A.3 of Knox, Lowe and Mummolo (2020). In our running analogy between selection bias and omitted variable bias, the derivation below is analogous to the general omitted-variable-bias formula of Section 3.2, $\frac{\gamma_{(1)}\alpha_{(1)}}{\alpha_{(1)}^2+\alpha_{(2)}^2+1} + \frac{\gamma_{(2)}\alpha_{(2)}}{\alpha_{(1)}^2+\alpha_{(2)}^2+1}$. In as-if experimental settings, invoking subset ignorability is equivalent to assuming that there exists no selection bias, i.e. that the naïve regression recovers the $\text{CDE}_{M=1}$. The direct analogy in Section 3.2 would be the assumption that a regression of Y_i on X_i will recover the causal quantity of interest, β—i.e., that there is no omitted variable bias. Though this "no-omitted-variable-bias assumption" is compact and easy to state, formally deriving the logical implications reveals its implausibility. For there to be no omitted variable bias in the presence of these unmeasured confounders, it must be precisely true that $\gamma_{(1)}\alpha_{(1)} = -\gamma_{(2)}\alpha_{(2)}$, a condition that only holds along an infinitesimally narrow region in the model space of all possible $\gamma_{(1)}$, $\alpha_{(1)}$, $\gamma_{(2)}$, and $\alpha_{(2)}$ values. If these parameters were randomly drawn from any smooth distribution, there would be zero probability of the no-omitted-variable-bias assumption holding.

This implausible condition is directly analogous to the knife-edge balancing condition presented in Proposition 1, the logical implication of the subset ignorability assumption. Much like the $\gamma_{(1)}\alpha_{(1)} = -\gamma_{(2)}\alpha_{(2)}$ condition, the Proposition 1 conditions are merely a special case that follows from the more general bias derivation. We now reexamine that derivation in depth.

We follow Appendix A.3 of Knox, Lowe and Mummolo (2020), which derives the bias of the naïve estimator when targeting the $\text{CDE}_{M=1,x}$, the conditional analog of the $\text{CDE}_{M=1}$ for the subset of encounters with $X_i = x$. (For clarity of exposition, we will implicitly condition on $X_i = x$ throughout.) The paper states, "The derivation is almost identical to that of the $\text{ATE}_{M=1,x}$ [Appendix A.1], differing only in that all individuals are held at $M_i = 1$ instead of…vary[ing] with civilian race, $M_i(D_i)$."

Literally,

$$
\begin{aligned}
&\mathbb{E}[\hat{\Delta}] - \text{CDE}_{M=1} \;=\; \\
&\Big\{ \mathbb{E}[Y_i(1,1)|D_i=1, M_i(1)=1, M_i(0)=1] \\
&\qquad \Pr(M_i(0)=1|D_i=1, M_i(D_i)=1)\Pr(D_i=0|M_i(D_i)=1) \\
&\quad + \mathbb{E}[Y_i(1,1)|D_i=1, M_i(1)=1, M_i(0)=0] \\
&\qquad \Pr(M_i(0)=0|D_i=1, M_i(D_i)=1)\Pr(D_i=0|M_i(D_i)=1) \\
&\quad - \mathbb{E}[Y_i(1,1)|D_i=0, M_i(1)=1, M_i(0)=1] \\
&\qquad \Pr(M_i(1)=1|D_i=0, M_i(D_i)=1)\Pr(D_i=0|M_i(D_i)=1) \\
&\quad - \mathbb{E}[Y_i(1,1)|D_i=0, M_i(1)=0, M_i(0)=1] \\
&\qquad \Pr(M_i(1)=0|D_i=0, M_i(D_i)=1)\Pr(D_i=0|M_i(D_i)=1) \Big\} (\alpha) \\
&\Big\{ - \mathbb{E}[Y_i(0,1)|D_i=0, M_i(1)=1, M_i(0)=1] \\
&\qquad \Pr(M_i(1)=1|D_i=0, M_i(D_i)=1)\Pr(D_i=1|M_i(D_i)=1) \\
&\quad - \mathbb{E}[Y_i(0,1)|D_i=0, M_i(1)=0, M_i(0)=1] \\
&\qquad \Pr(M_i(1)=0|D_i=0, M_i(D_i)=1)\Pr(D_i=1|M_i(D_i)=1) \\
&\quad + \mathbb{E}[Y_i(0,1)|D_i=1, M_i(1)=1, M_i(0)=1] \\
&\qquad \Pr(M_i(0)=1|D_i=1, M_i(D_i)=1)\Pr(D_i=1|M_i(D_i)=1) \\
&\quad + \mathbb{E}[Y_i(0,1)|D_i=1, M_i(1)=1, M_i(0)=0] \\
&\qquad \Pr(M_i(0)=0|D_i=1, M_i(D_i)=1)\Pr(D_i=1|M_i(D_i)=1) \Big\} (\omega)
\end{aligned}
$$

per Knox, Lowe and Mummolo (2020) Appendix pp. 1–2 and p. 6. It immediately follows that (i) the knife-edge condition of Proposition 2 achieves unbiasedness in general, and (ii) the knife-edge condition of Proposition 1 achieves unbiasedness if treatment ignorability is satisfied. To verify, observe that the first four terms are proportional to

$$
\begin{aligned}
\alpha \propto\; & \mathbb{E}[Y_i(1,1)|D_i=1, M_i(1)=1, M_i(0)=1]\ \Pr(M_i(0)=1|D_i=1, M_i(1)=1) \\
& + \mathbb{E}[Y_i(1,1)|D_i=1, M_i(1)=1, M_i(0)=0]\ \Pr(M_i(0)=0|D_i=1, M_i(1)=1) \\
& - \mathbb{E}[Y_i(1,1)|D_i=0, M_i(1)=1, M_i(0)=1]\ \Pr(M_i(1)=1|D_i=0, M_i(0)=1) \\
& - \mathbb{E}[Y_i(1,1)|D_i=0, M_i(1)=0, M_i(0)=1]\ \Pr(M_i(1)=0|D_i=0, M_i(0)=1). \qquad (3)
\end{aligned}
$$

Rearranging terms, it can be seen that Proposition 2 (plugging $d = 1$ into the proposition) is logically equivalent to the statement that $\alpha = 0$.[10] If treatment ignorability holds, this reduces to

$$\begin{aligned}\alpha \propto\ & \mathbb{E}[Y_i(1,1)|M_i(1)=1, M_i(0)=1]\ \Pr(M_i(0)=1|M_i(1)=1) \\ & + \mathbb{E}[Y_i(1,1)|M_i(1)=1, M_i(0)=0]\ \Pr(M_i(0)=0|M_i(1)=1) \\ & - \mathbb{E}[Y_i(1,1)|M_i(1)=1, M_i(0)=1]\ \Pr(M_i(1)=1|M_i(0)=1) \\ & - \mathbb{E}[Y_i(1,1)|M_i(1)=0, M_i(0)=1]\ \Pr(M_i(1)=0|M_i(0)=1), \end{aligned} \tag{4}$$

and the Proposition 1 knife-edge balancing statement (again plugging $d = 1$ into Proposition 1) is logically equivalent to the statement that $\alpha = 0$. To reiterate, *these two statements are mathematically identical*; to see this, set $\alpha = 0$ in Equation 4, move the latter two terms to the left-hand side, and expand the conditional probabilities. Similarly, when setting $d = 0$, the Proposition 1 and 2 statements are logically equivalent to $\omega = 0$.

More broadly, subset ignorability is logically equivalent to the assumption that $\alpha = \omega = 0$. As Knox, Lowe and Mummolo (2020) showed, the naïve estimator is unbiased for the $\text{CDE}_{M=1}$ when this holds and treatment ignorability is satisfied.

Knox, Lowe and Mummolo (2020) does not remark on the point that exact cancellation of opposing terms can produce zero bias. Such observations are simultaneously (i) applicable in virtually every formal analysis of causal identification, (ii) almost never satisfied, in a measure-theoretic sense, and (iii) therefore unproductive for applied policing scholars. (For the same reason, Knox, Lowe and Mummolo (2020) also did not remark on the fact that bias can be zero when $\alpha = -\omega$.)

[10] Specifically, add $\mathbb{E}[Y_i(1,1)|D_i = 0, M_i(1)=1, M_i(0)=1]\ \Pr(M_i(0)=1|D_i=1, M_i(1)=1)$ to both sides, then subtract $\mathbb{E}[Y_i(1,1)|D_i=1, M_i(1)=1, M_i(0)=1]\ \Pr(M_i(0)=1|D_i=1, M_i(1)=1)$ from both sides.

B Proof that Subset Ignorability Can Only Hold if Post-treatment Bias is Equal in Magnitude and Opposite in Sign to Omitted Variable Bias

First, define PTB as the bias that arises from post-treatment selection alone, i.e. when treatment ignorability is satisfied. Applying this property to the first equation in Appendix A and simplifying comparable terms, we obtain

$$
\begin{aligned}
\text{PTB} = \Bigg\{ & \mathbb{E}[Y_i(1,1)|D_i=0, M_i(1)=1, M_i(0)=1] \\
& \quad \big[\Pr(M_i(0)=1|D_i=1, M_i(D_i)=1) - \Pr(M_i(1)=1|D_i=0, M_i(D_i)=1)\big] \\
& \quad \Pr(D_i=0|M_i(D_i)=1) \\
& + \mathbb{E}[Y_i(1,1)|D_i=1, M_i(1)=1, M_i(0)=0] \\
& \quad \Pr(M_i(0)=0|D_i=1, M_i(D_i)=1)\Pr(D_i=0|M_i(D_i)=1) \\
& - \mathbb{E}[Y_i(1,1)|D_i=0, M_i(1)=0, M_i(0)=1] \\
& \quad \Pr(M_i(1)=0|D_i=0, M_i(D_i)=1)\Pr(D_i=0|M_i(D_i)=1) \Bigg\} (\alpha_{\text{PTB}}) \\
\Bigg\{ & - \mathbb{E}[Y_i(0,1)|D_i=0, M_i(1)=1, M_i(0)=1] \\
& \quad \big[\Pr(M_i(1)=1|D_i=0, M_i(D_i)=1) - \Pr(M_i(0)=1|D_i=1, M_i(D_i)=1)\big] \\
& \quad \Pr(D_i=1|M_i(D_i)=1) \\
& - \mathbb{E}[Y_i(0,1)|D_i=0, M_i(1)=0, M_i(0)=1] \\
& \quad \Pr(M_i(1)=0|D_i=0, M_i(D_i)=1)\Pr(D_i=1|M_i(D_i)=1) \\
& + \mathbb{E}[Y_i(0,1)|D_i=1, M_i(1)=1, M_i(0)=0] \\
& \quad \Pr(M_i(0)=0|D_i=1, M_i(D_i)=1)\Pr(D_i=1|M_i(D_i)=1). \Bigg\} (\omega_{\text{PTB}})
\end{aligned}
$$

Next, recall that the bias arising when treatment is nonignorable is

$$
\begin{aligned}
\text{Total Bias} \;=\; & \Big\} (\alpha) \\
& \mathbb{E}[Y_i(1,1)|D_i=1, M_i(1)=1, M_i(0)=1] \\
& \qquad \Pr(M_i(0)=1|D_i=1, M_i(D_i)=1)\Pr(D_i=0|M_i(D_i)=1) \\
& +\mathbb{E}[Y_i(1,1)|D_i=1, M_i(1)=1, M_i(0)=0] \\
& \qquad \Pr(M_i(0)=0|D_i=1, M_i(D_i)=1)\Pr(D_i=0|M_i(D_i)=1) \\
& -\mathbb{E}[Y_i(1,1)|D_i=0, M_i(1)=1, M_i(0)=1] \\
& \qquad \Pr(M_i(1)=1|D_i=0, M_i(D_i)=1)\Pr(D_i=0|M_i(D_i)=1) \\
& -\mathbb{E}[Y_i(1,1)|D_i=0, M_i(1)=0, M_i(0)=1] \\
& \qquad \Pr(M_i(1)=0|D_i=0, M_i(D_i)=1)\Pr(D_i=0|M_i(D_i)=1) \\
& \Big\} (\omega) \\
& -\mathbb{E}[Y_i(0,1)|D_i=0, M_i(1)=1, M_i(0)=1] \\
& \qquad \Pr(M_i(1)=1|D_i=0, M_i(D_i)=1)\Pr(D_i=1|M_i(D_i)=1) \\
& -\mathbb{E}[Y_i(0,1)|D_i=0, M_i(1)=0, M_i(0)=1] \\
& \qquad \Pr(M_i(1)=0|D_i=0, M_i(D_i)=1)\Pr(D_i=1|M_i(D_i)=1) \\
& +\mathbb{E}[Y_i(0,1)|D_i=1, M_i(1)=1, M_i(0)=1] \\
& \qquad \Pr(M_i(0)=1|D_i=1, M_i(D_i)=1)\Pr(D_i=1|M_i(D_i)=1) \\
& +\mathbb{E}[Y_i(0,1)|D_i=1, M_i(1)=1, M_i(0)=0] \\
& \qquad \Pr(M_i(0)=0|D_i=1, M_i(D_i)=1)\Pr(D_i=1|M_i(D_i)=1).
\end{aligned}
$$

We now proceed to decompose the total bias:

$$
\begin{aligned}
&\text{Total Bias} \;=\; \text{PTB} + \text{additional bias} \\
&\text{Total Bias} - \text{PTB} \;= \\
&\left.\begin{aligned}
&\Big\{\mathbb{E}[Y_i(1,1)|D_i=1, M_i(1)=1, M_i(0)=1] - \mathbb{E}[Y_i(1,1)|D_i=0, M_i(1)=1, M_i(0)=1]\Big\} \\
&\qquad \Pr(M_i(0)=1|D_i=1, M_i(D_i)=1)\Pr(D_i=0|M_i(D_i)=1)
\end{aligned}\right\} \begin{aligned}&\alpha - \alpha_{\text{PTB}} \\ &= \alpha_{\text{OVB}}\end{aligned} \\
&\left.\begin{aligned}
&+\Big\{\mathbb{E}[Y_i(0,1)|D_i=1, M_i(1)=1, M_i(0)=1] - \mathbb{E}[Y_i(0,1)|D_i=0, M_i(1)=1, M_i(0)=1]\Big\} \\
&\qquad \Pr(M_i(0)=1|D_i=1, M_i(D_i)=1)\Pr(D_i=1|M_i(D_i)=1)
\end{aligned}\right\} \begin{aligned}&\omega - \omega_{\text{PTB}} \\ &= \omega_{\text{OVB}}\end{aligned}
\end{aligned}
$$

so that Total Bias $= \alpha + \omega$ and PTB $= \alpha_{\text{PTB}} + \omega_{\text{PTB}}$. Finally, notice that the remaining terms take the form

$$\mathbb{E}[\text{ potential outcome} \mid D_i = 1,\ \text{subset}\ ] - \mathbb{E}[\text{ potential outcome} \mid D_i = 0,\ \text{subset}\ ],$$

which is the classic structure of omitted variable bias rendering average potential outcomes within the treated subset ($D_i = 1$) non-comparable to average potential outcomes within the control subset ($D_i = 0$). The severity of this bias within the treated and control subgroups is then weighted and averaged to yield what is straightforwardly interpretable as an overall omitted variable bias. Thus, the bias decomposition can be expressed

$$\text{Total Bias} = \text{PTB} + \text{OVB}$$

where OVB $= \alpha_{\text{OVB}} + \omega_{\text{OVB}}$. As we show in Proposition 2 and Appendix A, the subset ignorability assumption is logically equivalent to the assumption that $\alpha = \omega = 0$. This directly implies $\alpha_{\text{PTB}} = -\alpha_{\text{OVB}}$ and $\omega_{\text{PTB}} = -\omega_{\text{OVB}}$, which in turn implies PTB $= -$OVB. Thus, for subset ignorability to not be falsified, post-treatment bias must be precisely equal in magnitude and opposite in sign to omitted variable bias. However, because it is possible that $\alpha_{\text{PTB}} + \omega_{\text{PTB}} = -\alpha_{\text{OVB}} - \omega_{\text{PTB}}$ without $\alpha_{\text{PTB}} = -\alpha_{\text{OVB}}$ and $\omega_{\text{PTB}} = -\omega_{\text{OVB}}$, exactly cancelling bias is merely *necessary*, but not *sufficient*, for the subset ignorability assumption to hold. □