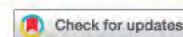


ARTICLES

<https://doi.org/10.1038/s41562-020-0858-1>nature
human behaviour

A large-scale analysis of racial disparities in police stops across the United States

Emma Pierson¹, Camelia Simoiu², Jan Overgoor², Sam Corbett-Davies¹, Daniel Jenson², Amy Shoemaker², Vignesh Ramachandran², Phoebe Barghouty², Cheryl Phillips³, Ravi Shroff⁴ and Sharad Goel^{1,2}✉

We assessed racial disparities in policing in the United States by compiling and analysing a dataset detailing nearly 100 million traffic stops conducted across the country. We found that black drivers were less likely to be stopped after sunset, when a 'veil of darkness' masks one's race, suggesting bias in stop decisions. Furthermore, by examining the rate at which stopped drivers were searched and the likelihood that searches turned up contraband, we found evidence that the bar for searching black and Hispanic drivers was lower than that for searching white drivers. Finally, we found that legalization of recreational marijuana reduced the number of searches of white, black and Hispanic drivers—but the bar for searching black and Hispanic drivers was still lower than that for white drivers post-legalization. Our results indicate that police stops and search decisions suffer from persistent racial bias and point to the value of policy interventions to mitigate these disparities.

More than 20 million Americans are stopped each year for traffic violations, making this one of the most common ways in which the public interacts with the police^{1–4}. Due to the decentralized nature of policing in the United States—and a corresponding lack of comprehensive and standardized data—it is difficult to rigorously assess the manner and extent to which race plays a role in traffic stops⁵. The most widely cited national statistics come from the Police–Public Contact Survey (PPCS)¹, which is based on a nationally representative sample of approximately 50,000 people who report having been recently stopped by the police. In addition to such survey data, some local and state agencies have released periodic reports on traffic stops in their jurisdictions, and have also made their data available to outside researchers for analysis^{6–20}. While useful, these datasets provide only a partial picture. For example, there is concern that the PPCS, like nearly all surveys, suffers from selection bias and recall errors. Data released directly by police departments are potentially more complete but are available only for select agencies, are typically limited in what is reported and are inconsistent across jurisdictions.

To address these challenges, we compiled and analysed a dataset detailing nearly 100 million traffic stops carried out by 21 state patrol agencies and 35 municipal police departments over almost a decade. This dataset was built through a series of public records requests filed in all 50 states. To facilitate future analysis, we have redistributed these records in a standardized form.

Our statistical analysis of these records proceeds in three steps. First, we assess potential bias in stop decisions by applying the 'veil of darkness' test developed by Grogger and Ridgeway²¹. This test is based on a simple observation: because the sun sets at different times throughout the year, one can examine the racial composition of stopped drivers as a function of sunlight while controlling for time of day. In particular, we use the discontinuity created by the start (and end) of daylight-saving time (DST), comparing the racial distribution of drivers stopped in the evenings immediately before

DST begins, when it is dark, to the distribution after DST begins, when it is light at the same time of day. If black drivers comprise a smaller share of stopped drivers when it is dark and accordingly difficult to determine a driver's race, that suggests black drivers were stopped during daylight hours in part because of their race. In both state patrol and municipal police stops, we find that black drivers comprise a smaller proportion of drivers stopped after sunset, suggestive of discrimination in stop decisions.

Second, we investigate potential bias in the post-stop decision to search drivers for contraband. To do so, we apply the threshold test recently developed by Simoiu et al.⁷ and refined by Pierson et al.²². The threshold test incorporates both the rate at which searches occur, as well as the success rate of those searches, to infer the standard of evidence applied when determining whom to search. This approach builds on traditional outcome analysis^{23,24}, in which a lower search success rate for one group relative to another is seen as evidence of bias against that group, as it suggests that a lower evidentiary bar was applied when making search decisions. Applied to our data, the threshold test indicates that black and Hispanic drivers were searched on the basis of less evidence than white drivers, both on the subset of searches carried out by state patrol agencies and on those carried out by municipal police departments.

Finally, we examine the effects of drug policy on racial disparities in traffic stop outcomes. We specifically compare patterns of policing in Colorado and Washington—two states that legalized recreational marijuana at the end of 2012—to those in 12 states in which recreational marijuana remained illegal. Using a difference-in-differences strategy, we find that legalization reduced both search rates and misdemeanor rates for drug offences for white, black and Hispanic drivers—though a gap in search thresholds persists.

In the process of collecting and analysing millions of traffic stop records across the country, we encountered many logistical and statistical challenges. Based on these experiences, we conclude by offering suggestions to improve data collection, analysis and reporting

¹Computer Science, Stanford University, Stanford, CA, USA. ²Management Science & Engineering, Stanford University, Stanford, CA, USA.

³Communication, Stanford University, Stanford, CA, USA. ⁴Applied Statistics, Social Science, and Humanities, New York University, New York, NY, USA.

✉e-mail: scgoel@stanford.edu



Fig. 1 | Geographic coverage of compiled traffic stop data. We analysed data on approximately 95 million stops from 21 state patrol agencies (blue) and 35 municipal police departments (red) across the country.

by law enforcement agencies. Looking forward, we hope this work provides a road map for measuring racial disparities in policing, and facilitates future empirical research into police practices.

Results

Compiling a national database of traffic stops. To assemble a national dataset of traffic stops, we filed public records requests with all 50 state patrol agencies and over 100 municipal police departments. The municipal police departments include those that serve the largest 100 cities in the nation; to achieve geographic coverage, we also included departments that serve some of the largest cities in each state.

To date, we have collected (and released) data on approximately 221 million stops carried out by 33 state patrol agencies, and 34 million stops carried out by 56 municipal police departments, for a total of 255 million records. In many cases, however, the data we received were insufficient to assess racial disparities (for example, the race of the stopped driver was not regularly recorded, or only a non-representative subset of stops was provided). For consistency in our analysis, we further restrict to stops occurring in 2011–2018, as many jurisdictions did not provide data on earlier stops. Finally, we limit our analysis to drivers classified as white, black or Hispanic, as there were relatively few recorded stops of drivers in other race groups. Our primary dataset thus consists of approximately 95 million stops from 21 state patrol agencies and 35 municipal police departments, as shown in Fig. 1 and described in more detail in Supplementary Table 2.

Because each jurisdiction provided stop data in idiosyncratic formats with varying levels of specificity, we used a variety of automated and manual procedures to create the final dataset. For each recorded stop, we attempted to extract and normalize the date and time of the stop; the county (for state patrol agencies) or police subdivision (for example, beat, precinct or zone, for municipal police departments) in which the stop took place; the race, gender and age of the driver; the stop reason (for example, speeding); whether a search was conducted; the legal justification for the search (for example, 'probable cause' or 'consent'); whether contraband was found during a search; and the stop outcome (for example, a citation or an arrest). As indicated in Supplementary Table 2, the information we received varies markedly across states. We therefore restricted each of our specific analyses to the corresponding subset of jurisdictions for which we have the required fields. Our complete data-cleaning pipeline is extensive and required subjective decisions, which we describe in more detail in Methods. For transparency and reproducibility, we have released the raw data, the standardized data and code to clean and analyse the records at <https://openpolicing.stanford.edu>.

Assessing bias in traffic stop decisions. Relative to their share of the residential population, we found that black drivers were, on average,

stopped more often than white drivers. In particular, among state patrol stops, the annual per-capita stop rate for black drivers was 0.10 compared to 0.07 for white drivers; and among municipal police stops, the annual per-capita stop rate for black drivers was 0.20 compared to 0.14 for white drivers. For Hispanic drivers, however, we found that stop rates were lower than for white drivers: 0.05 for stops conducted by state patrol (compared to 0.07 for white drivers) and 0.09 for those conducted by municipal police departments (compared to 0.14 for white drivers). In all cases, these numbers are the unweighted average annual per-capita stop rates across the states and cities we analysed, and all estimates have corresponding 95% confidence intervals (CIs) with radius <0.001. We note that these results are consistent with self-reported stop rates by white, black and Hispanic drivers who participated in the national PPCS¹.

These numbers are a starting point for understanding racial disparities in traffic stops, but they do not, per se, provide strong evidence of racially disparate treatment. In particular, per-capita stop rates do not account for possible race-specific differences in driving behaviour, including amount of time spent on the road and adherence to traffic laws. For example, if black drivers, hypothetically, spend more time on the road than white drivers, that could explain the higher stop rates we see for the former, even in the absence of discrimination. Moreover, drivers may not live in the jurisdictions where they were stopped, further complicating the interpretation of population benchmarks.

Quantifying potential bias in stop decisions is a statistically challenging problem, in large part because one cannot readily measure the racial distribution of those who actually violated traffic laws because the data contain information only on those stopped for such offences. To mitigate this benchmarking problem, Grogger and Ridgeway²¹ proposed a statistical approach known as the veil-of-darkness test. Their method starts from the idea that officers who engage in racial profiling are less able to identify a driver's race after dark than during the day. As a result, if officers are discriminating against black drivers—all else being equal—one would expect black drivers to comprise a smaller share of stopped drivers at night, when a veil-of-darkness masks their race. To account for patterns of driving and police deployment that may vary throughout the day, the test leverages the fact that the sun sets at different times during the year. For example, although it is typically dark at 19:00 during the winter months, it is often light at that time during the summer.

To illustrate the intuition behind this method, in Fig. 2 we examine the demographic composition of drivers stopped by the Texas State Patrol at various times of day. Each panel in the plot shows stops occurring in a specific 15-min window (for example, 19:00–19:15), and the horizontal axis indicates minutes since dusk. We restrict to white and black drivers—because the ethnicity of Hispanic drivers is not always apparent, even during daylight hours—and, following Grogger and Ridgeway, we filter out stops that occurred in the ~30-min period between sunset and dusk, when it is neither 'light' nor 'dark'. For each time period, the plot shows a marked drop in the proportion of drivers stopped after dusk who are black, suggestive of discrimination in stop decisions.

In this basic form, darkness—after adjusting for time of day—is a function of the date. As such, to the extent that driver behaviour changes throughout the year, and that these changes are correlated with race, the test can suggest discrimination where there is none. To account for these potential seasonal effects, we applied a more robust variant of the veil-of-darkness test that restricts to two 60-day windows centred on the beginning and end of DST²⁵. On this subset of the data, we fit the following logistic regression model:

$$\Pr(\text{black}|t, g, p, d, s, c) = \text{logit}^{-1}(\alpha_s \times s \times d + \alpha_c \times c \times d + \beta^T \times \text{ns}_6(t) + \gamma[g] + \delta[p]),$$

where $\Pr(\text{black}|t, g, p, d, s, c)$ is the probability that a stopped driver is black at a certain time t , location g and period p (with two periods

ARTICLES

NATURE HUMAN BEHAVIOUR

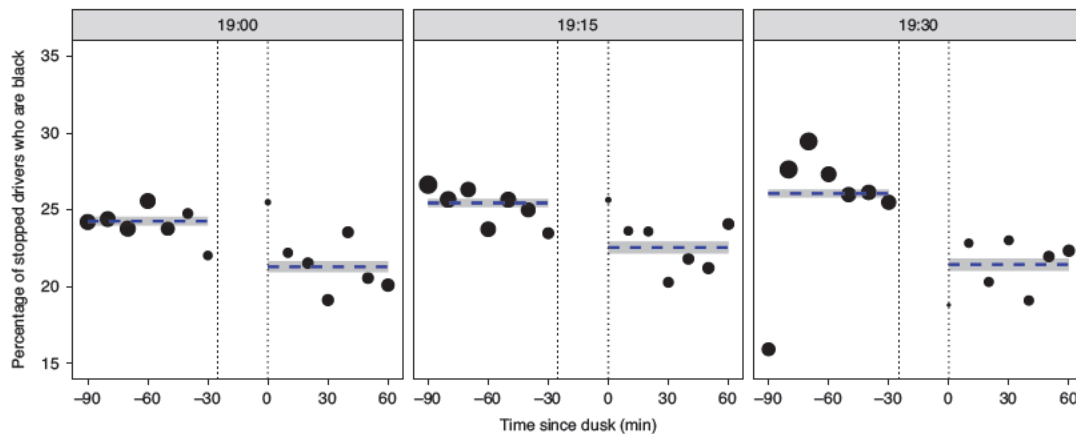


Fig. 2 | An illustration of the veil-of-darkness test for stops occurring in three short time windows in a single state, Texas. For each window (19:00–19:15, 19:15–19:30 and 19:30–19:45), we compute the percentage of stops that involved black drivers for a series of 10-min periods before and after dusk. The figure is based on 112,938 stops of black and white drivers (35,270 during 19:00–19:15, 38,726 during 19:15–19:30 and 38,942 during 19:30–19:45), with points sized according to the total number of stops in each bin. The vertical line at $t=0$ indicates dusk, at which point it is generally considered 'dark'; we remove stops in the ~30-min period between sunset (indicated by the left-most vertical line in each panel) and dusk, as this period is neither 'light' nor 'dark'. The dashed horizontal lines show the overall proportion of stops involving black drivers before and after dark, with 95% CI. For all three depicted time windows, black drivers comprise a smaller share of stopped drivers after dark, when a veil of darkness masks their race, suggestive of racial profiling.

per year, corresponding to the start and end of DST), with darkness status $d \in \{0, 1\}$ indicating whether a stop occurred after dusk ($d=1$) or before sunset ($d=0$), and with the enforcement agency being either state patrol, $s \in \{0, 1\}$ or city police department, $c \in \{0, 1\}$. In this model, $ns_g(t)$ is a natural spline over time with six degrees of freedom, $\gamma[g]$ is a fixed effect for location g and $\delta[p]$ is a fixed effect for period p . The location $g[i]$ for stop i corresponds to either the county (for state patrol stops) or city (for municipal police department stops), with $\gamma[g[i]]$ the corresponding coefficient. Finally, $p[i]$ captures whether stop i occurred in the spring (within a month of beginning DST) or the fall (within a month of ending DST) of each year, and $\delta[p[i]]$ is the corresponding coefficient for this period. For computational efficiency, we rounded time to the nearest 5-min interval when fitting the model.

The main terms of interest in our model are α_s and α_c , which describe differences in the composition of stopped drivers between daylight and dark, after adjusting for time and location, with $\alpha_s, \alpha_c < 0$ suggesting discrimination against black drivers for state patrols and city departments, respectively. We find $\alpha_s = -0.033$ (95% CI $[-0.039, -0.027]$, $P < 0.001$) and $\alpha_c = -0.039$ (95% CI $[-0.045, -0.033]$, $P < 0.001$), suggesting discrimination among both state patrols and municipal police departments. To gauge the robustness of our results, we fit multiple variations of the veil-of-darkness model described above (for example, using different degrees for the natural spline). In all cases, we found qualitatively similar results that are suggestive of racial bias against black drivers in stop decisions, as described in Methods and Supplementary Table 1.

The veil-of-darkness test is a popular technique for assessing disparate treatment but, like all statistical methods, it comes with caveats. Results could be skewed if race-specific driving behaviour is related more to lighting than time of day, leading the test to suggest discrimination where there is none. Conversely, artificial lighting (for example, from street lamps) can weaken the relationship between sunlight and visibility, and so the method may underestimate the extent to which stops are predicated on perceived race. Finally, if violation type is related to lighting, the test could give an inaccurate measure of discrimination. For example, broken tail lights are more likely to be detected at night and could potentially be more common among black drivers¹⁷, which could in turn mask

discrimination. To address this last limitation, one could exclude stops prompted by such violations but our data, unfortunately, do not consistently indicate stop reasons. Despite these shortcomings, we believe the veil-of-darkness test provides a useful, if imperfect, measure of bias in stop decisions.

Assessing bias in search decisions. After stopping a driver, officers may carry out a search of the driver or vehicle if they suspect more serious criminal activity. We next investigate potential bias in these search decisions. Among stopped drivers, we found that black and Hispanic individuals were, on average, searched more often than white individuals. However, as with differences in stop rates, the disparities we see in search rates are not necessarily the product of discrimination. Black and Hispanic drivers might, hypothetically, carry contraband at higher rates than white drivers, and so elevated search rates may result from routine police work even if no racial bias were present. To measure the role of race in search decisions, we apply two statistical strategies: outcome analysis and threshold analysis. To do so, we limit to the eight state patrol agencies and six municipal police departments for which we have sufficient data on the location of stops, whether a search occurred and whether those searches yielded contraband. We specifically consider state patrol agencies in Connecticut, Illinois, North Carolina, Rhode Island, South Carolina, Texas, Washington and Wisconsin; and municipal police departments in Nashville, TN, New Orleans, LA, Philadelphia, PA, Plano, TX, San Diego, CA and San Francisco, CA. We defer to each department's characterization of 'contraband' when carrying out this analysis.

In the eight state patrol agencies we consider, search rates were 4.3% (95% CI 4.2–4.4%) for black drivers, 4.1% (95% CI 4.0–4.1%) for Hispanic drivers and 1.9% (95% CI 1.9–1.9%) for white drivers. In particular, the gap in search rates between black and white drivers was 2.4% (95% CI 2.3–2.5%, $P < 0.001$) and the gap in search rates between Hispanic and white drivers was 2.2% (95% CI 2.1–2.2%, $P < 0.001$). Similarly, in the six municipal police departments we consider, the search rates were 9.5% (95% CI 9.4–9.6%) for black drivers, 7.2% (95% CI 7.0–7.3%) for Hispanic drivers and 3.9% (95% CI 3.8–3.9%) for white drivers. The gap in municipal search rates between black and white drivers was 5.6% (95% CI

5.5–5.7%, $P < 0.001$) and the gap in search rates between Hispanic and white drivers was 3.3% (95% CI 3.1–3.5%, $P < 0.001$). As above, these numbers are unweighted average search rates across our states and cities, respectively.

In these jurisdictions, stopped black and Hispanic drivers were searched about twice as often as stopped white drivers. To assess whether this gap resulted from biased decision-making, we apply the outcome test, originally proposed by Becker^{23,24}, to circumvent omitted variable bias in traditional tests of discrimination. The outcome test is based not on the search rate but on the ‘hit rate’: the proportion of searches that successfully turn up contraband. Becker argued that even if minority drivers are more likely to carry contraband, in the absence of discrimination, searched minorities should still be found to have contraband at the same rate as searched whites. If searches of minorities are successful less often than searches of whites, it suggests that officers are applying a double standard, searching minorities on the basis of less evidence. Implicit in this test is the assumption that officers exercise discretion in whom to search; therefore, when possible, we exclude non-discretionary searches, such as vehicle impound searches and searches incident to arrest, as those are often required as a matter of procedure, even in the absence of individualized suspicion. We note that outcome tests gauge discrimination only at one specific point in the decision-making process—in this case, the decision to search a driver who has been stopped. In particular, this type of analysis does not capture bias in the stop decision itself.

In Fig. 3 (top row), we plot hit rates by race and location for the states (left) and for the cities (right) for which we have the necessary information. Across jurisdictions, we consistently found that searches of Hispanic drivers were less successful than those of white drivers. However, searches of white and black drivers had more comparable hit rates. The outcome test thus indicates that search decisions may be biased against Hispanic drivers, but the evidence is more ambiguous for black drivers. Aggregating across state patrol stops, contraband was found in 32.0% (95% CI 31.6–32.4%) of searches of white drivers compared to 24.3% (95% CI 23.5–25.2%) of searches of Hispanic drivers and 29.4% (95% CI 28.7–30.0%) of searches of black drivers. In particular, the gap in hit rates between white and Hispanic drivers was 7.6% (95% CI 6.7–8.6%, $P < 0.001$), and the gap in hit rates between white and black drivers was 2.6% (95% CI 1.9–3.4%, $P < 0.001$). Similarly, aggregating across municipal police departments, contraband was found in 18.2% (95% CI 17.8–18.7%) of searches of white drivers compared to 11.0% (95% CI 10.6–11.5%) of searches of Hispanic drivers and 13.9% (95% CI 13.7–14.2%) of searches of black drivers. In this case, the gap in hit rates between white and Hispanic drivers was 7.2% (95% CI 6.6–7.8%, $P < 0.001$) and the gap in hit rates between white and black drivers was 4.3% (95% CI 3.8–4.8%, $P < 0.001$). These numbers all indicate unweighted averages across our cities and states, respectively.

The outcome test is intuitively appealing, but it is an imperfect barometer of bias; in particular, it suffers from the problem of infra-marginality^{8,26,27}. To illustrate this shortcoming, suppose that there are two, easily distinguishable, types of white driver: those who have a 5% chance of carrying contraband and those who have a 75% chance of carrying contraband. Likewise assume that black drivers have either a 5 or 50% chance of carrying contraband. If officers search drivers who are at least 10% likely to be carrying contraband, then searches of white drivers will be successful 75% of the time whereas searches of black drivers will be successful only 50% of the time. Thus, although the search criterion is applied in a race-neutral manner, the hit rate for black drivers is lower than that for white drivers and the outcome test would (incorrectly) conclude that searches are biased against black drivers. The outcome test can similarly fail to detect discrimination when it is present.

Addressing this possibility, Knowles et al.²⁷ suggested an economic model of behaviour—known as the KPT model—in which drivers balance their utility for carrying contraband with the risk of getting caught, while officers balance their utility of finding contraband with the cost of searching. Under equilibrium behaviour in this model, the hit rate of searches is identical to the search threshold and so one can reliably detect discrimination with the standard outcome test. However, Engel and Tillyer²⁸ argue that the KPT model of behaviour requires strong assumptions, including that drivers and officers are rational actors and that every driver has perfect knowledge of the likelihood that he or she will be searched.

To mitigate the limitations of outcome tests (as well as limitations of the KPT model), the threshold test has been proposed as a more robust means for detecting discrimination^{7,22}. This test aims to estimate race-specific probability thresholds above which officers search drivers—for example, the 10% threshold in the hypothetical situation above. Even if two race groups have the same observed hit rate, the threshold test may find that one group is searched on the basis of less evidence, indicative of discrimination. To accomplish this task, the test uses a Bayesian model to simultaneously estimate race-specific search thresholds and risk distributions that are consistent with the observed search and hit rates across all jurisdictions. The threshold test can thus be seen as a hybrid between outcome and benchmark analysis, as detailed in Methods.

As shown in Fig. 3 (bottom row), the threshold test indicates that the bar for searching black and Hispanic drivers is generally lower than that for searching white drivers across the municipal police departments and states we consider. In aggregate across cities, the inferred threshold for white drivers is 10.0% compared to 5.0 and 4.6% for black and Hispanic drivers, respectively. The estimated gaps in search thresholds between white and non-white drivers are large and statistically significant: the 95% credible interval for the Hispanic–white difference is (–6.4, –4.4%) and the corresponding interval for the black–white difference is (–6.1, –4.0%). Similarly across states, the inferred threshold for white drivers is 20.9% compared to 16.0% for black drivers and 13.9% for Hispanic drivers. These differences are again large and statistically significant: the 95% credible interval for the Hispanic–white gap is (–8.4, –5.6%); and the analogous interval for the black–white gap is (–6.5%, –3.1%). As with our outcome results, aggregate thresholds are computed by taking an unweighted average of the city- and state-specific thresholds, respectively.

Compared to by-location hit rates, the threshold test more strongly suggests discrimination against black drivers, particularly for municipal stops. Consistent with past work⁷, this difference appears to be driven by a small but disproportionate number of black drivers who have a high inferred likelihood of carrying contraband. Thus, even though the threshold test finds that the bar for searching black drivers is lower than that for white drivers, these groups have more similar hit rates.

The threshold test provides evidence of racial bias in search decisions. However, as with all tests of discrimination, it is important to acknowledge limits in what one can conclude from such statistical analysis per se. For example, if search policies differ not only across, but also within, the geographic subdivisions we consider, then the threshold test might mistakenly indicate discrimination where there is none. Additionally, if officers disproportionately suspect more serious criminal activity when searching black and Hispanic drivers compared to white drivers (for example, possession of larger quantities of contraband), then lower observed thresholds may stem from non-discriminatory police practices. Finally, we note that thresholds cannot be identified by the observed data alone⁷, and so inferences are dependent on the specific functional form of the underlying Bayesian model, including the prior distributions. (In Methods, however, we show that our main results are robust to relatively large changes to prior distributions).

ARTICLES

NATURE HUMAN BEHAVIOUR

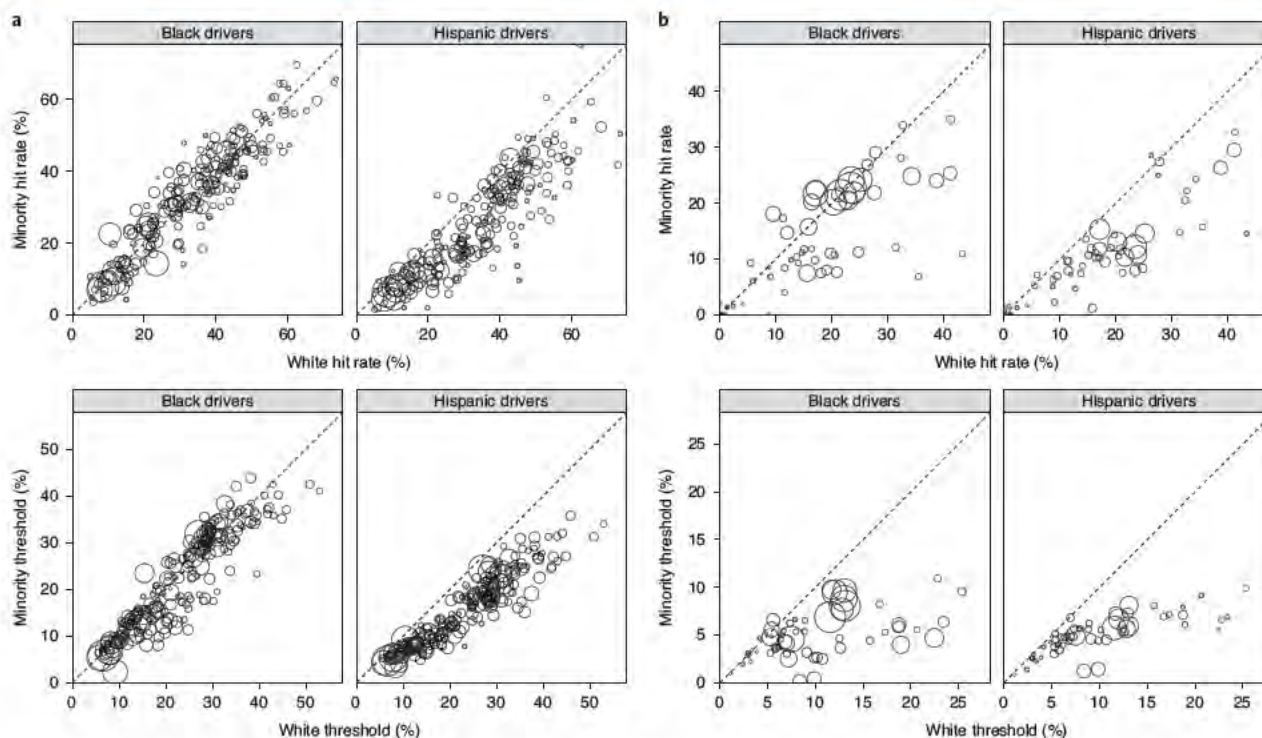


Fig. 3 | Hit rates and inferred search thresholds by race and location. **a**, Data for state patrol are based on 647,251 searches in eight states, with points corresponding to counties. **b**, Data for municipal police are based on 187,145 searches in six cities, with points corresponding to police precincts. Points are sized by number of searches. The top panels indicate search hit rates, and the bottom panels indicate search thresholds. Across locations, the inferred thresholds for searching black and Hispanic drivers are typically lower than those for searching white drivers, suggestive of bias. Despite these lower inferred search thresholds, hit rates for black drivers are comparable to those for white drivers, ostensibly due to the problem of infra-marginality in outcome tests.

The effects of legalization of marijuana on stop outcomes. We conclude our analysis by investigating the effects of legalization of recreational marijuana on racial disparities in stop outcomes. We specifically examine Colorado and Washington, two states in which marijuana was recently legalized and for which we have detailed data before and after legalization. In line with expectations, we find that the proportion of stops that resulted in either a drug-related infraction or misdemeanor fell substantially in both states after marijuana was legalized at the end of 2012 (see Supplementary Fig. 2). Notably, since black drivers were more likely to be charged with such offences previous to legalization, these drivers were also disproportionately impacted by the policy change. This finding is consistent with past work showing that marijuana laws disproportionately affect minorities²⁹.

Because the policy change decriminalized an entire class of behaviour (that is, possession of small amounts of marijuana), it is not surprising that drug offences correspondingly decreased. However, it is less clear, *a priori*, how the change might have affected officer behaviour more broadly. Investigating this issue, we found that after the legalization of marijuana the number of searches fell substantially in Colorado and Washington (Fig. 4), ostensibly because the policy change removed a common reason for conducting searches.

Because black and Hispanic drivers were more likely to be searched before legalization, the policy change reduced the absolute gap in search rates between race groups; however, the relative gap persisted, with black and Hispanic drivers still more likely to be searched than white drivers post-legalization. We further note that marijuana legalization had secondary impacts for law-abiding drivers, because fewer searches overall also meant fewer searches of

those without contraband. In the year after legalization in Colorado and Washington, about one-third fewer drivers were searched with no contraband found than in the year before legalization.

As further evidence that the observed drop in search rates in Colorado and Washington was due to marijuana legalization, we note that in the 12 states where marijuana was not legalized—and for which we have the necessary data—search rates did not exhibit sharp drops at the end of 2012 (Fig. 5). To add quantitative detail, we computed a difference-in-differences estimate³⁰; our precise model specification and the fitted coefficients are described in Methods. We found that the race-specific coefficients are large and negative for white, black and Hispanic drivers, which again suggests that the observed drop in searches in Colorado and Washington was caused by the legalization of marijuana in those states.

Despite the legalization of marijuana decreasing search rates across race groups, Fig. 4 shows that the relative disparity between whites and minorities remained. We applied the threshold test to assess the extent to which this disparity in search rates may reflect bias. Examining the inferred thresholds (shown in Supplementary Fig. 3), we found that white drivers faced consistently higher search thresholds than minority drivers, both before and after marijuana legalization. The data thus suggest that, although overall search rates dropped in Washington and Colorado, black and Hispanic drivers still faced discrimination in search decisions.

Discussion

Analysing nearly 100 million traffic stops across the country, we have worked to quantify the often complex relationship between race and policing in the United States. Our analysis provides evidence that decisions about whom to stop and, subsequently, whom to search are

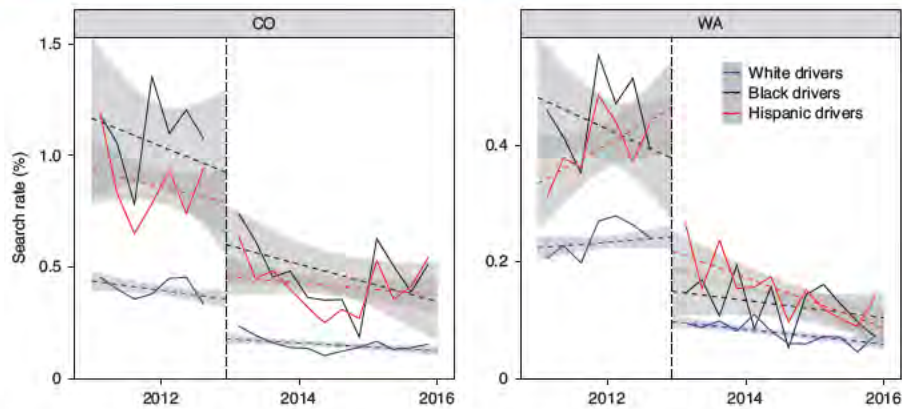


Fig. 4 | Percentage of stops that resulted in a search before and after recreational marijuana was legalized in Colorado and Washington at the end of 2012. After legalization, there was a substantial drop in search rates. The vertical dashed lines indicate the timing of legalization; the other dashed lines show fitted linear trends pre- and post-legalization. The left panel (Colorado) represents 1,674,619 stops and the right panel (Washington) represents 3,985,677 stops. We excluded data for the fourth quarter of 2012, since that period includes stops both before and after legalization. Additionally, in both states, we excluded searches incident to an arrest and other searches that are conducted as a procedural matter, irrespective of any suspicion of drug possession.

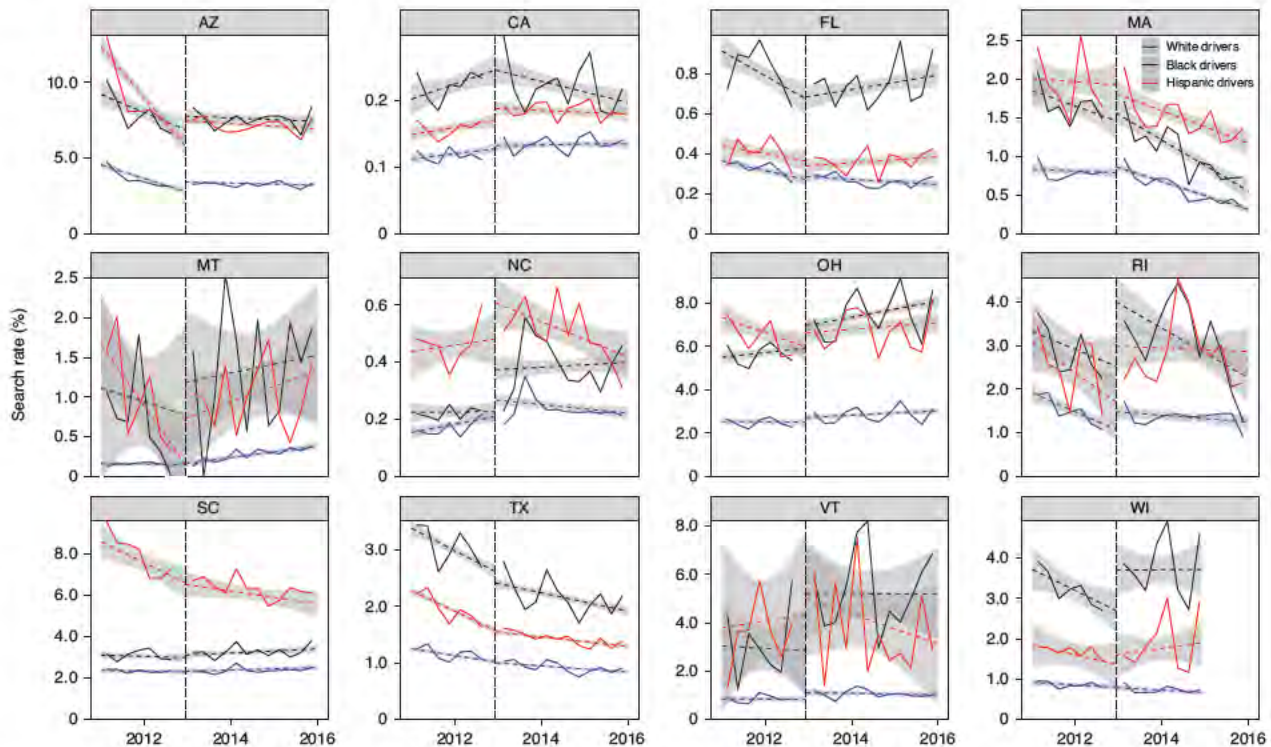


Fig. 5 | Percentage of stops that resulted in a search in 12 states in which recreational marijuana was not legalized. In the 12 states where marijuana was not legalized—and for which we have the necessary data—search rates did not exhibit sharp declines at the end of 2012, which further suggests that marijuana legalization caused the observed drop in search rates in Colorado and Washington. The plots are based on the following number of stops: 2,038,850 in AZ, 19,012,414 in CA, 3,686,757 in FL, 1,773,546 in MA, 547,115 in MT, 3,500,180 in NC, 4,660,935 in OH, 229,691 in RI, 3,909,950 in SC, 20,494,642 in TX, 250,949 in VT and 741,245 in WI. Dashed lines are as in Fig. 4.

biased against black and Hispanic drivers. Our results, however, also point to the power of policy interventions—specifically, legalization of recreational marijuana—to reduce these racial disparities.

These findings lend insight into patterns of policing, but they only partially capture the wider impacts of law enforcement on

communities of colour. If, for example, officers disproportionately patrol black and Hispanic neighbourhoods, the downstream effects can be injurious even if individual stop decisions are not directly affected by the colour of one's skin. Similarly, enforcement of minor traffic violations, like broken tail lights—even if conducted

ARTICLES

NATURE HUMAN BEHAVIOUR

uniformly and without animus—can place heavy burdens on black and Hispanic drivers without improving public safety¹⁷. We hope the data we have collected and released are useful for measuring, and in turn addressing, these broader effects.

In the course of carrying out this study, we encountered many challenges working with large-scale policing data. We conclude by offering several recommendations for future data collection, release and analysis. As a minimum, we encourage jurisdictions to collect individual-level stop data that include the date and time of the stop; the location of the stop; the race, gender and age of the driver; the stop reason; whether a search was conducted; the search type (for example, 'probable cause' or 'consent'); whether contraband was found during a search; the stop outcome (for example, a citation or an arrest); and the specific violation with which the driver was charged. Most jurisdictions collect only a subset of this information. There are also variables that are currently rarely collected but would be useful for analysis, such as indicia of criminal behaviour, an officer's rationale for conducting a search and short narratives written by officers describing the incident. New York City's UF-250 form for pedestrian stops is an example of how such information can be efficiently collected^{31,32}.

Equally important to data collection is ensuring the integrity of the recorded information. We frequently encountered missing values and errors in the data (for example, implausible values for a driver's age and invalid racial categorizations). Automated procedures can be put in place to help detect and correct such problems. For example, the recorded race of the driver is often based on the officer's perception rather than a driver's self-categorization. While there are perhaps sound reasons for this practice, it increases the likelihood of errors. To quantify and correct for this issue, police departments might regularly audit their data, possibly by comparing an officer's perception of race to a third party's judgement based on driver's licence photos for a random sample of stopped drivers.

Despite the existence of public records laws, several jurisdictions failed to respond to our repeated requests for information. We hope that law enforcement agencies consider taking steps to make data more accessible to both external researchers and the public. Connecticut and North Carolina are at the forefront of opening up their data, providing online portals for anyone to download and analyse this information.

We also hope that police departments start to analyse their data regularly and report the results of their findings. Such analyses might include estimates of stop, search and hit rates stratified by race, age, gender and location; distribution of stop reasons by race; and trends over time. More ambitiously, departments could use their data to design statistically informed guidelines that encourage more consistent, efficient and equitable decisions^{31,33–35}. Many of these analyses can be automated and rerun regularly with little marginal effort. In conjunction with releasing the data underlying these analyses, we recommend that the analysis code also be released to ensure reproducibility.

Finally, it bears emphasis that the type of large-scale data analysis we have carried out in this paper is but one of many complementary ways to gauge and rectify bias in police interactions with the public. Just as critical, for example, are conversations with both officers and community members, who can often provide more nuance than is possible with our aggregate statistical approach. Collecting, releasing and analysing police data are important steps for increasing the effectiveness and equity of law enforcement practices, and for improving relations with the public through transparency. Ultimately, though, data collection and analysis are not enough. We must act on the results of such efforts if we are to reduce the persistent, discriminatory impacts of policing on communities of colour.

Methods

Data collection and standardization. Our primary dataset includes 94,778,505 stops from 21 state patrol agencies and 35 municipal police departments.

For more detail on all jurisdictions whose data were used in our analyses, Supplementary Table 2 lists the total number of stops and the range of years in which these stops took place. The table also indicates whether each of several important covariates was available in each jurisdiction: the date and time of the stop; more granular geographic information; the race, age and gender of the stopped driver; and whether a search was conducted and, if so, whether contraband was found. Our data-processing pipeline was extensive. Below we describe some of the key steps in this process, and also note that the complete code required to process the data is available at <https://openpolicing.stanford.edu>.

For each state and city, we requested individual-level records for traffic stops conducted since 2005, under the state's public records law and filed with the agency responsible for traffic stop data collection. The 33 states and 56 cities that complied provided the data in various formats, including raw text files, Microsoft Excel spreadsheets and Microsoft Access databases. We converted all data we received to a standard comma-separated, text-file format. The states and cities varied widely in terms of the availability of the different data fields, the manner and detail of how the data were recorded and in recording consistency from year to year, even within the same location.

We standardized available fields when possible. Often, locations provided dictionaries to map numeric values to human-interpretable ones. Aggregated data—as provided by Missouri, Nebraska and Virginia—were disaggregated by expanding the number of rows by the reported count. For some locations, we had to manually map the provided location data to a county or district value. For example, in Washington, counties were mapped by first computing the latitude and longitude of the highway post that was recorded for the stop, and then those coordinates were mapped to a county using a shapefile. The raw data we received from the states and cities, the processed data we used in this analysis and the code to clean and analyse the data are all available for public inspection and reuse.

In many cases, more than one row in the raw data appeared to refer to the same stop. For example, in several jurisdictions each row in the raw data referred to one violation, not one stop. We detected and reconciled such duplicates by matching on a location-specific set of columns. For example, in Colorado we counted two rows as duplicates if they had the same officer identification code, officer first and last name, driver first and last name, driver birth date, stop location (precise to the milepost marker) and stop date and time. This type of de-duplication was a common procedure that we applied to many states and cities.

The raw data provided to us by state and municipal police agencies often contained clear errors. We ran numerous automated checks to detect and correct these where possible, although some errors probably remain due to the complex nature of the data. For example, after examining the distribution of recorded values in each jurisdiction, we discovered a spurious density of stops in North Carolina listed as occurring at precisely midnight. As the value '00:00' was probably used to indicate missing information, we treated it as such. In Pittsburgh, PA, recorded values for 'sex' and 'gender' did not match 73% of the time in the pedestrian stop data, suggesting data corruption. (Note that we did not include pedestrian stop data in our analysis, but we have released those records for other researchers to use.)

In another example, past work revealed that Texas State Patrol officers incorrectly recorded many Hispanic drivers as white, an error the agency subsequently corrected³⁶. To investigate and adjust for this issue, we imputed Hispanic ethnicity from surnames. To carry out this imputation, we used a dataset from the U.S. Census Bureau that estimates the racial and ethnic distribution of people with a given surname for surnames occurring at least 100 times³⁷. To increase the matching rate, we performed minor string edits to the names, including removal of punctuation and suffixes (for example, 'Jr.' and 'II'), and considered only the longest word in multi-part surnames. Following previous studies^{38,39}, we defined a name as 'typically' Hispanic if at least 75% of people with that name identified as Hispanic, and we note that 90% of those with typically Hispanic names identified as Hispanic in the 2000 Census.

Among drivers with typically Hispanic names, the proportion labelled as Hispanic in the raw data was quite low in Texas (37%), corroborating past results. For comparison, we considered Arizona and Colorado, the two other states that included driver name in the raw data. The proportion of drivers with typically Hispanic names labelled as Hispanic in the raw data was 70% in Colorado and 79% in Arizona, much higher than in Texas. Because of this known issue in the Texas data, we re-categorized as 'Hispanic' all drivers in Texas with Hispanic names who were originally labelled 'white' or who had missing race data; this method adds about 1.9 million stops of Hispanic drivers over the period 2011–2015. We did not re-categorize drivers in any other jurisdiction.

Veil-of-darkness analysis. In our veil-of-darkness analysis, we compared stop rates before sunset and after dusk—as is common when applying this test. Specifically, sunset is defined as the point in time where the sun dips below the horizon, and dusk (also known as the end of civil twilight) is the time when the sun is six degrees below the horizon and when it is generally considered to be 'dark'. As recommended by Grogger and Ridgeway²¹, we further restricted to stops that occurred during the 'inter-twilight period': the range from the earliest to the latest time that dusk occurs in the year. This range is approximately 17:00–22:00, although the precise values differ by location and year. All times in

the inter-twilight period are, by definition, light at least once in the year and dark at least once in the year. In Supplementary Table 1, we report the results of several variations of our primary veil-of-darkness model (for example, varying the degree of the natural spline from 1 to 6). The results were statistically significant and qualitatively similar in all cases.

Threshold test. The threshold test for discrimination was introduced by Simoiu et al.⁷ to mitigate the most serious shortcomings of benchmark and outcome analysis. The test is informed by a stylized model of officer behaviour. During each stop, officers observe a myriad of contextual factors—including the age, gender and race of the driver, and behavioural indicators of nervousness or evasiveness. We imagine that officers distil all these complex signals down to a single number, p , that represents their subjective estimate of the likelihood that the driver is carrying contraband. Based on these factors, officers are assumed to conduct a search if, and only if, their subjective estimate of finding contraband, p , exceeds a fixed, race-specific search threshold for each location (for example, county or district). Treating both the subjective probabilities and the search thresholds as latent, unobserved quantities, our goal is to infer them from data. The threshold test takes a Bayesian approach to estimating the parameters of this process, with the primary goal of inferring race-specific search thresholds for each location.

Under this model of officer behaviour, we interpret lower search thresholds for one group relative to another as evidence of discrimination. If, for example, officers have a lower threshold for searching black drivers than white drivers, that would indicate that they are willing to search black drivers on the basis of less evidence than for white drivers—and we would conclude that black drivers are being discriminated against. In the economics literature, this type of behaviour is often called taste-based discrimination³⁴ as opposed to statistical discrimination^{40,41}, in which officers might use a driver's race to improve their estimate that the driver is carrying contraband. Regardless of whether such information increases the efficiency of searches, officers are legally barred from using race to inform search decisions outside of circumscribed situations (for example, when acting on specific and reliable suspect descriptions that include race among other factors). The threshold test, however, aims to capture only taste-based discrimination, as is common in the empirical literature on discrimination.

In our work, we applied a computationally fast variant of the threshold test developed by Pierson et al.²³, which we fit separately on both state patrol stops and municipal police stops. As described below, we modified the Pierson et al. test to include an additional hierarchical component. For example, while the original Pierson et al. model considered one state, and allowed parameters to vary by county (and race) within that state, our state patrol model considers multiple states, allowing parameters to vary by county (and race) within each state. Relevant information is partially pooled across both counties and states. Similarly, our municipal model considers multiple police departments, allowing parameters to vary by police district (and race) within each department.

To run the threshold test, we assume the following information is observed for each stop, i :

1. the race of the driver, r_i
2. the region—state (for the state patrol model) or city (for the municipal police model)—where the stop occurred, g_i (for example, Texas or Nashville)
3. the specific county (for the state patrol model) or district (for the municipal police model) where the stop occurred, d_i (for example, Harris County, TX or Hermitage Precinct, Nashville)
4. whether the stop resulted in a search, indicated by $S_i \in \{0, 1\}$, and
5. whether the stop resulted in contraband recovery, indicated by $H_i \in \{0, 1\}$

To formally describe the threshold test, we need to specify the parametric process of search and recovery, as well as the priors on those parameters. For ease of exposition, we present the model for state patrol stops, where stops occur in counties that are nested within states. The municipal stop model has the same structure, with districts corresponding to counties and cities corresponding to states. We start by describing the latent signal distributions on which officers base their search decisions, which we parameterize by the race and location of drivers. As detailed in Pierson et al.²³, these signal distributions are modelled as homoskedastic discriminant distributions, a class of logit-normal mixture distributions supported on the unit interval $[0, 1]$. Discriminant distributions can closely approximate beta distributions, but they have properties that are computationally attractive. Each race- and county-specific signal distribution can be described using two parameters, which we denote by $\phi_{r,d}$ and $\delta_{r,d}$. In our setting, $\phi_{r,d} \in (0, 1)$ is the proportion of drivers of race r in county d that carry contraband; and $\delta_{r,d} > 0$ characterizes how difficult it is to identify drivers with contraband³⁴.

We impose additional structure on the signal distributions by assuming that $\phi_{r,d}$ and $\delta_{r,d}$ can be decomposed into additive race and location terms. Specifically, given parameters $\phi_{r,g}$ for each race group r and state g , and parameters ϕ_d for each county d , we set

$$\phi_{r,d} = \text{logit}^{-1}(\phi_{r,g[d]} + \phi_d)$$

where $g[d]$ denotes the state g in which county d lies. Similarly, for parameters $\delta_{r,g}$ and δ_d , we set

$$\delta_{r,d} = \exp(\delta_{r,g[d]} + \delta_d)$$

Following Pierson et al.²³, we use hierarchical priors⁴² on the signal distributions that restrict geographical heterogeneity while allowing for base rates to differ across race groups. This choice substantially accelerates model fitting. In particular, we use the following priors:

$$\mu_\phi, \mu_\delta \sim N(0, 1)$$

$$\phi_d, \delta_d \sim N(0, 0.1^2)$$

$$\phi_{r,g} \sim N(\mu_\phi, 0.1^2)$$

$$\delta_{r,g} \sim N(\mu_\delta, 0.1^2)$$

Next we detail the structure of our search thresholds. In our informal description above, officers search drivers when their estimated likelihood of possessing contraband exceeds a specific race- and county-specific value $t_{r,d} \in (0, 1)$. For computational reasons, we map these thresholds from the unit interval to the real line via a monotonic transformation outlined in Pierson et al.²². For simplicity, we continue using the notation $t_{r,d}$ to denote these transformed values. As above, we set a hierarchical prior on the threshold values. Specifically, we have

$$t_{r,d} \sim N(t_r, \sigma_t^2)$$

$$t_r \sim N(0, 1),$$

$$\sigma_t \sim N(0, 1)$$

Finally, given this parametric model of signals and thresholds, we describe the underlying data-generating process, which mirrors our informal description above. For each stop i , we draw a signal p_i from the associated race- and location-specific discriminant distribution. If the signal p_i exceeds the race- and location-specific threshold t_{r_i,d_i} , then a search is conducted and $S_i = 1$; otherwise there is no search and $S_i = 0$. If a search is conducted, contraband is found with probability p_n in which case $H_i = 1$; otherwise, $H_i = 0$.

The hierarchical structure we employ allows us to make reasonable inferences even for locations with a relatively small number of stops. However, to ensure more statistically robust estimates, we limit our analysis to counties and districts with at least 50 searches per race group. To gauge the sensitivity of our results, we repeated our analysis with priors having 0.5× and 1.5× s.d. of the priors in our main analysis, and found nearly identical results under these transformations.

For both the state patrol and municipal department model, we perform posterior predictive checks to ensure that the models fit the data well. Specifically, for each county (state patrol model) or district (municipal department model) and race group, we compare the observed search and recovery rates to their expected values under the assumed data-generating process, with parameters drawn from the inferred posterior distribution. Such posterior predictive checks are a common approach for identifying and measuring systematic differences between a fitted Bayesian model and the data^{42,43}. Supplementary Fig. 1 shows the results of these posterior predictive checks. For both the state patrol model and the municipal department model, the prediction errors are minimal and similar across race groups.

Given the inferred race- and location-specific thresholds, $\hat{t}_{r,d}$, we define overall race-specific thresholds \hat{t}_r for both states and municipal departments in two steps. First, for each state (municipal department) g , we compute an average $\hat{t}_{r,g}$ over all its counties (districts), weighting each threshold $\hat{t}_{r,d}$ by the proportion of stops in location d . Next, we calculate \hat{t}_r as the unweighted average of all state thresholds $\hat{t}_{r,g}$ (or municipal thresholds) for race group r . In the second step, we take the unweighted average to account for differences in reporting practices across states, although our results are qualitatively similar under alternative averaging schemes.

Effects of legalization of marijuana. Measures legalizing recreational marijuana took effect on 9 December 2012 in Washington, and on 10 December 2012 in Colorado. In Colorado, an additional Senate bill was passed on 28 May 2013 ('Inferences for Marijuana and Driving Offenses', HB 13-1325) limiting the amount of tetrahydrocannabinol a person can have in their bloodstream to 5 ng ml⁻¹. Below we describe the data-processing choices we made when examining the effects of marijuana legalization, and then provide additional detail to supplement the analyses reported in Results.

When calculating the proportion of stops that resulted in drug-related offences, we excluded all alcohol-related violations or those relating to drug paraphernalia or drug-related felonies. More specifically, we excluded the following violation

ARTICLES

NATURE HUMAN BEHAVIOUR

Table 1 | Effects of legalization of recreational marijuana on search rates, as estimated using a difference-in-difference model. All race groups experienced a large drop in search rate

	Coefficient	s.e.	95% CI	P value
Effect of legalization on white drivers	-0.96	0.02	(-1.01, -0.92)	<0.001
Effect of legalization on black drivers	-0.98	0.06	(-1.09, -0.87)	<0.001
Effect of legalization on Hispanic drivers	-0.69	0.03	(-0.76, -0.63)	<0.001
Time (years)	-0.04	0.00	(-0.04, -0.04)	<0.001
Black driver	0.78	0.00	(0.77, 0.79)	<0.001
Hispanic driver	0.59	0.00	(0.58, 0.59)	<0.001

codes in Washington: 'Drugs paraphernalia—misdemeanor', 'DUI—Drugs W/Test', 'DUI—Drugs No Test', 'Veh Hom—DUI/Drug', 'Veh Assault—DUI/Drug', 'Drugs paraphernalia—felony' and 'Drugs—felony'. And, in Colorado, we excluded the following: 'Drove Vehicle While Under the Influence of Alcohol or Drugs or Both', 'Possession of Drug Paraphernalia' and 'Drove (Motor/Off-highway) Vehicle upon Highway When (License/Privilege to Drive) was Restrained for Express Consent or Alcohol/Drug Related Offense'.

Furthermore, when analysing search rates, we excluded impound, incident to arrest, protective frisk and warrant searches, as these types of search are not usually motivated by drug recovery. In Colorado, this means that we include both consent and probable cause searches. However, in Washington, we note that probable cause searches are barred because the state does not have an automobile exception for warrantless searches. For further context, it is acceptable in most states for police officers to conduct a (limited) warrantless search of a vehicle when officers have probable cause to believe that contraband or other evidence of a crime is in a vehicle, a policy known as the 'automobile exception' to the warrant requirement for searches. Washington, however, does not recognize this exception. Even if an officer has well-founded probable cause to believe that contraband or other evidence of a crime is in a vehicle, the officer still cannot search the vehicle without a warrant and must instead conduct the search pursuant to another exception to the warrant requirement, usually a driver's consent.

As discussed in Results, the proportion of stops resulting in a search fell substantially in both Colorado and Washington, while in 12 states where marijuana was not legalized, search rates did not show sharp drops at the end of 2012. To assess this pattern quantitatively, we fit the following difference-in-differences search model on the set of stops in the 14 states we consider here (Colorado, Washington and the 12 non-legalization states):

$$\Pr(Y_i = 1) = \text{logit}^{-1} \left(\beta_{state}^{state} + \beta_{race}^{race} + \beta_{time}^{time} t_i + \alpha_{r|j}^{r|j} Z_i \right)$$

where Y indicates whether a search was conducted, β_{state}^{state} and β_{race}^{race} are state and race fixed effects and β_{time}^{time} is a time trend, with t a continuous variable in units of years since legalization (for example, $t=0.5$ means 6 months post-legalization). The Z term indicates 'treatment' status—that is, $Z_i = 1$ in Colorado and Washington for stops carried out during the post-legalization period, and $Z_i = 0$ otherwise. Thus the key parameters of interest are the race-specific coefficients $\alpha_{r|j}^{r|j}$. Table 1 lists coefficients for the fitted model. We found that $\alpha_{r|j}^{r|j}$ is large and negative for white, black and Hispanic drivers, suggesting that the observed drop in searches in Colorado and Washington was due to marijuana legalization in those states.

Despite marijuana legalization decreasing search rates for white, black and Hispanic drivers, Fig. 4 shows that the relative disparity between white and non-white drivers remains. In addition, Supplementary Fig. 3 shows that white drivers faced consistently higher search thresholds than black and Hispanic drivers, both before and after marijuana legalization. Supplementary Fig. 3 also shows that the threshold faced by all groups decreases after legalization (although not all drops are statistically significant). There are several possible explanations for this decrease. Officers may not have fully internalized the change of policy, searching people who would have been at risk of carrying contraband before legalization but who were no longer high risk after marijuana became legal. Alternatively, or in addition, officers may have become focused on more serious offences (such as drug trafficking), applying a lower threshold commensurate with the increase in the severity of the suspected crime. Finally, officers may have had more resources after being relieved of the task of policing marijuana possession, freeing them to make searches with a lower chance of finding contraband.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All data to reproduce the findings of this study are available at <https://openpolicing.stanford.edu>.

Code availability

All code to reproduce the findings of this study is available at <https://openpolicing.stanford.edu>.

Received: 23 August 2017; Accepted: 13 March 2020;

Published online: 4 May 2020

References

- Davis, E., Whyde, A. & Langton, L. *Contacts between Police and the Public, 2015* (Bureau of Justice Statistics, 2018); <http://www.bjs.gov/index.cfm?ty=pbdetail&id=6406>
- Langton, L. & Durose, M. *Police Behavior during Traffic and Street Stops, 2011* <https://www.bjs.gov/content/pub/pdf/pbtss11.pdf> (US Department of Justice, 2013).
- Baumgartner, E.R., Epp, D.A. & Shoub, K. *Suspect Citizens: What 20 Million Traffic Stops Tell Us about Policing and Race* (Cambridge Univ. Press, 2018).
- Glaser, J. *Suspect Race: Causes and Consequences of Racial Profiling* (Oxford Univ. Press, 2015).
- Epp, C., Maynard-Moody, S. & Haider-Markel, D. *Pulled Over: How Police Stops Define Race and Citizenship* (Univ. of Chicago Press, 2014).
- Antonovics, K. & Knight, B. A new look at racial profiling: evidence from the Boston Police Department. *Rev. Econ. Stat.* **91**, 163–177 (2009).
- Simolu, C., Corbett-Davies, S. & Goel, S. The problem of infra-marginality in outcome tests for discrimination. *Ann. Appl. Stat.* **11**, 1193–1216 (2017).
- Anwar, S. & Fang, H. An alternative test of racial prejudice in motor vehicle searches: theory and evidence. *Am. Econ. Rev.* **96**, 127–151 (2006).
- Ridgeway, G. & MacDonald, J. Doubly robust internal benchmarking and false discovery rates for detecting racial bias in police stops. *J. Am. Stat. Assoc.* **104**, 661–668 (2009).
- Ridgeway, G. Assessing the effect of race bias in post-traffic stop outcomes using propensity scores. *J. Quant. Criminol.* **22**, 1–29 (2006).
- Ryan, M. Frisky business: race, gender and police activity during traffic stops. *Eur. J. Law Econ.* **41**, 65–83 (2016).
- Rojek, J., Rosenfeld, R. & Decker, S. The influence of driver's race on traffic stops in Missouri. *Police Q.* **7**, 126–147 (2004).
- Smith, M. & Petrocelli, M. Racial profiling? A multivariate analysis of police traffic stop data. *Police Q.* **4**, 4–27 (2001).
- Warren, P., Tomaskovic-Devey, D., Smith, W., Zingraff, M. & Mason, M. Driving while black: bias processes and racial disparity in police stops. *Criminology* **44**, 709–738 (2006).
- Hetey, R., Monin, B., Maitreyi, A. & Eberhardt, J. *Data for Change: A Statistical Analysis of Police Stops, Searches, Handcuffs, and Arrests in Oakland, Calif., 2013–2014* (Stanford Univ., SPARQ, 2016); <https://sparq.stanford.edu/data-for-change>
- Voigt, R. et al. Language from police body camera footage shows racial disparities in officer respect. *Proc. Natl Acad. Sci. USA* **114**, 6521–6526 (2017).
- Chohlas-Wood, A., Goel, S., Shoemaker, A. & Shroff, R. *An Analysis of the Metropolitan Nashville Police Department's Traffic Stop Practices* <https://policylab.stanford.edu/media/nashville-traffic-stops.pdf> (Stanford Computational Policy Lab, 2018).
- Gelman, A., Fagan, J. & Kiss, A. An analysis of the New York City Police Department's 'stop-and-frisk' policy in the context of claims of racial bias. *J. Am. Stat. Assoc.* **102**, 813–823 (2007).
- Baumgartner, E. R., Epp, D. A., Shoub, K. & Love, B. Targeting young men of color for search and arrest during traffic stops: evidence from North Carolina, 2002–2013. *Polit. Groups Identities* **5**, 107–131 (2017).
- Legewie, J. Racial profiling and use of force in police stops: how local events trigger periods of increased discrimination. *Am. J. Sociol.* **122**, 379–424 (2016).
- Grogger, J. & Ridgeway, G. Testing for racial profiling in traffic stops from behind a veil of darkness. *J. Am. Stat. Assoc.* **101**, 878–887 (2006).
- Pierson, E., Corbett-Davies, S. & Goel, S. In *International Conference on Artificial Intelligence and Statistics* (eds Storkey, A. & Fernando Perez-Cruz, F.) 96–105 (PMLR, 2018).
- Becker, G. Nobel Lecture: the economic way of looking at behavior. *J. Polit. Econ.* **101**, 385–409 (1993).
- Becker, G. *The Economics of Discrimination* (Univ. of Chicago Press, 1957).
- Ridgeway, G. *Cincinnati Police Department Traffic Stops: Applying RAND's Framework to Analyze Racial Disparities* (RAND Corporation, 2009); <https://www.rand.org/pubs/monographs/MG914.html>
- Ayres, I. Outcome tests of racial disparities in police practices. *Justice Res. Policy* **4**, 131–142 (2002).
- Knowles, J., Persico, N. & Todd, P. Racial bias in motor vehicle searches: theory and evidence. *J. Polit. Econ.* **109**, 203–229 (2001).

28. Engel, R. S. & Tillyer, R. Searching for equilibrium: the tenuous nature of the outcome test. *Justice Q.* **25**, 54–71 (2008).
29. Mitchell, O. & Caudy, M. Examining racial disparities in drug arrests. *Justice Q.* **32**, 288–313 (2015).
30. Angrist, J.D. & Pischke, J.-S. *Mostly Harmless Econometrics: An Empiricist's Companion* (Princeton Univ. Press, 2008).
31. Goel, S., Rao, J. & Shroff, R. Precinct or prejudice? Understanding racial disparities in New York City's stop-and-frisk policy. *Ann. Appl. Stat.* **10**, 365–394 (2016).
32. Mummolo, J. Modern police tactics, police–citizen interactions, and the prospects for reform. *J. Polit.* **80**, 1–15 (2018).
33. Goel, S., Rao, J. & Shroff, R. Personalized risk assessments in the criminal justice system. *Am. Econ. Rev.* **106**, 119–123 (2016).
34. Jung, J., Concannon, C., Shroff, R., Goel, S. & Goldstein, D.G. Simple rules to guide expert classifications. *J. R. Stat. Soc. Ser. A* (in the press).
35. Corbett-Davies, S., Pierson, E., Feller, A., Goel, S. & Huq, A. Algorithmic decision making and the cost of fairness. In *Proc. 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2017).
36. Friberg, B. et al. *Texas Troopers Ticketing Hispanic Drivers as White* (KXAN, 2015); <https://www.kxan.com/investigations/texas-troopers-ticketing-hispanic-drivers-as-white/>
37. Word, D., Coleman, C., Nunziata, R. & Kominski, R. *Demographic Aspects of Surnames from Census 2000* <http://www2.census.gov/topics/genealogy/2000surnames/surnames.pdf> (US Census Bureau, 2008).
38. Word, D. & Perkins, C. *Building a Spanish Surname List for the 1990s: A New Approach to an Old Problem* (Population Division, US Census Bureau, 1996).
39. *Melendres v Arpaio* (2009). 598 F. Supp. 2d 1025 (D. Ariz., 2009).
40. Arrow, K. in *Discrimination in Labor Markets* (eds Ashenfelter, O. & Rees, A.) 3–33 (Princeton Univ. Press, 1973).
41. Phelps, E. The statistical theory of racism and sexism. *Am. Econ. Rev.* **62**, 659–661 (1972).
42. Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B. *Bayesian Data Analysis* 2nd edn (Taylor & Francis, 2014).
43. Gelman, A., Meng, X.-L. & Stern, H. Posterior predictive assessment of model fitness via realized discrepancies. *Stat. Sin.* **6**, 733–760 (1996).

Acknowledgements

We thank B. Bonilla, W. Kim, J. Nudell, S. Robertson and E. Sagara for their assistance throughout this project; we also thank A. Chohlas-Wood and A. Feller for their helpful feedback. This work was supported in part by the John S. and James L. Knight Foundation and by the Hellman Foundation. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

E.P., C.S., J.O., S.C.-D., D.J., A.S., V.R., P.B., C.P., R.S. and S.G. designed research, performed research, analyzed data, and wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41562-020-0858-1>.

Correspondence and requests for materials should be addressed to S.G.

Peer review information Primary handling editor: Aisha Bradshaw

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- | | |
|-----|-----------|
| n/a | Confirmed |
|-----|-----------|
- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
 - ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
 - ☐ ☒ The statistical test(s) used AND whether they are one or two sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
 - ☐ ☒ A description of all covariates tested
 - ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
 - ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
 - ☐ ☒ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
 - ☐ ☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
 - ☐ ☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
 - ☐ ☒ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection All code to reproduce the findings of this study is available at: <https://openpolicing.stanford.edu>.

Data analysis All code to reproduce the findings of this study is available at: <https://openpolicing.stanford.edu>.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data to reproduce the findings of this study are available at: <https://openpolicing.stanford.edu>.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- ☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	A quantitative study of traffic stop data.
Research sample	We analyzed data on traffic stops from 21 state patrol agencies and 35 municipal police departments in the United States.
Sampling strategy	We used all available data that were sufficiently complete to conduct our analysis.
Data collection	We collected the data in digital form through public records requests filed with the appropriate police agencies.
Timing	We continually collected data during the period 2014 -2019.
Data exclusions	We did not exclude any data that were deemed sufficient to conduct our analysis.
Non participation	N/A
Randomization	N/A

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging