1 Data analysis

We examine data on police behavior and give 3 analyses leading to 3 different relationships between the population and personalized AUCs: the population AUC 1) significantly more than, 2) significantly less than, and 3) not significantly different from the personalized AUC.

The data consists of Terry stops in New York City and Boston. A Terry stop is a policing procedure whereby an officer briefly detains an individual based on a reasonable suspicion that a crime has been committed, which is a lower evidentiary bar than required to arrest the individual. Terry stops are colloquially referred to as "stop and frisks" though the suspect need not be frisked or searched. The analysis here focuses on the relationship between the duration of the stop and race of the suspect. We cluster the stops according to precinct, in the case of NYC, and according to the officer conducting the stop, in the case of Boston. There is an extensive literature examining the relationship between race and Terry stops. Duration of the stop in particular is examined in, e.g., Ridgeway (2006), clustering at the precinct level in, e.g., Goel et al. (2016), and clustering at the officer level in, e.g., Ridgeway and MacDonald (2009).

The NYC data consists of measurements on 54,587 stops carried out between 2017 and 2021. The Boston data consists of 6,591 stops carried out between 2019 and 2021. The stop durations range between 0 minutes and 1–2 hours, with modes at multiples of 5 minutes, and 15 minutes being the most commonly recorded duration. While data is available for years prior to the cutoffs used here, key covariates used in the analysis were either missing or coded differently in the earlier data. So that the personalized AUC could be estimated, the data was further restricted to those clusters with at least 1 control and 1 case observation, where the interpretation of "control" or "case" depends on the racial classification under analysis below. The final number of clusters and cluster sizes are given in Table 2.

The racial classifications we consider are Black, White, and Hispanic, where Black and White are taken to include Black Hispanic and White Hispanic; see Table 2 for breakdowns.

1. $\theta_{12} < \theta_{11}$. With Black race as the binary classification, the AUC analysis looks for a difference in location between the distribution of stop durations of non-Black ("control") and Black ("case") suspects. For the NYC data, the population AUC estimate is $\hat{\theta}_{12} = 0.46$ with 95% CI 0.45—0.47, significantly different from the null value of 1/2. The personalized AUC estimate is $\hat{\theta}_{11} = 0.50$ with a 95% CI 0.47—0.53. A test of equality $H_0 : \theta_{12} = \theta_{11}$ against $\theta_{12} < \theta_{11}$ returns a p-value of .05%. The Boston data is similar. The population AUC estimate is 0.46 [0.42, 0.50] and the personalized AUC estimate is 0.52 [0.46, 0.58]. A test of equality $H_0 : \theta_{12} = \theta_{11}$ against $\theta_{12} < \theta_{11}$ returns the p-value .91%. Confidence ellipses are plotted in Figure 1. The data recalls the situation depicted in Fig. 1b, though of course the difference between the two AUCs is less dramatic here than in the artificial example constructed there.

2. $\theta_{11} < \theta_{12}$. We next consider differences in duration of stop between non-White ("control") or White ("case") suspect status. As Table 1 indicates, the vast majority of suspects are either Black or White, when those categories are taken inclusive of Hispanics, so one might expect that the analysis for non-White/White status to be nearly the same as the analysis for Black/non-Black status, therefore simply reversing the direction of the results just given, i.e., reflecting the AUCs across 1/2. That expectation largely holds for the NYC data, where the population and personalized AUCs are 0.53 [0.52, 0.54] and 0.50 [0.48, 0.53], and the population AUC remains the only one of the two significantly different from the null value 1/2. For the Boston estimates, however, the personalized AUC, 0.46 [0.40, 0.53], is more informative than the population AUC, 0.52 [0.48, 0.55], with the test of equality versus $\theta_{11} < \theta_{12}$ returning a p-value of 2.5%. This analysis therefore corresponds to the situation in Fig. 1a.

3. No significant difference between θ_{12} and θ_{11}. Finally, we consider duration of the stop between non-Hispanic ("control") and Hispanic ("case") suspects. For both the NYC and Boston data, neither the population AUC nor personalized AUC is significantly

different from the null value 1/2, and the test of equality of the two AUCs fails to reject. As a second example, in Boston, whether one takes the case status to be non-Hispanic Black or non-Hispanic White, the two AUCs are statistically indistinguishable from each other and each is indistinguishable from the null value 1/2.

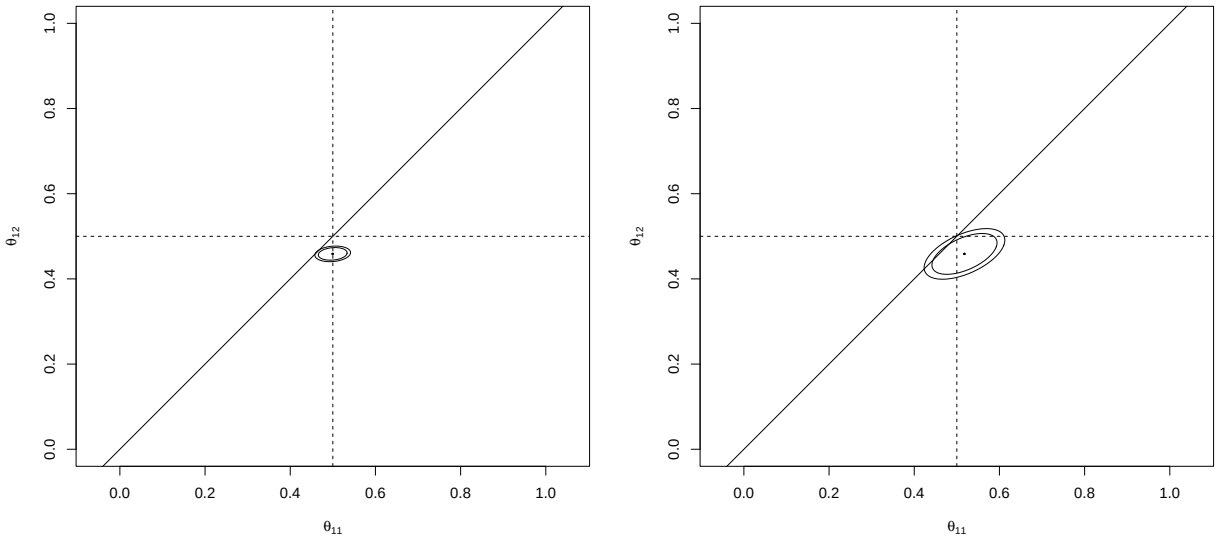
The decision to cluster at the officer or precinct level, as opposed to, say, the time of day of the stop, age of the suspect, or other partition of the data, is in part arbitrary. For the application of the definitions and results given in the previous sections, the decision amounts to the idealization that the officers' or precincts' data are drawn independently from a universe of officers or precinct Terry stop data. At the same time, many current analyses, such as cited above, besides this IID assumption further impose modeling assumptions such as linear random effects or logistic links. The approach here has the advantage of being otherwise nonparametric.

References

Goel, S., Rao, J. M., and Shroff, R. (2016). Precinct or prejudice? Understanding racial disparities in new york city's stop-and-frisk policy. *The Annals of Applied Statistics*, 10(1):365–394.

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

Ridgeway, G. and MacDonald, J. M. (2009). Doubly robust internal benchmarking and false discovery rates for detecting racial bias in police stops. *Journal of the American Statistical Association*, 104(486):661–668.



(a) NYC (b) Boston

Figure 1: Level 95% and 99% Confidence ellipses for the estimates of $(\theta_{11}, \theta_{12})$ for duration of Terry stop by non-Black/Black status.

	NYC			Boston		
group	mean duration (SD)	count	freq.	mean duration (SD)	count	freq.
Asian	14.24 (21.16)	1139	0.02	25.00 (24.22)	53	0.01
Black Hispanic	11.01 (17.12)	4675	0.09	15.28 (18.73)	391	0.06
Black non-Hispanic	10.99 (16.78)	31588	0.58	19.06 (28.93)	3448	0.55
White Hispanic	11.21 (15.15)	11486	0.21	15.63 (15.96)	578	0.09
White non-Hispanic	12.85 (16.18)	4854	0.09	21.74 (33.01)	1760	0.28
other	11.84 (17.70)	261	0.00	20.89 (23.90)	93	0.01

Table 1: Summary estimates on the duration of Terry stops by racial group.

case group	data set	I	ΣM_i	ΣN_i	θ_{12}	θ_{11}	$H_0 : \theta_{12} = \theta_{11}$
Black	NYC	187	17698	36152	0.46 [0.45, 0.47]	0.50 [0.47, 0.53]	0.00
	Boston	112	418	585	0.46 [0.42, 0.50]	0.52 [0.46, 0.58]	0.02
Black non-Hispanic	NYC	185	22348	31490	0.47 [0.46, 0.48]	0.51 [0.48, 0.53]	0.01
	Boston	117	464	569	0.48 [0.44, 0.51]	0.50 [0.44, 0.56]	0.30
Black Hispanic	NYC	154	48847	4672	0.48 [0.47, 0.49]	0.49 [0.47, 0.52]	0.42
	Boston	41	494	62	0.44 [0.37, 0.51]	0.49 [0.40, 0.59]	0.09
White	NYC	185	37547	16298	0.53 [0.52, 0.54]	0.50 [0.48, 0.53]	0.04
	Boston	109	614	385	0.52 [0.48, 0.55]	0.46 [0.40, 0.53]	0.05
White non-Hispanic	NYC	148	48327	4838	0.56 [0.55, 0.58]	0.52 [0.49, 0.55]	0.00
	Boston	106	631	324	0.52 [0.47, 0.56]	0.49 [0.43, 0.56]	0.39
White Hispanic	NYC	176	42333	11463	0.51 [0.50, 0.52]	0.49 [0.47, 0.52]	0.30
	Boston	62	631	89	0.48 [0.41, 0.55]	0.47 [0.39, 0.56]	0.81
Hispanic	NYC	180	37693	16125	0.50 [0.49, 0.51]	0.49 [0.46, 0.52]	0.41
	Boston	85	706	151	0.46 [0.41, 0.50]	0.48 [0.41, 0.55]	0.51

Table 2: Estimates of the population and personalized AUCs of the duration of Terry stops by racial group.