Administrative Records Mask Racially Biased Policing

Online Appendix

Contents

A Detailed proofs

A.1 Bias for $\text{ATE}_{M=1}$

We first derive the bias of the local difference in means (that is, among encounters with $X_i = x$, $\hat{\Delta}_x = \overline{Y_i|D_i = 1, M_i = 1, X_i = x} - \overline{Y_i|D_i = 0, M_i = 1, X_i = x}$), in estimating the local average treatment effect among stops, $\text{ATE}_{M=1,x} = \mathbb{E}[Y_i(1, M_i(1))|M_i = 1, X_i = x] - \mathbb{E}[Y_i(0, M_i(0))|M_i = 1, X_i = x]$. The overall bias is then given by $\sum_x \left(\mathbb{E}[\hat{\Delta}_x] - \text{ATE}_{M=1,x}\right) \Pr(X_i = x|M_i = 1)$.

$$
\begin{aligned}
&\mathbb{E}[\hat{\Delta}_x] - \text{ATE}_{M=1,x} \\
&\quad = (\mathbb{E}[\overline{Y_i|D_i = 1, M_i = 1, X_i = x} - \overline{Y_i|D_i = 0, M_i = 1, X_i = x}]) \\
&\qquad - (\mathbb{E}[Y_i(1, M_i(1))|M_i = 1, X_i = x] - \mathbb{E}[Y_i(0, M_i(0))|M_i = 1, X_i = x]) \\
&\quad = \mathbb{E}[Y_i|D_i = 1, M_i(D_i) = 1, X_i = x] - \mathbb{E}[Y_i|D_i = 0, M_i(D_i) = 1, X_i = x] \\
&\qquad - \mathbb{E}[Y_i(1, M_i(1))|M_i(D_i) = 1, X_i = x] + \mathbb{E}[Y_i(0, M_i(0))|M_i(D_i) = 1, X_i = x] \\
&\quad = \mathbb{E}[Y_i|D_i = 1, M_i(D_i) = 1, X_i = x] - \mathbb{E}[Y_i(1, M_i(1))|M_i(D_i) = 1, X_i = x] \\
&\qquad - \mathbb{E}[Y_i|D_i = 0, M_i(D_i) = 1, X_i = x] + \mathbb{E}[Y_i(0, M_i(0))|M_i(D_i) = 1, X_i = x] \\
&\quad = \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(D_i) = 1, X_i = x] \\
&\qquad - \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(D_i) = 1, X_i = x]\Pr(D_i = 1|M_i(D_i) = 1, X_i = x) \\
&\qquad - \mathbb{E}[Y_i(1, M_i(1))|D_i = 0, M_i(D_i) = 1, X_i = x]\Pr(D_i = 0|M_i(D_i) = 1, X_i = x) \\
&\qquad - \mathbb{E}[Y_i(0, M_i(0))|D_i = 0, M_i(D_i) = 1, X_i = x] \\
&\qquad + \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i(D_i) = 1, X_i = x]\Pr(D_i = 1|M_i(D_i) = 1, X_i = x) \\
&\qquad + \mathbb{E}[Y_i(0, M_i(0))|D_i = 0, M_i(D_i) = 1, X_i = x]\Pr(D_i = 0|M_i(D_i) = 1, X_i = x) \\
&\quad = \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(D_i) = 1, X_i = x]\Pr(D_i = 0|M_i(D_i) = 1, X_i = x) \\
&\qquad - \mathbb{E}[Y_i(1, M_i(1))|D_i = 0, M_i(D_i) = 1, X_i = x]\Pr(D_i = 0|M_i(D_i) = 1, X_i = x) \\
&\qquad - \mathbb{E}[Y_i(0, M_i(0))|D_i = 0, M_i(D_i) = 1, X_i = x]\Pr(D_i = 1|M_i(D_i) = 1, X_i = x) \\
&\qquad + \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i(D_i) = 1, X_i = x]\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)
\end{aligned}
$$

under mediator monotonicity, $\Pr(M_i(1) = 0|D_i = 0, M_i(D_i) = 1, X_i = x) = 0$ and $\Pr(M_i(1) = 1|D_i = 0, M_i(D_i) = 1, X_i = x) = 1$,

$$
\begin{aligned}
&= \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x] \\
&\qquad\qquad \Pr(M_i(0) = 1|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 0|M_i(D_i) = 1, X_i = x) \\
&\qquad + \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]
\end{aligned}
$$

$$\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 0|M_i(D_i) = 1, X_i = x)$$

$$- \mathbb{E}[Y_i(1, M_i(1))|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x]$$

$$\cancel{\Pr(M_i(1) = 1|D_i = 0, M_i(D_i) = 1, X_i = x)}\Pr(D_i = 0|M_i(D_i) = 1, X_i = x)$$

$$- \cancel{\mathbb{E}[Y_i(1, M_i(1))|D_i = 0, M_i(1) = 0, M_i(0) = 1, X_i = x]}$$

$$\cancel{\Pr(M_i(1) = 0|D_i = 0, M_i(D_i) = 1, X_i = x)\Pr(D_i = 0|M_i(D_i) = 1, X_i = x)}$$

$$- \mathbb{E}[Y_i(0, M_i(0))|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x]$$

$$\cancel{\Pr(M_i(1) = 1|D_i = 0, M_i(D_i) = 1, X_i = x)}\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)$$

$$- \cancel{\mathbb{E}[Y_i(0, M_i(0))|D_i = 0, M_i(1) = 0, M_i(0) = 1, X_i = x]}$$

$$\cancel{\Pr(M_i(1) = 0|D_i = 0, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)}$$

$$+ \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]$$

$$\Pr(M_i(0) = 1|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)$$

$$+ \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]$$

$$\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)$$

$$= \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]$$

$$\Pr(M_i(0) = 1|D_i = 1, M_i(D_i) = 1, X_i = x)$$

$$- \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]$$

$$\Pr(M_i(0) = 1|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)$$

$$+ \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]$$

$$\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)$$

$$- \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]$$

$$\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)$$

$$- \mathbb{E}[Y_i(1, M_i(1))|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x]$$

$$+ \mathbb{E}[Y_i(1, M_i(1))|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x]$$

$$\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)$$

$$- \mathbb{E}[Y_i(0, M_i(0))|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x]$$

$$\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)$$

$$+ \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]$$

$$\Pr(M_i(0) = 1|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)$$

$$+ \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]$$

$$\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)$$

adding and subtracting $\mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)$,

$$
\begin{aligned}
&= \mathbb{E}[Y_i(1, M_i(1)) - Y_i(0, M_i(0))|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x]\Pr(D_i = 1|M_i(D_i) = 1, X_i = x) \\
&\quad - \mathbb{E}[Y_i(1, M_i(1)) - Y_i(0, M_i(0))|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x] \\
&\qquad \Pr(M_i(0) = 1|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x) \\
&\quad - \mathbb{E}[Y_i(1, M_i(1)) - Y_i(0, M_i(0))|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x] \\
&\qquad \Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x) \\
&\quad - \mathbb{E}[Y_i(1, M_i(1))|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x] \\
&\quad + \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]\Pr(M_i(0) = 1|D_i = 1, M_i(D_i) = 1, X_i = x) \\
&\quad + \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x) \\
&\quad + \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x) \\
&\quad - \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)
\end{aligned}
$$

substituting potential mediators based on principal strata,

$$
\begin{aligned}
&= \mathbb{E}[Y_i(1, 1) - Y_i(0, 1)|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x]\Pr(D_i = 1|M_i(D_i) = 1, X_i = x) \\
&\quad - \mathbb{E}[Y_i(1, 1) - Y_i(0, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x] \\
&\qquad \Pr(M_i(0) = 1|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x) \\
&\quad - \mathbb{E}[Y_i(1, 1) - Y_i(0, 0)|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x] \\
&\qquad \Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x) \\
&\quad + \mathbb{E}[Y_i(1, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x] \\
&\quad - \mathbb{E}[Y_i(1, 1)|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x] \\
&\quad + \mathbb{E}[Y_i(1, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x) \\
&\quad - \mathbb{E}[Y_i(1, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)
\end{aligned}
$$

under mandatory reporting, $Y_i(d, 0) = 0$,

$$
\begin{aligned}
&= \mathbb{E}[Y_i(1, 1) - Y_i(0, 1)|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x]\Pr(D_i = 1|M_i(D_i) = 1, X_i = x) \\
&\quad - \mathbb{E}[Y_i(1, 1) - Y_i(0, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x] \\
&\qquad \Pr(M_i(0) = 1|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x) \\
&\quad - \mathbb{E}[Y_i(1, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x] \\
&\qquad \Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)
\end{aligned}
$$

$$
\begin{aligned}
&+ \mathbb{E}[Y_i(1,1)|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x] \\
&- \mathbb{E}[Y_i(1,1)|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x] \\
&+ \mathbb{E}[Y_i(1,1)|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x) \\
&- \mathbb{E}[Y_i(1,1)|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)
\end{aligned}
$$

invoking assumption 4(b) (treatment ignorability),

$$
\begin{aligned}
= &\Big(\mathbb{E}[Y_i(1,1) - Y_i(0,1)|M_i(1) = 1, M_i(0) = 1, X_i = x] \\
&\quad - \mathbb{E}[Y_i(1,1) - Y(0,0)|M_i(1) = 1, M_i(0) = 0, X_i = x] \\
&\quad \Big) \Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x) \\
&- (\mathbb{E}[Y_i(1,1)|M_i(1) = 1, M_i(0) = 1, X_i = x] \\
&\quad - \mathbb{E}[Y_i(1,1)|M_i(1) = 1, M_i(0) = 0, X_i = x])\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)
\end{aligned}
$$

which is Equation 6.

A.2 Bias for $\text{ATT}_{M=1}$

Next, we consider the bias that results when the local difference in means is used as an estimator for the local average racial effect among stopped minorities, $\text{ATT}_{M=1,x} = \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i = 1, X_i = x] - \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i = 1, X_i = x]$. Again, overall bias is found by the weighted average of local biases, $\sum_x \left(\mathbb{E}[\hat{\Delta}_x] - \text{ATT}_{M=1,x}\right) \Pr(X_i = x|D_i = 1, M_i = 1)$.

$$
\begin{aligned}
\mathbb{E}[\hat{\Delta}_x] - \text{ATT}_{M=1,x} &= (\mathbb{E}[\overline{Y_i|D_i = 1, M_i = 1, X_i = x} - \overline{Y_i|D_i = 0, M_i = 1, X_i = x}]) \\
&\quad - (\mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i = 1, X_i = x] - \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i = 1, X_i = x]) \\
&= \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(D_i) = 1, X_i = x] \\
&\quad - \mathbb{E}[Y_i(0, M_i(0))|D_i = 0, M_i(D_i) = 1, X_i = x] \\
&\quad - \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(D_i) = 1, X_i = x] \\
&\quad + \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i(D_i) = 1, X_i = x] \\
&= - \mathbb{E}[Y_i(0, M_i(0))|D_i = 0, M_i(D_i) = 1, X_i = x] \\
&\quad + \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i(D_i) = 1, X_i = x] \\
&= - \mathbb{E}[Y_i(0,1)|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x]\Pr(M_i(1) = 1|D_i = 0, M_i(D_i) = 1, X_i = x) \\
&\quad - \mathbb{E}[Y_i(0,1)|D_i = 0, M_i(1) = 0, M_i(0) = 1, X_i = x]\Pr(M_i(1) = 0|D_i = 0, M_i(D_i) = 1, X_i = x) \\
&\quad + \mathbb{E}[Y_i(0,1)|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]\Pr(M_i(0) = 1|D_i = 1, M_i(D_i) = 1, X_i = x) \\
&\quad + \mathbb{E}[Y_i(0,0)|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)
\end{aligned}
$$

by mediator monotonicity

$$
\begin{aligned}
= & -\mathbb{E}[Y_i(0,1)|D_i=0, M_i(1)=1, M_i(0)=1, X_i=x] \\
& +\mathbb{E}[Y_i(0,1)|D_i=1, M_i(1)=1, M_i(0)=1, X_i=x]\Pr(M_i(0)=1|D_i=1, M_i(D_i)=1, X_i=x) \\
& +\mathbb{E}[Y_i(0,0)|D_i=1, M_i(1)=1, M_i(0)=0, X_i=x]\Pr(M_i(0)=0|D_i=1, M_i(D_i)=1, X_i=x)
\end{aligned}
$$

by mandatory reporting

$$
\begin{aligned}
= & -\mathbb{E}[Y_i(0,1)|D_i=0, M_i(1)=1, M_i(0)=1, X_i=x] \\
& +\mathbb{E}[Y_i(0,1)|D_i=1, M_i(1)=1, M_i(0)=1, X_i=x]\Pr(M_i(0)=1|D_i=1, M_i(D_i)=1, X_i=x)
\end{aligned}
$$

by treatment ignorability

$$
= -\mathbb{E}[Y_i(0,1)|M_i(1)=1, M_i(0)=1, X_i=x]\Pr(M_i(0)=0|M_i(1)=1, X_i=x)
$$

A.3 Bias for $\text{CDE}_{M=1}$

The $\text{CDE}_{M=1}$ is defined as

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

It is a somewhat contrived estimand because it necessarily involves counterfactuals that, for racial-stop encounters, could never realize even if researchers could somehow randomize civilian race in police-civilian encounters. For example, when a minority civilian is racially stopped for a "furtive movement" and reaches for their wallet, it makes little sense to consider an officer's potential use of force if the civilian suddenly became white at that moment: had police observed a white civilian from the onset, a stop would never have occurred. Moreover, the assumptions required for such counterfactuals are fundamentally unverifiable (Robins and Greenland, 1992), unless the experimentalist can somehow also manipulate officer stopping decisions without distorting outcomes. (We note that always-stop encounters are not subject to this issue, but analysts cannot hope to estimate the conditional CDE in this group because it is impossible to identify which minority encounters belong to this group.)

In the previous bias expression for the $\text{ATE}_{M=1}$, white individuals in the data—necessarily belonging to the always-stop group, $M_i(1) = M_i(0) = 1$—were used to estimate the $Y_i(0, M_i(0))$ potential outcomes of minority encounters in the data. Unavoidable bias arose as long as any minority individuals in the data belonged to the racial-stop group—had these individuals been white,

they would never have been stopped, and hence would not be subject to force. Changing the target estimand to the $\text{CDE}_{M=1}$ conceptually sidesteps this specific issue by considering a different counterfactual, $Y_i(0,1)$ instead of $Y_i(0, M_i(0))$. But as we note above, for encounters in which only minority civilians would be stopped—that is, encounters with $M_i(1) = 1$ and $M_i(0) = 0$—this new counterfactual represents an impossible cross-world scenario. The $\text{CDE}_{M=1}$ asks whether force would have been used if officers were forced, against nature, to stop a white individual in this encounter as if they were a minority.

We now demonstrate that the local difference in means remains biased for the local controlled direct effect, $\text{CDE}_{M=1,x} = \mathbb{E}[Y_i(1,1)|M_i = 1, X_i = x] - \mathbb{E}[Y_i(0,1)|M_i = 1, X_i = x]$, unless officers are as violent toward minorities in always-stop encounters (where they are forced to intervene) as they are in racially discriminatory stops (where they are free to exercise discretion). In other words, for the naïve estimator to recover the $\text{CDE}_{M=1}$, Assumption 5 must hold. The derivation is almost identical to that of the $\text{ATE}_{M=1,x}$, differing only in that all individuals are held at $M_i = 1$ instead of allowed stops to vary with civilian race, $M_i(D_i)$. Bias for $\text{CDE}_{M=1}$ is then given by the weighted average of local biases, $\sum_x \left(\mathbb{E}[\hat{\Delta}_x] - \text{CDE}_{M=1,x} \right) \Pr(X_i = x|M_i = 1)$.

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

under assumption 4(b),

$$
\begin{aligned}
&= \big(\mathbb{E}[Y_i(1,1) - Y_i(0,1)|M_i(1) = 1, M_i(0) = 1, X_i = x] \\
&\qquad - \mathbb{E}[Y_i(1,1) - Y_i(0,1)|M_i(1) = 1, M_i(0) = 0, X_i = x]
\end{aligned}
$$

$$
\begin{aligned}
&) \Pr(M_i(0) = 0 | D_i = 1, M_i(D_i) = 1, X_i = x) \Pr(D_i = 1 | M_i(D_i) = 1, X_i = x) \\
&- (\mathbb{E}[Y_i(1,1) | M_i(1) = 1, M_i(0) = 1, X_i = x] - \mathbb{E}[Y_i(1,1) | M_i(1) = 1, M_i(0) = 0, X_i = x]) \\
&\quad \Pr(M_i(0) = 0 | D_i = 1, M_i(D_i) = 1, X_i = x)
\end{aligned}
$$

which reproduces Equation 7.

Finally, we demonstrate that this bias is weakly negative. Rearranging terms yields

$$
\begin{aligned}
&\mathbb{E}[Y_i(1,1) - Y_i(0,1) | M_i(1) = 1, M_i(0) = 1, X_i = x] \\
&\quad \times \Pr(M_i(0) = 0 | D_i = 1, M_i(D_i) = 1, X_i = x) \Pr(D_i = 1 | M_i(D_i) = 1, X_i = x) \\
&- \mathbb{E}[Y_i(1,1) - Y_i(0,1) | M_i(1) = 1, M_i(0) = 0, X_i = x] \\
&\quad \times \Pr(M_i(0) = 0 | D_i = 1, M_i(D_i) = 1, X_i = x) \Pr(D_i = 1 | M_i(D_i) = 1, X_i = x) \\
&- \mathbb{E}[Y_i(1,1) | M_i(1) = 1, M_i(0) = 1, X_i = x] \\
&\quad \times \Pr(M_i(0) = 0 | D_i = 1, M_i(D_i) = 1, X_i = x) \\
&+ \mathbb{E}[Y_i(1,1) | M_i(1) = 1, M_i(0) = 0, X_i = x] \\
&\quad \times \Pr(M_i(0) = 0 | D_i = 1, M_i(D_i) = 1, X_i = x) \\
&= -\mathbb{E}[Y_i(0,1) | M_i(1) = 1, M_i(0) = 1, X_i = x] \\
&\qquad \times \Pr(M_i(0) = 0 | D_i = 1, M_i(D_i) = 1, X_i = x) \Pr(D_i = 1 | M_i(D_i) = 1, X_i = x) \\
&\quad + \mathbb{E}[Y_i(0,1) | M_i(1) = 1, M_i(0) = 0, X_i = x] \\
&\qquad \times \Pr(M_i(0) = 0 | D_i = 1, M_i(D_i) = 1, X_i = x) \Pr(D_i = 1 | M_i(D_i) = 1, X_i = x) \\
&\quad - \mathbb{E}[Y_i(1,1) | M_i(1) = 1, M_i(0) = 1, X_i = x] \\
&\qquad \times \Pr(M_i(0) = 0 | D_i = 1, M_i(D_i) = 1, X_i = x)(1 - \Pr(D_i = 1 | M_i(D_i) = 1, X_i = x)) \\
&\quad + \mathbb{E}[Y_i(1,1) | M_i(1) = 1, M_i(0) = 0, X_i = x] \\
&\qquad \times \Pr(M_i(0) = 0 | D_i = 1, M_i(D_i) = 1, X_i = x)(1 - \Pr(D_i = 1 | M_i(D_i) = 1, X_i = x))
\end{aligned}
$$

The sum of the first and second terms is weakly negative by Assumption 3, because the magnitude of the first term is greater than that of the second, and similarly the sum of the third and fourth terms is also weakly negative. Therefore, bias is weakly negative when the naïve estimator is used to estimate the $\text{CDE}_{M=1}$.

A.4 Nonparametric sharp bounds for $\text{ATE}_{M=1}$

In this section, we derive nonparametric sharp bounds for the $\text{ATE}_{M=1,x}$. We begin with the case when the proportion of racially discriminatory stops among reported minority encounters, $\Pr(M_i(0) = 0 | D_i = 1, M_i(D_i) = 1, X_i = x)$, is known or can be assumed. Rearrangement of Equa-

tion 6 (within levels of X) yields

$$
\begin{aligned}
\text{ATE}_{M=1,x} &= \mathbb{E}[\hat{\Delta}_x] \\
&\quad + \mathbb{E}[Y_i(1,1)|M_i(1)=1, M_i(0)=1, X_i=x] \\
&\quad \Pr(M_i(0)=0|D_i=1, M_i(D_i)=1, X_i=x)(1-\Pr(D_i=1|M_i(D_i)=1, X_i=x)) \\
&\quad - \mathbb{E}[Y_i(1,1)|M_i(1)=1, M_i(0)=0, X_i=x] \\
&\quad \Pr(M_i(0)=0|D_i=1, M_i(D_i)=1, X_i=x)(1-\Pr(D_i=1|M_i(D_i)=1, X_i=x)) \\
&\quad + \mathbb{E}[Y_i(0,1)|M_i(1)=1, M_i(0)=1, X_i=x] \\
&\quad \Pr(M_i(0)=0|D_i=1, M_i(D_i)=1, X_i=x)\Pr(D_i=1|M_i(D_i)=1, X_i=x) \\
&= \mathbb{E}[\hat{\Delta}_x] \\
&\quad + \frac{\mathbb{E}[Y_i(1,1)|D_i=1, M_i(1)=1, X_i=x]}{\Pr(M_i(0)=1|D_i=1, M_i(1)=1, X_i=x)}\Pr(M_i(0)=0|D_i=1, M_i(D_i)=1, X_i=x)\Pr(D_i=0|M_i(D_i)=1, X_i=x) \\
&\quad - \frac{\mathbb{E}[Y_i(1,1)|D_i=1, M_i(1)=1, M_i(0)=0, X_i=x]}{\Pr(M_i(0)=1|D_i=1, M_i(1)=1, X_i=x)}\Pr(M_i(0)=0|D_i=1, M_i(D_i)=1, X_i=x)^2 \qquad (2) \\
&\qquad \Pr(D_i=0|M_i(D_i)=1, X_i=x) \\
&\quad - \mathbb{E}[Y_i(1,1)|M_i(1)=1, M_i(0)=0, X_i=x]\Pr(M_i(0)=0|D_i=1, M_i(D_i)=1, X_i=x)\Pr(D_i=0|M_i(D_i)=1, X_i=x) \\
&\quad + \mathbb{E}[Y_i(0,1)|M_i(1)=1, M_i(0)=1, X_i=x]\Pr(M_i(0)=0|D_i=1, M_i(D_i)=1, X_i=x)\Pr(D_i=1|M_i(D_i)=1, X_i=x) \\
&= \mathbb{E}[\hat{\Delta}_x] \\
&\quad + \frac{\mathbb{E}[Y_i|D_i=1, M_i=1, X_i=x]}{\Pr(M_i(0)=1|D_i=1, M_i(1)=1, X_i=x)}\Pr(M_i(0)=0|D_i=1, M_i(D_i)=1, X_i=x)\Pr(D_i=0|M_i=1, X_i=x) \\
&\quad - \frac{\mathbb{E}[Y_i(1,1)|D_i=1, M_i(1)=1, M_i(0)=0, X_i=x]}{\Pr(M_i(0)=1|D_i=1, M_i(1)=1, X_i=x)}\Pr(M_i(0)=0|D_i=1, M_i(D_i)=1, X_i=x)^2 \qquad (3) \\
&\qquad \Pr(D_i=0|M_i=1, X_i=x) \\
&\quad - \mathbb{E}[Y_i(1,1)|M_i(1)=1, M_i(0)=0, X_i=x]\Pr(M_i(0)=0|D_i=1, M_i(D_i)=1, X_i=x)\Pr(D_i=0|M_i=1, X_i=x) \\
&\quad + \mathbb{E}[Y_i|D_i=0, M_i=1, X_i=x]\Pr(M_i(0)=0|D_i=1, M_i(D_i)=1, X_i=x)\Pr(D_i=1|M_i=1, X_i=x) \qquad (4)
\end{aligned}
$$

We then construct bounds on $\mathbb{E}[Y_i(1,1)|D_i=1, M_i(1)=1, M_i(0)=0, X_i=x]$ based on Fréchet inequalities for the joint distribution, $\Pr(Y_i(1,1)=1, M_i(0)=0|D_i=1, M_i(1)=1, X_i=x)$, which incorporate marginal information about $Y_i(1,1)$ and $M_i(0)$.

$$
\begin{aligned}
&\frac{\max\{0, \Pr(M_i(0)=0|D_i=1, M_i(1)=1, X_i=x) + \mathbb{E}[Y_i|D_i=1, M_i=1, X_i=x] - 1\}}{\Pr(M_i(0)=0|D_i=1, M_i(1)=1, X_i=x)} \\
&\qquad \leq \ \mathbb{E}[Y_i(1,1)|D_i=1, M_i(1)=1, M_i(0)=0, X_i=x] \ \leq \qquad (5) \\
&\frac{\min\{\Pr(M_i(0)=0|D_i=1, M_i(1)=1, X_i=x), \mathbb{E}[Y_i|D_i=1, M_i=1, X_i=x]\}}{\Pr(M_i(0)=0|D_i=1, M_i(1)=1, X_i=x)}
\end{aligned}
$$

These bounds are sharp given only marginal information, $\Pr(Y_i(1,1)=1|D_i=1, M_i(1)=1, X_i=x)$

and $\Pr(M_i(0) = 0|D_i = 1, M_i(1) = 1, X_i = x)$. However, the upper bound can be tightened further under Assumption 3, which implies $\mathbb{E}[Y_i(1,1)|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x] \leq \mathbb{E}[Y_i|D_i = 1, M_i = 1, X_i = x]$; this is at least as small as the upper Fréchet bound.

Finally, note that the reported data contain no information that can be used to constrain the proportion of racially discriminatory minority stops, $\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)$. If this proportion were zero, then the distribution of civilian race in police reports would reflect that of all police encounters (within levels of X). The reported data cannot distinguish between this possibility and an alternative population in which $\rho_x = \Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)$ is large, but white encounters are also larger by the proportion $1/(1 - \rho_x)$. Without side information about the total number of encounters, this proportion can take on any value in $[0, 1)$. Therefore, sharp bounds on $\text{ATE}_{M=1}$ alone are obtained by substituting Equation 5 into Equation 4 and setting the proportion of racial stops to unity. The bivariate bounds define the region in which $(\text{ATE}_{M=1}, \rho_x)$ pairs are consistent with the observed data. When ρ_x is set to zero or one, these respectively recover the difference in reported means and the marginal upper bounds on $\text{ATE}_{M=1}$. For $\rho_x \in (0, 1)$,

$$\begin{gathered}
\mathbb{E}[\hat{\Delta}_x] + \rho_x \, \mathbb{E}[Y_i|D_i = 0, M_i = 1, X_i = x](1 - \Pr(D_i = 0|M_i = 1, X_i = x)) \\
\leq \ \text{ATE}_{M=1,x} \ \leq \\
\mathbb{E}[\hat{\Delta}_x] \\
+\frac{\rho_x}{1 - \rho_x} \left(\mathbb{E}[Y_i|D_i = 1, M_i = 1, X_i = x] - \max\left\{ 0, 1 + \frac{1}{\rho_x}\mathbb{E}[Y_i|D_i = 1, M_i = 1, X_i = x] - \frac{1}{\rho_x} \right\} \right) \Pr(D_i = 0|M_i = 1, X_i = x) \\
+\rho_x \, \mathbb{E}[Y_i|D_i = 0, M_i = 1, X_i = x] \, (1 - \Pr(D_i = 0|M_i = 1, X_i = x)),
\end{gathered}$$

which reduces to Proposition 1 in the no-covariate case. Otherwise, bounds on $\text{ATE}_{M=1}$ are given by $\sum_x \underline{\text{ATE}}_{M=1,x} \Pr(X_i = x|M_i = 1) \leq \text{ATE}_{M=1} \leq \sum_x \overline{\text{ATE}}_{M=1,x} \Pr(X_i = x|M_i = 1)$, where $\underline{\text{ATE}}_{M=1,x}$ ($\overline{\text{ATE}}_{M=1,x}$) denote the lower (upper) bounds on the local average treatment effect.

Finally, we note that per Equation 3, the $\text{ATT}_{M=1,x}$ can be written

$$\begin{aligned}
\text{ATT}_{M=1,x} &= \mathbb{E}[Y_i(1, M_i(1)) - Y_i(0, M_i(0))|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x] \Pr(M_i(0) = 1|D_i = 1, M_i = 1, X_i = x) \\
&\quad + \mathbb{E}[Y_i(1, M_i(1)) - Y_i(0, M_i(0))|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x] \Pr(M_i(0) = 0|D_i = 1, M_i = 1, X_i = x) \\
&= \mathbb{E}[Y_i(1,1)|D_i = 1, M_i = 1, X_i = x] \\
&\quad - \mathbb{E}[Y_i(0,1)|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x] \Pr(M_i(0) = 1|D_i = 1, M_i = 1, X_i = x) \\
&\quad - \mathbb{E}[Y_i(0,0)|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x] \Pr(M_i(0) = 0|D_i = 1, M_i = 1, X_i = x)
\end{aligned}$$

under Assumption 1,

$$\begin{aligned}
&= \mathbb{E}[Y_i(1,1) - |D_i = 1, M_i = 1, X_i = x] \\
&\quad - \mathbb{E}[Y_i(0,1)|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x] \Pr(M_i(0) = 1|D_i = 1, M_i = 1, X_i = x)
\end{aligned}$$

and under Assumption 4,

$$= \mathbb{E}[Y_i|D_i = 1, M_i = 1, X_i = x] - \mathbb{E}[Y_i|D_i = 0, M_i = 1, X_i = x]\,(1 - \Pr(M_i(0) = 0|D_i = 1, M_i = 1, X_i = x))$$

which can be estimated from observed data if the proportion of racial stops is known. It then follows that

$$\begin{aligned}\text{ATT}_{M=1} &= \sum_x (\mathbb{E}[Y_i|D_i = 1, M_i = 1, X_i = x] - \mathbb{E}[Y_i|D_i = 0, M_i = 1, X_i = x] + \rho_x \mathbb{E}[Y_i|D_i = 0, M_i = 1, X_i = x]) \\ &= \sum_x (\mathbb{E}[\Delta_x] + \rho_x \mathbb{E}[Y_i|D_i = 0, M_i = 1, X_i = x])\,.\end{aligned}$$

A.5 Uncertainty of bounds

Here, we describe our approach for constructing confidence intervals for the bounds on these causal quantities. We take X_i, D_i and M_i as fixed, so that uncertainty in the bounds arises strictly from the estimation of the conditional expectations, $\mathbb{E}[Y_i|D_i = d, M_i = 1, X_i = x]$. The asymptotic distribution of the estimated lower and upper bounds endpoints, $(\underline{\hat{\text{ATE}}}_{M=1}, \overline{\hat{\text{ATE}}}_{M=1})$, then follows directly from the asymptotic joint distribution of $\hat{\mathbb{E}}[Y_i|D_i = d, M_i = 1, X_i = x]$ for all d and x. We approximate this through a Monte Carlo simulation in which parameters of the logistic regression models described in Section 5 are sampled from a multivariate normal distribution centered on the parameter estimates and with the estimated covariance matrix. For each parameter sample θ^*, the corresponding bounds endpoint pair $(\underline{\text{ATE}}^*_{M=1}, \overline{\text{ATE}}^*_{M=1})$ is computed deterministically; after drawing a sufficient number of such samples, we numerically obtain the shortest range that fully contains 95% of all simulated bounds intervals. Closely related alternatives to this approach are the bootstrap-based method of Horowitz and Manski (2000) and the fully Bayesian approach taken in Knox et al. (2019). For the analysis in Section 5, we follow Fryer (2019) in using a cluster-robust covariance estimator, clustering on precinct, and 5,000 samples were drawn for each force threshold and model specification.

A.6 Point identification of ATE

First, we note that strata sizes are identified with information on the total count of encounters by race (both reported and unreported).

$$
\begin{aligned}
\Pr(M_i(1)=1, M_i(0)=1, X_i=x) &= \Pr(M_i(1)=1, M_i(0)=1 | D_i=0, X_i=x) \\
&= \Pr(M_i=1 | D_i=0, X_i=x) \\
\Pr(M_i(1)=1, M_i(0)=0, X_i=x) &= \Pr(M_i(1)=1, X_i=x) \\
&\quad - \Pr(M_i(1)=1, M_i(0)=1, X_i=x)\Pr(M_i=1 | D_i=1, X_i=x) \\
&\quad - \Pr(M_i=1 | D_i=0, X_i=x) \\
\Pr(M_i(1)=0, M_i(0)=1, X_i=x) &= 0 \\
\Pr(M_i(1)=0, M_i(0)=0, X_i=x) &= 1 - \Pr(M_i(1)=0, M_i(0)=1, X_i=x) \\
&\quad - \Pr(M_i(1)=1, M_i(0)=0, X_i=x) \\
&\quad - \Pr(M_i(1)=1, M_i(0)=1, X_i=x) \\
&= 1 - \Pr(M_i=1 | D_i=1, X_i=x)
\end{aligned}
$$

We then reexpress the ATE in terms of strata-specific mean potential outcomes and simplify.

$$
\begin{aligned}
\text{ATE} = \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x] \\
\quad \Pr(M_i(1) = 1, M_i(0) = 1|D_i = 1, X_i = x) \Pr(D_i = 1, X_i = x) \\
+ \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x] \\
\quad \Pr(M_i(1) = 1, M_i(0) = 0|D_i = 1, X_i = x) \Pr(D_i = 1, X_i = x) \\
+ \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 0, M_i(0) = 1, X_i = x] \\
\quad \Pr(M_i(1) = 0, M_i(0) = 1|D_i = 1, X_i = x) \Pr(D_i = 1, X_i = x) \\
+ \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 0, M_i(0) = 0, X_i = x] \\
\quad \Pr(M_i(1) = 0, M_i(0) = 0|D_i = 1, X_i = x) \Pr(D_i = 1, X_i = x) \\
+ \mathbb{E}[Y_i(1, M_i(1))|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x] \\
\quad \Pr(M_i(1) = 1, M_i(0) = 1|D_i = 0, X_i = x) \Pr(D_i = 0, X_i = x) \\
+ \mathbb{E}[Y_i(1, M_i(1))|D_i = 0, M_i(1) = 1, M_i(0) = 0, X_i = x] \\
\quad \Pr(M_i(1) = 1, M_i(0) = 0|D_i = 0, X_i = x) \Pr(D_i = 0, X_i = x) \\
+ \mathbb{E}[Y_i(1, M_i(1))|D_i = 0, M_i(1) = 0, M_i(0) = 1, X_i = x] \\
\quad \Pr(M_i(1) = 0, M_i(0) = 1|D_i = 0, X_i = x) \Pr(D_i = 0, X_i = x) \\
+ \mathbb{E}[Y_i(1, M_i(1))|D_i = 0, M_i(1) = 0, M_i(0) = 0, X_i = x] \\
\quad \Pr(M_i(1) = 0, M_i(0) = 0|D_i = 0, X_i = x) \Pr(D_i = 0, X_i = x) \\
- \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x] \\
\quad \Pr(M_i(1) = 1, M_i(0) = 1|D_i = 1, X_i = x) \Pr(D_i = 1, X_i = x) \\
- \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x] \\
\quad \Pr(M_i(1) = 1, M_i(0) = 0|D_i = 1, X_i = x) \Pr(D_i = 1, X_i = x) \\
- \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i(1) = 0, M_i(0) = 1, X_i = x] \\
\quad \Pr(M_i(1) = 0, M_i(0) = 1|D_i = 1, X_i = x) \Pr(D_i = 1, X_i = x) \\
- \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i(1) = 0, M_i(0) = 0, X_i = x] \\
\quad \Pr(M_i(1) = 0, M_i(0) = 0|D_i = 1, X_i = x) \Pr(D_i = 1, X_i = x) \\
- \mathbb{E}[Y_i(0, M_i(0))|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x] \\
\quad \Pr(M_i(1) = 1, M_i(0) = 1|D_i = 0, X_i = x) \Pr(D_i = 0, X_i = x) \\
- \mathbb{E}[Y_i(0, M_i(0))|D_i = 0, M_i(1) = 1, M_i(0) = 0, X_i = x] \\
\quad \Pr(M_i(1) = 1, M_i(0) = 0|D_i = 0, X_i = x) \Pr(D_i = 0, X_i = x) \\
- \mathbb{E}[Y_i(0, M_i(0))|D_i = 0, M_i(1) = 0, M_i(0) = 1, X_i = x] \\
\quad \Pr(M_i(1) = 0, M_i(0) = 1|D_i = 0, X_i = x) \Pr(D_i = 0, X_i = x) \\
- \mathbb{E}[Y_i(0, M_i(0))|D_i = 0, M_i(1) = 0, M_i(0) = 0, X_i = x]
\end{aligned}
$$

$$\Pr(M_i(1) = 0, M_i(0) = 0 | D_i = 0, X_i = x) \Pr(D_i = 0, X_i = x)$$

under mandatory reporting

$$\begin{aligned}
&= \mathbb{E}[Y_i(1, M_i(1)) | D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x] \\
&\quad \Pr(M_i(1) = 1, M_i(0) = 1 | D_i = 1, X_i = x) \Pr(D_i = 1, X_i = x) \\
&\quad + \mathbb{E}[Y_i(1, M_i(1)) | D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x] \\
&\quad \Pr(M_i(1) = 1, M_i(0) = 0 | D_i = 1, X_i = x) \Pr(D_i = 1, X_i = x) \\
&\quad + \mathbb{E}[Y_i(1, M_i(1)) | D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x] \\
&\quad \Pr(M_i(1) = 1, M_i(0) = 1 | D_i = 0, X_i = x) \Pr(D_i = 0, X_i = x) \\
&\quad + \mathbb{E}[Y_i(1, M_i(1)) | D_i = 0, M_i(1) = 1, M_i(0) = 0, X_i = x] \\
&\quad \Pr(M_i(1) = 1, M_i(0) = 0 | D_i = 0, X_i = x) \Pr(D_i = 0, X_i = x) \\
&\quad - \mathbb{E}[Y_i(0, M_i(0)) | D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x] \\
&\quad \Pr(M_i(1) = 1, M_i(0) = 1 | D_i = 1, X_i = x) \Pr(D_i = 1, X_i = x) \\
&\quad - \mathbb{E}[Y_i(0, M_i(0)) | D_i = 1, M_i(1) = 0, M_i(0) = 1, X_i = x] \\
&\quad \Pr(M_i(1) = 0, M_i(0) = 1 | D_i = 1, X_i = x) \Pr(D_i = 1, X_i = x) \\
&\quad - \mathbb{E}[Y_i(0, M_i(0)) | D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x] \\
&\quad \Pr(M_i(1) = 1, M_i(0) = 1 | D_i = 0, X_i = x) \Pr(D_i = 0, X_i = x) \\
&\quad - \mathbb{E}[Y_i(0, M_i(0)) | D_i = 0, M_i(1) = 0, M_i(0) = 1, X_i = x] \\
&\quad \Pr(M_i(1) = 0, M_i(0) = 1 | D_i = 0, X_i = x) \Pr(D_i = 0, X_i = x)
\end{aligned}$$

under mediator monotonicity

$$\begin{aligned}
&= \mathbb{E}[Y_i(1, M_i(1)) | D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x] \\
&\quad \Pr(M_i(1) = 1, M_i(0) = 1 | D_i = 1, X_i = x) \Pr(D_i = 1, X_i = x) \\
&\quad + \mathbb{E}[Y_i(1, M_i(1)) | D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x] \\
&\quad \Pr(M_i(1) = 1, M_i(0) = 0 | D_i = 1, X_i = x) \Pr(D_i = 1, X_i = x) \\
&\quad + \mathbb{E}[Y_i(1, M_i(1)) | D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x] \\
&\quad \Pr(M_i(1) = 1, M_i(0) = 1 | D_i = 0, X_i = x) \Pr(D_i = 0, X_i = x) \\
&\quad + \mathbb{E}[Y_i(1, M_i(1)) | D_i = 0, M_i(1) = 1, M_i(0) = 0, X_i = x] \\
&\quad \Pr(M_i(1) = 1, M_i(0) = 0 | D_i = 0, X_i = x) \Pr(D_i = 0, X_i = x) \\
&\quad - \mathbb{E}[Y_i(0, M_i(0)) | D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x] \\
&\quad \Pr(M_i(1) = 1, M_i(0) = 1 | D_i = 1, X_i = x) \Pr(D_i = 1, X_i = x) \\
&\quad - \mathbb{E}[Y_i(0, M_i(0)) | D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x]
\end{aligned}$$

$$\Pr(M_i(1) = 1, M_i(0) = 1 | D_i = 0, X_i = x) \Pr(D_i = 0, X_i = x)$$

under treatment ignorability

$$\begin{aligned}
&= \mathbb{E}[Y_i(1, M_i(1)) | D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x] \Pr(M_i(1) = 1, M_i(0) = 1, X_i = x) \\
&\quad + \mathbb{E}[Y_i(1, M_i(1)) | D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x] \Pr(M_i(1) = 1, M_i(0) = 0, X_i = x) \\
&\quad - \mathbb{E}[Y_i(0, M_i(0)) | D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x] \Pr(M_i(1) = 1, M_i(0) = 1, X_i = x)
\end{aligned}$$

which can be recovered from observed data

$$\begin{aligned}
&= \mathbb{E}[Y_i | D_i = 1, M_i(D_i) = 1, X_i = x] \Pr(M_i = 1 | D_i = 1, X_i = x) \\
&\quad - \mathbb{E}[Y_i | D_i = 0, M_i(D_i, X_i = x) = 1, X_i = x] \Pr(M_i = 1 | D_i = 0, X_i = x)
\end{aligned}$$

which reduces to Proposition 2 in the no-covariate case.

A.7 Derivation of outcome test bounds on ρ

Our paper focuses on the difficulty of estimating a race effect on post-stop police behavior such as the use of force. However, another popular approach, the outcome test, focuses on establishing whether there exists any bias in the decision to stop a civilian (Becker, 1971; Goel, Rao and Shroff, 2016; Engel, 2008; Knowles, Perisco and Todd, 2001; Ridgeway and MacDonald, 2010). Because the degree of the statistical bias we explore is a function of racial discrimination in stopping decisions, it is useful to clarify the assumptions undergirding outcome tests. In the process, we demonstrate that the principal stratification framework sheds light on the precise interpretation of outcome tests, and we prove that the outcome test can be used to establish a lower bound on the share of police stops of racial minorities that are racially discriminatory.

Outcome tests compare the rates of finding evidence of a crime—conditional on a suspect being stopped by police—across racial groups. The logic behind the test is that if the decision to stop a civilian is unbiased, the rate of discovering evidence of a crime ("hit rates") should be identical across groups. Proponents of outcome tests thus claim that differences in hit rates amount to evidence of racially biased policing. The empirical observation that hit rates are lower among minority stops can be written as $\mathbb{E}[Y_i | D_i = 0, M_i = 1] > \mathbb{E}[Y_i | D_i = 1, M_i = 1]$, where Y_i is an indicator, say, for finding contraband on a suspect. However, interpreting the above inequality as evidence of racial discrimination in fact requires assumptions that closely mirror those we describe above.

To see this, first observe that the overall hit rate among minority stops can be decomposed into

the weighted average of the hit rate among always-stop encounters and the hit rate among the (possibly nonexistent) set of racially discriminatory stops. In contrast, if we invoke Assumption 2 (which states that there are no white civilians stopped in circumstances where a minority civilian would be allowed to pass), then stops involving white civilians belong exclusively to the always-stop group.[1] In this case, the empirical difference in hit rates can be rewritten in the potential outcomes framework as

$$\begin{aligned}
&\mathbb{E}[Y_i(0,1)|M_i(1)=1, M_i(0)=1, D_i=0] \\
&\quad -\mathbb{E}[Y_i(1,1)|M_i(1)=1, M_i(0)=1, D_i=1]\Pr(M_i(1)=1, M_i(0)=1|D_i=1, M_i=1) \\
&\quad -\mathbb{E}[Y_i(1,1)|M_i(1)=1, M_i(0)=0, D_i=1]\Pr(M_i(1)=1, M_i(0)=0|D_i=1, M_i=1) > 0 \quad (6)
\end{aligned}$$

A major critique of the outcome test is that observed racial disparities in hit rates alone do not constitute evidence of racially discriminatory stops because of the problem of "infra-marginality" (Ayres, 2002; Simoiu, Corbett-Davies and Goel, 2017). This critique suggests that the above inequality may hold simply because white civilians in always-stop encounters engage in more criminal conduct than minority suspects. In other words, the analyst might observe $\mathbb{E}[Y_i|D_i=1, M_i=1] < \mathbb{E}[Y_i|D_i=0, M_i=1]$ even if $\Pr(M_i(1)=1, M_i(0)=0)=0$—that is, with no discrimination in stops—as long as $\mathbb{E}[Y_i(1,1)|M_i(1)=1, M_i(0)=1, D_i=1] < \mathbb{E}[Y_i(0,1)|M_i(1)=1, M_i(0)=1, D_i=0]$. Some analysts employing the outcome test cast this scenario as unlikely, arguing that absent racial bias in stopping, "it would be difficult to explain why…whites for some reason had a systematically higher chance of possessing evidence of illegality" (Ayres, 2002) (137) and "there are not compelling reasons to suspect" this to be the case (138). Indeed, the validity of the outcome test hinges on the assumption that white and minority civilians in always-stop encounters commit crimes at the same rates, or that $\mathbb{E}[Y_i(1,1)|M_i(1)=1, M_i(0)=1, D_i=1] = \mathbb{E}[Y_i(0,1)|M_i(1)=1, M_i(0)=1, D_i=0]$. This assumption closely parallels Assumption 4, which requires that treatment status is ignorable with respect to potential outcomes. (For simplicity, we suppose that this holds without conditioning on covariates, but the result also holds within levels of $X_i=x$.) In this case, the observed racial difference in hit rates can be rewritten as

$$\begin{aligned}
&\mathbb{E}[Y_i|D_i=0, M_i=1] - \mathbb{E}[Y_i|D_i=1, M_i=1] \\
&\quad = \Big(\mathbb{E}[Y_i(0,1)|M_i(1)=1, M_i(0)=1, D_i=0] \\
&\qquad -\mathbb{E}[Y_i(1,1)|M_i(1)=1, M_i(0)=0, D_i=1]\Big)\Pr(M_i(1)=1, M_i(0)=0|D_i=1, M_i=1). \quad (7)
\end{aligned}$$

This formulation makes clear that the observed evidence gap is due to the difference in hit

[1] If we do not assume mediator monotonicity, and allow for the presence of stops of white suspects that would not have occurred if the suspect was a racial minority, then the inequality used to estimate the outcome test becomes uninformative with respect to racial discrimination.

rates between always-stop minority encounters—in which officers would also have stopped a white civilian—and racially discriminatory minority stops. If the former is assumed to produce more evidence of criminal behavior (Assumption 3; this might hold if racially discriminatory stops are made under weaker standards of evidence), then it can be seen from Equation 7 that the empirical difference in hit rates implies that $\Pr(M_i(1) = 1, M_i(0) = 0) > 0$: that there must exist encounters in which minority civilians would be stopped but white civilians would not, precisely as proponents of the outcome test suggest.

Equation 7 also shows that outcome tests are unable to identify the exact prevalence of racial stops. Outcome tests allow the analyst to infer whether there is *any* racial bias in the decision to stop a suspect—but only if the analyst makes assumptions similar to those we outline above. However, we show that the outcome test can *partially* identify a range of possible proportions of racial stops. This clarification is useful, as it allows us later in this analysis to appeal to a published study of hit rates (Goel, Rao and Shroff, 2016) to help characterize the statistical bias in analyses of post-stop police behavior (e.g. Fryer, 2019).

By rearranging Equation 7 and substituting observed quantities, we arrive at

$$\Pr(M_i(1) = 1, M_i(0) = 0 | D_i = 1, M_i = 1) = \frac{\mathbb{E}[Y_i | D_i = 0, M_i = 1] - \mathbb{E}[Y_i | D_i = 1, M_i = 1]}{\mathbb{E}[Y_i | D_i = 0, M_i = 1] - \mathbb{E}[Y_i(1,1) | M_i(1) = 1, M_i(0) = 0, D_i = 1]}$$

Although the second term in the denominator is unknown, the implied proportion of racially discriminatory stops is smallest when this value is zero—if, hypothetically, searches of racially stopped minorities never produce evidence. Thus, the outcome test suggests that *at least* $(\mathbb{E}[Y_i | D_i = 0, M_i = 1] - \mathbb{E}[Y_i | D_i = 1, M_i = 1]) / \mathbb{E}[Y_i | D_i = 0, M_i = 1]$ of all minority stops are racially discriminatory, and to the extent that racially discriminatory searches result in any evidence of contraband, the proportion could potentially be much larger.

B Additional results

B.1 Coding schemes for dependent variables

In this section, we reanalyze the NYPD SQF data using both the original and revised coding schemes for dependent variables in Fryer (2019). In an analysis of the use of force by level of severity, Fryer (2019) codes binary outcomes indicating whether force is used at or above some threshold. However, rather than coding all encounters with lower levels of force than a given threshold as a zero, the analysis coded only encounters with no force at all as a zero, while levels of force between no force and the threshold level were dropped from the data.[2] This data dropping strategy, a form of selection on the dependent variable, is problematic. If the analyst suspects that civilian race affects which level of force is applied—the motivating hypothesis for this very analysis—then dropping data based on which level of force was applied is another form of post-treatment conditioning and will induce bias. Further, the amount of data lost under this coding scheme is substantial. In the case of the point-weapon threshold, for example, over one million encounters—over 20% of the data—appear to have been discarded despite containing sub-threshold force use, such as pushing a civilian to the ground. Table B1 displays the number of observations reported for various regressions in the original paper, our best attempts at replication, and the corrected procedure used in this paper.

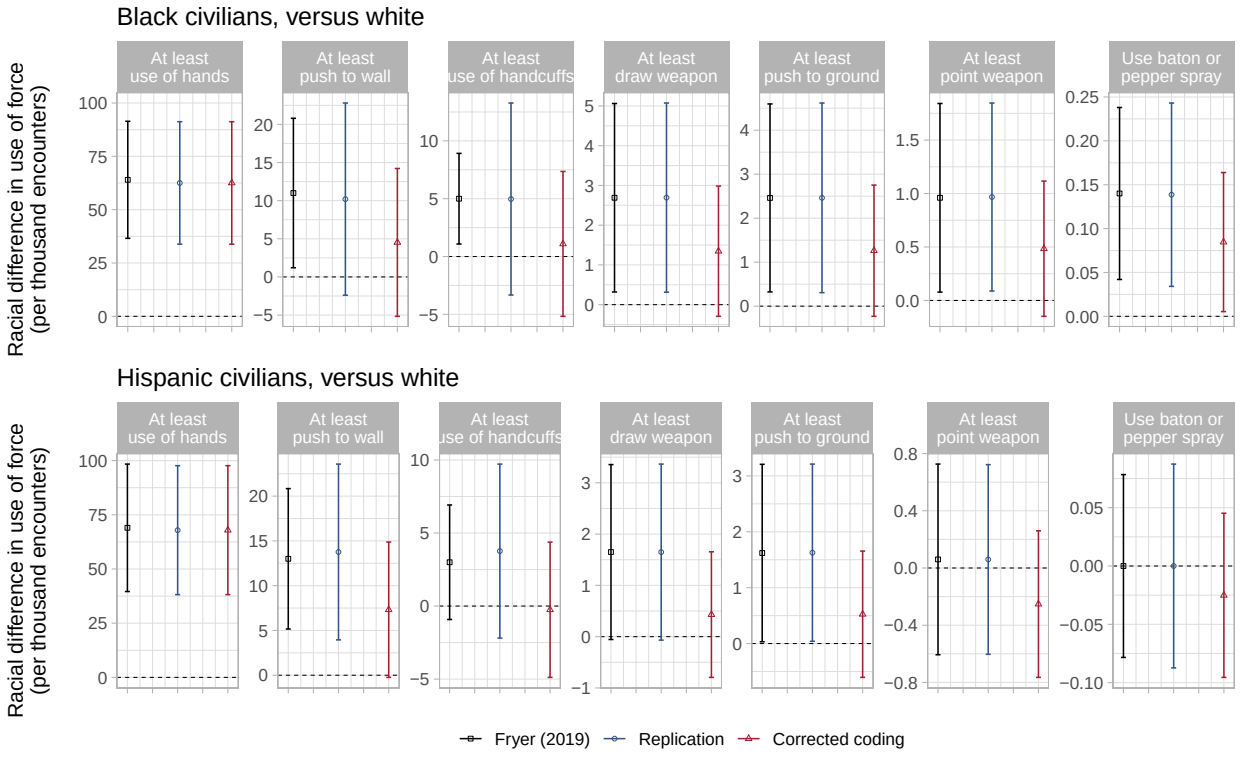
We present results of our replication study using the original coding scheme and our corrected version side by side below. As the results show, a corrected analysis generally depresses the naïve treatment effects relative to the inadvisable coding scheme in Fryer (2019) and in most cases renders the original results statistically insignificant. However, these discrepancies in results across coding decisions do not alter the central point of our paper: post-treatment conditioning exerts a large downward bias on estimates of racially discriminatory uses of force.

[2] Fryer (2019) acknowledges this data dropping strategy, writing "To be clear, an observation that records only hands would be in the hands regression but not the regression which restricts the sample to observations in which individuals were at least forced to the ground," (21, emphasis in original).

Table B1: **Comparison of SQF Data Dimensions Based on Outcome Coding.** The table displays the number of observations from bivariate analyses of the use of force by the NYPD using three coding procedures for force outcomes. The first column displays the number of observations as reported in results in Fryer (2019) (Appendix Tables 3A-3G in original). The second column reports the number of observations we recover when using the coding procedure in Fryer (2019) which drops observations where some level of force was used that was below a given threshold. The third column displays the number of observations we recover when using our corrected coding procedure, which codes outcomes as a 1 if a certain force threshold is reached and 0 otherwise.

	N (published)	*N* (replicated)	*N* (corrected coding)
At least use of hands	4,927,962	4,980,701	4,980,701
At least push to wall	4,152,918	4,245,091	4,980,701
At least use of handcuffs	4,017,783	4,122,329	4,980,701
At least draw weapon	3,957,687	3,965,721	4,980,701
At least push to ground	3,950,324	3,958,374	4,980,701
At least point weapon	3,918,741	3,926,805	4,980,701
Use baton or pepper spray	3,900,977	3,909,064	4,980,701

Figure B1: **Replication of Fryer (2019) using various outcome coding rules.** The figure displays odds ratios generated by OLS regressions that show the effect of suspect race without covariates on the use of force across all force types generated using three approaches: the published OLS results from the Appendix of Fryer (2019) (black points and bars), our best attempt at replication of these results (blue points and bars), and results using our corrected outcome coding scheme (red points and bars). Revising the coding scheme so as to retain data on sub-threshold uses of force generally deflates estimated treatment effects.



B.2 Varying levels of force

Figure B2: **Corrected ATE$_{M=1}$ and ATT$_{M=1}$ for encounters with Black and white civilians, varying levels of force.** This figure shows bounded effects comparing predicted levels of force when setting suspect race for all observations to black vs. white. These estimates use our corrected coding scheme for dependent variables (as described above). Results from regressions without covariates appear in the top panels and results from models with a full set of covariates appear in bottom panels.

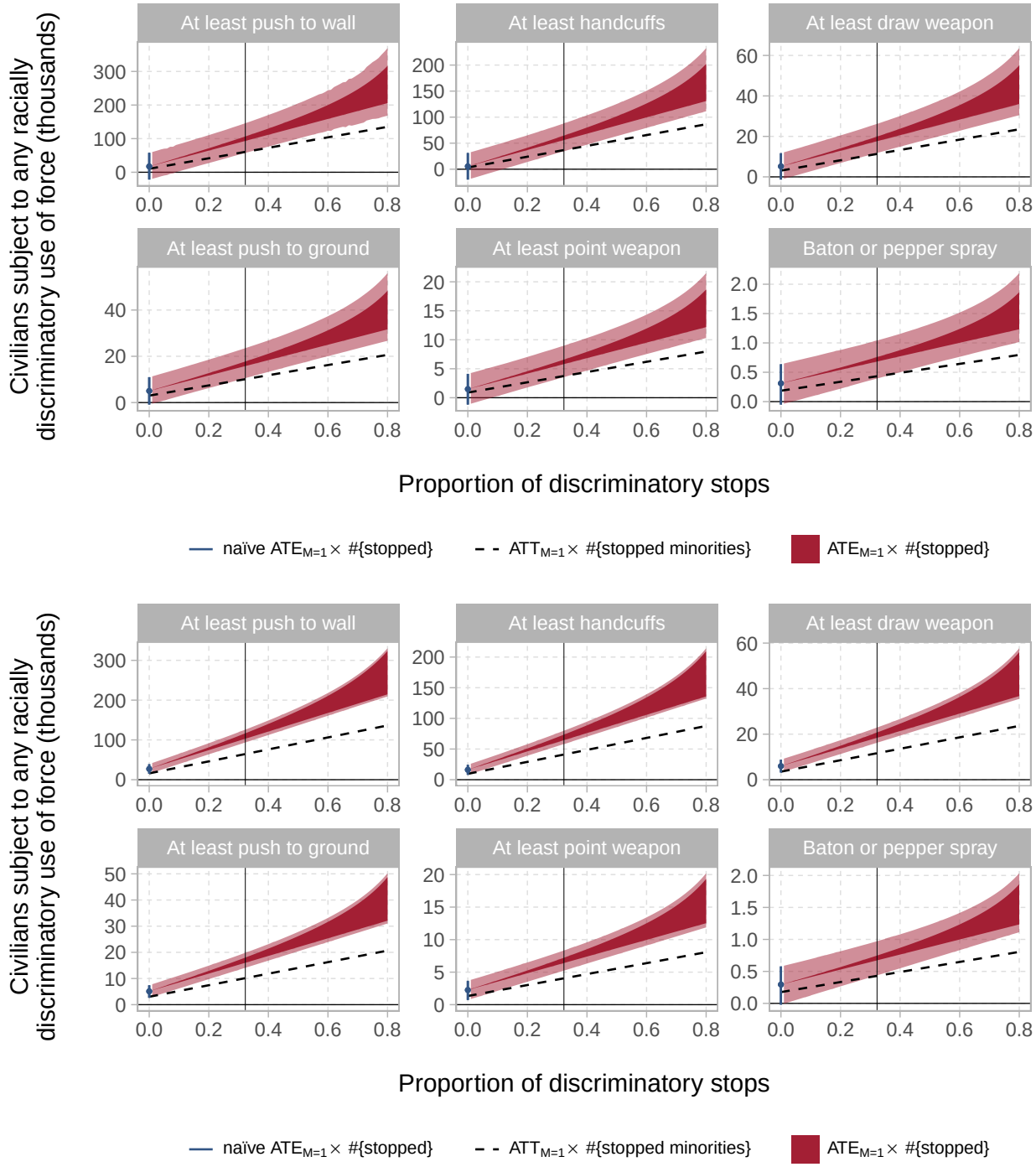
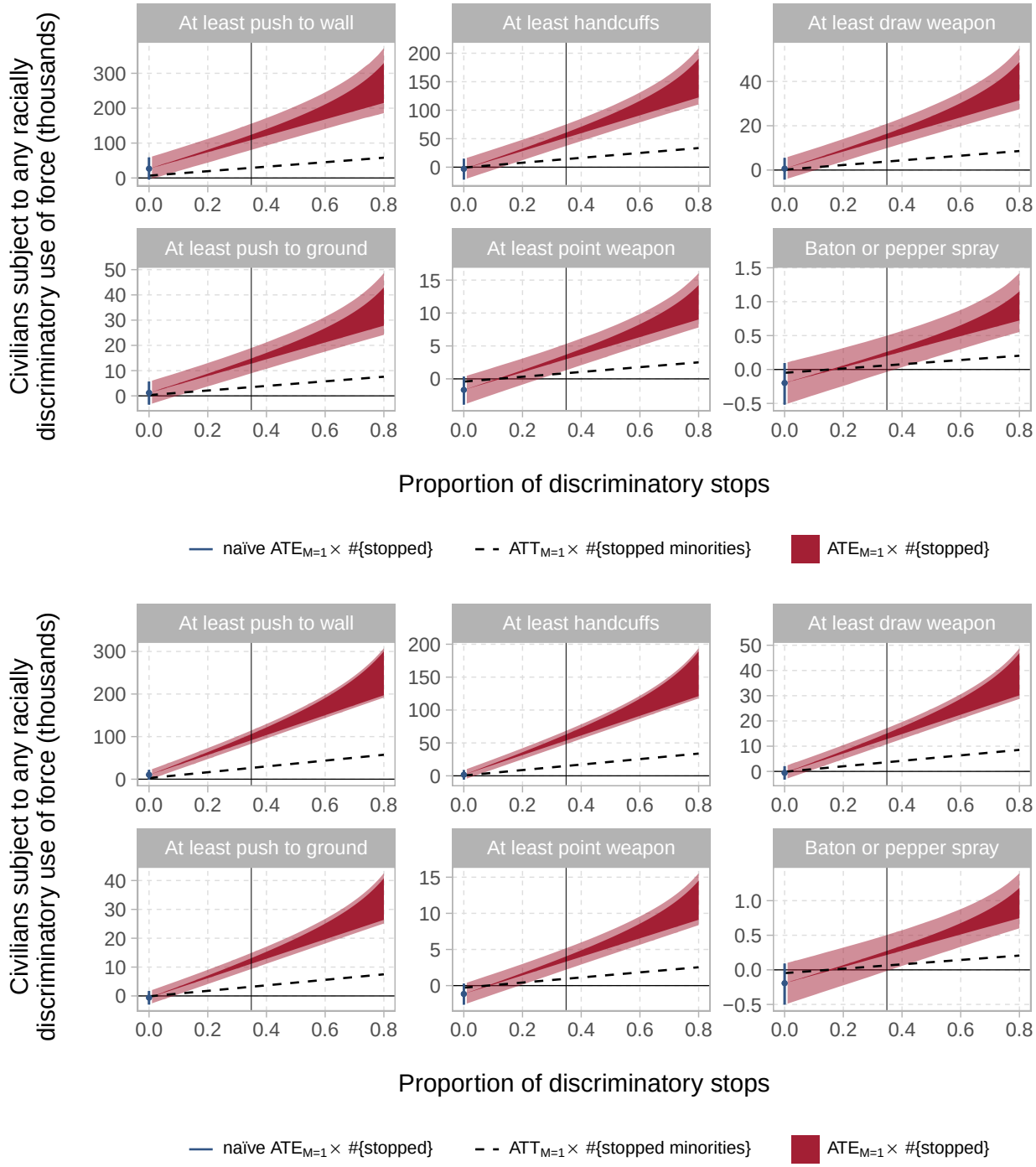
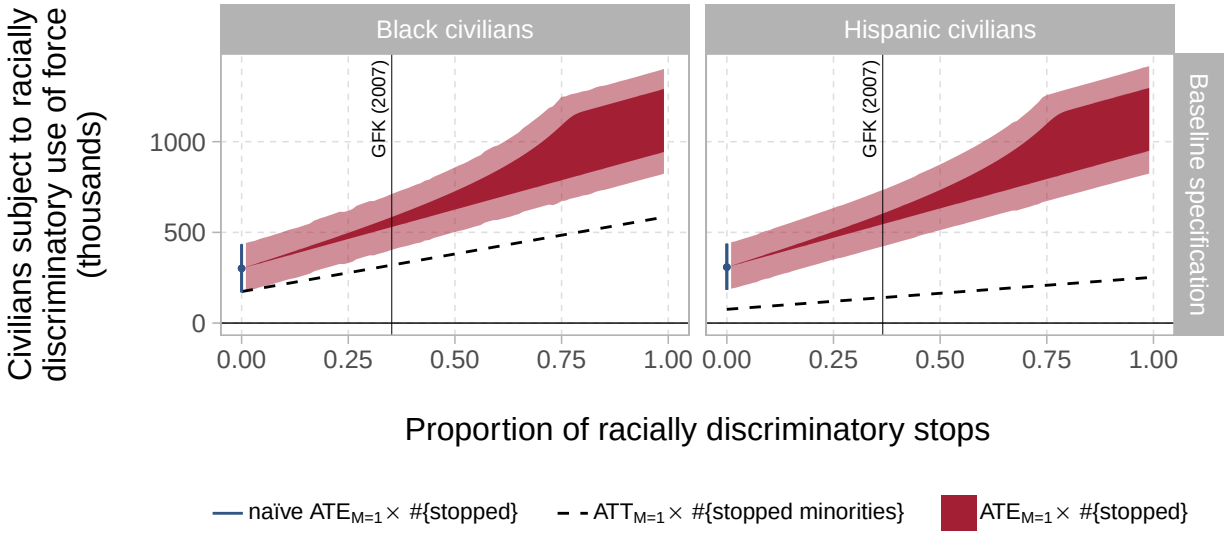


Figure B3: **Corrected ATE$_{M=1}$ and ATT$_{M=1}$ for encounters with Hispanic and white civilians, varying levels of force.** This figure shows bounded effects comparing predicted levels of force when setting suspect race for all observations to Hispanic vs. white. These estimates use our corrected coding scheme for dependent variables (as described above). Results from regressions without covariates appear in the top panels and results from models with a full set of covariates appear in bottom panels.



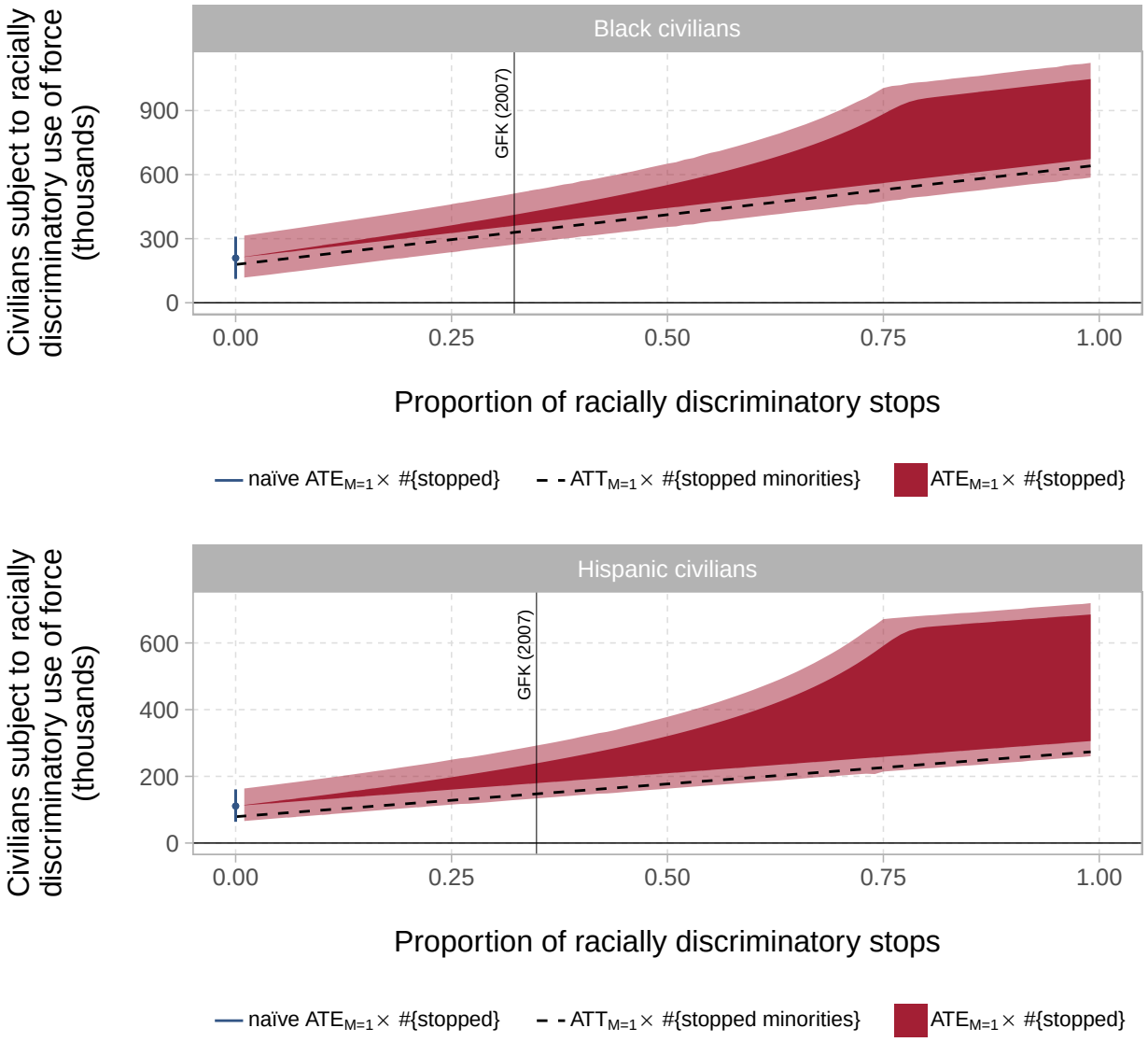
B.3 Excluding drug stops

Figure B4: **Bounds on race effect excluding drug stops.** This analysis replicates the analysis in Figure 4 in the main text excluding stops that were motivated by suspicion of a drug transaction, as such instances may violate the mediator monotonicity assumption. The results remain substantively similar.



B.4 Analysis of two races at a time

Figure B5: **Bounds on race effect limiting analysis to two racial groups of suspects.** Plots in the main text estimated bounds using data on multiple racial groups of suspects by predicting counterfactual values for every observation, regardless of a suspect's actual race, after model parameters were estimated. These figures reproduce the same analysis using only data on the two racial groups being compared, and exclude data on suspects who were not black, Hispanic or white entirely.



References

Ayres, Ian. 2002. "Outcome Tests of Racial Disparities in Police Practices." *Justice Research and Policy* 4(1-2):131–142.

Becker, Gary. 1971. *The Economics of Discrimination.* University of Chicago Press.

Engel, Robin. 2008. "A Critique of the "Outcome Test" in Racial Profiling Research." *Justice Quarterly* 25(1):1–36.

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

Goel, Sharad, Justin M. Rao and Ravi Shroff. 2016. "Precinct or Prejudice? Understanding Racial Disparities in New York City's Stop-And-Frisk Policy." *Annals of Applied Statistics* 10(1):365–394.

Horowitz, Joel L. and Charles F. Manski. 2000. "Nonparametric Analysis of Randomized Experiments With Missing Covariate and Outcome Data." *Journal of the Americal Statistical Association* 95(449):77–84.

Knowles, J., N. Perisco and P. Todd. 2001. "Racial bias in motor vehicle searches: Theory and evidence." *Journal of Political Economy* 109(1):203–229.

Knox, Dean, Teppei Yamamoto, Matthew A. Baum and Adam J. Berinsky. 2019. "Design, Identification, and Sensitivity Analysis for Patient Preference Trials." *Journal of the American Statistical Association* 00(0):1–15.

Ridgeway, Greg and John MacDonald. 2010. *Race, Ethnicity, and Policing: New and Essential Readings.* NYU Press chapter Methods for Assessing Racially Biased Policing.

Robins, J.M. and S. Greenland. 1992. "Identifiability and exchangeability for direct and indirect effects." *Epidemiology* 3(2):143–155.

Simoiu, Camelia, Sam Corbett-Davies and Sharad Goel. 2017. "The problem of infra-marginality in outcome tests for discrimination." *The Annals of Applied Statistics* 11(3):1193–1216. https://arxiv.org/abs/1706.05678.