# Monitoring Racial Bias in Traffic Enforcement with Noisy Proxies of Driver Behavior

**Kai Cooper**[1] Gregory Lanzalotto[1] Haosen Ge[1] Jacob Kaplan[2] Scott Desposato[3] Dean Knox[1] Jonathan Mummolo[2]

[1]Wharton [2]Princeton [3]UCSD

## Fair Comparisons

Traffic stops represent the most common type of encounter in which civilians interact with police, and thus offer a critical lens into racial disparities in police enforcement (The New York Times, 2024; U.S. Department of Justice, 2024). Evidence-based debates in this area frequently rely on case-specific **benchmarks** to evaluate the racial distribution of police stops, e.g. per-capita demographics. However, these measures only approximate the desired comparison because they fail to (i) accurately describe the nature of drivers visible to officers, (ii) explain disparities due to potentially biased selection decisions by officers. **Problem**: these shortcomings have not been formalized → improvement attempts fall short. Enter this work.

## Legitimacy Crisis in Benchmark Analysis

How is discrimination measured? We want to know if the prevalence of minorities in police data is *unusually* large, relative to white civilians. But what defines "unusually"?

### Benchmark

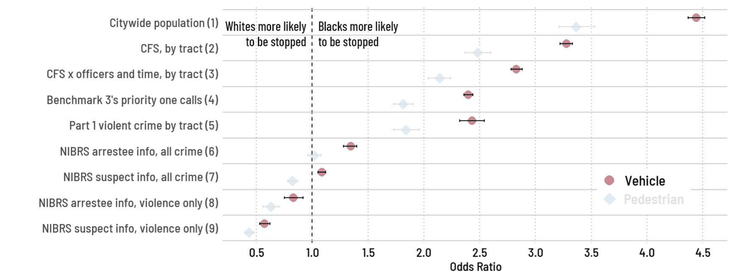
"[A benchmark] estimates the racial distribution of the individuals who would be stopped if the police were racially unbiased" (Ridgeway and MacDonald, 2011)

To quantify the extent of a violation, typically an odds ratio is used

$$\frac{\#\{\text{minorities in stops}\}}{\#\{\text{whites in stops}\}} \Big/ \frac{\#\{\text{minorities in bench.}\}}{\#\{\text{whites in bench.}\}}.$$

But this is problematic.

**Conclusions highly sensitive to choice of benchmark.**



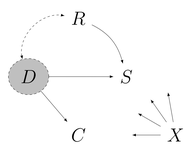

Unchecked distributional comparisons lead to wildly inconsistent results, taken from Ratcliffe and Hyland (2025)
Notice also the false equivalence between the odds ratio and the risk ratio.

*Reason:* (i) Benchmarks noncompliant with definition (ii) odds **absent theory** is too crude:

- Confounded comparisons
- Mismeasurement
- Inherent selection bias
- Unspecified data gen. process
- Inconsistent unit-of-analysis
- Undefined causal estimand

## A Formal Framework

$R$ : minority status; $D$ : dangerous driving; $C$ : benchmark flag (running example: camera citation); $S$ : officer decision; $X$ : context (spatiotemporal features).




Data generating process

### Estimand

"How much more likely is an officer to ticket a minority civilian, compared to a white civilian, for the same behavior?"

i.e. **causal risk ratio (CRR)**:

$$\psi^\dagger(x) = \frac{\Pr(S(r=1)=1 \mid D=1, X=x)}{\Pr(S(r=0)=1 \mid D=1, X=x)}$$

Interestingly, the setup of the problem means that confounding is not the primary causal issue. **Problem:** $D$, criminal behavior, is latent. **Solution:** treat $D$ as a confounder of race–officer decision endowed with an outcome-proxy, or benchmark: $C$ (Kuroki and Pearl, 2014; Wang et al., 2018). **Key intuition:** benchmark is independent of race conditional on $D$. **Examples:** traffic cameras, at-fault in vehicle collision, surpassing blood-alcohol threshold at DUI checkpoints.

## Identification

### ODDS RATIO: EVER CAUSAL?

Well, for this to be true, as we condition on actual dangerous drivers ($D=1$), we need cameras and officers to be honest about that. Both the benchmark and the officer data must reflect the population justifiably at-risk (cf. benchmark definition). First assume:

A1 (Cameras don't lie): $\Pr(D=1 \mid C=1)=1$ A2 (Officers don't lie): $\Pr(D=1 \mid S=1)=1$,

then $\psi^\dagger(x) = \text{odds}[R \mid S=1, X=x] / \text{odds}[R \mid C=1, X=x]$.

Non-collapsibility of the odds ratio → cannot average cond. odds to get marginal. Need more:

A3 For each $r$, $\Pr(S=1 \mid D=1, R=r, X=x)$, A4 $\Pr(C=1 \mid D=1, X=x)$ both constant in $x$.

Of course, w/ possible exception of A4, these assumptions are too strong.

### PARTIAL IDENTIFICATION OF CAUSAL RISK RATIO

Let $\theta_x = \Pr(D=1 \mid C=1, X=x)$ and $\eta_x = \Pr(D=1 \mid S=1, X=x)$. Under Assumption A4:

$$\frac{\sum_x \Pr(S=1,X=x)\left\{\Pr(R=1 \mid S=1,X=x)+\eta_x-1\right\}\frac{\theta_x}{\Pr(R=1|C=1,X=x)}}{\sum_x \Pr(S=1,X=x)\Pr(R=0 \mid S=1,X=x)\frac{\theta_x}{\Pr(R=0|C=1,X=x)+\theta_x-1}} \le \psi^\dagger \le \frac{\sum_x \Pr(S=1,X=x)\Pr(R=1 \mid S=1,X=x)\frac{\theta_x}{\Pr(R=1|C=1,X=x)+\theta_x-1}}{\sum_x \Pr(S=1,X=x)\left\{\Pr(R=0 \mid S=1,X=x)+\eta_x-1\right\}\frac{\theta_x}{\Pr(R=0|C=1,X=x)}}$$

for $1-\theta_x < \min_r \Pr(R=r \mid C=1, X=x)$ and $1-\eta_x < \min_r \Pr(R=r \mid S=1, X=x)$.

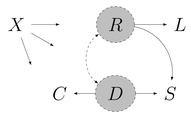
**Intuition:** think IPW.

- Count drivers of given $r$ stopped by police, correcting for $1-\eta_x$ miscapture rate
- To get counterfactual stops, weight by prevalence of $r$ in bad drivers, who we observe using cameras, but they make miscaptures fraction $1-\theta_x$ of the time
- Assign miscaptures adversarially to produce bounds: e.g. lower bound, all safe drivers stopped by police are black and all safe camera captures are white

Core theory is widely applicable, but the reality of police data makes practicalities challenging.

- Police officers misreport race
- Cameras issue tickets to owners
- Police presence distorts proxy capture
- Dangerous driving can be non-binary

## Race is Missing or Mismeasured




Modified DGP, $R$ is unobserved; $L$ surname.

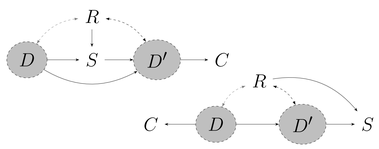
Race is not available, but last name could be. By leveraging the core proxy principle (notice $L \perp\!\!\!\perp C \mid R$), $\Pr(R=r \mid C=1)$ is identified:

$$\underbrace{\Pr(L=l \mid C=1)}_{\text{have}} = \sum_r \underbrace{\Pr(L=l \mid R=r)}_{\text{known from census}} \underbrace{\Pr(R=r \mid C=1)}_{\text{want}},$$

(We can swap $C$ with $S$ here). Solve with maximum likelihood and EM algorithm for computational efficiency. Note this approach is vital for **preventing prediction errors from selection dangerous driving**. Close overlap with recent BiRDIE idea (McCartan et al., 2024) and 2-sample IV (Inoue and Solon, 2010).

## Police Presence Distorts Proxy Capture

Observation or encounters with police may (differentially) distort camera captures.




Top left: first see the officer. Bottom right: first see camera. If officers are racially biased against minorities (whites), weighting by camera stops inflates (deflates) CRR → sign preservation. *Approximate solution: remove one camera stop per officer stop.*

## Dangerous Driving is not Always Binary

Cameras, checkpoints → $D$ is binary, e.g. speeding, blood-alcohol too high. Crashes, not quite. Binary at-fault in a crash too coarse to describe high-dimensional $\vec{D}$ bad driving. **Trick:** use crash *severities* $l$, e.g. property, injury, fatal. **Define dangerousness** $D := \Pr(C > 0 \mid \vec{D}, X)$, prob. of causing a crash of any severity. Assume officers police w/ aim of protecting public safety

$$\Pr(S=1 \mid R=r, D=d, X=x) = \sum_l \alpha_l(x) \Pr(C \ge l \mid D=d, X=x).$$

### A FALSIFICATION TEST FOR RACIAL BIAS

Under this assumption we derive a falsification test of racial bias via the implication:

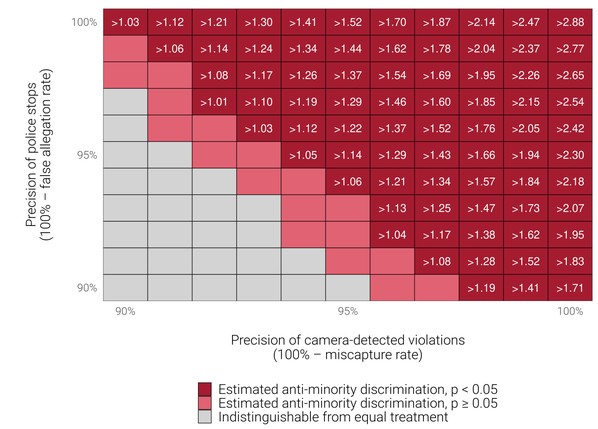
$$\Pr(R=r \mid S=1, X=x) \le \max_l \Pr(R=r \mid C \ge l, X=x).$$

If the disparity manifests linearly, so that $\alpha_l(x) \to \alpha_l(x)\gamma_r(x)$, then $\psi^\dagger(x)$ is identified by the standard odds ratio. Otherwise, two explanations: (i) disparate treatment by race enters non-linearly, (ii) disparate impact by pathological basis functions $\Pr(C \ge l \mid D, X)$ e.g. a step/bump. We show that the fraction of stops caused by (ii) is strongly limited by the presence of (i).

## Applications

### RED-LIGHT VIOLATIONS IN CHICAGO (2017–2024)

100s of cameras. 1M camera violations. 100k stops. Coarsened exact matching on $X$.




Lower confidence interval on lower bound of causal risk ratio. Focal comparison: black vs white.

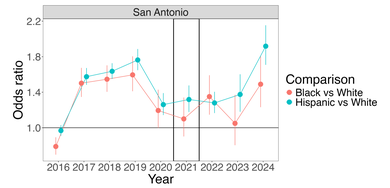
"Out of place": black drivers 40x more likely to be stopped in black areas; 30x for hispanics.

### CRASHES IN TEXAS (2016-2024)

Millions of stops; tens of millions of crashes.

| Race | Crash-based max | Prop in stops |
|---|---|---|
| Black | 0.068 | 0.077 |
| Hispanic | 0.485 | 0.550 |

San Antonio, 2021. Falsification test results




Models fit using **double debiased machine learning** for an odds ratio target (Tchetgen Tchetgen et al., 2010; Chernozhukov et al., 2018). Find **heterogeneity** over time and place; trends for different minority groups similar.

Odds ratio interpretable as a risk ratio under linearity.

Falsification test can fail but statistical power can limit detection of discrimination.