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

Dean Knox, Will Lowe and Jonathan Mummolo*

August 24, 2020

Abstract

Observational studies of racial bias in policing often rely on stop or arrest records—which may themselves be a product of racial bias—to estimate discrimination in subsequent actions like police violence. This contaminated data raises the well-known threat of post-treatment selection bias, which recent work shows can lead to drastic underestimates of discrimination. However, Gaebler, Cai, Basse, Shroff, Goel and Hill (GCBSGH*a*, 2020) proposes new theoretical arguments aimed at "clarifying the statistical foundations of discrimination analysis," and concludes, "concerns about post-treatment bias may be misplaced." GCBSGH*a*'s proposal merits close study, as it posits a massive methodological breakthrough which, if confirmed, would undermine over 40 years of research on selection bias. We analyze the proposal formally and find its key underlying assumption, "subset ignorability," corresponds to the measure-zero (i.e., knife-edge) conditions in which differing biases happen to sum to zero. That is, rather than developing improved research designs or deriving better estimation techniques, GCBSGH*a* advocates assuming that even with imperfect controls, biases from multiple sources will happen to perfectly offset one another. When treatment is confounded, as in several GCBSGH examples, we prove "subset ignorability" holds *only if* post-treatment selection bias exactly offsets omitted variable bias. And in ideal experimental settings, this approach is unbiased *if and only if* the following knife-edge assumption holds: *despite bias in detainment*—e.g. stopping minority civilians for as little as jaywalking but white civilians only for assault—the groups are *nonetheless exactly comparable in potential for police violence.* We conclude the "subset ignorability" assumption is unlikely to be defensible in real-world settings, and we emphasize the need for caution and increased rigor in high-stakes analyses of discriminatory policing with contaminated data.

*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.

Contents

1 Summary

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 (KLM, 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 biasing subsequent estimates. Despite prominent claims to the contrary (Gelman, 2020; Fryer, forthcoming), KLM proves this selection bias contaminates estimates of racial bias regardless of whether the target estimand is the *total effect* of civilian race or the *controlled direct effect* of civilian race after detainment. It further shows that by failing to account for selection, studies of racial bias in police force that employ standard regression approaches using post-treatment-selected data (e.g. Fryer, 2019) may drastically understate the severity of police discrimination. KLM is not alone in noting the challenge that selection poses in this setting. Heckman and Durlauf (forthcoming) notes 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).

In a recent paper, Gaebler, Cai, Basse, Shroff, Goel, and Hill (2020) (GCBSGH*a*) tackle this challenge head-on, developing new statistical theory aimed at "clarifying the statistical foundations of discrimination analysis" (p. 4), which often relies on contaminated data—e.g., by "estimat[ing] discrimination... based only on data describing those who were arrested" (p. 7). Specifically, GCBSGH*a* formalizes the "often unstated assumptions in studies of discrimination," and seeks "to put that research on more solid theoretical footing" (p. 22). GCBSGH*a*'s proposed approach is described as broadly applicable, not only for the study of police violence, but potentially "help[ing] to advance" the "entire enterprise of quantitative discrimination studies" (p. 4)—even when researchers cannot plausibly assume the as-if randomness of race, because such a condition is "unlikely to hold" (p. 21).[1] The broad scope of GCBSGH*a*'s technique suggests a major advance; in contrast, past work has "emphasize[d] 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). But under the statistical theory of GCBSGH*a*, the paper claims, "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, GCBSGH*a* argues that analysts 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.

To be clear, the proposed method of GCBSGH*a* does not entail new estimation techniques or research designs to counteract these well-known sources of statistical bias. Rather, GCBSGH*a* develops *assumptions* to justify the long-standing empirical practice of applying "standard difference-in-means estimator[s]" or "common... regression model[s]" (p. 8) to contaminated data. GCBSGH*a* thus rejects the methodological points of KLM and Heckman and Durlauf (forthcoming), instead advocating research designs like that of Fryer (2019) as wholly suitable for assessing racial bias in police violence.[2]

[1] Specifically, GCBSGH*a* 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). KLM agrees, noting "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).

[2] Equation 5 in GCBSGH*a* invokes the same estimation strategy as the one used in Fryer (2019), namely, a comparison of the conditional probabilities of the outcome of interest, e.g. use of force, across white and nonwhite encounters, controlling for observable features of police stops.

If credible, this analytic strategy represents a massive methodological breakthrough, undermining decades of research on the challenges of analyzing post-treatment-selected data. It therefore merits close investigation. What does the proposed method of GCBSGH*a* 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 GCBSGH*a*'s assumption 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. Rather than providing a research design that avoids statistical bias, or developing a partial-identification approach to quantify the range of possible statistical bias, GCBSGH*a* advocates that researchers instead *assume* omitted variable bias and post-treatment selection bias perfectly offset one another. In Propositions 1–3, we show this formally, then provide examples of the hyper-specific assumptions that analysts would need to articulate and defend to use GCBSGH*a*'s method. 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 Section 3, we show how Gaebler, Cai, Basse, Shroff, Goel, and Hill's arguments—not only GCBSGH*a*, but also subsequent revisions and amendments, GCBSGH*b*, GCBSGH*c*, and GCBSGH*d*—are built on a foundation of precisely such assumptions.

Our findings reveal that the key identifying assumption in GCBSGH*a*—innocuously termed "subset ignorability," but shown in Position 3 to be even stronger than a zero-bias assumption—is far less plausible than traditional ignorability assumptions. These traditional assumptions are about groups being comparable *given as-if-random assignment of treatment*. In contrast, GCBSGH*a*'s approach works only 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.[3] Critically, GCBSGH*a*'s assumption is an assertion about the world that cannot be guaranteed *even by gold-standard experimental*

[3]As GCBSGH*a* states, "we imagine that the perception of race is counterfactually determined after the first-stage decision but before the second-stage decision," (7).

designs that randomize actors into police-civilian encounters. As we show in Section 3, even in such ideal settings, the proposed approach rests on a confusion between (i) observable features of police encounters and (ii) the unobservable "principal strata" (Frangakis and Rubin, 2002)—latent groupings in the data defined by unobserved counterfactual profiles—to which they belong. This proposal essentially assumes away the core problem of post-treatment selection in this setting, which is 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): 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 core assumption of GCBSGH*a* 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).

GCBSGH*a* formalizes an assumption that discrimination researchers often make implicitly, allowing scholars to carefully assess its logical implications and precisely debate the scenarios in which it might be valid. This is a valuable contribution. However, we show the conditions under which GCBSGH*a*'s proposal will obtain unbiased results are exceedingly unlikely to hold in many real-world policing settings where discrimination constitutes a major policy concern. These differences of opinion, at their core, stem from a differing philosophical approach to the quantitative analysis of racial discrimination. GCBSGH*a* advocates researchers assume biases will happen to precisely cancel one another, without any substantive explanation of why such perfect cancellation might occur. In contrast, given the high stakes in this policy arena, we argue for an alternative: cautious bounding approaches that describe the range of possible causal effects without such assumptions, accounting for all possible severities of statistical bias and guarding against the very real possibility that the "subset ignorability" assumption is false (KLM). Careful research design, using quasi-experimental scenarios that *justify* assumptions and mitigate sources of bias using expert

knowledge and case selection (e.g. West, 2018) offers a second alternative for securing reliable estimates. 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 GCBSGH*a*'s proposed approach, deriving its logical implications and clarifying its applicability to applied research. In addition, because GCBSGH*a* was centrally motivated by a critique of KLM, and because those critiques speak directly to the inferential issues we examine, we then respond to those critiques in Section 4. After carefully probing the newly developed theory and weighing the plausibility of its implications, we stand by our original assertion that "existing empirical work in this area is producing a misleading portrait of evidence as to the severity of racial bias in police behavior" (KLM, p. 620). 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

GCBSGH*a* employs a general causal framework and proposes a broadly applicable statistical approach for "discriminatory decision making in a variety of real-world situations" (p. 5). Their theoretical arguments and estimation procedures are described as "clarifying the statistical foundations of discrimination analysis" (p. 4) and the "theoretical underpinnings of [discrimination] research" (p. 22) broadly conceived, including "police violence stemming from discrimination in initial stop decisions" (p. 4). In what follows, we utilize notation and terminology for the police-violence setting following KLM, the core motivating study for GCBSGH*a* (abstract). (In a stylized example, GCBSGH*a* 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, 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.)

We consider the data-generating process in the directed acyclic graph (DAG) in Figure 1, reproduced from KLM. 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 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 an individual with a different racial identity into the encounter, while holding the encounter's objective context—location, time of day, criminal activity, etc.—fixed" (Knox, Lowe and Mummolo, 2020, 621).

This counterfactual deserves special note, as it is critical to conceptualizing a feasible causal exercise. Here, the unit of analysis, the police-civilian encounter, is clearly defined, and specifically chosen to avoid well-known issues regarding "immutable, and hence non-manipulable, characteristic[s]" KLM (p. 621); thus, the "ideal experiment" does not entail the difficult-to-imagine manipulation of an individual's race, as described in GCBSGH*a*, but rather the substitution of comparable actors into pre-existing scenes.[4] As KLM notes on pp. 3–4, this approach could never hope to capture 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*.[5]

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

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 "prin-

[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.

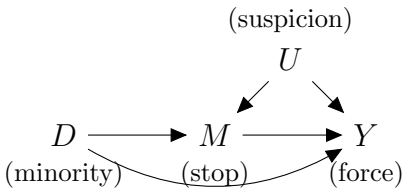
[5]In spite of this explicit discussion of a feasible counterfactual, GCBSGH*a* reintroduces confusion over units of analysis and ideal experiments when critiquing the notion of "intermediate outcomes" like whether a civilian is stopped. It states,

> The very idea of "intermediate outcomes"—a concept central to the Knox et al. critique—is a slippery notion in the context of discrimination studies, where there is no clear point in time where one can imagine that race is 'assigned.' Even birth cannot be considered the ultimate starting point since, in theory, one might include, at the least, the race of a child's parents, determined at an earlier stage, when assessing discrimination. (p. 11)

However, KLM's ideal experiment revolves around substituting *different* individuals into encounters; at no point does it suggest "assigning" an individual's race.

[6]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."

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.



cipal 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, where police would 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.

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)

Acknowledging the existence of these principal strata illustrates the core challenge with making inferences from post-treatment-selected data: given racial bias in stopping ($D \to M$), 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 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 they capture the full severity of racial bias during police encounters, both direct and indirect, KLM focuses on estimating various conditional total effects ($\text{ATE}_{M=1}$ and $\text{ATT}_{M=1}$, defined in KLM). However, KLM also analyzes another causal estimand in brief asides and an appendix: the controlled direct effect among the detained, $\text{CDE}_{M=1}$ (denoted CDE_{Ob} in GCBSGH*a*), expressed $\text{CDE}_{M=1} = \mathbb{E}[Y_i(1,1)|M_i(D_i) = 1] - \mathbb{E}[Y_i(0,1)|M_i(D_i) = 1]$. As KLM explains, this quantity is extraordinarily difficult to estimate with observational data, because it 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*. However, because it is the target quantity in GCBSGH*a*, we focus on it exclusively below.

3 A Formal Analysis of GCBSGH's Proposal

We begin by formally analyzing the core claim of GCBSGH*a*—that the $\text{CDE}_{M=1}$ can be estimated without bias, after selecting on detainment, as long as GCBSGH*a*'s "subset ignorability" assumption (Definition 1, below) holds. Because the steps of this proposed approach—first invoking the "subset ignorability" assumption, then using standard regression or differences-in-means estimators—are described as "clarifying the statistical foundations of discrimination analysis" (p. 4) and the "theoretical underpinnings of [discrimination] research" (p. 22), it is important to understand precisely what GCBSGH*a*'s method entails. To do so, 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. Even in this ideal case, we find that the "subset ignorability" assumption is logically equivalent to acknowledging *selection*, but assuming away selection *bias*. Specifically, Proposition 1 shows that the GCBSGH*a* assumption can be satisfied *if and only if* an extraordinarily difficult knife-edge balancing condition holds: that even though 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 GCBSGH*a* describe treatment ignorability as "unlikely to hold" (p. 21), we then turn to the general case: when analysts must also grapple with baseline differences in encounters due to omitted variables. It is here that the theoretical arguments of GCBSGH*a* are most provocative. 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). But GCBSGH*a* disputes this characterization, arguing that using their 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).

How can this be? In Proposition 2, we formally analyze the proposed method in full generality. We show that under confounding, GCBSGH*a*'s advocated assumption logically corresponds to an even more implausible knife-edge condition. Specifically, we prove that "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 GCBSGH*a*'s core assumption, "subset ignorability." 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, "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 3 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 GCBSGH*a*'s "subset ignorability" assumption requires that officers to 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 3, to find cases where the subset ignorability was not violated, we carefully hand-tuned 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 almost never satisfied, in a measure-theoretic sense. Thus, knife-edge balancing is essentially a blind hope the analyst expresses about the world.

Proof. The proof follows Appendices A.1 and A.3 of KLM; a detailed walkthrough 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 ($\iff$)

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

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

[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 $\text{leniency}_{\text{minority}} \cdot \frac{2}{3} = \text{leniency}_{\text{white}} \cdot \frac{1}{3}$.

Figure 3: **What would it take for "subset ignorability" to hold in experiments? Three hypothetical scenarios.** Each panel presents a hypothetical composition of police stops. GCBSGH*a*'s no-selection-bias assumption 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)

Stop if minority?	**Stop if white?** Yes	**Stop if white?** No
Yes	**assault: 1/3 potential stops** if minority, 100% force if white, 50% force	**jaywalk: 1/3 potential stops** if minority, 25% force if white, 10% force
No	**anti-white: 1/3 potential stops** if minority, 25% force if white, 10% force	

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 GCBSGH*a*'s assumption: requires *exact equality between jaywalking and anti-white encounters* (whatever those may be) in terms of potential officer force.

(b)

Stop if minority?	**Stop if white?** Yes	**Stop if white?** No
Yes	**assault: 2/7 potential stops** if minority, 100% force if white, 50% force	**jaywalk: 4/7 potential stops** if minority, 75% force if white, 37.5% force
No	**anti-white: 1/7 potential stops** if minority, 50% force if white, 25% force	

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 GCBSGH*a*'s assumption: 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)

Stop if minority?	**Stop if white?** Yes	**Stop if white?** No
Yes	**assault: 1/2 potential stops** if minority, 100% force if white, 50% force	**jaywalk: 1/2 potential stops** if minority, 100% force if white, 50% force
No	**anti-white: nonexistent** if minority, *NA* if white, *NA*	

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 GCBSGH*a*'s assumption: requires that there is *absolutely no difference between assaults and jaywalking* in terms of potential officer force.

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. 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.[8] 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).

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 errors $\varepsilon_i^{(*)} \sim \mathcal{N}(0,1)$. In these circumstances, a typical causal inference 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, the argument in GCBSGH*a* 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 that of GCBSGH*a*.

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

[8] For example, KLM 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.

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 GCBSGH*a* 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. KLM 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, GCBSGH*a* nonetheless asserts that in spite of confounding *and* post-treatment selection, their 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 the proposed method of GCBSGH*a* require the analyst to believe? In Proposition 2, we analyze the proposed method formally, and find that under confounding, GCBSGH*a*'s advocated assumption 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:*

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

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

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

(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 so 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 "subset ignorability" restatement: $\mathbb{E}[Y_i(d,1) \mid M_i(1)=1] = \mathbb{E}[Y_i(d,1) \mid D_i=0,\ M_i(0)=1]$.

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

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 omitted here. Beginning with the first step of Appendix A.3 in KLM, the result follows immediately. Interested readers are referred to footnote 11 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 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 Reply to Critiques of Knox, Lowe and Mummolo (2020)

GCBSGH*a* motivates its theoretical contributions with a critique of KLM, a recent paper that takes some steps toward formalizing discrimination research—developing a causal framework, defining candidate estimands, and enumerating possible assumptions—that were subsequently extended by GCBSGH*a*. Both studies share a common goal, improving statistical practice in the study of racial bias, and we appreciate GCBSGH*a*'s rigor in carefully outlining the "theoretical underpinnings of [discrimination] research" (p. 22) using what KLM calls the "naïve estimator," but which GCBSGH*a* clarify is in fact statistically justified (i.e., not naïve) if their advocated "subset ignorability" assumption is satisfied. In Section 3, we built on GCBSGH*a*'s work to unpack the precise meaning of this proposed approach.

Despite this common goal, the two analyses diverge sharply in their philosophical approach to the quantitative analysis of racial discrimination. KLM advocates a cautious partial-identification approach for quantifying racial bias in light of the challenges examined above, develops new empirical techniques for assessing best- and worst-case levels of discrimination that are consistent with data contaminated by post-treatment selection, and proves that these bounds are *sharp*—i.e., cannot be narrowed without additional information or further assumptions that are deemed indefensible. In contrast, GCBSGH*a* advocate stronger assumptions that, if satisfied, would eliminate the need for bounds entirely and justify the long-standing empirical practice of directly comparing minority and white encounters within the contaminated data.

Beyond this philosophical divergence, GCBSGH*a* also levies a number of sharp criticisms against KLM, which we briefly respond to here in order to resolve any lingering confusion.

4.1 Clarifying Statements on Necessary vs. Sufficient Identifying Assumptions

GCBSGH*a* claims that KLM is "flawed," and suffers from a "mathematical error" (abstract),

but fails to identify any computational mistakes in derivations. Rather, the critique states that KLM erred in describing identifying assumptions as "necessary" when, GCBSGH*a* claim, the option of invoking "subset ignorability" means they were merely sufficient. This charge appears to hinge on a misreading of our work. GCBSGH*a*'s critique, as stated in their Footnote 3, claims that KLM assert Assumptions 2, 4, and 5[9] in KLM are necessary to point identify the $\text{CDE}_{M=1}$—a quantity that KLM says "makes little sense" in policing given its physically impossible counterfactuals (KLM Appendix p. 5), but is briefly referred to in passing asides and one appendix. We are unable to find a specific textual basis for this critique in KLM. Rather, in Appendix A.3, KLM shows that, *conditional on Assumptions 1–4*, Assumption 5 is necessary to point identify this quantity.

The imprecision in this critique not withstanding, we acknowledge that at various points, KLM uses the term "necessary" to describe identifying assumptions in ways that, if read out of context, may be misleading. For example, KLM prefaces its statement of Assumptions 1–4 by stating, "Without these assumptions, causal quantities of interest in this substantive area cannot be identified in data" (p. 7). This bears clarification. Taken in isolation, this language is of course imprecise—analysts are free to assert all manner of assumptions to render quantities identifiable. The key question, as always, is whether assumptions are credible. We regret any confusion this language may have caused.

However, the extensive surrounding discussion makes clear the inferential goals for which these assumptions *are* required, and that our assertions hold "*except in the implausible edge cases described in the Online Appendix*" (p. 628, emphasis added), like exact cross-principal-strata balancing in potential outcomes (p. 626 and Appendix p. 6–7). In other words, read in context, our claim was that certain identifying assumptions were needed *if analysts wish to obtain informative bounds* on the severity of racial bias in the use of force: (i) using "only data on stopped individuals" (p. 620); (ii) when there exists "racial bias in stops" (p. 627), $D \rightarrow M$, *and* "unobserved subjective aspects" (p. 623) that are common causes of detainment and the use of force, $M \leftarrow U \rightarrow Y$; and (iii) without appealing to untenable assumptions or "implausible edge cases" (p. 628)—conditions which are explicitly and prominently outlined in KLM. These conditions reflect what we believe, for strong substantive reasons, is a close approximation to the reality of police-civilian interactions. Assumptions 1–4 in KLM allow analysts to not only sign the statistical bias of standard approaches, but also construct nonparametric sharp bounds, allowing for credible and informative conclusions—"focus[ing] on. . . average treatment [total] effects" (p. 625).[10] We are

[9] These assumptions are mediator monotonicity, treatment ignorability, and mediator-outcome ignorability, respectively. We refer readers to KLM for detailed discussions of each assumption.

[10] KLM also discusses potential methods of data collection and improved research design should analysts find Assumptions 1–4 implausible.

unaware of alternatives for informatively bounding causal effects under these circumstances that rest on weaker assumptions, and as we demonstrated above, GCBSGH*a* supplies no such alternative.

4.2 Claimed Counterexamples Mirror Previously Stated Edge Cases

In their original paper and in a series of amendments and revisions, GCBSGH also propose several claimed counterexamples that purportedly invalidate the approach of KLM. However, upon closer inspection, it can be seen that every proposed counterexample in GCBSGH*a*, GCBSGH*b*, and GCBSGH*d* merely mirrors arguments and edge cases from KLM. In other words, these scenarios—despite being presented as critiques—simply echo the same scenarios that KLM considered but rejected due to their implausibility.

The earliest claimed counterexample appeared in GCBSGH*a*, comprising the entirety of Section 3 there: a toy example that entirely omitted the U node in Figure 1 (though this crucial omission was not emphasized for the reader). GCBSGH demonstrated by simulation that the $\text{CDE}_{M=1}$ could, in this case, be estimated without bias. However, this no-U scenario mirrors p. 625 of KLM: "We show that [the $\text{CDE}_{M=1}$] cannot be recovered in this setting unless analysts make the untenable assumption that no mediator-outcome confounding exists," where "no mediator-outcome confounding" refers to the absence of $U \to M$, $U \to Y$, or both (like in KLM Assumption 5, already rejected as untenable). In fact, Figure 3 of KLM considered two possible graphs containing no U. However, KLM continued,

> We find mediator ignorability to be highly implausible in the context of policing. Subjective factors such as an officer's suspicion and sense of threat—depicted as U in Figure 3(c)—can not only lead to investigation (stopping) but also a heightened willingness to use force. These mediator-outcome confounders must be captured in X for this assumption to hold, but they are notoriously difficult to capture in officers' self-reported accounts (p. 626).

After we alerted GCBSGH to this issue, a second claimed counterexample was developed in GCBSGH*b*. However, a close examination of Eq. 2 in that paper revealed the new counterexample had been constructed in a way that was later acknowledged to "not capture the $[D] \to M$ dependence" (GCBSGH*c*), though this important design decision was also not initially conveyed. To reiterate: the procedure of GCBSGH*b* effectively assumed away police discrimination in stops, *in a study of police discrimination* (much like in KLM Assumption 6, also rejected as untenable). The potential discrimination encoded in $D \to M$ is so foundational to KLM that it was addressed in the very second sentence of the problem statement: "police-civilian encounters inherently involve a mediating variable that may be affected by

race: whether an individual is stopped by police" (p. 623). Shortly after we publicly disclosed this design decision, the second counterexample was retracted in GCBSGH*c*.

It has been well known since at least Pearl (1995) that the sort of knife-edge balancing conditions described above will be trivially satisfied by estimation approaches that achieve *d*-separation for a data-generating process. If analysts can guarantee that any of the $D \rightarrow M$, $U \rightarrow M$, or $U \rightarrow Y$ dependencies are nonexistent—thereby eliminating collider bias, the technical source of post-treatment selection bias in this setting—unbiasedness will directly follow. *Design*-based approaches that break one of these dependencies by randomized interventions are thus a credible way to achieve the knife-edge conditions required for unbiased inference. In the examples from GCBSGH*a* and GCBSGH*b*, the naïve estimator works precisely for the reasons described in KLM: because it mimics these design-based approaches. However, the present task is to make credible inferences in an observational setting. In such circumstances, *assumption*-based approaches are simply not a credible route to ensuring the nonexistence of these causal channels.

Finally, GCBSGH*d* produces yet another claimed counterexample, with numeric values precisely constructed to satisfy the knife-edge condition of Proposition 2. Like its two predecessors, this third attempt is merely another edge case contained in the detailed bias derivations in KLM Appendices A.1 and A.3, and it rests entirely on knife-edge cancellation. Appendix A describes in detail how this edge case was covered in KLM, along with the exact mapping between GCBSGH*a*'s critique of KLM's use of the term "necessary" and the accidentally cancelling omitted variable bias illustration of Section 3.2.

In sum, all the supposed counterexamples offered by GCBSGH hinge on implausible conditions that KLM considered, but explicitly rejected. GCBSGH are of course free to advocate for whatever alternative identifying assumptions they prefer—however untenable—but their claims that KLM rests on a "flawed" framework or "mathematical error" are unsupported.

4.3 The Need to Scrutinize New Methods, Including Ours

Despite our intellectual disagreements, we are grateful for GCBSGH*a*'s work in clarifying the theoretical underpinnings of research using what KLM refers to as the "naïve" estimator. After carefully probing the newly developed theory, and weighing the plausibility of the knife-edge assumptions outlined in Propositions 1 and 2, we stand by our original assertion: "existing empirical work in this area is producing a misleading portrait of evidence as to the severity of racial bias in police behavior" (KLM, p. 620). We nonetheless thank GCBSGH*a* for their role in bringing attention to the need for continued statistical innovation in this

critically important policy arena. Close scrutiny of newly proposed methods is always needed, and it is all the more necessary when bearing on life-and-death issues like racial bias in policing.

We are also thankful that other scholars have continued to press on open methodological questions, and we refer interested readers to this recent work. Zhao et al. (n.d.) conducts a thorough examination of causal estimands in KLM, showing that it may be difficult to extrapolate from the $\text{ATE}_{M=1}$ and $\text{ATT}_{M=1}$ to the ATE; it develops an approach to estimate risk ratios that sidestep problems relating to the unknown magnitude of $\Pr(M_i = 1)$. Clark et al. (2020) devises a formal theory of racial bias in police-involved shootings and tests it using an analytic result from KLM proving that so-called "outcome tests" (Becker, 1957) in fact imply a lower bound on racial discrimination in intermediate events. In addition, Humphreys (n.d.) explores how relaxing or modifying the substantively motivated assumptions in KLM has implications for the nature and severity of selection bias in this setting. As these continued innovations make clear, KLM is by no means the last word on racial bias in policing. Further research on these issues, as well as other difficult open questions such as the role of treatment ignorability, is needed to make progress.

5 Conclusion

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. Much work still opts to ignore these challenges. But 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 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.

A KLM Appendices A.1 & A.3 Included Proposition 1 & 2 Statements

In this appendix, we show that KLM characterized the full set of conditions under which the naïve estimator is an unbiased estimator of the $\text{CDE}_{M=1}$. This general bias result, an early step in the KLM Appendix A.3 derivations, is shown below. 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, $\frac{\gamma_{(1)}\alpha_{(1)}}{\alpha_{(1)}^2+\alpha_{(2)}^2+1}+\frac{\gamma_{(2)}\alpha_{(2)}}{\alpha_{(1)}^2+\alpha_{(2)}^2+1}$. The approach advocated by GCBSGH*a*, in as-if experimental settings, is to assume 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 the 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 GCBSGH*a*'s advocated "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 our more general bias derivation. We now reexamine that derivation in depth. For clarity of exposition, we will implicitly condition on $X_i = x$ throughout and drop the distinction made in Appendix A.3 of KLM between the $\text{CDE}_{M=1,x}$ and the $\text{CDE}_{M=1}$. (Aggregating over the former easily recovers the latter.)

Appendix A.3 of KLM 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 $X_i = x$. We write, "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,

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

per KLM 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 $\alpha = 0$.[11] If treatment ignorability holds, this

[11] 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.

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, the GCBSGH*a* "subset ignorability" assumption is logically equivalent to the assumption that $\alpha = \omega = 0$. As we showed, the naïve estimator is unbiased for the $\text{CDE}_{M=1}$ when this holds and treatment ignorability is satisfied.

KLM does not remark on this point because it is scarcely worth noting 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, KLM also did not remark on the fact that bias can be zero when $\alpha = -\omega$.) The remainder of the derivation in Appendix A.3 expands on these opening steps to characterize, substantively, how this bias contaminates causal inferences.

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

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

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

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

We now proceed to decompose the total bias:

$$\text{Total Bias} \;=\; \text{PTB} + \text{additional bias}$$

$$\text{Total Bias} - \text{PTB} \;=$$

$$\left.\begin{array}{r}\Big\{\mathbb{E}[Y_i(1,1)|D_i=1,M_i(1)=1,M_i(0)=1]-\mathbb{E}[Y_i(1,1)|D_i=0,M_i(1)=1,M_i(0)=1]\Big\}\\ \Pr(M_i(0)=1|D_i=1,M_i(D_i)=1)\Pr(D_i=0|M_i(D_i)=1)\end{array}\right\}\;\begin{array}{r}\alpha-\alpha_{\text{PTB}}\\ =\alpha_{\text{OVB}}\end{array}$$

$$+\left.\begin{array}{r}\Big\{\mathbb{E}[Y_i(0,1)|D_i=1,M_i(1)=1,M_i(0)=1]-\mathbb{E}[Y_i(0,1)|D_i=0,M_i(1)=1,M_i(0)=1]\Big\}\\ \Pr(M_i(0)=1|D_i=1,M_i(D_i)=1)\Pr(D_i=1|M_i(D_i)=1)\end{array}\right\}\;\begin{array}{r}\omega-\omega_{\text{PTB}}\\ =\omega_{\text{OVB}}\end{array}$$

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. □

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.

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

Clark, Tom S., Elisha Cohen, Adam Glynn, Michael Leo Owens, Anna Gunderson and Kaylyn Jackson. 2020. "Are Police Racially Biased in the Decision to Shoot?" *Proceedings of the*

National Academy of Sciences . Working Paper. https://static1.squarespace.com/static/58d3d264893fc0bdd12db507/t/5ed6859f3c31fe420713f58b/1591117221040/Racial_Bias_in_Shootings.pdf.

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.

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. forthcoming. "A Response to Steven Durlauf and James Heckman." *Journal of Political Economy* . https://www.journals.uchicago.edu/doi/pdfplus/10.1086/710977?casa_token=Iq96FTyHUx0AAAAA:cde57PrCR0uMJz9iC8suepiI7pxbnxUsq3NGedUGfZxc8s-Jd4g_4kYNTuS_GcL-mBdsdEZ4eg.

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*a*. "Deconstructing Claims of Post-Treatment Bias in Observational Studies of Discrimination." 23 June 2020 working paper, https://5harad.com/papers/post-treatment-bias.pdf.

Gaebler, Johann, William Cai, Guillaume Basse, Ravi Shroff, Sharad Goel and Jennifer Hill. 2020*b*. "Under confounding, KLM conditions are not necessary for consistency (Version 1)." 25 June 2020 working paper, https://www.dropbox.com/s/u3zlkpjc2j0j5q7/klm-example-v1.pdf?dl=0.

Gaebler, Johann, William Cai, Guillaume Basse, Ravi Shroff, Sharad Goel and Jennifer Hill. 2020*c*. "Under confounding, KLM conditions are not necessary for consistency (Version 1 Correction)." 26 June 2020 working paper, https://5harad.com/papers/klm-example.pdf.

Gaebler, Johann, William Cai, Guillaume Basse, Ravi Shroff, Sharad Goel and Jennifer Hill. 2020*d*. "Under confounding, KLM conditions are not necessary for consistency (Version 2)." 26 June 2020 working paper, https://5harad.com/papers/klm-example-v2.pdf.

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.

Gelman, Andrew. 2020. "Statistical controversy on estimating racial bias in the criminal justice system." *Statistical Modeling, Causal Inference, and Social Science (Blog)* . Posted July 6, 2020, https://statmodeling.stat.columbia.edu/2020/07/06/statistical-controversy-on-racial-bias-in-the-criminal-justice-system/.

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* .

Humphreys, Macartan. n.d. "Ambiguous implications of collider bias for the estimation of discrimination.". https://macartan.github.io/i/notes/bias_both_ways.html.

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.

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.

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

Pearl, Judea. 1995. "Causal diagrams for empirical research." *Biometrika* 82(4):669–710.

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 non-randomized studies." *Journal of Educational Psychology* 66(5):688–701.

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

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

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. n.d. "A note on post-treatment selection in studying racially biased policing." 22 July 2020 working paper.