Below is the Online Supplement for "A note on post-treatment 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: $D = f_D(\epsilon_D), M = f_M(D, \epsilon_M), Y = f_Y(D, M, \epsilon_Y)$ where $\epsilon_D, \epsilon_M, \epsilon_Y$ are mutually independent (Pearl 2009). Potential outcomes for M and Y can be defined by replacing random variables in the functions by fixed values; for example, $M(d) = f_M(d, \epsilon_M),\ d = 0, 1$. 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 $\text{ATE}_{M=1}$ and $\text{ATT}_{M=1}$ 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 $\mathcal{S} = \{\text{al}, \text{mi}, \text{ma}, \text{ne}\}$ be all possible values for S. Using this notation, we have

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

By using the independence between $M(d)$ and $Y(d, m)$ and Assumption 1, it is easy to show that

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

Average treatment effects, whether conditional on M or D or not, can be written as weighted averages of the entries of $\boldsymbol{\theta}$.

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 $= \mathbb{E}[Y(1)-Y(0) \mid D=1]$ can be written as weighted averages $(\boldsymbol{w}^T\boldsymbol{\theta})/(\boldsymbol{w}^T\mathbf{1})$ ($\mathbf{1}$ is the all-ones vector) with weights given by, respectively,*

$$\boldsymbol{w}(\mathit{ATE}_{M=1}) = \begin{pmatrix} \mathbb{P}(S=\mathit{al}) \\ \left[\mathbb{P}(S=\mathit{ma}) + \beta_M\right]\mathbb{P}(D=1) \\ \mathbb{P}(S=\mathit{ma})\,\mathbb{P}(D=0) \\ 0 \end{pmatrix}, \quad \boldsymbol{w}(\mathit{ATT}_{M=1}) = \begin{pmatrix} \mathbb{P}(S=\mathit{al}) \\ \mathbb{P}(S=\mathit{ma}) + \beta_M \\ 0 \\ 0 \end{pmatrix},$$

and

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

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:

$$\text{ATE}_{M=1} = \mathbb{E}[Y(1)-Y(0) \mid M=1] = \sum_{s \in \mathcal{S}} \mathbb{E}[Y(1)-Y(0) \mid M=1, S=s] \cdot \mathbb{P}(S=s \mid M=1).$$

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

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

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

$$\mathsf{ATE}_{M=1} = \boldsymbol{\theta}^T \boldsymbol{w}(\mathsf{ATE}_{M=1}), \text{ where } \boldsymbol{w}(\mathsf{ATE}_{M=1}) = \begin{pmatrix} \mathbb{P}(S = \mathsf{al} \mid M = 1) \\ \mathbb{P}(S = \mathsf{mi} \mid M = 1) \\ \mathbb{P}(S = \mathsf{ma} \mid M = 1) \\ \mathbb{P}(S = \mathsf{ne} \mid M = 1) \end{pmatrix}.$$

Similarly, $\mathsf{ATT}_{M=1}$, ATE, and ATT can also be written as weighted averages of the entries of $\boldsymbol{\theta}$, where the weights are

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

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

$$\mathbb{P}(M(0) = m_0, M(1) = m_1 \mid M = 1)$$

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

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

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

From this we can derive the form of $\boldsymbol{w}(\mathsf{ATT}_{M=1})$ 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,

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

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* ≥ 0*. Conversely, if* $\beta_M \leq 0$ *and* $\beta_Y \leq 0$*, then ATE* $=$ *ATT* ≤ 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 $\text{ATE}_{M=1}$ and $\text{ATT}_{M=1}$. Here are some concrete counterexamples:

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

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

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

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

$$\text{ATE} = \text{ATE}_{M=1}\,\mathbb{P}(M = 1) + \mathbb{E}[Y(1) - Y(0) \mid M = 0]\,\mathbb{P}(M = 0).$$

In this decomposition, the second term may be nonzero and have the opposite sign of $\text{ATE}_{M=1}$. 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 $\mathbb{E}[Y(d) \mid M = 0] = 0$ from $Y(d,0) = 0$, because $\mathbb{E}[Y(d) \mid M = 0] = \mathbb{E}[Y(d, M(d)) \mid M = 0]$ is not necessarily equal to $\mathbb{E}[Y(d,0) \mid M = 0]$.

The fundamental problem driving this paradox is that conditioning on the post-treatment variable M alters the weights on the principal strata, as shown in Proposition 1. $\text{ATE}_{M=1}$ and $\text{ATT}_{M=1}$ 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, $\text{ATE}_{M=1}$ becomes mostly determined by the smaller bias (captured by $\mathbb{P}(S = \text{ma}) = \mathbb{P}(M(0) = 1, M(1) = 0)$) 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 $\text{ATE}_{M=1} < 0$ whenever $\beta_M, \beta_Y < 0$. Furthermore, it can be shown that $\text{ATT}_{M=1} < 0$ whenever $\beta_M, \beta_Y > 0$. So in a very rough sense we might say that as causal estimands, $\text{ATE}_{M=1}$ is unfavorable for the minority group (because $\text{ATE}_{M=1}$ can be negative even if both $\beta_M, \beta_Y > 0$) and $\text{ATT}_{M=1}$ is unfavorable for the majority group (because $\text{ATT}_{M=1}$ 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 $\mathbb{P}(S = \text{ma}) = 0$, 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 $\text{ATE}_{M=1}$ and $\text{ATT}_{M=1}$. So a supporter of the estimand $\text{ATE}_{M=1}$ may argue that if one is willing to assume mediator monotonicity, there is no paradox regarding $\text{ATE}_{M=1}$. However, it is worthwhile to point out that under mediator monotonicity, the pure indirect effect is guaranteed to be nonnegative because $\beta_M = \mathbb{P}(S = \text{mi}) - \mathbb{P}(S = \text{ma}) = \mathbb{P}(S = \text{mi}) \geq 0$. 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 non-zero $\mathbb{P}(S = \text{ma})$ 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 Assumption 1, $\mathbb{E}[Y(d) \mid M(d) = 0] = \mathbb{E}[Y(d, 0) \mid M(d) = 0] = 0$. Therefore

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

The third equality above uses treatment ignorability: $D \perp\!\!\!\perp Y(d, 1) \mid M(d)$ (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 $D \perp\!\!\!\perp M(d)$, we have $\mathbb{P}(M(d) = 1) = \mathbb{P}(M(d) = 1 \mid D = d) = \mathbb{P}(M = 1 \mid D = d)$. Plugging this into the last display equation, we have

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

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 (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 by their age and gender. The same analysis 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.

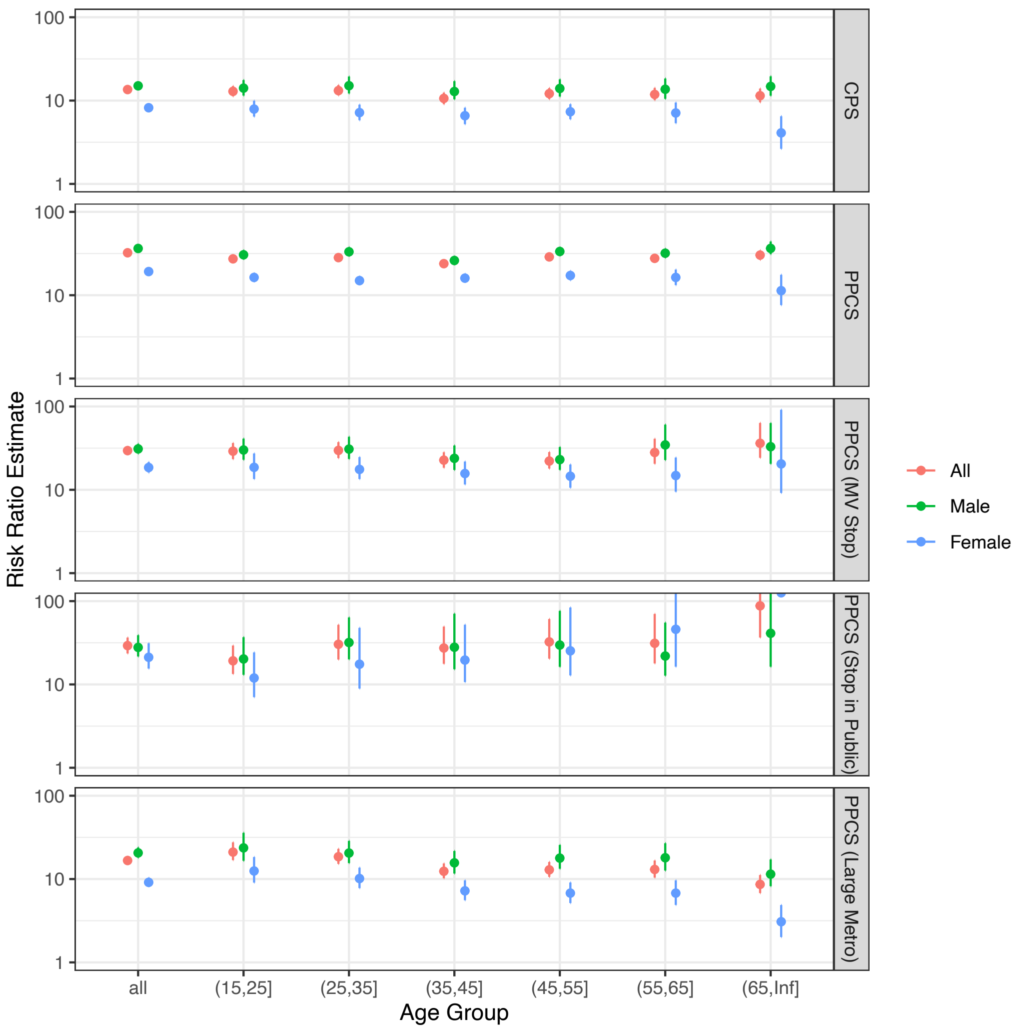

Figure D.1: Results of the stratified analysis of the NYPD Stop-and-Frisk dataset by age and gender. The estimated risk ratio is truncated at 100.