PhD Preliminary Exam Report on “A Note on Posttreatment Selection in Studying Racial Discrimination in Policing”

Shreya Prakash

Abstract

This report examines the causal estimands that [Zhao et al., 2022] uses to study racial discrimination in policing. We first introduce statistical discrimination and causal inference to motivate racial discrimination in policing as a causal inference problem. Then we will review previous literature to understand how posttreatment selection bias is a problem when studying racial discrimination in policing and how some methods may be used to help alleviate this selection bias. After, we will detail the methods presented in this paper to correct for this posttreatment selection bias. We demonstrate these methods by reproducing the authors' analysis of their causal risk ratio estimator for the NYPD Stop and Frisk data. We conclude by assessing the contributions and shortcomings of their estimator and some extensions of the authors' results.

1 Introduction

Researchers have long been interested in quantifying racial discrimination in policing. This is a difficult problem. To learn about racial discrimination in policing, one often uses administrative data on police detainments. However, as pointed out by [Knox et al., 2020], which we will denote as KLM from here on, this results in posttreatment selection bias. This is because administrative data only records certain police-civilian encounters where civilians are already detained, which itself is likely racially biased. Since people often use this data to estimate the racial discrimination in policing, their estimates are biased. As done by KLM, Zhao et al. shows that when there is any racial discrimination in the decision to detain civilians, current naive estimates of the effect of race on police behavior are biased. This paper also aims to clarify causal estimands of interest, since this is often what previous literature like [Fryer, 2018] aim to implicitly estimate but never quite explicitly do. This has its challenges in that it is often hard to identify or estimate causal quantities with the data on hand. The aim of both this paper and the KLM paper is to represent this problem as a causal question since it is often implicitly treated as one. They also explicitly discuss the assumptions needed to identify their estimands and the challenges with getting unbiased estimates using their estimands.

As mentioned previously, racial discrimination in policing is naturally a causal question. Researchers often want to know if police behavior would differ based on one's race. To setup this question for causal analysis, Zhao et al., similarly to KLM, use a police-civilian encounter, or the sighting of civilians by a police officer as the unit of analysis. They chose to use race as the treatment. However, the race causal effect is undefined because race is immutable or is a very complex construct since it affects outcomes through many channels like skin tone, dialect and clothing. To avoid this complication, Zhao et al., similarly to KLM, think of the subset of comparable situations in which minority and majority civilians are observed by the police. Specifically, they define the counterfactual as an encounter with a comparable person who is participating in comparable behavior but is of a different race. They characterize race as D, where $D = d$, $d = 0, 1$; 0 represents someone who is White and 1 represents someone who is Black. They use M_i to indicate a police detainment or a stop of civilian; $M_i = 1$ represents when encounter i results in detainment. The administrative data contains only encounters that result in detainment. They characterize Y_i as the outcome, where $Y_i = 1$ indicates police use of force in encounter i. They characterize U_i to capture the unobservable subjective aspects of the encounter like the officer's suspicion or sense of threat. Lastly, X_i is the collection of covariates that describe the aspects of the encounter. The paper, in most of its analysis, assumes conditioning on X_i. Please refer to Table 1 for easy access of these variables of interest.

Variable	Variable Meaning
D	Race
M	Police Stops/Detainments
Y	Police use of Force
X	Covariates
U	Unobserved Variables

Table 1: Variables of Interest when studying Racial Discrimination in Policing

They introduce the potential outcomes M_i and Y_i. $M_i(d)$ represents whether encounter i would have resulted in a stop if the civilian is of race d. $Y(d)$ is the use of force when D is set to race d, and $Y(d, m)$ is the use of force when D is set to race d and whether the encounter resulted in detainment is set to m. KLM constructs the DAG shown in Figure 1 to characterize the police encounters. This

DAG presents posttreatment selection bias. In this DAG, M is a posttreatment variable since it is a consequence of race D. Now the data being used in the analysis of racial discrimination is administration data which only includes entries for civilians who have been stopped by the police, $M = 1$. By conditioning on $M = 1$, we essentially open up the path $D \to M \leftarrow U \to Y$. So instead of just studying the direct discrimination of race on use of force through the direct path $D \to Y$, we are unintentionally introducing bias by conditioning on $M = 1$ and also including information from the indirect path $D \to M \leftarrow U \to Y$.

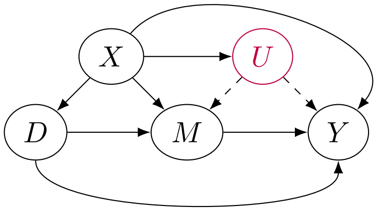

Figure 1: DAG assumed in analysis

Using this DAG structure, KLM focuses on several local causal estimands to represent the racial discrimination while correcting for the selection bias. However Zhao et al. claims that these local estimands are not ideal to study because: (1) they require several assumptions/additional data to identify and (2) even when identified, they cannot be used to understand global causal affects like the Average Treatment Effect (ATE). KLM derives a point estimate of the ATE, but this point estimate also requires several assumptions and additional data that may be hard collect or are not given.

Zhao et al. sets up their analysis in a similar manner to KLM. In their analysis, Zhao et al.

1. Shows that KLM's local causal estimands cannot give any information about the global estimands

2. Introduces the global causal risk ratio which helps identify the global causal effect:

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

This estimand requires less assumptions than KLM's ATE point estimate identification. Here is the form when identified after making some assumptions which will be detailed in Section 3:

$$CRR(x) = \frac{E[Y(1)|X=x]}{E[Y(0)|X=x]} = \underbrace{\frac{E[Y|D=1,M=1,X=x]}{E[Y|D=0,M=1,X=x]}}_{\text{naive risk ratio}} \times \underbrace{\frac{\frac{P[D=1|M=1,X=x]}{P[D=0|M=1,X=x]}}{\frac{P[D=1|X=x]}{P[D=0|X=x]}}}_{\text{bias factor}} \tag{2}$$

It also illuminates and corrects for the posttreatment selection bias that is present when just using a naive risk ratio. However, for this estimator, it is hard to estimate the denominator of the bias term because it requires data that is not easily accessible. Despite there not being good data to accurately estimate the CRR, they claim that as seen the re-analysis of the NYPD Stop and Frisk dataset, that using this estimator with external data sources can illuminate the probable bias of the naive estimator and can serve as a baseline for a sensitivity analysis.

The rest of the report is organized as follows: in Section 2 we will talk about other related works. In Section 3 we will discuss their methods and some issues with their methods. In Section 4, we present results from applying their methods to real data. In Section 5, we present some possible extensions to their work. We conclude our discussion in Section 6.

2 Previous Literature

In this section we will discuss some of the previous literature for studying racial discrimination in policing and spend majority of the section focusing on the methods used by KLM since this paper informs a lot of the decisions made by Zhao et al.

2.1 Other Prior Work

Most of the previous literature studying racial discrimination in policing is on adjusting for omitted variables that may correlate with the suspect race. Even if the omitted variable bias problem is solved, the posttreatment selection bias that results from studying racial discrimination using records that are the product of racial discrimination still remains ([Angrist and Pischke, 2008]; [Elwert and Winship, 2014]; [Rosenbaum, 1984]). At first, this posttreatment selection bias may look like classic example of sample selection bias, which is already a well studied field ([Elwert and Winship, 2014]; [Heckman, 1979]). But upon closer inspection, these methods are unsuitable for this problem [Fryer, 2018]. For instance methods for bounding or estimating population average treatment effects while accounting for selection bias will be unsuitable since policing scholars seek to estimate among people who actually interact with police rather than the entire population [Lee, 2009]. There is also a large literature on posttreatment bias, but these methods are also unsuitable for this problem. That is because these methods require either complete data on the posttreatment variable [Acharya et al., 2016] or knowledge of the scale of the missing data [Nyhan et al., 2017]. However in policing data, the administrative data only conditions on one level of the posttreatment variable and there is not information about of the number of police-civilian encounters.

Racial discrimination in policing has been studied by many policy researchers. For instance [Grogger and Ridgeway, 2006] leverage a strategy called the "veil of darkness" that compares traffic stop patterns before and after the sun sets assuming that the race of the driver is somewhat hidden to police officers after the sun sets, and thus aims to look at a sample of police-civilian interactions that were initiated in a race-blind manner. Most studies in the literature focus on mitigating the omitted variable bias issue. For instance [Fryer Jr, 2019] aims to estimate racial discrimination in police-civilian encounters by using regression controlling for several other variables relating to civilians, officers and the circumstance. [Fryer, 2018] claims that regression can recover the race effect if race is as good as randomly assigned conditioned on the other covariates. [Fryer Jr, 2019] finds that there is racial bias in sublethal force but not lethal force. [Johnson et al., 2019] attempts to estimate the racial bias in police shootings by only examining the cases where a fatal shooting occurred. They incorrectly conclude that there is no antiminority bias by wrongly assuming that police encounters with White and minority civilians happen in equal numbers [Knox and Mummolo, 2020].

There is also prior work examining racial bias in traffic enforcement. For instance [Ridgeway, 2006] uses propensity score weighting to estimate the racial bias in traffic stops in Oakland, CA. They conclude that there is little evidence of racial bias on most outcomes. All of these examples of previous literature in racial discrimination in policing fails to capture the unobserved selection process through which police choose to engage with civilians, which has been shown to be a function of one's race [Gelman et al., 2007]. Since this is the case, even if one controls for a complete set of possible confounders so that it is as if race were randomly assigned to police encounters, there will still be statistical bias. In addition the previous studies implicitly make causal claims but leave the precise estimand of interest ambiguous (i.e. $ATE, ATE_{M=1}, CDE_{M=1}$). KLM claims that it is the first to (1) make these causal quantities of interest explicit, and (2) address this posttreatment selection bias.

2.2 Prior Work by Knox. et al. (KLM)

As previously mentioned, the KLM paper assumes the DAG presented in Figure 1. Traditionally policy researchers use the following naive estimator to measure the racial discrimination in policing. Here, conditioning on confounders X is implicit.

$$\hat{\Delta} = E[Y_i|D_i = 1, M_i = 1] - E[Y_i|D_i = 0, M_i = 1] \tag{3}$$

This naive estimator has no causal interpretation as long as the treatment affects the mediator without further assumptions. This is because this estimator is comparing two different groups. For instance, if officers exhibit racial bias in that they detain White civilians who commit serious crimes but detain minority civilians regardless of the behavior, the two groups that the naive estimator is comparing are fundamentally different. So in this case, the analysts would observe that the use of force used against detained White civilians is larger than the use of force used against detained minority civilians. Thus erroneously concluding antiwhite bias. KLM formalizes this by introducing principle strata with respect to the mediator:

$$S = \begin{cases} \text{always stop (al)}, & \text{if } M(0) = M(1) = 1 \\ \text{minority stop (mi)}, & \text{if } M(0) = 0, M(1) = 1 \\ \text{majority stop (ma)}, & \text{if } M(0) = 1, M(1) = 0 \\ \text{never stop (ne)}, & \text{if } M(0) = M(1) = 0 \end{cases} \tag{4}$$

Using the principle strata, KLM further breaks down the naive estimator:

$$\begin{aligned} E[\hat{\Delta}] &= E[Y_i|D_i = 1, M_i = 1] - E[Y_i|D_i = 0, M_i = 1] \\ &= E[Y_i(1,1)|D_i = 1, M_i(1) = 1, M_i(0) = 1]P(M_i(0) = 1|D_i = 1, M_i(1) = 1) \\ &+ E[Y_i(1,1)|D_i = 1, M_i(1) = 1, M_i(0) = 0]P(M_i(0) = 0|D_i = 1, M_i(1) = 1) \\ &- E[Y_i(0,1)|D_i = 0, M_i(1) = 1, M_i(0) = 1]P(M_i(1) = 1|D_i = 0, M_i(0) = 1) \\ &- E[Y_i(0,1)|D_i = 0, M_i(1) = 0, M_i(0) = 1]P(M_i(1) = 0|D_i = 0, M_i(0) = 1) \end{aligned}$$

Here we explicitly see that the naive estimator compares groups with different potential outcomes. Since these quantities are unobservable, the resulting bias is hard to address without making further assumptions.

KLM defines the average treatment effect (ATE), the average treatment effect on the treated (ATT), the average treatment effect conditional on the mediator ($ATE_{M=1}$), ($ATT_{M=1}$), and the average treatment effect on the treated conditional on the mediator. In this context, the ATE measures the extent to which minority civilians face a greater risk of police use of force compared to White civilians because of their race. It measures whether a minority civilian is differentially stopped and if minority civilians are differentially subject to violence. $ATE_{M=1}$ is the ATE restricted to subset of people who were detained. $ATT_{M=1}$ is the number of minority detainments in which force would not have been employed if the civilian had been White.

$$ATE = E[Y(1) - Y(0)] \tag{5}$$

$$ATT = E[Y(1) - Y(0)|D = 1] \tag{6}$$

$$ATE_{M=1} = E[Y(1) - Y(0)|M = 1] \tag{7}$$

$$ATT_{M=1} = E[Y(1) - Y(0)|M = 1, D = 1] \tag{8}$$

Using assumptions that will be detailed later on, KLM derives the bias and nonparametric sharp bounds on the ATE. KLM also claims that the ATE can be point identified if researchers also collect information about the total number of White and minority encounters. KLM also derives nonparametric sharp bounds on $ATE_{M=1}$ and $ATT_{M=1}$. KLM claims that the $ATT_{M=1}$ can be point identified if they have access to the rate of detainment. For their analysis, KLM makes the following assumptions:

(i) Mandatory Reporting: $Y_i(d, 0) = 0$ for all i and for $d \in \{0, 1\}$. They are assuming that all encounters that escalate to using force are observed in the administrative data.

(ii) Mediator Monotonicity: $M_i(1) \geq M_i(0)$ for all i, but the reverse is never true. This essentially means that they are assuming that the minority group is being discriminated against

(iii) Relative Nonseverity of Racial Stops: $E[Y_i(d, m)|D_i = d', M_i(1) = 1, M_i(0) = 1, X_i = x] \geq E[Y_i(d, m)|D_i = d', M_i(1) = 1, M_i(0) = 0, X_i = x]$. For example this assumption would imply that a police officer would be more likely to use force against a White individual committing assault compared to a White individual jaywalking on average.

(iv) Treatment Ignorability:

(a) With respect to a potential mediator M, $M_i(d) \perp D_i|X$

(b) With respect to a potential outcome Y, $Y_i(d,m) \perp D_i | M_i(0) = m', M_i(1) = m'', X_i$. This essentially means that conditional on the covariates X, the individual's race is as good as randomly assigned to the encounters and officers encounter minority individuals in circumstances that are no different than the the circumstances of the White encounters.

(iv) SUTVA: Assumes that each encounter is unaffected by a civilian's race in other encounters. This might be violated in the case that a group of individuals were stopped simultaneously

KLM shows in their analysis that the naive estimator underestimates the bias regardless of the causal quantity one wants to estimate. KLM does an analysis of the NYPD Stop-and-Frisk data to study the $ATT_{M=1}$ and $ATE_{M=1}$. They compare these quantities to the naive estimate. KLM also provides a research plan for future research. They suggest using traffic cameras or police body-worn cameras to collect data on total encounters. They suggest using videos or photos to document license plate numbers which then can be used to merge this data with the administrative data. With this merged data, the ATE and $ATT_{M=1}$ can be identified. However a tradeoff is that they may not be able to study rare events such as shootings.

3 Methods

Zhao et al. shows that KLM's local causal estimands cannot give any information about the global estimands like the ATE. They show that this is true even in the simplest case where there is no unobserved confounding U and thus no posttreatment selection bias. Because the local effect does not inform the global effect, Zhao et al. introduce the global causal risk ratio (CRR) to estimate the global causal effect:

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

This estimate requires less assumptions for identification than KLM's ATE point estimate identification. It also corrects for selection bias when compared to naive risk ratio.

3.1 Local vs Global Estimands

Policy-makers have focused more on the local estimator conditional on police detainment because never-stopped data is not available. However one might be interested in a global estimand to understand when unreported White encounters would have escalated to a stop if the involved civilian was a minority and vise-versa. Zhao et al. show that the sign of the local estimand does not inform the sign of the global estimand. They show that this is true even in the simpler case where there is no unobserved confounding U and thus no posttreatment selection bias. They claim that this still holds when unobserved confounding, U, exists. But for simplicity's sake, they chose to omit U in their analysis of the local and global effects. To conduct this analysis, they break down the ATE into the pure indirect effect (PIE) and pure direct effect (PDE):

$$ATE = E[Y(1) - Y(0)] = \underbrace{E[Y(1, M(1)) - Y(1, M(0))]}_{\text{PIE}} + \underbrace{E[Y(1, M(0)) + Y(0, M(0))]}_{\text{PDE}} \tag{9}$$

The pure indirect effect measures the effect of race on use of force through indirect pathways, such as through the mediator police stops. The pure direct effect measures the direct effect of race on use of force.

They simplify the interpretation of this analysis of local vs global effects by introducing a new variable to denote the principal stratum:

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

Zhao et al. also makes the following assumptions:

1. Variables (X, D, M, Y) are generated from a nonparametric structural equation model (SEM):

$$\begin{aligned} X &= f_X(\epsilon_X) \\ D &= f_D(X, \epsilon_D) \\ M &= f_M(X, D, \epsilon_M) \\ Y &= f_Y(X, D, M, \epsilon_Y) \end{aligned}$$

   where $\epsilon_X, \epsilon_D, \epsilon_M, \epsilon_Y$ are mutually independent, which implies that $D, \{M(0), M(1)\}$, and $\{Y(0,0), Y(0,1), Y(1,0), Y(1,1)\}$ are mutually independent. This SEM assumes that there is no unobserved confounding U and corresponds to the DAG in Figure 2.

2. Mandatory reporting assumption:

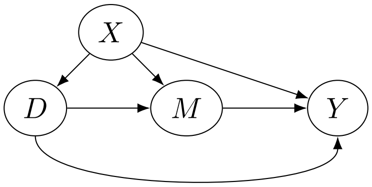

Figure 2: DAG assumed in local and global estimand analysis

(a) $Y(0,0) = Y(1,0) = 0$

(b) administrative data contains all detainments/stops of civilians by the police

to show that

$$PIE = \beta_M \times E[(Y(1,1)], PDE = \beta_Y \times E[M(0)]$$

We have that $\beta_M = E[M(1)-M(0)]$ is the average effect of race on detainment and $\beta_Y = E[Y(1,1)-Y(0,1)]$ is the controlled direct effect (CDE) of race on police violence. The full proof this can be found in the Appendix. Using the same assumptions they also show that the $ATE \geq 0$ if $\beta_M, \beta_Y \geq 0$ and the $ATE \leq 0$ if $\beta_M, \beta_Y \leq 0$. So the global ATE and ATT are nonnegative whenever both the direct and indirect effects are non-negative and vise versa. We have that this is not true for $ATE_{M=1}$ and the second property is not true for $ATT_{M=1}$ as shown through some counterexamples included in the Appendix.

The reason why this does not hold is because conditioning on the post-treatment variable M alters the principal strata weights. The local estimands depend on the racial bias in detainment on use of force (in β_M and β_Y) but also $E[Y(0,1)]$, the baseline rate of violence, and $P(D = 1)$, the composition of race.

They show that $ATE_{M=1}$ and $ATT_{M=1}$ are not guaranteed to inherit the sign of β_M, β_Y. This means that if there is evidence of antiminority bias in the local effects, which would mean that local effect would be positive, that does not guarantee that there would be antiminority bias globally.

3.2 The Causal Risk Ratio

KLM notes that ATE estimation requires estimating the magnitude of the rate of detainments $P(M = 1)$. This quantity is hard to estimate because we don't have a good way to quantify the frequency of stops. This paper shows that by formulating this estimator on the relative scale by using a ratio, one can avoid this problem.

Zhao et al. gets the causal risk ratio at the covariate level x,

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

When this ratio is greater than 1, we see that the risk of violence is greater for minorities. Similar to KLM, this paper compares this CRR to a naive ratio estimator that does not take into account the posttreatment selection bias. This naive estimator conditions on the mediator which is a posttreatment variable, so it will have posttreatment selection bias.

$$NaiveRR(x) = \frac{E[Y|D = 1, M = 1, X = x]}{E[Y|D = 0, M = 1, X = x]} \tag{10}$$

To identify the CRR, they make the following assumptions, note that these assumptions are a subset of the assumptions made by KLM:

1. SUTVA: Requires that the response of a particular unit depends only on the treatment to which he himself was assigned, not the treatments of others around him
2. Treatment Ignorability: Potential outcomes are independent of treatment assignment conditioned on the mediator and the covariates $D \perp Y(d, 1)|M(d), X$
3. The DAG presented in Figure 1 which has posttreatment selection bias due to U

Using these assumptions and similar methods to KLM, they get identifiability:

$$E[Y(d)|X = x] = E[Y|M = 1, D = d, X = x]P(M = 1|D = d, X = x) \tag{11}$$

To see the entire derivation, reference the Appendix. Then they use Bayes rule to cancel out $P(M = 1, X = x)$, which as discussed by KLM is hard to estimate. They get the following identification form:

$$CRR(x) = \frac{E[Y(1)|X = x]}{E[Y(0)|X = x]} = \underbrace{\frac{E[Y|D = 1, M = 1, X = x]}{E[Y|D = 0, M = 1, X = x]}}_{\text{naive risk ratio}} \times \underbrace{\frac{\frac{P[D=1|M=1,X=x]}{P[D=0|M=1,X=x]}}{\frac{P[D=1|X=x]}{P[D=0|X=x]}}}_{\text{bias factor}}$$

The first term represents the naive risk ratio, which ignores the the possible bias resulting from the selection process of the administrative data. The second term represents the bias correcting factor. The numerator measures the relative probability of a detainment being with a minority conditional on the covariates X. This can be measured with administrative data. The denominator is the odds of an encounter being with a minority conditional on the covariates. These need to be approximated using a separate data source. The ratio of these two terms will correct the bias since if minorities are overrepresented in the administrative data, the bias will correct for that overrepresentation and increase the magnitude of the risk ratio.

If we assume stochastic mediator monotonicity $E[M(1)|X = x] \geq E[M(0)|X = x]$, the bias factor can be lower bounded by 1 and the naive risk ratio provides a lower bound for the causal risk ratio.

3.3 Some problems

The main issue with this estimator is that it comes with the complication of needing two data sources for estimates. Specifically, most data sources contain information on stops rather than encounters. Because of that they plan to use population level data on police stops to approximate encounter rates by racial group. If these quantities are proportional, then the method will be accurate. Also, there might be some measurement differences between the external data source and the administrative data. Currently, the authors do not have a good data source to accurately measure rates of encounters.

Another issue is that the CRR is conditional on the covariates X, which is needed for identification due to treatment ignorability. However, it is nontrivial to control for all possible discrete and continuous covariates in practice. Due to this difficulty, this paper ignored conditioning on more than 2 covariates in their real data example.

Despite there not being good data to accurately estimate the CRR, they claim that using this estimator with external data sources can illuminate the probable bias of the naive estimator. This can be as seen in the re-analysis of the NYPD Stop and Frisk dataset. They also claim this estimator can serve as a baseline for a sensitivity analysis.

4 Results

Using the identification form of the CRR, this paper estimated the CRR using the New York Police Department's (NYPD) Stop-and-Frisk dataset (2003-13), which is the same dataset used by KLM. The NYPD's Stop-and-Frisk dataset contains about 5 million records of pedestrian stops. The majority of these stops are of nonwhite suspects. This dataset records various levels of use of force, ranging from laying hands on a suspect, to pointing a weapon on a suspect. Zhao et al. removed all races in the dataset other than Black and White. This leaves the dataset to having around 3 million records. They also focused on all forms of force rather than individual types of force as KLM did. Previous literature like [Fryer Jr, 2019] used the naive approach that did not account for posttreatment selection bias as shown in equation 10. To replicate this naive approach for the risk ratio, this paper compares the CRR to the naive risk ratio. To estimate the CRR, additional data sources must be used to estimate the denominator of the bias factor. They use 3 different data sources to estimate this bias factor:

1. 2013 Current Population Survey (CPS)
2. 2011 Police-Public Contact Survey (PPCS)
3. 2010 Census Data constructed with census blocks by (Keefe 2020)

The CPS dataset measures for race and geographic information to narrow down information down to New York City. However, this data source has no measures for police encounters or stops so it would just be an approximation for the denominator of the bias term. The PPCS dataset has measures for police stops, however it does not have any information about the police encounters. Further, this dataset has no geographic identifiers to narrow it to the New York area. For this reason, the authors also looked at the racial distributions for different subsets of the PPCS data. They looked at subgroups of the survey subjects that experienced a motor vehicle stop, any other kind of police stop, and individuals in a large metropolitan area, or regions with a population more than 1 million people. They also weighed the PPCS survey subjects by their reported number of face-to-face contacts with the police. They excluded respondents with larger than 30 reported contacts. Keefe 2020 used census blocks and the 2010 census data to construct a population breakdown for each of the NYPD precincts. This data allows the authors to compare the make up of the population of Black civilians in each precinct compared to the proportion of detainments of Black civilians for each precinct. This dataset is again used to estimate the denominator of the bias term. Similarly to the CPS data source, the census data is population level data and does not have any information about the police encounters. Thus, it would only serve as an approximation of the bias term. A sensitivity analysis was performed using this census dataset to in part determine whether the racial distribution in the police-civilian encounters can be well approximated by the racial distribution in the census data.

The author reports the risk ratios using the naive and causal risk ratios and different datasets in Table 2. In this analysis they did not control for any confounders. This would affect the validity their estimates, since their causal risk ratio is conditional on X. Based on their estimates, the naive estimator indicates that Black people have a 29% higher risk of receiving police force than white people in an encounter. The causal risk ratio for all the different supplementary data used all have estimates that are almost or greater than 10 times that of the naive risk ratio.

External Dataset	Estimated risk ratio	95% Confidence Interval
None (Naive estimator)	1.291	(1.284 – 1.299)
CPS	13.566	(12.812 – 14.375)
PPCS	32.300	(31.289 – 33.402)
PPCS (MV Stop)	29.549	(26.726 – 32.903)
PPCS (Other Stop)	29.241	(23.446 – 37.201)
PPCS (Large Metro)	16.688	(15.237 – 18.180)
PPCS*	31.131	(28.203 – 34.736)
PPCS* (Large Metro)	19.873	(14.147 – 28.607)

Table 2: Estimates of the Causal Effect of Minority Race (Black) on Police Violence
Estimates of the causal effect of minority race (Black) on police violence without controlling for confounding. Confidence intervals were computed using nonparametric bootstrap.

The author also conducted the same analysis while conditioning on one or two confounders. They conducted a stratified analysis by age and gender and found that the estimates more or less match those found in Table 2, with the exception that female minorities are more likely to have a smaller risk ratio than male minorities. They did not find age to be that important of an effect modifier. These results can be found in the Appendix in Figure 8.

They also used modified census data to use police precinct as a confounder. Using the census block data, the authors visualize how Black civilians make up less than half of the population but more than half of the detainments/stops. This is shown by Figures 6 and 7 in the Appendix. This would indicate that that the bias factor of the causal risk ratio can be pretty large for this analysis.

In their analysis by precinct, they found that for all the neighborhoods except for those where Black civilians account for more than 90% of the population, the CRR is much larger than the naive risk ratio. This analysis can be seen by Figure 3. The reasoning for those cases for where it is not larger is most likely due to the residential distribution in the census data poorly approximating the racial distribution in the police-civilian encounters because the civilians could be those from other precincts or visitors from other areas. Their analysis shows that most of the highest CRR's are in wealthy Manhattan/Brooklyn neighborhoods, which may be due in part to increased suspicion of minorities in areas where their presence is not as commonplace. There is a strong negative correlation between the CRR and the percentage of Black residents in the precinct as seen in Figure 4. This indicates that racial discrimination in police use of force may be strongly moderated by characteristics associated with geographic location such as income, racial composition and average crime rate.

They also conduct a sensitivity analysis to in part determine whether the racial distribution in the police-civilian encounters can be well approximated by the racial distribution in the census or survey datasets. This can be seen in Figure 5. The 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. From the 2010 census data, 36.7% of the population in NYC was Black. So if the proportion of Black residents is lower than 36.7%, then in the sensitivity analysis, the presumed proportion of encounters with Black civilians is higher than the proportion of encounters with Black residents in that precinct. This shrunk the CRR towards a more centralized value, especially for precincts that are predominantly White or predominately Black.

Something to also note is that when asked about their reasoning for using the numbers that they chose for their sensitivity analysis, the authors claim that that the particular choice of the 90% and 10% were arbitrary. However, one thing that was noted was that when the chance of encountering a resident from another precinct is too high, the trend between race and police force was reversed (i.e the more Black the precinct, the larger the CRR).

These figures and tables presented are from my implementation of their methods. My results are very similar to Zhao et al.'s results. The only differences are due to the randomness of bootstrapping procedure for generating confidence intervals.

5 Extensions

5.1 Local vs Global Estimands

As seen in the methods section, Zhao et al. had shown that the sign of the local estimands do not tell you the sign of the global estimands. However, they claim that this can be relaxed with a few additional assumptions. They actually make this claim without any proof, and after contacting Dr. Zhao, he was not sure if the claim actually stood. I studied the claims that he made and proved

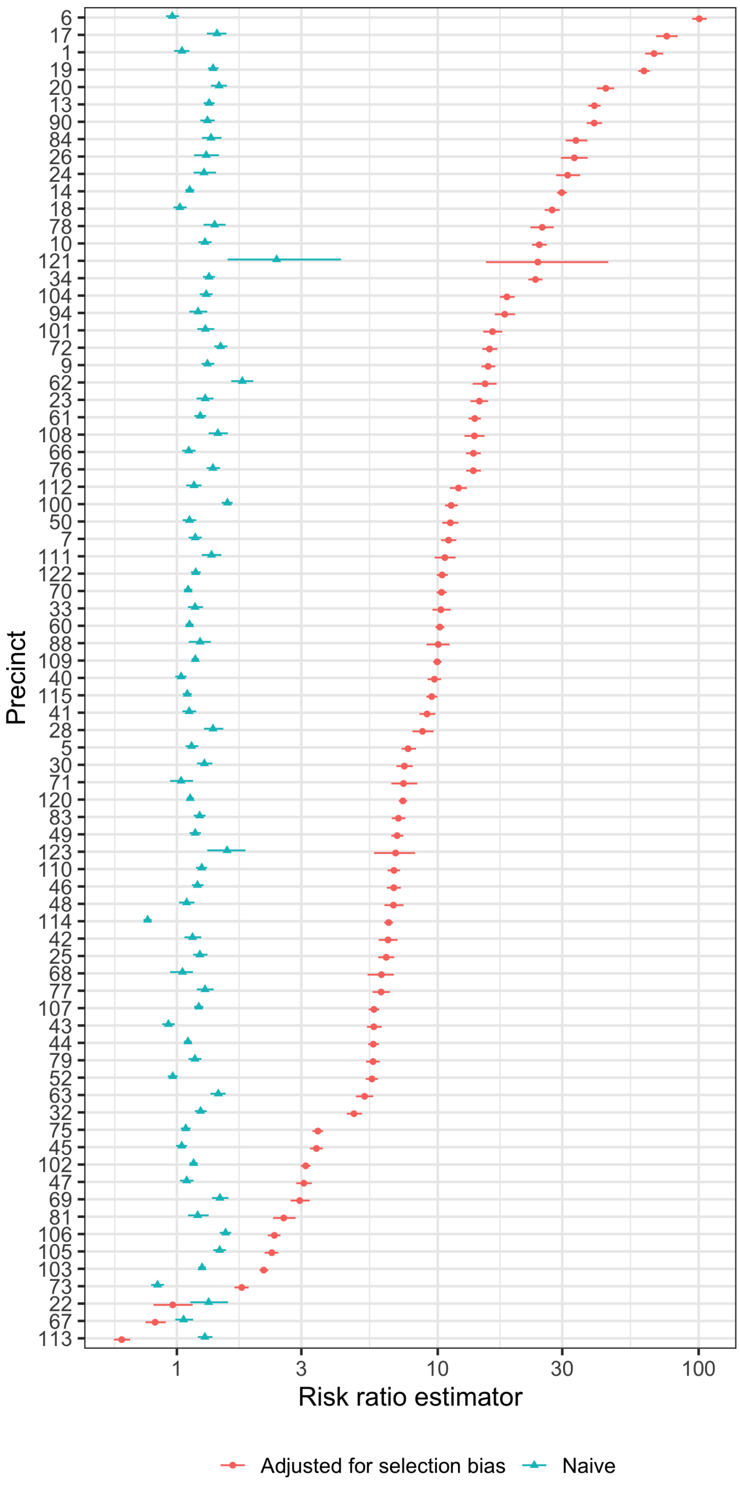

Figure 3: Risk Ratio Estimates for Every NYPD Precinct

Note that the error bars are correspond to 95% confidence intervals computed by nonparametric bootstrap. The census data was not resampled because it is a residential distribution rather than a statistical estimate. The estimates in blue are obtained using the naive estimator and the ones in red are using the CRR.

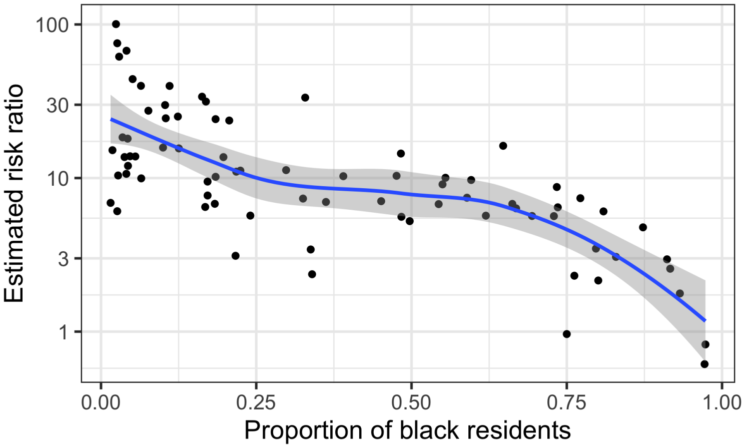

Figure 4: Estimated Risk Ratio versus the proportion of Black Residents in each Precinct

This is the estimated risk ratio versus the proportion of Black residents in each precinct

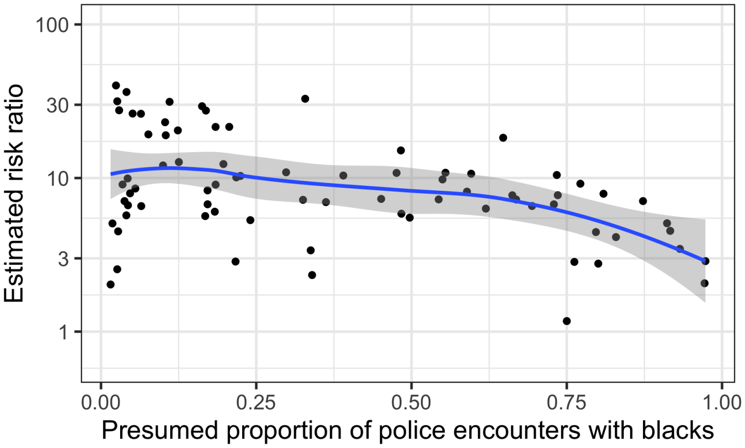

Figure 5: Estimated Risk Ratio versus the proportion of Black Residents in each Precinct

This is the estimated risk ratio in a sensitivity analysis versus the proportion of Black residents in each precinct. Assume that police encounters a mixture of 90 % local residents and 10% New York city wide residents.

that they did in fact hold.

Zhao et al. claims that when looking at local estimators, certain assumptions can eliminate particular counterexamples. The paper claims that you can eliminate the case where β_Y, β_M are negative but the $ATE_{M=1}$ is positive if one assumes that $D = 1$ in fact represents the minority group, (i.e. $P(D = 1) < 0.5$). They also claim that the case where β_Y, β_M are positive but $ATE_{M=1}$ is negative can be eliminated if one assumes mediator monotonicity ($P(S = ma) = 0$). This essentially means that majority race group is never discriminated against in any encounter between a civilian and the police. Note that KLM used this assumptions to get bounds on the local estimands $ATE_{M=1}$ and $ATT_{M=1}$. The consequence of this assumption is that $\beta_M \geq 0$, which is saying that the effect of race on detainments will always be positive. If researchers make this assumption, it should be clearly communicated.

I verified and generalized the authors claims and provide proofs of these claims in the Appendix. The generalized claims I made are:

1. If one assumes that $D = 1$ is in fact the minority group, then we have that $\beta_Y < 0, \beta_M < 0$ implies that $ATE_{M=1} < 0$

2. If one assumes mediator monotonicity $P(S = ma) = 0$, then we have that the sign of the local estimands $ATE_{M=1}, ATT_{M=1}$ are consistent with the sign of the global estimand.

Note that in the KLM paper, they make the mediator monotonicity assumption, so in their case, the signs of the local and global estimands are consistent with each other.

5.2 Other Global Estimands

When looking at antiminority discrimination as a whole, KLM chose to focus on the ATE, but was only able to get bounds since this was not identified without further information. This paper looks at the causal risk ratio, which is identified. Both of these estimands help policy researchers understand whether there is antiminority discrimination but it does not give a good sense of the gravity of the antiminority discrimination. An estimand that helps us understand the gravity of the antiminority discrimination is the following:

$$P(Y(1) = 1 | Y(0) = 0)$$

This essentially equates to the probability of an officer using force if an officer perceives that person as Black in an encounter where they would have not have otherwise used force if the police officer had perceived them as White. This quantity will help us understand the probability of the police officers engaging in antiminority discrimination. To get non-parametric sharp bounds on this quantity in the NYPD Stop and Frisk example, it would suffice to use the algorithm presented by [Duarte et al., 2021], since all the quantities are discrete. This algorithm essentially uses principal stratification to rewrite the causal query into a polynomial expression. They then rewrite the modeling assumptions into polynomial constraints. Then they have transformed the problem into a computationally tractable constrained optimization problem. The code for this algorithm is not publicly available yet, but applying this data and estimand to this algorithm can help policy researchers better understand the gravity of the discrimination in policing.

6 Discussion

The paper makes the following main contributions:

1. A local causal estimator for ATE conditioned on the administrative data does not tell you anything about the global ATE

2. A novel global causal risk ratio can be used to get the global effects and has been used in practice on the NYPD Stop and Frisk dataset

Zhao et al. found in their analysis of the NYPD Stop and Frisk data that for Black civilians, the risk of experiencing force is much higher on average than for White civilians.

Some limitations that came from their estimator is that it is hard to find data to properly estimate the bias term. So the analysis done for NYPD does not have any information about police encounters and should only be taken as very crude estimates of the estimator. Second, the CRR is conditional on covariates X, which is needed for identification due to the treatment ignorability conditional on confounders included in X assumption. So ideally one would want to condition simultaneously on confounders like time, location, and other relevant characteristics of the police-civilian encounters. However, this is usually not available in datasets. Most of Zhao et al.'s analysis focuses on only conditioning on at most two confounders (age and gender, or precinct). Their method does not have a way to summarize over multiple covariate strata even if the risk ratios are identified and estimated so that it is easier to use in practice. Because they did not account for multiple confounders that are associated with race such as criminal activity, they may have overestimated the effect of race on use of force. This lack of being able to average over multiple confounders and the lack of high quality supporting data affects both the estimates as well as inference. The authors, specifically Marshall Joffe, were considering how to estimate a marginal risk ratio but due to some circumstances, they haven't been able to push forward this work. Finally, since there is a lot of movement in New York, the census data may be a bad estimator of the police-civilian encounters, especially for extreme residential distributions.

KLM detailed a research plan where they encourage the use of traffic cameras to allow all police encounters to be measured. With this additional data they claim that their analysis provides a guideline for how to measure racial discrimination in policing while correcting for posttreatment selection bias. Zhao et al. say that their analysis will help supplement this research plan outlined by KLM with their global risk ratio and sensitivity analysis.

References

Avidit Acharya, Matthew Blackwell, and Maya Sen. Explaining causal findings without bias: Detecting and assessing direct effects. *American Political Science Review*, 110(3):512–529, 2016.

Joshua D Angrist and Jörn-Steffen Pischke. Mostly harmless econometrics. In *Mostly Harmless Econometrics*. Princeton university press, 2008.

Guilherme Duarte, Noam Finkelstein, Dean Knox, Jonathan Mummolo, and Ilya Shpitser. An automated approach to causal inference in discrete settings. *arXiv preprint arXiv:2109.13471*, 2021.

Felix Elwert and Christopher Winship. Endogenous selection bias: The problem of conditioning on a collider variable. *Annual review of sociology*, 40:31, 2014.

Jr. Fryer, Roland G. Reconciling results on racial differences in police shootings. *AEA Papers and Proceedings*, 108:228–33, May 2018. doi: 10.1257/pandp.20181004. URL `https://www.aeaweb.org/articles?id=10.1257/pandp.20181004`.

Roland G Fryer Jr. An empirical analysis of racial differences in police use of force. *Journal of Political Economy*, 127(3):1210–1261, 2019.

Andrew Gelman, Jeffrey Fagan, and Alex Kiss. An analysis of the new york city police department's "stop-and-frisk" policy in the context of claims of racial bias. *Journal of the American statistical association*, 102(479):813–823, 2007.

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

James J Heckman. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pages 153–161, 1979.

David J Johnson, Trevor Tress, Nicole Burkel, Carley Taylor, and Joseph Cesario. Officer characteristics and racial disparities in fatal officer-involved shootings. *Proceedings of the National Academy of Sciences*, 116(32):15877–15882, 2019.

Dean Knox and Jonathan Mummolo. Making inferences about racial disparities in police violence. *Proceedings of the National Academy of Sciences*, 117(3):1261–1262, 2020.

Dean Knox, Will Lowe, and Jonathan Mummolo. Administrative records mask racially biased policing. *American Political Science Review*, 114(3):619–637, 2020. doi: 10.1017/S0003055420000039.

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

Brendan Nyhan, Christopher Skovron, and Rocío Titiunik. Differential registration bias in voter file data: A sensitivity analysis approach. *American Journal of Political Science*, 61(3):744–760, 2017.

Greg Ridgeway. Assessing the effect of race bias in post-traffic stop outcomes using propensity scores. *Journal of quantitative criminology*, 22(1):1–29, 2006.

Paul R Rosenbaum. 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): 656–666, 1984.

Qinguan Zhao, Luke J Keele, Dylan S Small, and Marshall M Joffe. A note on posttreatment selection in studying racial discrimination in policing. *American Political Science Review*, 116(1): 337–350, 2022. doi: 10.1017/S0003055421000654.

7 Appendix

7.1 Proofs for showing that no relationship between local and global estimator

Let (D, M, Y) be generated from a nonparametric structural equation model (SEM):

$$D = f_D(\epsilon_D), M = f_M(D, \epsilon_M), Y = f_Y(D, M, \epsilon_Y)$$

where $\epsilon_D, \epsilon_M, \epsilon_Y$ are mutually independent

Since $\epsilon_D, \epsilon_M, \epsilon_Y$ are mutually independent, we have

$$\begin{aligned} D = f_D(\epsilon_D) &\perp \{M(d=0) = f_M(d=0, \epsilon_m), M(d=1) = f_M(d=1, \epsilon_m)\} \\ &\perp \{Y(d=0, m=0) = f_Y(d=0, m=0, \epsilon_Y), Y(d=0, m=1) = f_Y(d=0, m=1, \epsilon_Y), \\ &\quad Y(d=1, m=0) = f_Y(d=1, m=0, \epsilon_Y), Y(d=1, m=1) = f_Y(d=1, m=1, \epsilon_Y)\} \end{aligned}$$

Also make the Mandatory reporting assumption:

(a) $Y(0,0) = Y(1,0) = 0$

(b) administrative data contains all detainments/stops of civilians by the police

Will derive expressions for $ATE_{M=1}$ and $ATE_{M=1}$ using the following two basic causal effects:

- $\beta_M = E[M(1) - M(0)]$, the average effect of race on detainment
- $\beta_Y = E[Y(1,1) - Y(0,1)]$, the controlled direct effect (CDE) of race on police violence

Also we will introduce a new variable S to introduce the principle stratum:

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

Using this notation we have that

$$\begin{aligned} \beta_M &= E[M(1) - M(0)] \\ &= E[E[M(1) - M(0)|S = s]] \end{aligned}$$

$$
\begin{aligned}
&= \sum_{s \in S} E[M(1) - M(0)|S = s]P(S = s) \\
&= \overbrace{E[M(1) - M(0)|S = al]}^{=0} P(S = al) + \overbrace{E[M(1) - M(0)|S = mi]}^{=1} P(S = mi) \\
&+ \overbrace{E[M(1) - M(0)|S = ma]}^{=-1} P(S = ma) + \overbrace{E[M(1) - M(0)|S = ne]}^{=0} P(S = ne) \\
&= P(S = mi) - P(S = ma)
\end{aligned}
$$

Let

$$
\theta = \begin{pmatrix} E[Y(1) - Y(0)|S = al] \\ E[Y(1) - Y(0)|S = mi] \\ E[Y(1) - Y(0)|S = ma] \\ E[Y(1) - Y(0)|S = ne] \end{pmatrix} \tag{12}
$$

Then we have that

$$
\begin{aligned}
\theta = &\begin{pmatrix} E[Y(1) - Y(0)|S = al] \\ E[Y(1) - Y(0)|S = mi] \\ E[Y(1) - Y(0)|S = ma] \\ E[Y(1) - Y(0)|S = ne] \end{pmatrix} = \begin{pmatrix} E[Y(1,1) - Y(0,1)|S = al] \\ E[Y(1,1) - Y(0,0)|S = mi] \\ E[Y(1,0) - Y(0,1)|S = ma] \\ E[Y(1,0) - Y(0,0)|S = ne] \end{pmatrix} \\
&= \begin{pmatrix} E[Y(1,1) - Y(0,1)] \\ E[Y(1,1) - Y(0,0)] \\ E[Y(1,0) - Y(0,1)] \\ E[Y(1,0) - Y(0,0)] \end{pmatrix} \text{because } M(d) \perp Y(d,m) \\
&= \begin{pmatrix} \beta_Y \\ E[Y(1,1)] + 0 \\ 0 - E[Y(0,1)] \\ 0 - 0 \end{pmatrix} \text{by Assumption 1} \\
&= \begin{pmatrix} \beta_Y \\ E[Y(1,1)] + \beta_Y - \beta_Y \\ -E[Y(0,1)] \\ 0 \end{pmatrix} = \begin{pmatrix} \beta_Y \\ E[Y(1,1)] + \beta_Y - E[Y(1,1)] + E[Y(0,1)] \\ -E[Y(0,1)] \\ 0 \end{pmatrix}
\end{aligned}
$$

$$= \begin{pmatrix} \beta_Y \\ \beta_Y + E[Y(0,1)] \\ -E[Y(0,1)] \\ 0 \end{pmatrix}$$

Proposition 7.1. *For simplification, assume there is no unmeasured mediator-outcome confounder in the DAG (no U). Under Assumption 1 (Mandatory Reporting), the estimands* $ATE_{M=1}, ATT_{M=1}, ATE = E[Y(1) - Y(0)]$*, and* $ATT = E[Y(1) - Y(0)|D = 1]$ *can be written as weighted averages* $w^T\theta$ *with weights given by:*

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

,

$$w(ATT_{M=1}) = \begin{pmatrix} P(S = al) \\ P(S = ma) + \beta_M \\ 0 \\ 0 \end{pmatrix}$$

, and

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

Proof:

First consider $ATE_{M=1}$

$$\begin{aligned} ATE_{M=1} &= E[Y(1) - Y(0)|M = 1] \\ &= E[E[Y(1) - Y(0)|M = 1, S = s] \\ &= \sum_{s \in S} E[Y(1) - Y(0)|M = 1, S = s]P(S = s|M = 1) \end{aligned}$$

$$= E[Y(1) - Y(0)|M = 1, S = al]P(S = al|M = 1)$$

$$+ E[Y(1) - Y(0)|M = 1, S = mi]P(S = mi|M = 1)$$

$$+ E[Y(1) - Y(0)|M = 1, S = ma]P(S = ma|M = 1)$$

$$+ E[Y(1) - Y(0)|M = 1, S = ne]P(S = ne|M = 1)$$

$$= E[Y(1) - Y(0)|M = 1, M(0) = M(1) = 1]P(M(0) = M(1) = 1|M = 1)$$

$$+ E[Y(1) - Y(0)|M = 1, M(0) = 0, M(1) = 1]P(M(0) = 0, M(1) = 1|M = 1)$$

$$+ E[Y(1) - Y(0)|M = 1, M(0) = 1, M(1) = 0]P(M(0) = 1, M(1) = 0|M = 1)$$

$$+ E[Y(1) - Y(0)|M = 1, M(0) = M(1) = 0]P(M(0) = M(1) = 0|M = 1)$$

We can simplify the principle stratum effects using recursive substitution of the potential outcomes.

For $m_0, m_1 \in \{0, 1\}$,

$E[Y(1) - Y(0)|M = 1, M(0) = m_0, M(1) = m_1]$

$= E[Y(1, M(1)) - Y(0, M(0))|M = 1, M(0) = m_0, M(1) = m_1]$

$= E[Y(1, m_1) - Y(0, m_0)|M = 1, M(0) = m_0, M(1) = m_1]$

$= E[Y(1, m_1) - Y(0, m_0)|M(0) = m_0, M(1) = m_1]$

since $M \perp \{Y(1, m_1), Y(0, m_0)\}|\{M(0), M(1)\}$

Only random term in $M = D \cdot M(1) + (1 - D) \cdot M(0)$ is D and have mutual independence so have

$= E[Y(1, m_1) - Y(0, m_0)]$ because mutually independence between $D, Y(d, m)$, and $M(d)$

So have

$$\begin{aligned} ATE_{M=1} &= E[Y(1,1) - Y(0,1)|M(0) = M(1) = 1]P(S = al|M = 1) \\ &+ E[Y(1,1) - Y(0,0)|M(0) = 0, M(1) = 1]P(S = mi|M = 1) \\ &+ E[Y(1,0) - Y(0,1)|M(0) = 1, M(1) = 0]P(S = ma|M = 1) \\ &+ E[Y(1,0) - Y(0,0)|M(0) = M(1) = 0]P(S = ne|M = 1) \end{aligned}$$

$$= \begin{pmatrix} E[Y(1,1) - Y(0,1)|M(0) = M(1) = 1] \\ E[Y(1,1) - Y(0,0)|M(0) = 0, M(1) = 1] \\ E[Y(1,0) - Y(0,1)|M(0) = 1, M(1) = 0] \\ E[Y(1,0) - Y(0,0)|M(0) = M(1) = 0] \end{pmatrix}^T \begin{pmatrix} P(S = al|M = 1) \\ P(S = mi|M = 1) \\ P(S = ma|M = 1) \\ P(S = ne|M = 1) \end{pmatrix}$$

$$= \theta^T w(ATE_{M=1})$$

Second consider $ATT_{M=1}$

$$\begin{aligned} ATT_{M=1} &= E[Y(1) - Y(0)|D = 1, M = 1] \\ &= E[E[Y(1) - Y(0)|D = 1, M = 1, S = s] \\ &= \sum_{s \in S} E[Y(1) - Y(0)|D = 1, M = 1, S = s] P(S = s|D = 1, M = 1) \end{aligned}$$

We can simplify the principle stratum effects using recursive substitution of the potential outcomes.

For $m_0, m_1 \in \{0, 1\}$,

$E[Y(1) - Y(0)|D = 1, M = 1, M(0) = m_0, M(1) = m_1]$

$= E[Y(1, M(1)) - Y(0, M(0))|D = 1, M = 1, M(0) = m_0, M(1) = m_1]$

$= E[Y(1, m_1) - Y(0, m_0)|D = 1, M = 1, M(0) = m_0, M(1) = m_1]$

$= E[Y(1, m_1) - Y(0, m_0)|M(0) = m_0, M(1) = m_1]$

since $M, D \perp \{Y(1, m_1), Y(0, m_0)\}|\{M(0), M(1)\}$

Only random term in $M = D \cdot M(1) + (1 - D) \cdot M(0)$ is D and have mutual independence so have

$= E[Y(1, m_1) - Y(0, m_0)]$ because mutually independence between $D, Y(d, m)$, and $M(d)$

So have

$$ATT_{M=1} = \sum_{s \in S} E[Y(1) - Y(0)|D = 1, M = 1, S = s] P(S = s|D = 1, M = 1)$$

$$= \begin{pmatrix} E[Y(1,1) - Y(0,1)|D = 1, M = 1, S = al] \\ E[Y(1,1) - Y(0,0)|D = 1, M = 1, S = mi] \\ E[Y(1,0) - Y(0,1)|D = 1, M = 1, S = ma] \\ E[Y(1,0) - Y(0,0)|D = 1, M = 1, S = ne] \end{pmatrix}^T \begin{pmatrix} P(S = al|M = 1, D = 1) \\ P(S = mi|M = 1, D = 1) \\ P(S = ma|M = 1, D = 1) \\ P(S = ne|M = 1, D = 1) \end{pmatrix}$$

$$= \begin{pmatrix} E[Y(1,m_1) - Y(0,m_0)|S = al] \\ E[Y(1,m_1) - Y(0,m_0)|S = mi] \\ E[Y(1,m_1) - Y(0,m_0)|S = ma] \\ E[Y(1,m_1) - Y(0,m_0)|S = ne] \end{pmatrix}^T \begin{pmatrix} P(S = al|M = 1, D = 1) \\ P(S = mi|M = 1, D = 1) \\ P(S = ma|M = 1, D = 1) \\ P(S = ne|M = 1, D = 1) \end{pmatrix}$$

$$= \theta^T w(ATT_{M=1})$$

Third consider ATE

$$\begin{aligned} ATE &= E[Y(1) - Y(0)] \\ &= E[E[Y(1) - Y(0)|S = s]] \\ &= \sum_{s \in S} E[Y(1) - Y(0)|S = s]P(S = s) \\ &= \begin{pmatrix} E[Y(1) - Y(0)|S = al] \\ E[Y(1) - Y(0)|S = mi] \\ E[Y(1) - Y(0)|S = ma] \\ E[Y(1) - Y(0)|S = ne] \end{pmatrix}^T \begin{pmatrix} P(S = al) \\ P(S = mi) \\ P(S = ma) \\ P(S = ne) \end{pmatrix} \\ &= \begin{pmatrix} E[Y(1) - Y(0)|S = al] \\ E[Y(1) - Y(0)|S = mi] \\ E[Y(1) - Y(0)|S = ma] \\ E[Y(1) - Y(0)|S = ne] \end{pmatrix}^T \begin{pmatrix} P(S = al) \\ P(S = ma) + \beta_M \\ P(S = ma) \\ P(S = ne) \end{pmatrix} \\ &= \theta^T w(ATE) \end{aligned}$$

Finally consider ATT

$$\begin{aligned} ATT &= E[Y(1) - Y(0)|D = 1] \\ &= E[E[Y(1) - Y(0)|D = 1, S = s]] \\ &= \sum_{s \in S} E[Y(1) - Y(0)|S = s, D = 1]P(S = s) \end{aligned}$$

Have that $E[Y(1) - Y(0)|D = 1] = E[Y(1) - Y(0)]$ because of mutual independence, so

$$= \begin{pmatrix} E[Y(1)-Y(0)|D=1,S=al] \\ E[Y(1)-Y(0)|D=1,S=mi] \\ E[Y(1)-Y(0)|D=1,S=ma] \\ E[Y(1)-Y(0)|D=1,S=ne] \end{pmatrix}^T \begin{pmatrix} P(S=al) \\ P(S=mi) \\ P(S=ma) \\ P(S=ne) \end{pmatrix}$$

$$= \begin{pmatrix} E[Y(1)-Y(0)|S=al] \\ E[Y(1)-Y(0)|S=mi] \\ E[Y(1)-Y(0)|S=ma] \\ E[Y(1)-Y(0)|S=ne] \end{pmatrix}^T \begin{pmatrix} P(S=al) \\ P(S=maa0+\beta_M \\ P(S=ma) \\ P(S=ne) \end{pmatrix}$$

$$= \theta^T w(ATT)$$

Next will compute the conditional probabilities for the principle strat in $w(ATE_{M=1})$ and $w(ATT_{M=1})$

Have using Bayes Rule that for any $m_0, m_1 \in \{0,1\}$

$$P(M(0)=m_0, M(1)=m_1|M=1)$$

$$\propto P(M(0)=m_0, M(1)=m_1)P(M=1|M(0)=m_0, M(1)=m_1)$$

$$= P(M(0)=m_0, M(1)=m_1)\sum_{d=0}^{1} P(M=1, D=d|M(0)=m_0, M(1)=m_1)$$

$$= P(M(0)=m_0, M(1)=m_1)\sum_{d=0}^{1} P(M(D)=1, D=d|M(0)=m_0, M(1)=m_1)$$

$$= P(M(0)=m_0, M(1)=m_1)$$

$$\sum_{d=0}^{1} P(M(D)=1|D=d, M(0)=m_0, M(1)=m_1)P(D=d|M(0)=m_0, M(1)=m_1)$$

$$= P(M(0)=m_0, M(1)=m_1)\sum_{d=0}^{1} I_{m_d=1}P(D=d|M(0)=m_0, M(1)=m_1)$$

since $M(D)=M$ and $D \perp M(d)$

$$= P(M(0)=m_0, M(1)=m_1)\sum_{d=0}^{1} I_{m_d=1}P(D=d)$$

since $D \perp M(d)$

Then we can obtain the form of $w(ATE_{M=1})$

$w(ATE_{M=1})$

$$= \begin{pmatrix} P(S = al|M = 1) \\ P(S = mi|M = 1) \\ P(S = ma|M = 1) \\ P(S = ne|M = 1) \end{pmatrix}$$

$$= \begin{pmatrix} P(M(0) = M(1) = 1) \sum_{d=0}^{1} P(D = d) = P(M(0) = M(1) = 1) = P(S = al) \\ P(M(0) = 0, M(1) = 1)P(D = 1) = P(S = mi)P(D = 1) = (P(S = ma) + \beta_M)P(D = 1) \\ P(M(0) = 1, M(1) = 0)P(D = 0) = P(S = ma)P(D = 0) \\ P(M(0) = M(1) = 0) - 0 = 0 \end{pmatrix}$$

Similarly for $ATT_{M=1}$

$$\begin{aligned}
&P(M(0)=m_0, M(1)=m_1|D=1, M=1)\\
&\propto P(M(0)=m_0, M(1)=m_1)P(M=1|M(0)=m_0, M(1)=m_1, D=1)\\
&= P(M(0)=m_0, M(1)=m_1)\sum_{d=0}^{1} P(M(D)=1|M(0)=m_0, M(1)=m_1, D=1)\\
&= I_{m_1=1}P(M(0)=m_0, M(1)=m_1)\\
&\text{since } M(D)=M \text{ and } D \perp M(d)
\end{aligned}$$

$$\begin{aligned}
&w(ATT_{M=1})\\
&= \begin{pmatrix} P(S=al|M=1,D=1)\\ P(S=mi|M=1,D=1)\\ P(S=ma|M=1,D=1)\\ P(S=ne|M=1,D=1)\end{pmatrix}\\
&= \begin{pmatrix} P(M(0)=M(1)=1)=P(S=al)\\ P(M(0)=0, M(1)=1)=P(S=mi)=P(S=ma)+\beta_M\\ 0\\ 0\end{pmatrix}
\end{aligned}$$

□

Proposition 7.2. *Under the same assumptions as above, $PIE = \beta_M \times E[(Y(1,1)]$ and $PDE = \beta_Y \times E[M(0)]$*

Proof:

For any $d, d' \in \{0,1\}$:

$$\begin{aligned}
E[Y(d, M(d'))] &= E[E[Y(d, M(d'))|M(d')=1]]\\
&= E[E[Y(d,1)|M(d')=1]]\\
&= E[Y(d,1)|M(d')=1]P(M(d')=1)\\
&= E[Y(d,1)]P(M(d')=1) && \text{because } Y(d,m) \perp M(d)
\end{aligned}$$

So we have that

$$\begin{aligned} PDE &= E[Y(1, M(0)) - Y(0, M(0))] \\ &= E[Y(1, M(0))] - E[Y(0, M(0))] \\ &= E[Y(1,1)]P(M(0) = 1) - E[Y(0,1)]P(M(0) = 1) \\ &= \beta_Y \times P(M(0) = 1) \end{aligned}$$

and

$$\begin{aligned} PIE &= E[Y(1, M(1)) - Y(1, M(0))] \\ &= E[Y(1, M(1))] - E[Y(1, M(0))] \\ &= E[Y(1,1)]P(M(1) = 1) - E[Y(1,1)]P(M(0) = 1) \\ &= E[(Y(1,1)]\{P(M(1) = 1) - P(M(0) = 1)\} \\ &= E[Y(1,1)]\{E[M(1) - M(0)]\} \\ &= \beta_M \times E[Y(1,1)] \end{aligned}$$

Corollary 7.1. *Let the assumptions from Proposition 5.1 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 $ATT_{M=1}$.*

Will first show that the ATT and ATE would have the same sign as β_M when β_M and β_Y have the same sign:

$$ATE = ATT = PIE + PDE$$

if $\beta_M \geq 0$ and $\beta_Y \geq 0$, then

$$ATE = ATT = \underbrace{\beta_M}_{\geq 0} \underbrace{E[Y(1,1)]}_{\geq 0} + \underbrace{\beta_Y}_{\geq 0} \underbrace{E[M(0)]}_{\geq 0} \geq 0 \qquad \text{by Corollary 7.1}$$

if $\beta_M \leq 0$ and $\beta_Y \leq 0$, then

$$ATE = ATT = \underbrace{\beta_M}_{\leq 0} \underbrace{E[Y(1,1)]}_{\geq 0} + \underbrace{\beta_Y}_{\leq 0} \underbrace{E[M(0)]}_{\geq 0} \leq 0 \qquad \text{by Corollary 7.1}$$

However this property doesn't hold for $ATE_{M=1}$ and $ATT_{M=1}$. We will show this through the following counterexamples. Explicit computation of these counterexamples are available with the other available code for this project.

(i) When $\beta_M = \beta_Y = 0.01, P(S = al) = 0.1, P(S = ma) = 0.05, E[Y(0,1)] = 0.1$ and $P(D = 1) = 0.01$, we have that $ATE_{M=1} = -0.003884$

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

(iii) When $\beta_M = \beta_Y = -0.01, P(S = al) = 0.1, P(S = ma) = 0.05, E[Y(0,1)] = 0.1$ and $P(D = 1) = 0.01$, we have that $ATE_{M=1} = 0.0026$

Intuitively this is because all of the causal estimands above including $\beta_M, \beta_Y, ATE, ATE_{M=1}$, and $ATT_{M=1}$, measure some weighted average treatment effect for the use of force or police detainment. So conditioning on the post-treatment mediator M may result in unintuitive weights.

We see that $ATE_{M=1}$ and ATE can have different signs from the following expression:

$$\begin{aligned}
ATE &= E[Y(1) - Y(0)] \\
&= E[E[Y(1) - Y(0)|M = m]] \\
&= \sum_{m=1}^{1} E[Y(1) - Y(0)|M = m]P(M = m) \\
&= E[Y(1) - Y(0)|M = 1]P(M = 1) + E[Y(1) - Y(0)|M = 0]P(M = 0) \\
&= ATE_{M=1}P(M = 1) + \qquad \underbrace{E[Y(1) - Y(0)|M = 0]P(M = 0)}_{\text{may be non-zero and have opposite sign of } ATE_{M=1}}
\end{aligned}$$

Note that we cannot drop the second term due to Assumption 1 (i.e $Y(0,0) = Y(1,0) = 0$ with probability 1). This is because $Y(d, M(d))$ and $Y(d, 0)$ are two different random variables. So we do not have that even after conditioning on $M = 0$, that $M(d) = 0$ necessarily since $M = 0$ is observed data and $M(d)$ is the counterfactual.

The main issue is that condition on a posttreatment variable M alters the principle strata weights which were derived in Proposition 7.1. The local estimands depend on the racial bias in detainment on use of force (in β_M and β_Y) but also $E[Y(0,1)]$, the baseline rate of violence, and $P(D=1)$, the composition of race. For the first counterexample, we have that the minority group is discriminated against in both detainment and use of force but since the baseline violence is high and the minority group is small, we have that $ATE_{M=1}$ is mostly determined by the smaller bias experienced by the larger majority group, which is captured by $P(S=ma)$.

□

Corollary 7.2. *Let the assumptions from Proposition 7.1 be given.*

1. *If one assumes that $D=1$ is in fact the minority group, $P(D=1)<0.5$, then we have that $\beta_Y<0, \beta_M<0$ implies that $ATE_{M=1}<0$*
2. *If one assumes mediator monotonicity $P(S=ma)=0$, then we have that the local and global estimands signs are consistent with one another*

We have that

$$\beta_M = P(S=mi) - P(S=ma)$$

and

$$\beta_Y = E[Y(1,1) - Y(0,1)]$$

Proof of (1):

$$\begin{aligned} w(ATE_{M=1})^T\theta = {} & P(S=al)\beta_Y + \beta_M P(D=1)\beta_Y + \beta_M P(D=1)E[Y(0,1)] + P(S=ma)P(D=1)\beta_Y \\ & + \underbrace{2P(S=ma)P(D=1)E[Y(0,1)] - P(S=ma)E[Y(0,1)]}_{\text{if } D<0.5, \text{ then this term } <0} \end{aligned}$$

Then we have that when we plug in β_M, β_Y into the top term that

$$\begin{aligned} & P(S=al)\beta_Y + \beta_M P(D=1)\beta_Y + \beta_M P(D=1)E[Y(0,1)] + P(S=ma)P(D=1)\beta_Y \\ & = P(S=al)\{E[Y(1,1)] - E[Y(0,1)]\} + P(D=1)\{P(S=mi)E[Y(1,1)] - P(S=ma)E[Y(0,1)]\} \end{aligned}$$

Since $\beta_Y<0$, we have that

1. $P(S=al)\{E[Y(1,1)] - E[Y(0,1)]\} < 0$

2. $P(S = mi)E[Y(1,1)] < P(S = ma)E[Y(0,1)]$

3. $P(D = 1)\{P(S = mi)E[Y(1,1)] - P(S = ma)E[Y(0,1)]\} < 0$

When we assume $P(D = 1) < 0.5$, we have that $w(ATE_{M=1})^T\theta < 0$, since it is the sum of two negative terms. So with this assumption we have that if $\beta_Y < 0, \beta_M < 0$, this implies that $ATE_{M=1} < 0$ and that ATE, ATT are negative as shown in Corollary 7.1.

Proof of (2):

Assume that $P(S = ma) = 0$, then we have that

$$\beta_M = P(S = mi) - P(S = ma) = P(S = mi)$$

and

$$\beta_Y = E[Y(1,1) - Y(0,1)]$$

We have that

$$\begin{aligned} w(ATE_{M=1})^T\theta &= P(S = al)\beta_Y + \beta_M P(D = 1)\beta_Y + \beta_M P(D = 1)E[Y(0,1)] + P(S = ma)P(D = 1)\beta_Y \\ &\quad + 2P(S = ma)P(D = 1)E[Y(0,1)] - P(S = ma)E[Y(0,1)] \\ &= P(S = al)\beta_Y + \beta_M P(D = 1)\beta_Y + \beta_M P(D = 1)E[Y(0,1)] \end{aligned}$$

if $P(S = ma) = 0$, then it is not possible for $\beta_M < 0$ so then we will only have to worry about the case when $\beta_M > 0$ and $\beta_Y > 0$. We can clearly see that all the terms above will still be positive.

Now we will look at the $ATT_{M=1}$

$$w(ATT_{M=1})^T\theta = P(S = al)\beta_Y + \beta_M\beta_Y + \beta_M E[Y(0,1)] + P(S = ma)\beta_Y + P(S = ma)E[Y(0,1)]$$

If $\beta_M > 0, \beta_Y > 0$, we have that all the terms are positive. So if we assume $P(S = ma) = 0$, then we have that $\beta_M \geq 0$. So we will not need to worry about the case where $\beta_M < 0, \beta_Y < 0$

Thus we have that when we assume mediator monotonicity, then sign of the local estimands $ATE_{M=1}, ATT_{M=1}$ are consistent with the signs of the global estimands. This is true because with mediator monotonicity, we would only be looking at the case when $\beta_M > 0, \beta_Y > 0$ as reasoned above, and as shown above we have that $ATE_{M=1}, ATT_{M=1}$ are positive. And as shown in Corollary 7.1 if $\beta_M > 0, \beta_Y > 0$, then ATE, ATT are positive.

7.2 Derivation of the causal risk ratio (CRR)

$$
\begin{aligned}
E[Y(d)|X] &= E[E[Y(d)|M(d)=1, X=x]|X=x] \\
&= E[Y(d)|M(d)=1, X=x]P(M(d)=1|X=x) \\
&= E[Y(d,1)|M(d)=1, X=x]P(M(d)=1|X=x) \\
&= E[Y(d,1)|M(d)=1, D=d, X=x]P(M(d)=1|X=x) \qquad D \perp Y(d,1)|M(d), X
\end{aligned}
$$

(i.e conditional treatment ignorability)

$$
\begin{aligned}
&= E[Y|M=1, D=d, X=x]P(M(d)=1|X=x) \qquad \text{SUTVA/consistency} \\
&= E[Y|M=1, D=d, X=x]P(M(d)=1|D=d, X=x) \qquad D \perp M(d) \\
&= E[Y|M=1|D=d, X=x]P(M=1|D=d, X=x) \qquad d=0,1
\end{aligned}
$$

$$
\begin{aligned}
E[Y(1)|X=x] &= E[Y|M=1|D=1, X=x]P(M=1|D=1, X=x) \\
P(M=1|D=1, X=x) &= \frac{P(D=1|M=1, X=x)P(X=x, M=1)}{P(M=1|D=1, X=x)P(X=x)} \\
E[Y(0)|X=x] &= E[Y|M=1|D=0, X=x]P(M=1|D=0, X=x) \\
P(M=1|D=0, X=x) &= \frac{P(D=0|M=1, X=x)P(X=x, M=1)}{P(M=1|D=0, X=x)P(X=x)} \\
\frac{E[Y(1)|X=x]}{E[Y(0)|X=x]} &= \frac{E[Y|M=1|D=1, X=x]}{E[Y|M=1|D=0, X=x]} \times \frac{\frac{P(D=1|M=1,X=x)P(X=x,M=1)}{P(M=1|D=1,X=x)P(X=x)}}{\frac{P(D=0|M=1,X=x)P(X=x,M=1)}{P(M=1|D=0,X=x)P(X=x)}} \\
&= \frac{E[Y|M=1|D=1, X=x]}{E[Y|M=1|D=0, X=x]} \times \frac{\frac{P(D=1|M=1,X=x)}{P(M=1|D=1,X=x)}}{\frac{P(D=0|M=1,X=x)}{P(M=1|D=0,X=x}}
\end{aligned}
$$

□

7.3 Analysis by Precinct

Zhao et al. examined the causal risk ratio conditioned on location (precinct) of the civilian in the encounter. The results of analysis are presented in Figures 6-7 below. Figures 6 and 7 visualize the number of minority (Black) civilians per precinct and the number of minorities detained per precinct.

Racial Distributions (by fill color) in each NYPD Precinct

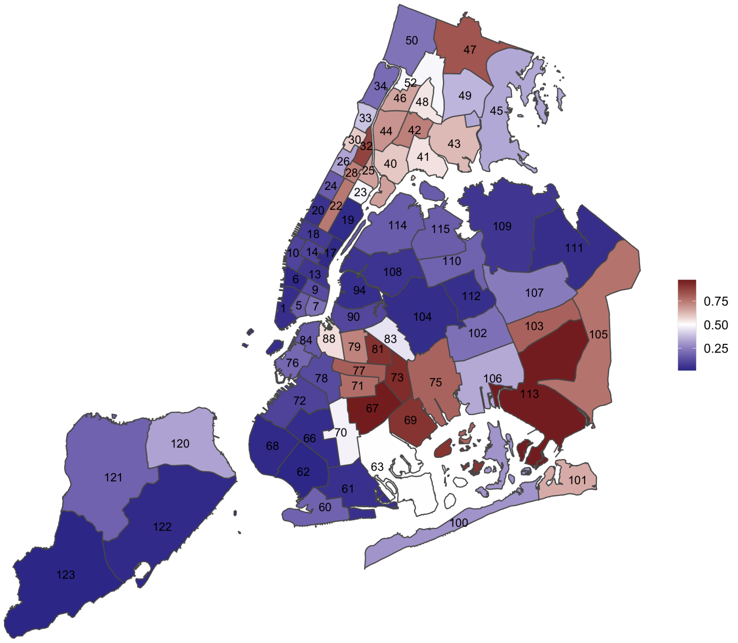

Figure 6: Proportion of Black residents in 2010 census data

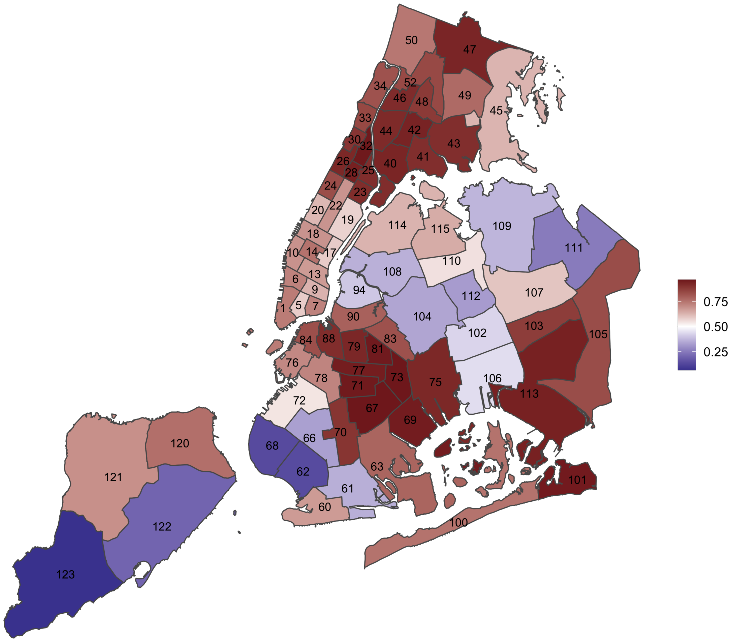

Figure 7: Proportion of detainments of Black civilians in NYPD Stop and Frisk dataset

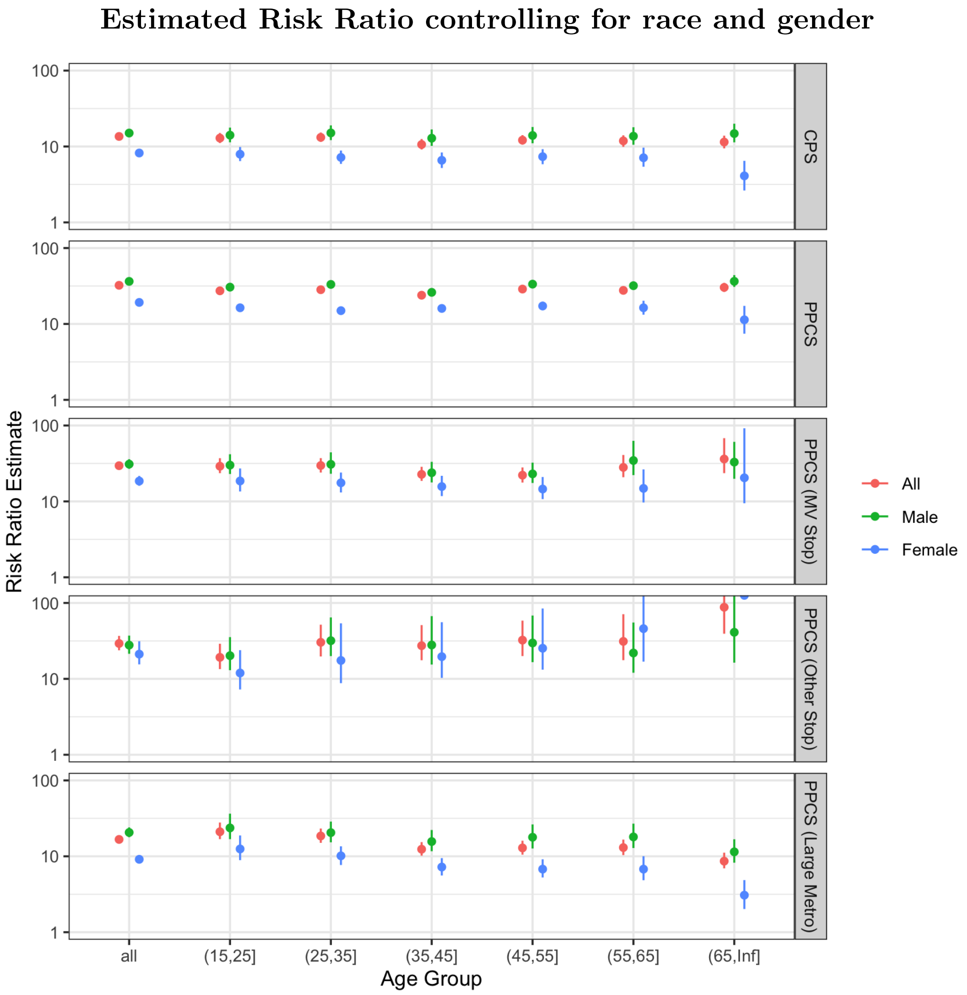

Figure 8: Results of the stratified analysis of the NYPD Stop-and-Frisk dataset by age and gender. The estimated risk ratio is truncated at 100.

7.4 Stratified Analysis by Age and Gender

Zhao et al. examined the causal risk ratio conditioned on age and gender of the civilian in the encounter. The results of analysis are presented in Figure 8. We see that gender is an important effect modifier but age does not seem to be so.

7.5 Algorithm for Computing CRR

Algorithm 1: Algorithm for Estimating CRR

```
1. Estimate naive risk ratio using the administrative data;
       $\frac{E[Y|D=1,M=1,X=x]}{E[Y|D=0,M=1,X=x]}$
   get the mean use of force of stops with Black individuals and divide by the mean use of force of stops with White individuals
2. Estimate the numerator of the bias term using administrative data; calculate the relative probability of detainment being with a minority
3. Estimate the denominator of the bias term using supplementary data (e.g census data); calculate the relative probability of having a Black encounter
```