

Letter

A Note on Posttreatment Selection in Studying Racial Discrimination in Policing

QINGYUAN ZHAO *University of Cambridge, United Kingdom*

LUKE J KEELE *University of Pennsylvania, United States*

DYLAN S SMALL *University of Pennsylvania, United States*

MARSHALL M JOFFE *University of Pennsylvania, United States*

We discuss some causal estimands that are used to study racial discrimination in policing. A central challenge is that not all police–civilian encounters are recorded in administrative datasets and available to researchers. One possible solution is to consider the average causal effect of race conditional on the civilian already being detained by the police. We find that such an estimand can be quite different from the more familiar ones in causal inference and needs to be interpreted with caution. We propose using an estimand that is new for this context—the causal risk ratio, which has more transparent interpretation and requires weaker identification assumptions. We demonstrate this through a reanalysis of the NYPD Stop-and-Frisk dataset. Our reanalysis shows that the naive estimator that ignores the posttreatment selection in administrative records may severely underestimate the disparity in police violence between minorities and whites in these and similar data.

INTRODUCTION

Evidence of racial disparities in policing is an urgent and highly relevant policy question in empirical research. A growing number of studies have focused on this critical topic (Baumgartner, Epp, and Shoub 2018; Christiani et al. 2021; Eckhouse 2017; Edwards, Lee, and Esposito 2019; Epp and Erhardt 2020; Shoub et al. 2020). However, studies of racial disparities are fraught with methodological challenges (Goel, Rao, and Shroff 2016; Ridgeway 2006; Ridgeway and MacDonald 2009). Recent work by Knox, Lowe, and Mummolo (2020, hereafter KLM) provides important new results on the difficulties of learning about racial disparities in policing from administrative data. One key point made by KLM is that such investigations have an intrinsic selection bias because administrative records only contain those encounters in which civilians are detained. If there is racial discrimination in police detention in the first place, any naive


analysis using the administrative data may then suffer from potentially severe selection bias.

Here, we present a research note on this important topic with two purposes. First, KLM focused on several local causal estimands that are being used in the empirical studies. We demonstrate that these local estimands—even when identified with observational data—cannot be used to make inferences about more global effects like the average treatment effect. Second, we introduce a global causal risk ratio estimand that is straightforward to interpret and requires fewer assumptions to identify than either the local effects considered by KLM or global risk differences. Although it still depends on some quantities that need to be estimated from external data, we demonstrate how we can use Bayes' formula to avoid the hard problem of estimating the probability of detainment in police–civilian encounters. We conclude this research note with a reanalysis of the New York City Police Department (NYPD) Stop-and-Frisk dataset and some further discussion. Our empirical results show that a naive analysis of police administrative datasets that ignores the selection bias can severely underestimate the risk of police force for minorities. We present results that suggest a naive approach may understate the effect of civilian race on risk of police violence by a factor of 10 or more.

REVIEW

We begin with a brief review of the key quantities in KLM. Following their work, the unit of analysis is an encounter between civilians and police, where an encounter is defined as all events in which the police

Qingyuan Zhao , Statistical Laboratory, Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, United Kingdom, qyzhao@statslab.cam.ac.uk.

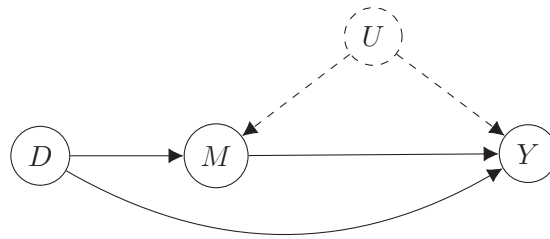
Luke J Keele , Department of Surgery, Perelman School of Medicine, University of Pennsylvania, United States, luke.keele@uphs.upenn.edu.

Dylan S Small, Department of Statistics, Wharton School, University of Pennsylvania, United States, dsmall@wharton.upenn.edu.

Marshall M Joffe, Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, United States, mjoffe@pennteam.upenn.edu.

Received: September 10, 2020; revised: February 08, 2021; accepted: June 14, 2021. First published online: July 26, 2021.

FIGURE 1. KLM’s Directed Acyclic Graph (DAG) Model for Racial Discrimination in Policing with an Unmeasured Mediator-Outcome Confounder U



Note: The treatment D is race of the civilian. The mediator M is an indicator for police detainment and the outcome Y is an indicator for police use of force. Administrative records only contain observations with $M = 1$.

sight a civilian, including those in which a civilian is allowed to pass undisturbed. There are n encounters indexed by $I = 1, \dots, n$. We denote the outcome with Y_i , where $Y_i = 1$ indicates the use of force by the police in encounter i . Next, D_i is a binary variable where $D_i = 1$ records the race of the civilian as a minority. While the race of the civilian is not manipulable, we adopt the approach in KLM where the counterfactual is the replacement of the civilian in an encounter with a separate, comparable civilian engaged in comparable behavior, but differing on race (Knox, Lowe, and Mummolo 2020, 621). We use M_i to indicate a police detainment or stop of a civilian. Critically, $M_i = 1$ for the subset of encounters that resulted in a stop by the police and are present in the administrative data. Finally, X_i represents a collection of covariates that describe aspects of the stops in the data. These could include measures for time of day, location, age, sex, and civilian behavior at the time when first encountered by police. Unless stated otherwise, conditioning on X is implicit.

For formal causal inference, we introduce the potential outcomes for M_i and Y_i . We have the potential mediator $M_i(d)$, which represents whether encounter i would have resulted in a stop if civilian race is d . Next, $Y_i(d, m)$ is the potential outcome for the use of force if race is d and the mediating variable is set to m ; similarly, $Y_i(d)$ is the potential outcome if race is d . Throughout this note we make the stable unit treatment assumption (SUTVA), so $M_i(D_i) = M_i$ and $Y_i(D_i, M_i) = Y_i(D_i) = Y_i$. This assumption means that the observed mediator (detainment) and outcome (use of force) are consistent with their corresponding counterfactual values. Hereafter, we assume the variables D_i, M_i , and Y_i and the potential outcomes of M_i and Y_i are drawn independently from the same unknown distribution. To simplify the exposition, we will drop the i subscript.

KLM studied the following “naive” treatment effect estimand:

$$\Delta = \mathbb{E}[Y|D = 1, M = 1] - \mathbb{E}[Y|D = 0, M = 1], \quad (1)$$

where \mathbb{E} denotes expectation over a random police-civilian encounter. Intuitively, Δ compares the average rates of force between different racial groups who are detained by police. KLM showed that, if there is racial discrimination in detainment and an unmeasured

confounder between detainment and use of force (see Figure 1), the naive treatment effect Δ can be quite misleading when used to represent the causal effect of race on police violence.

The key issue is that the structure of the data implies all estimates are conditional on M —a posttreatment variable, which often leads to biased estimators of the causal effect (Rosenbaum 1984). Bias of this type occurs in many applied problems in social science (Elwert and Winship 2014; Montgomery, Nyhan, and Torres 2018) and medicine (Paternoster, Tilling, and Davey Smith 2017).

Using the principal stratification framework of Frangakis and Rubin (2002), KLM showed that it is still possible to either identify or partially identify certain forms of average treatment effects using a set of tailored causal assumptions. These assumptions include mandatory reporting, mediator monotonicity, and treatment ignorability. Specifically, KLM derived nonparametric bounds for the average treatment effect of race on use of force among those who are detained by the police:

$$ATE_{M=1} = \mathbb{E}[Y(1) - Y(0) | M = 1].$$

They also derived a point identification formula for the average treatment effect among those who are minorities and detained by the police:

$$ATT_{M=1} = \mathbb{E}[Y(1) - Y(0) | D = 1, M = 1].$$

Notice that their results rely on an external estimate of the proportion of racially motivated detainments among all reported minority detainments—that is, $\mathbb{P}(M(0) = 0 | D = 1, M = 1)$. See KLM (631) for a discussion on estimating this quantity. Moreover, KLM also derived an identification formula for the average treatment effect $ATE = \mathbb{E}[Y(1) - Y(0)]$ given external estimates of the rate of detainments $\mathbb{P}(M = 1 | D = d)$ by race $d = 0, 1$.

The identification results in KLM depend crucially on the following assumption:

Assumption 1 (Mandatory reporting). (i) $Y(0, 0) = Y(1, 0) = 0$ and (ii) the administrative data contains all detainments/stops of civilians by the police.

The first part of this assumption assumes that there will be no police violence if the civilian is not stopped in

the first place. The second part assumes we observe a sample from the conditional distribution of the variables given $M = 1$, which is essential for statistical inference. We will make Assumption 1 throughout this note and further discuss its practical implications before the real data analysis.

AVERAGE TREATMENT EFFECTS CONDITIONAL ON THE MEDIATOR

In many causal analyses, investigators are focused on the sample average treatment effect (ATE), which is the average difference in potential outcomes averaged over the study population. At times, researchers define the ATE over specific subpopulations, which makes the ATE more local; for example, the average treatment effect might be defined for the subpopulation exposed to the treatment or the average treatment effect on the treated (ATT). Often the “global” ATE is the goal in many studies and is preferred over more local effects (Gerber and Green 2012, chap. 2). For example, IV studies have been strongly critiqued for identifying a local average treatment effect (LATE) instead of the global ATE (Deaton 2010; Swanson and Hernán 2014). Moreover, even some defenders of IV studies view the LATE as a “second choice” estimand compared with the global ATE (Imbens 2014).

As KLM outlined, the global ATE has not generally been the target causal estimand in this literature. Instead, researchers have focused on $ATE_{M=1}$ and $ATT_{M=1}$ which are both conditional on the mediator M . Notice that these estimands not only are more local than the global ATE but also condition on a posttreatment quantity. Nonetheless, they are not the first estimands in causal inference that condition on posttreatment quantities. Other examples of estimands that condition on posttreatment quantities include the survivor average treatment effect in Frangakis and Rubin (2002) (though conceptually the always survivor principal stratum can be thought as a pretreatment variable), effect modification by a posttreatment quantity (Ertefaie et al. 2018; Stephens, Keele, and Joffe 2016), and the probability of causation $\mathbb{P}[Y(0) = 0 | D = 1, Y = 1]$ (Dawid, Musio, and Murtas 2017; Pearl 1999; Robins and Greenland 1989).

The local effects in this context may have important policy relevance. As such, the preference for a global ATE may not always be warranted in this domain. However, an inexperienced researcher might think these local estimands are informative about the global ATE or even an estimand such as the controlled direct effect: $\mathbb{E}[Y(1,1) - Y(0,1)]$. Next, we build upon the population stratification framework in KLM and clarify the difference between the conditional estimands in KLM and estimands like the global ATE.

To simplify the illustration, we will consider the case where there is no mediator-outcome confounder (i.e., no variable U in the diagram in Figure 1). The issues we describe below will still occur if there is mediator-outcome confounding. In mediation analysis,

a standard way to decompose the average treatment effect is

$$\begin{aligned} \text{ATE} &= \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[Y(1, M(1)) - Y(1, M(0))] \\ &\quad + \mathbb{E}[Y(1, M(0)) - Y(0, M(0))]. \end{aligned}$$

The two terms on the right-hand side are called the pure indirect effect (PIE) and pure direct effect (PDE; Robins and Greenland 1992). Under the nonparametric structural equation model with independent errors model (Pearl 2009; Richardson and Robins 2013) and Assumption 1, they can be expressed as (See the Appendix section A)

$$\text{PIE} = \beta_M \cdot \mathbb{E}[Y(1, 1)], \text{PDE} = \beta_Y \cdot \mathbb{E}[M(0)],$$

where $\beta_M = \mathbb{E}[M(1) - M(0)]$ is the average effect of race on detainment and $\beta_Y = \mathbb{E}[Y(1, 1) - Y(0, 1)]$ is the controlled direct effect of race on police violence. An immediate consequence of the above expressions is that

$$\text{ATE} \geq 0 \text{ if } \beta_M, \beta_Y \geq 0 \text{ and } \text{ATE} \leq 0 \text{ if } \beta_M, \beta_Y \leq 0. \quad (2)$$

In words, the global ATE is nonnegative whenever both the direct and indirect effects are nonnegative, and vice versa. This property also holds for the ATT because in the simple setting here the treatment D is completely randomized.

In the Appendix, we use principal stratification to show that neither $ATE_{M=1}$ or $ATT_{M=1}$ is guaranteed to inherit the sign of β_M and β_Y and satisfy the property in Equation 2. Specifically, we outline concrete examples in which

- (i) The pure direct and indirect effects are both positive, but $ATE_{M=1} < 0$;
- (ii) The pure direct and indirect effects are both negative, but $ATE_{M=1} > 0$ and $ATT_{M=1} > 0$.

That is, when there is racial discrimination of the same direction in both police detainment and the use of force, it is still possible for $ATE_{M=1}$ and $ATT_{M=1}$ to have the opposite sign. We refer the reader to the Appendix for some concrete counterexamples and further comments on this phenomenon.

In sum, the local estimands $ATE_{M=1}$ and $ATT_{M=1}$ are generally different from the global estimands that are routinely the target in causal analyses. As such, we urge applied researchers to use caution when using these local estimands to infer anything about the global estimands.

A NEW ESTIMATOR FOR THE CAUSAL RISK RATIO

KLM also derived an identification formula for $ATE_{M=1}$ using external estimates of the rate of detainment $\mathbb{P}(M = 1 | D = d)$ for race $d = 0, 1$. Unfortunately, it is often difficult to quantify the frequency of stops among all police-civilian encounters, as noted in their paper. In particular, it can be difficult to determine the magnitude of $\mathbb{P}(M = 1 | D = d)$. Here, we show that by

formulating the estimand on a relative scale, we can avoid this difficulty and obtain point identification.

More specifically, we consider the following causal risk ratio (CRR) for covariate level x :

$$\text{CRR}(x) = \frac{\mathbb{E}[Y(1)|X = x]}{\mathbb{E}[Y(0)|X = x]}.$$

When this term is equal to one, the risk of police violence does not vary with the race of the civilian. When this term is greater than one, the risk of violence is higher for minorities. Risk ratios, while not commonly used in political science, have been used in the literature on policing (Christiani et al. 2021; Eckhouse 2017; Edwards, Lee, and Esposito 2019). However, previous researchers that use risk ratios have tended to present them as descriptive values rather than as causal quantities. Moreover, risk ratios can be a powerful rhetorical tool for understanding discussions of racial disparities. In the context of police violence, it may be tempting to use the following ratio to measure racial disparities:

$$\text{Naive risk ratio} = \frac{\mathbb{E}[Y|D = 1, M = 1, X = x]}{\mathbb{E}[Y|D = 0, M = 1, X = x]}.$$

This quantity divides the rates of police violence experienced by minorities and nonminorities, given that they have the same covariate x and are detained by the police. We will see below that the naive risk ratio is generally not the same as the causal risk ratio due to conditioning on the colliding variable M (detainment); in fact, these two quantities can be drastically different.

Expressing results in a relative fashion can be an effective way of communication, especially when the risk of police violence is fairly low among a specific population. For example, let's say in one specific locale, the risk of police violence for Black residents is 0.01% and is 0.001% for white residents. The difference in these risks is obviously very small. However, in relative terms, the risk of police violence for Black residents is 10 times that for white residents. As such, even if the absolute risk is low, a large increase in relative risk is likely to be of significant interest.

Using treatment ignorability (i.e., the DAG model in Figure 1 conditional on X) and Assumption 1, the causal effect of race can be identified based on the decomposition

$$\mathbb{E}[Y(d)|X = x] = \mathbb{E}[Y|M = 1, D = d, X = x] \cdot \mathbb{P}(M = 1|D = d, X = x), \text{ for } d = 0, 1.$$

The same result is derived in KLM and forms the basis of their identification of the ATE. We simplify their proof in the Appendix and show that some of KLM's identification assumptions can be relaxed. Specifically, we can arrive at the same result without invoking mediator monotonicity and relative nonseverity of racial stops (Assumptions 2 and 3 in KLM).

By using Bayes formula for the last term on the right hand side (see the Appendix), we obtain the following identification result:

$$\text{CRR}(x) = \underbrace{\frac{\mathbb{E}[Y|D = 1, M = 1, X = x]}{\mathbb{E}[Y|D = 0, M = 1, X = x]}}_{\text{naive risk ratio}} \underbrace{\left/ \frac{\mathbb{P}(D = 1|M = 1, X = x)}{\mathbb{P}(D = 0|M = 1, X = x)} \right/}_{\text{bias factor}} \frac{\mathbb{P}(D = 1|X = x)}{\mathbb{P}(D = 0|X = x)}. \tag{3}$$

Therefore, by targeting the causal risk ratio, we are able to avoid the difficulties associated with estimating the absolute rate of detainment $\mathbb{P}(M = 1)$ through cancellation.

The first term on the right-hand side of Equation 3 is the naive risk ratio estimand conditional on baseline covariates. It is the risk ratio counterpart to the naive risk difference in Equation 1, and both of them ignore the possible bias from the selection process into the administrative data. The second term inside the curly brackets is a ratio of probability ratios. The first ratio of probabilities measures the relative probability of a detainment being with a minority conditional on covariate $X = x$, which can be estimated from the administrative data. The second ratio measures the relative probability (odds) of an encounter being with a minority conditional on covariate $X = x$, but these probabilities need to be approximated or bounded with a second data source. This ratio between the last two terms is thus an odds ratio that characterizes the bias of the naive estimator; for this reason, we call it the "bias factor." That is, if minorities are overrepresented in the administrative data, the bias factor corrects that overrepresentation and so increases the magnitude of the risk ratio. For example, if the probability of a detainment being with a minority is 0.8 in the administrative data and 0.25 in a random police-civilian encounter, the bias factor would be $(0.8/0.2) / (0.25/0.75) = 12$, which would increase the magnitude of the naive risk ratio when it is larger than 1. All the terms in Equation 3 can be estimated using generalized linear models (such as logistic regression) or more flexible models. Confidence intervals can be estimated using the bootstrap or the delta method.

Note that if we are willing to assume stochastic mediator monotonicity: $\mathbb{E}[M(1)|X = x] \geq \mathbb{E}[M(0)|X = x]$ (i.e., there is racial bias against the minority in detainment), the bias factor can indeed be lower bounded by 1. In this case, the naive risk ratio (first term on the right hand side of Equation 3) provides a lower bound for the causal risk ratio $\text{CRR}(x)$.

While the risk ratio estimand does avoid Assumptions 2 and 3 in KLM critical complications are still present. That is, the constraints that tend to arise from the use of two data sources remain a significant source of complexity. In particular, the administrative dataset can only be used to estimate the first two terms on the right hand side of Equation 3. We must find an additional data source that allows us to estimate the racial distribution conditional on the covariates $-\mathbb{P}(D = 1|X = x)$ and $\mathbb{P}(D = 0|X = x)$ —since the administrative data only contain those encounters where $M = 1$. However, secondary data sources tend to also contain data on stops rather than encounters

(sightings of civilians by the police). As such, typically, we use population level data on police stops to approximate encounter rates by racial group. To the extent that these quantities are proportional, the method will be accurate. However, to the extent that these quantities differ, the measure will be biased. Moreover, there may be measurement inconsistencies between the secondary data and the administrative data. This can be partly addressed by a sensitivity analysis; see the next section for an example. See also Knox and Mummolo (2020) for further discussion on the usage of external datasets in this context.

Take the NYPD database of police stops as an example. This data source was used in KLM and will be reanalyzed in the next section. For a second data source, we will use the Current Population Survey (CPS), which contains measures for race and also has geographic information that allows us to restrict the data to the metro area in the state of New York (which is larger than the five boroughs of New York City). However, The CPS does not contain any more fine-grained geographic identifiers or any measures of police encounters or stops. Another data source we will use is the Police-Public Contact Survey (PPCS) collected by the U.S. Department of Justice. However, PPCS is a national survey and geographic identifiers are not available to researchers. As such, if we use the PPCS, we can do little to measure the prevalence of police–minority interactions in New York City. Additionally, the PPCS collects data on police stops and not encounters. As such, we cannot measure rates of encounters with either data source.

In other settings such as traffic stops, one may use the “veil of darkness” test (Grogger and Ridgeway 2006) and use nighttime police stops in the same dataset to estimate the bias factor, as police are less likely to know the race of a motorist. However, this still requires the assumption that the racial distribution of motorists is the same during the day and at night. Moreover, data sources on encounters are exceedingly rare, and despite the limitations, as we show next, the results using the risk ratio with different data sources can still be useful and illuminate the probable bias in the naive estimator. They can also serve as the baseline of a sensitivity analysis.

We conclude this section, with a final comment on data constraints. Identification of the risk ratio estimand as well as those derived in KLM depend on mandatory reporting (Assumption 1). It is important to note that this assumption is both a restriction on potential outcomes and a feature of the data collection. The first part of the assumption says that the potential outcome $Y(d,m)$ is equal to 0 whenever $m = 0$. This assumption is reasonable because, besides inadvertent collateral damage, there should be virtually no police violence if the civilian is not stopped by the police in the first place. The second part of the assumption is needed so that we can use the administrative dataset to get the conditional distribution of (D,Y,X) given $M = 1$. For a given administrative data source, it is possible that some police stops are unrecorded. If that is the case, any analysis relying on

Assumption 1 needs to be interpreted with care. This is not a major concern in the NYPD dataset reanalyzed below, as all NYPD police officers are required to report all the stops.

A REANALYSIS OF THE NYPD STOP-AND-FRISK DATASET

We used the identification formula in Equation 3 to estimate the causal risk ratio using the NYPD “Stop-and-Frisk” dataset analyzed in Fryer (2019) and KLM. Specifically, we use the replication data from KLM. As such, we followed KLM’s preprocessing of the dataset, with the one exception that we removed all races other than Black and white. We also focused on all forms of force rather than estimate the effects for different types of force. We used CPS 2013 and PPCS 2011 data to estimate the third term in Equation 3. See the end of this section for a sensitivity analysis where we perturb the estimates from census data. Because PPCS does not contain a geographic identifier, we also used the racial distributions for different subsets of the PPCS data. Specifically, we used subgroups for those in the survey that experienced a motor vehicle stop, any other kind of police stop, and those in a large metro area. We further explored weighting the PPCS respondents by their reported number of face-to-face contacts with the police. Respondents with more than 30 reported contacts with the police were excluded in that analysis. See the Appendix section C for details on the exact survey items we used in this analysis. As we noted above, neither CPS or PPCSD records police–civilian encounters per our definition (sighting of civilians), so they can only be regarded as approximations of the actual racial distribution in encounters.

Table 1 reports the estimated risk ratios using different estimators and external datasets. Using the naive estimator—the first term in Equation 3, we find a modest causal effect: Black people have 29% higher risk of the police using force than white people. Recall that we can view this as lower bound on the true causal risk ratio if we are willing to assume stochastic mediator monotonicity (i.e., there is discrimination against Black civilians in police detentions on average). The estimator from Equation 3 that adjusts for the selection bias shows a very different picture. No matter which external dataset we used, the estimated risk ratio for Black versus white is always greater than 10.

The estimates in Table 1 did not condition on any covariate that confounds the effect of race on police use of force. In the Appendix section D, we report the results of a stratified analysis by age and gender of the civilian. The estimates are broadly consistent with those reported in Table 1, but it appears that female minorities have a much smaller risk ratio (less discriminated against) than male minorities. Age does not appear to be an important effect modifier.

Another potentially important confounder is the location of the police–civilian encounter. However, detailed geographic information is not available in CPS or PPCS. The NYPD currently has 77 precincts

TABLE 1. Estimates of the Causal Effect of Minority Race (Black) on Police Violence

External dataset	Estimated risk ratio	95% Confidence interval
Naive estimator—First term in Equation 3		
None	1.29	1.28–1.30
Adjusted for selection bias by using Equation 3		
CPS	13.6	12.8–14.3
PPCS	32.3	31.3–33.3
PPCS (MV Stop)	29.5	26.9–32.7
PPCS (Stop in public)	29.2	23.5–36.5
PPCS (Large metro)	16.7	15.4–18.4
PPCS*	31.1	27.9–34.7
PPCS* (Large metro)	19.9	14.2–29.0

Note: CPS is the Current Population Survey. PPCS is Police–Public Contact Survey. PPCS* is PPCS, with the respondents weighted by their reported number of face-to-face contacts with the police. MV Stop is the subset of survey respondents that has been the passenger in a motor vehicle that was stopped by the police. Large Metro is the subset that lives in a region with more than 1 million population. Confidence intervals were computed using the nonparametric bootstrap.

that are responsible for the law enforcement within a designated geographic area. Using census blocks and the 2010 census data, Keefe (2020) constructed a population breakdown for each NYPD precinct. This allows us to compare the proportion of Black residents (among Black and white residents) with the proportion of detainments of Black civilians in each precinct (Figure 2). It is evident from this figure that in most of the precincts, Black civilians make up less than half of the population but more than half of the detainment records. This shows that the bias factor in Equation 3 can be quite large in this problem.

By using the census data to estimate the last term in Equation 3, Figure 3 compares the naive risk ratio estimator and selection-adjusted risk ratio estimator for each precinct. The selection-adjusted estimates are almost always much larger except for three outliers—precincts 67 and 113, where Blacks account for more than 90% of the population, and precinct 22 (Central Park), where only 25 residents were recorded and the majority of police–civilian encounters were likely with nonresidents. It is likely that in these precincts, the residential distribution in the census data poorly approximate the racial distribution in police–civilian encounters because the civilians could be visitors from other precincts or anywhere else in the world. Most of the precincts with the highest estimated risk ratios are wealthy neighborhoods in Manhattan and Brooklyn. In several precincts, our method estimated that the risk of police use of force for Blacks is more than 30 times higher than the risk for whites. This may be due in part to increased suspicion of minorities in areas where their presence is not common. Finally, Figure 4a shows a strong negative correlation between the estimated

risk ratios and the percentage of Black residents in the precinct. This indicates that the racial discrimination in police use of force may be strongly moderated by characteristics of the geographic location such as the racial composition, affluence, and average crime rate of the neighborhood.

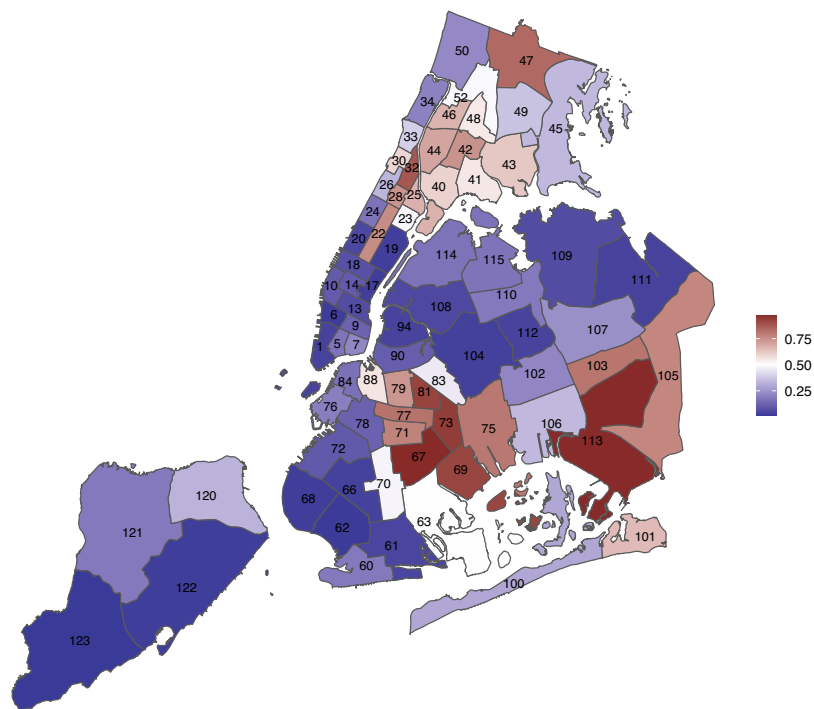
The above analysis relies on the assumption that the racial distribution in police–civilian encounters can be well approximated by the racial distribution in census or survey datasets. A sensitivity analysis can be useful to gauge the potential bias due to poor approximations of the racial distribution in police–civilian encounters. Figure 4b presents such a sensitivity analysis, where the civilians who encountered the police are assumed to be a mixture of local and citywide residents. More precisely, this sensitivity analysis assumes that in each precinct, there is a 90% chance of the police encountering a local resident and a 10% chance of the police encountering a resident from another precinct. According to the census data, 36.7% of the population in New York City (excluding races other than Black and white) was Black in 2010. Thus, in this sensitivity analysis, the presumed proportion of encounters with Black civilians is higher than the proportion of Black residents in the precinct, if the proportion of Black residents is lower than 36.7%. This shrinks the estimated causal risk ratio towards a common value, especially for precincts that are predominantly white or predominantly Black, as shown in Figure 4b.

CONCLUSIONS

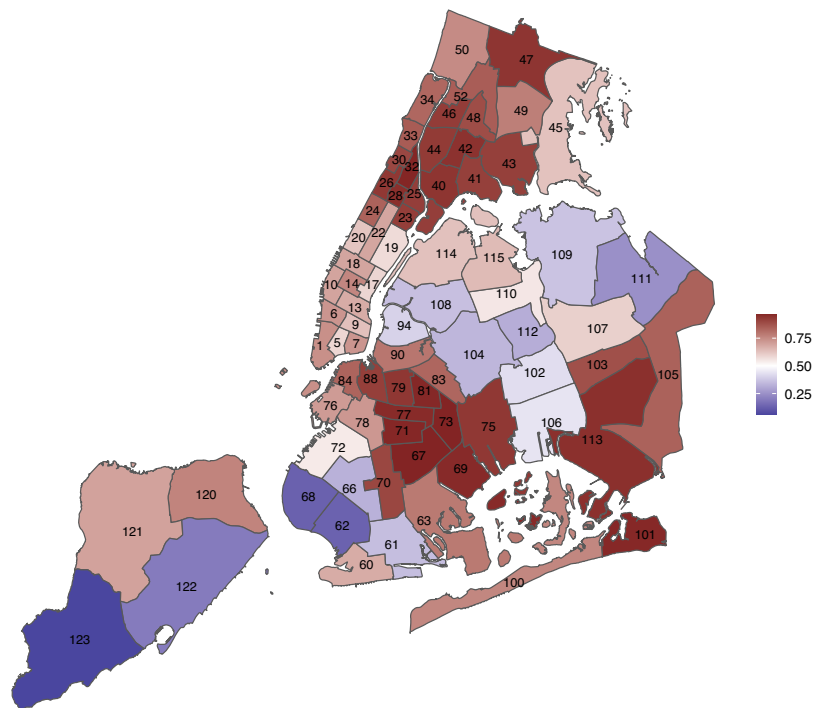
In this research note, we studied some causal estimands in the context of racial discrimination in policing. We found that the ATE that conditions on the mediator (police detainment) can differ in sign from the unconditional ATE and other routinely used causal estimands, so extra caution is needed when using these estimands and interpreting the results. We also proposed a new estimator for the causal risk ratio, which is straightforward to interpret and avoids the difficult task of discerning the percentage of stops in all police–civilian encounters. In a reanalysis of the NYPD Stop-and-Frisk dataset with causal risk ratio being the estimand, we found that for Blacks the risk of experiencing force is much higher than for whites.

When interpreting the results of our reanalysis, the reader should keep in mind its limitations. First, it is difficult to find a good external dataset to estimate the bias factor. The datasets we used should only be viewed as crude approximations to the racial distribution in police–civilian encounters in New York City. Second, our measure of the causal risk ratio is conditional on covariates X ; identification requires treatment ignorability conditional on confounders included in X . In principle, that would involve conditioning simultaneously on confounders like time, location, and other relevant characteristics of the police–civilian encounter. However, such covariates are not always available in external datasets and our analysis only conditions on

FIGURE 2. Racial Distributions (Indicated by the Filled Color) in Each NYPD Precinct



(a) Proportion of black residents in the census data.

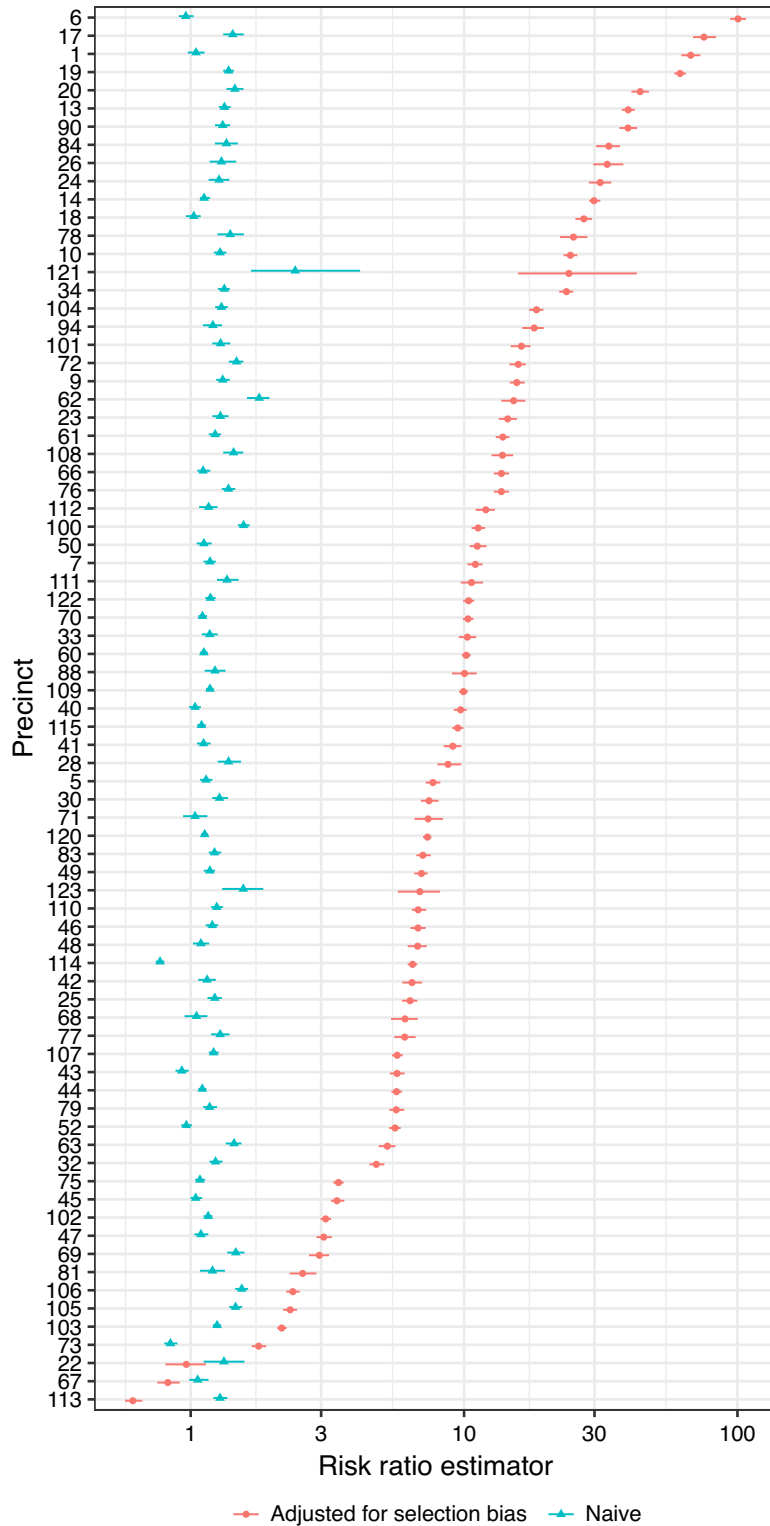


(b) Proportion of detainments of black civilians in the NYPD stop-and-frisk data.

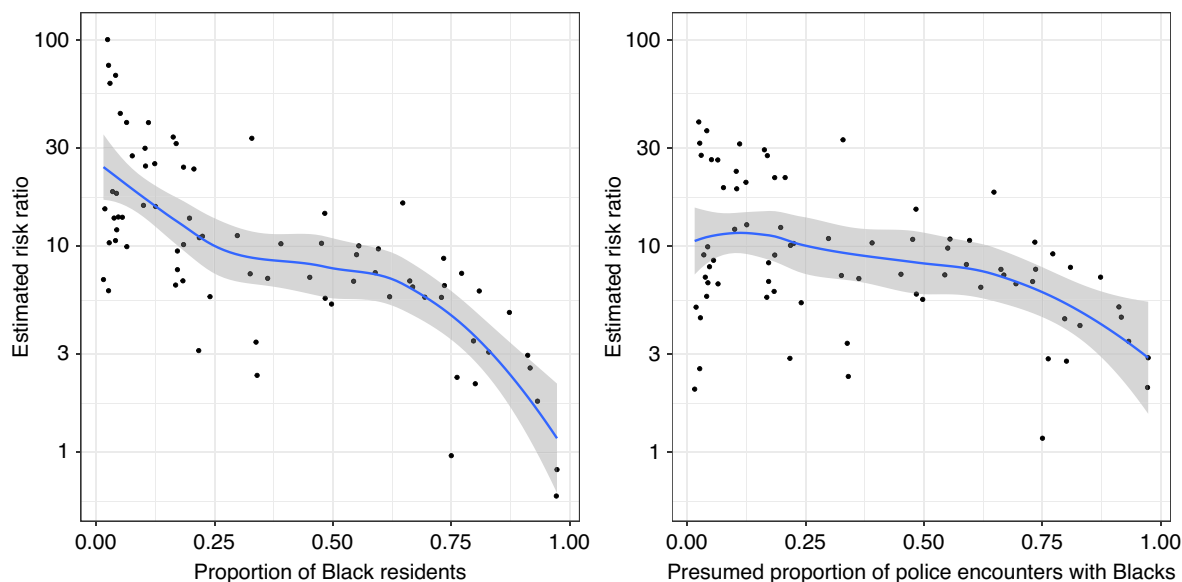
NYPD precinct. Additionally, our method does not yet have a way to summarize over multiple covariate strata even if the conditional risk ratios are identified and estimated. Since we did not use visible features of the

civilians that are associated with race and criminal activity (they are not available in the data), this may have led to overestimation of the effect of race on use of force. It is highly implausible that this bias could fully

FIGURE 3. Risk Ratio Estimates for Every NYPD Precinct



Note: The error bars correspond to 95% confidence intervals computed by the bootstrap. We did not resample the census data because that is already the residential distribution (instead of a statistical estimate). The blue estimates are obtained by using the naive estimator, the first term in Equation 3; the red estimates further take into account the bias factor due to sample selection in Equation 3.

FIGURE 4. Estimated Risk Ratio versus Proportion of Black Residents in Each Precinct

(a) Estimated risk ratio versus proportion of Black residents in each precinct.

(b) Estimated risk ratio in a sensitivity analysis versus proportion of Black residents. In each precinct, we assume the police encounters a mixture of 90% local residents in the precinct and 10% city-wide residents.

explain the large measures of association found here. Finally, since New York is a metropolitan in which people move around a great deal on a daily basis, the racial distribution of the residents in a precinct might poorly represent the racial distribution in police–civilian encounters, especially when the residential distribution is extreme, as demonstrated in our sensitivity analysis. In other words, Figure 4a may have exaggerated the effect modification by the racial distribution of the local residents. A further analysis on carefully selected precincts (e.g., residential areas with different racial compositions) is needed to better quantify the effect modification.

Nevertheless, our empirical results show that a naive analysis of police administrative datasets that ignores the selection bias can severely underestimate the risk of police force for minorities. This also highlights the importance of defining the causal estimand clearly in observational studies. Further careful analyses are needed to better quantify the racial discrimination in policing and understand the socioeconomic factors that moderate racial discrimination.

Finally, we offer a concrete suggestion for applied analysts based on our results. KLM conclude by outlining a feasible research design for policing studies. Our risk-ratio-based analysis and the associated sensitivity analysis are useful additions to their suggested research plan. Our methods provide useful complements to the analyses outlined by KLM. Any policing study will depend on strong assumptions and a broad set of results that agree will provide higher quality evidence.

SUPPLEMENTARY MATERIALS

To view supplementary material for this article, please visit <http://dx.doi.org/10.1017/S0003055421000654>.

DATA AVAILABILITY STATEMENT

Research data that support the findings of this study are openly available at the American Political Science Review Dataverse: <https://doi.org/10.7910/DVN/ZQMYII>.

ACKNOWLEDGMENTS

The authors thank Dean Knox, Joshua Loftus, Jonathan Mummolo, and four anonymous reviewers for their helpful suggestions.

CONFLICT OF INTEREST

The authors declare no ethical issues or conflicts of interest in this research.

ETHICAL STANDARDS

The authors affirm this research did not involve human participants.

REFERENCES

- Baumgartner, Frank R., Derek A. Epp, and Kelsey Shoub. 2018. *Suspect Citizens: What 20 Million Traffic Stops Tell Us about Policing and Race*. Cambridge: Cambridge University Press.
- Christiani, Leah, Kelsey Shoub, Frank R. Baumgartner, Derek A. Epp, and Kevin Roach. 2021. "Better for Everyone: Black Descriptive Representation and Police Traffic Stops." *Politics, Groups, and Identities* <https://doi.org/10.1080/21565503.2021.1892782>.
- Dawid, A. Philip, Monica Musio, and Rossella Murtas. 2017. "The Probability of Causation." *Law, Probability and Risk* 16 (4): 163–79.
- Deaton, Angus. 2010. "Instruments, Randomization, and Learning about Development." *Journal of Economic Literature* 48 (2): 424–55.
- Eckhouse, Laurel. 2017. "Descriptive Representation and Equal Access to the Law: Race, Representation, and Crime Policy in America." PhD diss. University of California, Berkeley.
- Edwards, Frank, Hedwig Lee, and Michael Esposito. 2019. "Risk of Being Killed by Police Use of Force in the United States by Age, Race–Ethnicity, and Sex." *Proceedings of the National Academy of Sciences* 116 (34): 16793–798.
- Elwert, Felix, and Christopher Winship. 2014. "Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable." *Annual Review of Sociology* 40: 31–53.
- Epp, Derek A., and Macey Erhardt. 2020. "The Use and Effectiveness of Investigative Police Stops." *Politics, Groups, and Identities* <https://doi.org/10.1080/21565503.2020.1724160>.
- Ertefaie, Ashkan, Jesse Y. Hsu, Lindsay C. Page, and Dylan S. Small. 2018. "Discovering Treatment Effect Heterogeneity through Post-Treatment Variables with Application to the Effect of Class Size on Mathematics Scores." *Journal of the Royal Statistical Society Series C* 67 (4): 917–38.
- Frangakis, Constantine E., and Donald B. Rubin. 2002. "Principal Stratification in Causal Inference." *Biometrics* 58 (1): 21–9.
- Fryer, Roland G., Jr. 2019. "An Empirical Analysis of Racial Differences in Police Use of Force." *Journal of Political Economy* 127 (3): 1210–61.
- Gerber, Alan S., and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. New York: Norton.
- Goel, Sharad, Justin M. Rao, and Ravi Shroff. 2016. "Precinct or Prejudice? Understanding Racial Disparities in New York City's Stop-and-Frisk Policy." *The Annals of Applied Statistics* 10 (1): 365–94.
- Grogger, Jeffrey, and Greg Ridgeway. 2006. "Testing for Racial Profiling in Traffic Stops from behind a Veil of Darkness." *Journal of the American Statistical Association* 101 (475): 878–87.
- Imbens, Guido. 2014. "Instrumental Variables: An Econometrician's Perspective." *Statistical Science* 29 (3): 323–58.
- Keefe, John. 2020. "Sharing NYC Police Precinct Data." Retrieved: August 31, 2020. <https://johnkeefe.net/nyc-police-precinct-and-census-data>.
- Knox, Dean, and Jonathan Mummolo. 2020. "Toward a General Causal Framework for the Study of Racial Bias in Policing." *Journal of Political Institutions and Political Economy* 1 (3): 341–78.
- Knox, Dean, Will Lowe, and Jonathan Mummolo. 2020. "Administrative Records Mask Racially Biased Policing." *American Political Science Review* 114 (3): 619–37.
- Montgomery, Jacob M., Brendan Nyhan, and Michelle Torres. 2018. "How Conditioning on Posttreatment Variables Can Ruin Your Experiment and What to Do about It." *American Journal of Political Science* 62 (3): 760–75.
- Paternoster, Lavinia, Kate Tilling, and George Davey Smith. 2017. "Genetic Epidemiology and Mendelian Randomization for Informing Disease Therapeutics: Conceptual and Methodological Challenges." *PLoS Genetics* 13 (10): e1006944.
- Pearl, Judea. 1999. "Probabilities of Causation: Three Counterfactual Interpretations and Their Identification." *Synthese* 121 (1–2): 93–149.
- Pearl, Judea. 2009. *Causality: Models, Reasoning and Inference*. Cambridge: Cambridge University Press.
- Richardson, Thomas S., and James M. Robins. 2013. "Single World Intervention Graphs (SWIGs): A Unification of the Counterfactual and Graphical Approaches to Causality." Technical Report 128. Seattle, WA: Center for the Statistics and the Social Sciences, University of Washington.
- Ridgeway, Greg. 2006. "Assessing the Effect of Race Bias in Post-Traffic Stop Outcomes Using Propensity Scores." *Journal of Quantitative Criminology* 22 (1): 1–29.
- Ridgeway, Greg, and John M. MacDonald. 2009. "Doubly Robust Internal Benchmarking and False Discovery Rates for Detecting Racial Bias in Police Stops." *Journal of the American Statistical Association* 104 (486): 661–68.
- Robins, James, and Sander Greenland. 1989. "The Probability of Causation under a Stochastic Model for Individual Risk." *Biometrics* 45 (4): 1125–38.
- Robins, James M., and Sander Greenland. 1992. "Identifiability and Exchangeability for Direct and Indirect Effects." *Epidemiology* 3 (2): 143–55.
- 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: Series A (General)* 147 (5): 656–66.
- Shoub, Kelsey, Derek A. Epp, Frank R. Baumgartner, Leah Christiani, and Kevin Roach. 2020. "Race, Place, and Context: The Persistence of Race Effects in Traffic Stop Outcomes in the Face of Situational, Demographic, and Political Controls." *Journal of Race, Ethnicity and Politics* 5 (3): 481–508.
- Stephens, Alisa, Luke J. Keele, and Marshall Joffe. 2016. "Generalized Structural Mean Models for Evaluating Depression as a Post-Treatment Effect Modifier of a Jobs Training Intervention." *Journal of Causal Inference* 4 (2): 20150032. <https://doi.org/10.1515/jci-2015-0032>.
- Swanson, Sonja A., and Miguel A. Hernán. 2014. "Think Globally, Act Globally: An Epidemiologist's Perspective on Instrumental Variable Estimation." *Statistical Science* 29 (3): 371–74.

APPENDIX

A NOTE ON POSTTREATMENT SELECTION IN STUDYING RACIAL DISCRIMINATION IN POLICING

A Average Treatment Effects Conditional on the Mediator

We assume the variables (D, M, Y) are generated from a nonparametric structural equation model: $D = f_D(c_D)$, $M = f_M(D, c_M)$, $Y = f_Y(D, M, c_Y)$, where c_D, c_M, c_Y are mutually independent (Pearl 2009). Potential outcomes for M and Y can be defined by replacing random variables in the functions by fixed values; for example, $M(d) = f_M(d, c_M)$, $d = 0, 1$. Because the errors are independent, D , $\{M(0), M(1)\}$, and $\{Y(0,0), Y(0,1), Y(1,0), Y(1,1)\}$ are mutually independent (Richardson and Robins 2013). We also make the mandatory assumption (Assumption 1). The derivations below do not need mediator monotonicity ($M(1) \geq M(0)$).

We next derive expressions of $ATE_{M=1}$ and $ATT_{M=1}$ using two basic causal effects: $\beta_M = \mathbb{E}[M(1) - M(0)]$, the racial bias in detainment, and $\beta_Y = \mathbb{E}[Y(1, 1) - Y(0, 1)]$, the controlled direct effect of race on police violence. To simplify the interpretation, we introduce a new variable to denote the the principal stratum (see Figure 2 in KLM):

$$S = \begin{cases} \text{always stop (al)}, & \text{if } M(0) = M(1) = 1, \\ \text{minority stop (mi)}, & \text{if } M(0) = 0, M(1) = 1, \\ \text{majority stop (ma)}, & \text{if } M(0) = 1, M(1) = 0, \\ \text{never stop (ne)}, & \text{if } M(0) = M(1) = 0, \end{cases}$$

Let $S = \{\text{al}, \text{mi}, \text{ma}, \text{ne}\}$ be all possible values for S . Using this notation, we have

$$\beta_M = \sum_{s \in S} \mathbb{E}[M(1) - M(0) | S = s] \mathbb{P}(S = s) = \mathbb{P}(S = \text{mi}) - \mathbb{P}(S = \text{ma}).$$

By using the independence between $M(d)$ and $Y(d, m)$ and assumption:m0y0, it is easy to show that

$$\theta = \begin{pmatrix} \mathbb{E}[Y(1) - Y(0) | S = \text{al}] \\ \mathbb{E}[Y(1) - Y(0) | S = \text{mi}] \\ \mathbb{E}[Y(1) - Y(0) | S = \text{ma}] \\ \mathbb{E}[Y(1) - Y(0) | S = \text{ne}] \end{pmatrix} = \begin{pmatrix} \mathbb{E}[Y(1, 1) - Y(0, 1)] \\ \mathbb{E}[Y(1, 1) - Y(0, 0)] \\ \mathbb{E}[Y(1, 0) - Y(0, 1)] \\ \mathbb{E}[Y(1, 0) - Y(0, 0)] \end{pmatrix} = \begin{pmatrix} \beta_Y \\ \beta_Y + \mathbb{E}[Y(0, 1)] \\ -\mathbb{E}[Y(0, 1)] \\ 0 \end{pmatrix}.$$

Average treatment effects, whether conditional on M or D or not, can be written as weighted averages of the entries of θ .

Proposition 1. *Suppose there is no unmeasured mediator-outcome confounder (i.e., no U) in Figure 1. Under Assumption 1, the estimands $ATE_{M=1}$, $ATT_{M=1}$, $ATE = \mathbb{E}[Y(1) - Y(0)]$, and $ATT = \mathbb{E}[Y(1) - Y(0) | D = 1]$ can be written as weighted averages ($\mathbf{w}^T \theta$) / ($\mathbf{w}^T \mathbf{1}$) ($\mathbf{1}$ is the all-ones vector) with weights given by, respectively,*

$$\mathbf{w}(ATE_{M=1}) = \begin{pmatrix} \mathbb{P}(S = \text{al}) \\ [\mathbb{P}(S = \text{ma}) + \beta_M] \mathbb{P}(D = 1) \\ \mathbb{P}(S = \text{ma}) \mathbb{P}(D = 0) \\ 0 \end{pmatrix}, \mathbf{w}(ATT_{M=1}) = \begin{pmatrix} \mathbb{P}(S = \text{al}) \\ \mathbb{P}(S = \text{ma}) + \beta_M \\ 0 \\ 0 \end{pmatrix},$$

and

$$\mathbf{w}(ATE) = \mathbf{w}(ATT) = \begin{pmatrix} \mathbb{P}(S = \text{al}) \\ \mathbb{P}(S = \text{mi}) \\ \mathbb{P}(S = \text{ma}) \\ \mathbb{P}(S = \text{ne}) \end{pmatrix} = \begin{pmatrix} \mathbb{P}(S = \text{al}) \\ \mathbb{P}(S = \text{ma}) + \beta_M \\ \mathbb{P}(S = \text{ma}) \\ \mathbb{P}(S = \text{ne}) \end{pmatrix}.$$

Proof. Let's first consider $ATE_{M=1}$. By using the law of total expectations, we can first decompose it into a weighted average of principal stratum effects:

$$ATE_{M=1} = \mathbb{E}[Y(1) - Y(0) | M = 1] = \sum_{s \in S} \mathbb{E}[Y(1) - Y(0) | M = 1, S = s] \cdot \mathbb{P}(S = s | M = 1).$$

We can simplify the principal stratum effects using recursive substitution of the potential outcomes and the assumption that D , $\{M(0), M(1)\}$, and $\{Y(0,0), Y(0,1), Y(1,0), Y(1,1)\}$ are mutually independent. For $m_0, m_1 \in \{0, 1\}$,

$$\begin{aligned} \mathbb{E}[Y(1) - Y(0) | M = 1, M(0) = m_0, M(1) = m_1] &= \mathbb{E}[Y(1, M(1)) - Y(0, M(0)) | M = 1, M(0) = m_0, M(1) = m_1] \\ &= \mathbb{E}[Y(1, m_1) - Y(0, m_0) | M = 1, M(0) = m_0, M(1) = m_1] \\ &= \mathbb{E}[Y(1, m_1) - Y(0, m_0) | M(0) = m_0, M(1) = m_1] \\ &= \mathbb{E}[Y(1, m_1) - Y(0, m_0)]. \end{aligned}$$

The third equality uses the fact that $M \perp \{Y(1, m_1), Y(0, m_0)\} | \{M(0), M(1)\}$ because given $\{M(0), M(1)\}$ the only random term in $M = D M(1) + (1 - D) M(0)$ is D . Thus $ATE_{M=1}$ can be written as

$$ATE_{M=1} = \theta^T \mathbf{w}(ATE_{M=1}), \text{ where } \mathbf{w}(ATE_{M=1}) = \begin{pmatrix} \mathbb{P}(S = \text{al} | M = 1) \\ \mathbb{P}(S = \text{mi} | M = 1) \\ \mathbb{P}(S = \text{ma} | M = 1) \\ \mathbb{P}(S = \text{ne} | M = 1) \end{pmatrix}.$$

Similarly, $ATT_{M=1}$, ATE , and ATT can also be written as weighted averages of the entries of θ , where the weights are

$$w(ATT_{M=1}) = \begin{pmatrix} \mathbb{P}(S = al|D = 1, M = 1) \\ \mathbb{P}(S = mi|D = 1, M = 1) \\ \mathbb{P}(S = ma|D = 1, M = 1) \\ \mathbb{P}(S = ne|D = 1, M = 1) \end{pmatrix}, w(ATE) = w(ATT) = \begin{pmatrix} \mathbb{P}(S = al) \\ \mathbb{P}(S = mi) \\ \mathbb{P}(S = ma) \\ \mathbb{P}(S = ne) \end{pmatrix}.$$

Next we compute the conditional probabilities for the principal strata in $w(ATE_{M=1})$ and $w(ATT_{M=1})$. By using Bayes' formula, for any $m_0, m_1 \in \{0,1\}$,

$$\begin{aligned} & \mathbb{P}(M(0) = m_0, M(1) = m_1 | M = 1) \\ & \propto \mathbb{P}(M(0) = m_0, M(1) = m_1) \cdot \mathbb{P}(M = 1 | M(0) = m_0, M(1) = m_1) \\ & = \mathbb{P}(M(0) = m_0, M(1) = m_1) \cdot \sum_{d=0}^1 \mathbb{P}(M = 1, D = d | M(0) = m_0, M(1) = m_1) \\ & = \mathbb{P}(M(0) = m_0, M(1) = m_1) \cdot \sum_{d=0}^1 1_{\{m_d=1\}} \mathbb{P}(D = d | M(0) = m_0, M(1) = m_1) \\ & = \mathbb{P}(M(0) = m_0, M(1) = m_1) \cdot \sum_{d=0}^1 1_{\{m_d=1\}} \mathbb{P}(D = d). \end{aligned}$$

The last two equalities used $M = M(D)$ and $D \perp \{M(0), M(1)\}$. For this, it is straightforward to obtain the form of $w(ATE_{M=1})$ in Proposition 1. Similarly,

$$\mathbb{P}(M(0) = m_0, M(1) = m_1 | D = 1, M = 1) \propto \mathbb{P}(M(0) = m_0, M(1) = m_1) \cdot 1_{\{m_1=1\}}.$$

From this we can derive the form of $w(ATT_{M=1})$ in Proposition 1.

Proposition 2. Under the same assumptions as above, $PIE = \beta_M \cdot \mathbb{E}[Y(1, 1)]$ and $PDE = \beta_Y \cdot \mathbb{E}[M(0)]$.

Proof. This follows from the definition of pure direct and indirect effects and the following identity,

$$\mathbb{E}[Y(d, M(d'))] = \mathbb{E}[Y(d, 1) | M(d') = 1] \cdot \mathbb{P}(M(d') = 1) = \mathbb{E}[Y(d, 1)] \cdot \mathbb{P}(M(d') = 1),$$

for any $d, d' \in \{0, 1\}$.

Using the forms of weighted averages in Proposition 1, we can make the following observation on the sign of the causal estimands when β_M and β_Y are both nonnegative or both nonpositive:

Corollary 1. Let the assumptions in Proposition 1 be given. If $\beta_M \geq 0$ and $\beta_Y \geq 0$, then $ATE = ATT \geq 0$. Conversely, if $\beta_M \leq 0$ and $\beta_Y \leq 0$, then $ATE = ATT \leq 0$. However, both of these properties are not true for $ATE_{M=1}$ and the second property is not true for $ATT_{M=1}$.

The fact that ATT and ATE would have the same sign as β_M when β_M and β_Y have the same sign follows immediately from Proposition 2. However, this important property does not hold for $ATE_{M=1}$ and $ATT_{M=1}$. Here are some concrete counterexamples:

- (i) When $\beta_M = \beta_Y = 0.01$, $\mathbb{P}(S = al) = 0.1$, $\mathbb{P}(S = ma) = 0.05$, $\mathbb{E}[Y(0, 1)] = 0.1$, and $\mathbb{P}(D = 1) = 0.01$, we have $ATE_{M=1} = -0.003884$.
- (ii) When $\beta_M = \beta_Y = -0.01$, $\mathbb{P}(S = al) = 0.1$, $\mathbb{P}(S = ma) = 0.05$, $\mathbb{E}[Y(0, 1)] = 0.1$, and $\mathbb{P}(D = 1) = 0.99$, we have $ATE_{M=1} = 0.002514$.
- (iii) When $\beta_M = \beta_Y = -0.01$, $\mathbb{P}(S = al) = 0.1$, $\mathbb{P}(S = ma) = 0.05$, $\mathbb{E}[Y(0, 1)] = 0.1$, and $\mathbb{P}(D = 1) = 0.01$, we have $ATT_{M=1} = 0.0026$.

Heuristically, this is due to the fact that all of the causal estimands above, including β_M , β_Y , ATE , $ATE_{M=1}$, and $ATT_{M=1}$ only measure some weighted average treatment effect for police detainment and/or use of force. Conditioning on the posttreatment M may correspond to unintuitive weights. The possibility that $ATE_{M=1}$ and ATE can have different signs can be understood from the following iterated expectation:

$$ATE = ATE_{M=1} \mathbb{P}(M = 1) + \mathbb{E}[Y(1) - Y(0) | M = 0] \mathbb{P}(M = 0).$$

In this decomposition, the second term may be nonzero and have the opposite sign of $ATE_{M=1}$. An inexperienced researcher might be tempted to drop the second term because of Assumption 1, as $Y(0,0) = Y(1,0) = 0$ with

probability 1. However, conditioning on $M = 0$ is not the same as the intervention that sets $M = 0$. This means that we cannot deduce $\mathbb{E}[Y(d)|M = 0] = 0$ from $Y(d,0) = 0$, because $\mathbb{E}[Y(d)|M = 0] = \mathbb{E}[Y(d, M(d))|M = 0]$ is not necessarily equal to $\mathbb{E}[Y(d, 0)|M = 0]$.

The fundamental problem driving this paradox is that conditioning on the posttreatment variable M alters the weights on the principal strata, as shown in Proposition 1. $ATE_{M=1}$ and $ATT_{M=1}$ then depend on not only the racial bias in detainment and use of force (captured by β_M and β_Y) but also the baseline rate of violence $\mathbb{E}[Y(0, 1)]$ and the composition of race $\mathbb{P}(D = 1)$. For instance, in the first counterexample above, even though the minority group $D = 1$ is discriminated against in both detainment and use of force, because the baseline violence is high and the minority group is extremely small, $ATE_{M=1}$ becomes mostly determined by the smaller bias (captured by $\mathbb{P}(S = ma) = \mathbb{P}(M(0) = 1, M(1) = 0)$) experienced by the much larger majority group.

We make some further comments on the above paradox. First of all, the second counterexample can be eliminated if we additionally assume $\mathbb{P}(D = 1) < 0.5$, that is $D = 1$ indeed represents the minority group. With this benign assumption, one can show that $ATE_{M=1} < 0$ whenever $\beta_M, \beta_Y < 0$. Furthermore, it can be shown that $ATT_{M=1} < 0$ whenever $\beta_M, \beta_Y > 0$. So in a very rough sense we might say that as causal estimands, $ATE_{M=1}$ is unfavorable for the minority group (because $ATE_{M=1}$ can be negative even if both $\beta_M, \beta_Y > 0$) and $ATE_{M=1}$ is unfavorable for the majority group (because $ATT_{M=1}$ can be positive even if both $\beta_M, \beta_Y < 0$).

Our second comment is about the first counterexample. We can eliminate such possibility by assuming mediator monotonicity $\mathbb{P}(S = ma) = 0$, or in other words, by assuming that the majority race group is never discriminated against in any police–civilian encounter. KLM indeed used mediator monotonicity to obtain bounds on $ATE_{M=1}$ and $ATT_{M=1}$. So a supporter of the estimand $ATE_{M=1}$ may argue that if one is willing to assume mediator monotonicity, there is no paradox regarding $ATE_{M=1}$. However, it is worthwhile to point out that under mediator monotonicity, the pure indirect effect is guaranteed to be nonnegative because $\beta_M = \mathbb{P}(S = mi) - \mathbb{P}(S = ma) = \mathbb{P}(S = mi) \geq 0$. Empirical researchers should be mindful of and clearly communicate the consequences of the mediator monotonicity assumption unless it is compelling in the specific application. See KLM’s discussion after their Assumption 2 on when mediator ignorability may be violated. This concern can be alleviated if future work can incorporate nonzero $\mathbb{P}(S = ma)$ as sensitivity parameters in KLM’s bounds.

B Derivation of the Causal Risk Ratio

To simplify the derivation, we will omit the conditioning on $X = x$ below. Fix a $d \in \{0, 1\}$. Using assump:m0y0 , $\mathbb{E}[Y(d)|M(d) = 0] = \mathbb{E}[Y(d, 0)|M(d) = 0] = 0$. Therefore,

$$\begin{aligned} \mathbb{E}[Y(d)] &= \mathbb{E}[Y(d)|M(d) = 1] \cdot \mathbb{P}(M(d) = 1) \\ &= \mathbb{E}[Y(d, 1)|M(d) = 1] \cdot \mathbb{P}(M(d) = 1) \\ &= \mathbb{E}[Y(d, 1)|M(d) = 1, D = d] \cdot \mathbb{P}(M(d) = 1) \\ &= \mathbb{E}[Y|M = 1, D = d] \cdot \mathbb{P}(M(d) = 1). \end{aligned}$$

The third equality above uses treatment ignorability: $D \perp Y(d, 1)|M(d)$ (this follows from the single world intervention graph corresponding to Figure 1); the last equality follows from the consistency (or stable unit value treatment) assumption for potential outcomes. By further using $D \perp M(d)$, we have $\mathbb{P}(M(d) = 1) = \mathbb{P}(M(d) = 1|D = d) = \mathbb{P}(M = 1|D = d)$. Plugging this into the last display equation, we have

$$\mathbb{E}[Y(d)] = \mathbb{E}[Y|M = 1, D = d] \cdot \mathbb{P}(M = 1|D = d), d = 0, 1.$$

Thus we have recovered KLM’s Proposition 2 (point identification of ATE) without assuming their Assumption 2 (mediator monotonicity) and Assumption 3 (relative nonseverity of racial stops). To get the causal risk ratio, we only needs to take a ratio between $\mathbb{E}[Y(1)]$ and $\mathbb{E}[Y(0)]$ and apply Bayes’ formula to cancel $\mathbb{P}(M = 1)$.

C Implementation Details of the Empirical Analysis

To estimate encounter rates in our empirical analysis using the PPCS data we used the following three survey questions:

The following are questions about any time in the last 12 months when police have initiated contact with you. In the last 12 months, have you:

- V11** Been stopped by the police while in a public place, but not a moving vehicle? This includes being in a parked vehicle.
- V13** Been stopped by the police while driving a motor vehicle?
- V21** Have you been stopped or approached by the police in the last 12 months for something I haven’t mentioned?

We created two binary measures as indicators of police encounters. The first measure (Stop in Public in Table 1) was 1 for being stopped by the police if the respondent answered Yes to either V11 or V21 and 0 otherwise. We used V13 as the measure for being stopped in a motor vehicle (MV Stop in Table 1).

In our alternative analysis (labelled as PPCS * in Table 1), the stop indicators are weighted by the responses to the following question:

V30 Thinking about the times you initiated contact with the police and the times they initiated contact with you, how many face-to-face contacts did you have with the police during the last 12 months?

In that analysis, we excluded outliers with more than 30 reported contacts with the police.

D Stratified Analysis by Age and Gender

Our identification in Equation 3 of the causal risk ratio depends on conditioning on all the confounders in X . Here we report the results of an additional analysis where the police-civilian encounters were stratified by the age and gender of the civilian. Similarly, the survey respondents were also stratified by their age and gender. The same analyses that generated Table 1 were repeated for each stratum, and the results are reported in Figure D.1. It appears that gender is an important effect modifier but age is not.

