Imperfect Inferences: A Practical Assessment

Aaron Rieke
Upturn

Vincent Southerland
New York University School of Law

Dan Svirsky
Uber

Mingwei Hsu
Upturn

ABSTRACT

Measuring racial disparities is challenging, especially when demographic labels are unavailable. Recently, some researchers and advocates have argued that companies should infer race and other demographic factors to help them understand and address discrimination. Others have been more skeptical, emphasizing the inaccuracy of racial inferences, critiquing the conceptualization of demographic categories themselves, and arguing that the use of demographic data might encourage algorithmic tweaks where more radical interventions are needed.

We conduct a novel empirical analysis that informs this debate, using a dataset of self-reported demographic information provided by users of the ride-hailing service Uber who consented to share this information for research purposes. As a threshold matter, we show how this data reflects the enduring power of racism in society. We find differences by race across a range of outcomes. For example, among self-reported African-American riders, we see racial differences on factors from iOS use to local pollution levels.

We then turn to a practical assessment of racial inference methodologies and offer two key findings. First, every inference method we tested has significant errors, miscategorizing people relative to their self-reports (even as the self-reports themselves suffer from selection bias). Second, and most importantly, we found that the inference methods worked: they reliably confirmed directional racial disparities that we knew were reflected in our dataset.

Our analysis also suggests that the choice of inference methods should be informed by the measurement task. For example, disparities that are geographic in nature might be best captured by inferences that rely on geography; discrimination based on a person's name might be best detected by inferences that rely on names.

In conclusion, our analysis shows that common racial inference methods have real and practical utility in shedding light on aggregate, directional disparities, despite their imperfections. While the recent literature has identified notable challenges regarding the collection and use of this data, these challenges should not be seen as dispositive.

CCS CONCEPTS

• **Social and professional topics → Race and ethnicity**; *Gender*.

KEYWORDS

inference, demographics, race, discrimination, civil rights, fairness

ACM Reference Format:
Aaron Rieke, Vincent Southerland, Dan Svirsky, and Mingwei Hsu. 2022. Imperfect Inferences: A Practical Assessment. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3531146.3533140

1 INTRODUCTION

Statistical models, especially those built with machine learning, can both reflect and exacerbate discrimination along demographic lines [26]. Discriminatory effects can occur even when demographic features are not used to train or employed as inputs to a model [36]. In recent years, these effects have been observed and debated in a variety of domains including healthcare, [33] credit, [40] employment, [16] and digital advertising [5].

In response, a growing number of advocates have urged technology companies to proactively measure and remediate discrimination — especially racial discrimination — using demographic data. For example, Color of Change, the United States' "largest racial justice organization," has asked technology companies to "measure racial and demographic differences regarding user experience of all products. [34]" Major civil rights groups endorsed a plan by Airbnb to measure racial discrimination on its platform by collecting demographic data about its users' race [3]. The NAACP Legal Defense and Educational Fund announced it is working with fintech lender Upstart to measure and remediate discrimination in its underwriting models [38]. And an independent civil rights audit of Meta (previously Facebook), which incorporated the feedback and perspectives of many civil rights advocates, urged the company to adopt strategies to assess its products for bias using demographic data [1].

These efforts are difficult when demographic labels, such as those for race and gender, are unavailable in relevant datasets. This is typically the case outside of limited domains such as employment, healthcare, and (occasionally) financial services [27]. Many practitioners are hesitant to collect demographic data, and there is little clear guidance on the topic [27, 28].

Nonetheless, a growing number of companies are turning to inferential methods in their efforts to measure discrimination, even in the absence of clear legal or organizational guidance [28]. For example, in 2020, Airbnb announced it would measure discrimination on its platform, relying on class labels inferred by a third-party partner (subject to stringent privacy protections) [6]. LinkedIn began inferring its users' gender in 2018, in order to show employers searching for new employees more representative search results [32]. Uber partnered with academic economists to study its platform,

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06.
https://doi.org/10.1145/3531146.3533140

sometimes using demographic inferences to understand socially important outcomes, like who benefits from tipping [8] or flexible work arrangements [21]. Recently, Meta announced a new race measurement program using inference methods [41]. And in the authors' experience, many more companies besides are actively exploring — or quietly using — racial inferences to assess their products and services.

Still, using proxies for unobserved protected class labels to measure discrimination remains controversial, and direct empirical research is limited. A recent meta-analysis, published by the Partnership on AI, characterized inference based on proxy information "largely inaccurate. [35]" McKane et. al. summarize concerns about the collection of demographic data, noting that it "raises a host of difficult questions, including how to balance privacy and fairness, how to define relevant social categories, how to ensure meaningful consent, and whether it is appropriate for private companies to infer someone's demographics. [27]"

We conduct a novel empirical analysis that informs this debate. Using a dataset of self-reported demographic information, provided by Uber users who consented to share this information for research purposes, we track a range of racial disparities and conduct a comparative analysis of racial inference methodologies. We conclude that common racial inference methods have real, practical utility in shedding light on aggregate, directional disparities. While the recent literature has identified real and important challenges regarding the collection and use of this data, we believe these challenges should not be dispositive.

2 RELATED WORK

There has been substantial scholarship in this space from technical, practical, policy, and critical perspectives. We briefly summarize each category below.

Quantitative researchers have conducted a range of analyses on the accuracy of racial inference methods. For example, Chen et. al. found that, under some conditions, Bayesian Improved Surname Geocoding (BISG), is more likely to overestimate demographic disparities than to underestimate them. The authors found that without ground truth labels, it is not possible to determine the optimal choice for a threshold and urged caution when using BISG with a threshold estimator method [25]. BISG may also lead to underestimation of demographic disparities given certain conditions. Baines and Courchane tested various thresholds for BISG and found that at an 80% threshold, "BISG fails to identify 3 out of 4 African American applicants and 4 out of 10 Hispanic applicants" in the auto lending context [9]. Additionally, Baines and Courchane found that "[f]alse negative rates are highest in tracts with the lowest shares of the group in question. For example, in tracts that are less than 10% African American, BISG at an 80% probability threshold fails to identify 98% of the actual African American applicants in such tracts." In testing for estimation bias, Chen et. al. found that using a weighted estimator for BISG underestimates demographic disparity. Expanding this research, Kallus et. al demonstrated that "it is generally impossible to identify impact disparities when only proxy information is available for protected class" in algorithms such as BISG. Kallus et. al. proposed methods that quantify "the fundamental ambiguity in disparities and the value of more informative proxies or assumptions" which allow for "credible, principled conclusions about disparities. [30]"

Dzifa Adjaye Gbewonyo et al. found that BISG appeared to perform better in older age groups, as well as men. The researchers advised that it is crucial to account for the age of the population when applying the BISG algorithm [17]. Researchers at Meta confirmed that BISG was more accurate for older individuals in their own validation tests and noted that Census surname and ZIP Code level data from the 2010 decennial census survey is "likely stale. [39]"

Ghosh et al. cautioned against using inferred demographics to try to intervene in the context of ranking algorithms, noting that errors in race inference can "dramatically undermine fair ranking algorithms," producing rankings that are closer to the unfair baseline than the optimal fair ranking. Using five different demographic inference algorithms, the authors found that the impact of using inferred demographic data as input to fair ranking algorithms are difficult to predict and often harm vulnerable groups [7].

From a practical and policy perspective, Andrus et. al. interviewed professionals either working in, or adjacent to, algorithmic fairness and analyzed how practitioners confront issues around demographic data procurement [27]. Bogen et. al. surveyed U.S. anti-discrimination law on the collection and usage of sensitive attribute data in employment, credit, and healthcare, and concluded that there are few consistent principles about how and why companies should collect demographic data [28].

Finally, many in the algorithmic fairness community have employed a critical perspective to the concept of demographic attributes themselves. For example, Benthall et. al. and Hanna et. al. have discussed race [4, 13], Hamidi et. al., Hu et. al., and Scheuerman et. al. gender [18, 24, 29, 44], and Bennett et. al. disability [12]. These authors critique the very notion of using these categories as a basis for assessing unfairness, and the potential harms relying on these categories might create [4].

3 RACE AS A SOCIAL CONSTRUCT

Before embarking on an empirical analysis of the accuracy or utility of racial proxies, it is critical to acknowledge that race is a social construct [15]. Because "the dimensions of racism transform over time as the political, legal, and social context change, it may not be possible to design a specific measure to capture its effects, [43]" much less speak with clarity about what an "accurate" racial inference even means. These issues are becoming even more fraught as racial labels are "falling behind the growing diversity within each racial and ethnic group and failing to capture mixed-race people. [2]"

However, these conceptual challenges cannot be the end of the conversation. Despite the constructed, artificial nature of race, race-based inequality has shaped society and institutions in ways that are undeniably real and deserving of ongoing attention. Race remains the "miner's canary," alerting us to "danger[s] that threaten us all. [22]" It operates as a signal of rank in America's caste system, which is itself "an artificial construction, a fixed and embedded ranking of human value that sets the presumed supremacy of one group against the presumed inferiority of other groups on the basis of ancestry and often immutable traits. . . . [46]" The enduring and endemic nature of racism itself [40], and its place as "an integral, permanent, and indestructible component of this society, [11]" demands that we work to address its implications by any and all means at our disposal.

In short, although racial labels and concepts are fraught, they are what we have today, and they will be with us for the foreseeable future. We cannot shy away from tackling race head on, but we are left with imperfect choices about how to do so.

These tensions are at the core of this paper. At the outset of our drafting, we had to make a series of sometimes uncomfortable decisions in our analysis, especially around wording, the demographic categories we studied, and what gets left out.

This paper is motivated by the continuing problem of racial disparities in the United States. (There is important future work to be done focused on other categories, like sexual orientation, immigration status, neurotypicality, and gender identity, to name just a few.) The survey data that informs our analysis asks users to self-categorize their "race/ethnicity." But these are two distinct ideas, with race referring to artificially constructed categories that largely track skin color, while ethnicity more commonly aligns with a person's culture. In this paper, we generally use "race", rather than "race/ethnicity", since that was the wording used in the survey materials. We use the term 'Hispanic,' to refer to people from or with ancestry from Latin American countries who now live in the United States. We choose this word, rather than Latine or Latinx, as this was the wording in the survey question offered to users. Our analysis also elides nuances and problems with the race category. We do not look at intersections of race and gender. We do not examine people who fall into more than one race category, or who do not fit neatly into any of the broad categories collected in the survey. We examine a person's self-reported race and largely treat it as a fixed feature, even though people's self-definitions can change over time.

4 METHODS

To conduct our comparative analysis of racial inference methods, we joined three datasets: self-reported data on race, inferred data on race, and various outcomes of interest. By joining the first two datasets, we can assess how well inferred race matched with self-reported race. By layering on the outcomes of interest, we can then assess whether disparities measured by self-reported racial labels are similarly measured by inferred labels.

The *self-reported data* comes from survey research that Uber conducted in 2019. The target population of the survey was active rideshare users in the United States, defined as those who took a trip with Uber in the 84 days leading up to the survey. The surveys were conducted on a weekly basis. A representative sample of 1,800 riders were selected from the total sample of active Uber riders for each weekly survey. We focus on a sample of 12,764 responses, evenly balanced among four racial categories: non-Hispanic white, non-Hispanic African-American, Hispanic, and Asian or Asian American. (Note: We use these categories because they track the language in the survey question and because these are the categories used by the inferences we test. This leaves out two categories that were collected in the survey: "Native American / Alaska Native" and "Other.") To recruit users, active riders were contacted via an emailed link and offered a $10 gift card to participate in a survey. The survey took roughly 10 minutes to complete. Potential respondents were informed that the survey responses would be used for research purposes and that the information provided "will not be used to try to sell you anything or for any other marketing purposes." All collected data was stored in encrypted ZIP files on local computers separated from Uber's internal data tables, so that data scientists who did not work with this survey data were not able to access it without pre-authorization from in-house counsel who focus on privacy compliance. Any use had to be in line with Uber's Privacy Policy and Terms and Conditions. Before publishing this paper, we were advised by New York University's institutional review board that our analyses do not fall under the purview of an IRB or human subject research because it involves secondary use of de-identified data without any practical way for outside researchers to re-identify the subjects.

The *inferences data* measures the extent to which a user's individual traits are associated with different races. For example, researchers can look at birth certificate data to assess the proportion of people with a given first and/or last name who identify with a specific racial group. If, in a sample of birth certificates, $X_{\text{name,Asian}}\%$ of babies named Jacquelyn were of Asian ancestry, $Y_{\text{name,Hispanic}}\%$ were of Hispanic ancestry, and so on, then the values $X_{\text{name,Asian}}\%$ and $Y_{\text{name,Hispanic}}\%$ could be attributed to all users with the first name Jacquelyn. We can use United States Census data on the demographics of different census tracts to conduct a similar analysis. If a user takes trips that start in census tracts where, on average, $X_{\text{geo,Asian}}\%$ of residents are of Asian ancestry and $Y_{\text{geo,Hispanic}}\%$ are of Hispanic ancestry, then the numbers $X_{\text{geo,Asian}}\%$ and $Y_{\text{geo,Hispanic}}\%$ are assigned to all users who take trips in those census tracts.

We used several name- and geography-based approaches in our analysis: Ethnicolr, Namsor, BIFSG, and a custom "Trips measure" which uses the Census demographics of the pickup and drop off destinations of the user's Uber trips. The inference approaches make different methodological choices in implementation. Here, we describe how they work at a high level but point the reader to underlying literature and source code to get a deeper understanding, where available. Ethnicolr predicts race from the sequence of characters in a name, relying on Florida voter registration data to train a Long Short Term Memory Network [20]. Namsor is a proprietary model that also uses machine learning to associate names with race [31]. BIFSG is an improved version of BISG, the oldest, most-tested inference method [45]. BISG was developed by the RAND Corporation and has been used by federal agencies like the Centers for Medicare and Medicaid Services and the Consumer Financial Protection Bureau [14]. BIFSG relies on birth certificate data to calculate the probability of being in a racial category given one's first name, surname, and location, and then performs a Bayesian combination of these three probabilities. Finally, the Trips methodology is the coarsest, inferring race through an analysis of patterns of racial segregation by imputing a category based on the most common race category across the trip pickup and locations that the user took in 2019. For example, suppose a user who opted into this research took two Uber trips in 2019. This corresponds to four locations: two pickups and two drop offs. We can use the demographics of the Census tracts of these four locations to calculate the average proportion of Asian residents, white residents, African-American residents, and Hispanic residents in each location. Whichever proportion is highest is then assigned to that user.

When we applied these inference methods, we did not retain any link between the demographic prediction and individual users' data, including location data, which in practice is often unique enough to be identifiable [47]. Our analyses relied only on aggregated distributions, from which there was no practical way to reidentify individual users.

For the *outcomes data*, we aimed to include a diversity of outcomes that were readily available at Uber to illustrate the relevance of demographic measurement. The goal was to be broad in scope. We are interested in how well demographic inferences recover underlying differences between groups, so we want to make sure we used a host of relevant outcome variables. Appendix Table 1 is a Data Dictionary with a description of each outcome. These include, but are not limited to, the number of Uber trips taken in 2019, the proportion of trip requests where a driver canceled the trip, the proportion of trips on weekend nights, and the proportion of trips taken to an airport. We also use public data sources to enrich the outcomes data. We merge in CDC data on the average level of Fine Particulate Matter (PM 2.5) in the census tracts of the user's pickup locations [19] and data from the University of Richmond's Mapping Inequality Project on the proportion of trips in a historically redlined area [42].

5 ANALYSES

This section describes the results of the empirical analyses and presents three findings. First, race categories are associated with significant variance in nearly every outcome we measured. This is an unsurprising finding, but still an important illustration of racism's effects in American society. It also shows how even innocuous decisions, like designing an app with iOS users in mind, can have racial dimensions, since different racial groups use iOS at different rates. Second, the racial inferences are, at best, imperfect but roughly accurate. Third, despite imperfect accuracy, the inferences we tested did a surprisingly good job of uncovering racial disparities in the outcomes data. That is, when we do see a racial disparity in outcomes, we still find such a disparity when we rely on inferred demographics instead of self-reported ones.

5.1 Summary Statistics on Racial Disparities

Table 1 presents summary statistics of the data set, focusing on the outcomes data, broken down by self-reported race. These summary statistics are as close as we can get to a ground truth measure of disparity, even as self-reports of race themselves suffer from sample selection bias and non-response. (These problems are of first-order importance, to the point that in some cases, inferring demographics may be a superior approach. If few people respond to self-identification prompts, then discrimination testing might be impossible. Similarly, if enough people respond, but only those who don't typically face discrimination, then again, relying on self-reporting will lead the tester astray.) With few exceptions, there are measurable, significant disparities in almost every outcome — geographic measures, public dataset outcomes, user behaviors from Uber's data, and so on. For example, Hispanic users take trips in areas with higher pollution levels. African-American users are less likely to request a trip from the iOS operating system. Hispanic users are more likely to take trips on weekend nights. Asian users are more likely to take trips in areas with a high density of restaurants.

These disparities are unsurprising, but illustrate two important points: race matters, and even innocuous decisions may have racial dimensions.

First, the table gives empirical backing to the arguments that have long been made by legal theorists who study race. When Derrick Bell, a leading scholar on race and the law, argued that racial disparities were ordinary, not an aberration, he drew on theoretical arguments and his own observations of American history and society [11]. But if one were looking for real-world data to support his arguments, the mix of disparities in Table 1 are exactly what one would expect.

Second, the results show that when a technology company makes even seemingly innocuous decisions, there is cause to consider racial impacts. For example, a software development team that prioritizes pushing out an update for an Apple iOS may have made this choice for thoughtful, defensible reasons, but the choice has racial dimensions: it will serve more white people than African-American people in the sample as a whole. A message targeted at weekend night users is more likely to be seen by self-reported Hispanic riders.

5.2 Comparative Accuracy of Inference Methods

Next, we turn to a simple measurement of how each inference method performs in predicting the self-reports. Figure 1 gives a visualization which depicts the accuracy of comparing the self-report data to the inference data for each of the four inference methods. The left side shows the proportion of users in each of four possible self-reported race categories. The right side shows the proportion of users in each of four possible inferred race categories. The bars between the sides show how many individuals in each self-reported category end up in each of the four inferred categories.

The visualization demonstrates several important facts. First, there are significant inaccuracies for all four inference methods. Second, inference accuracy varies by subgroup. For the three non-white categories, the inferences typically err by mistakenly categorizing the user as white. This happens most for self-reported African-American users, but also for Hispanic and Asian users. Third, the inferences err in different ways. BIFSG and Ethnicolr both make extensive errors for self-reported African-American users, misclassifying many as white. Namsor and the Trips inference make this mistake less often. In general, all three inferences which use name information classify more people as white than in the underlying data, but the Trips inference does not commit this mistake to the same degree.

5.3 Assessing Inferences' Ability to Identify Disparities

Racial inferences are bound to be imperfect. But are they useful for detecting disparities in real-world scenarios? Or does using

Table 1: Summary Statistics

	All Users	African American	Asian	Hispanic	White
Number of Respondents	12,764	2,999	2,999	2,999	2,999
More than 30 trips/year	0.61 (0.004)	0.66*** (0.009)	0.69*** (0.008)	0.6 (0.009)	0.51*** (0.009)
Prop'n of Trips with a Driver Cancel	0.0043 (0.0002)	0.004 (0.0004)	0.0034 (0.0003)	0.0045 (0.0005)	0.0046 (0.0005)
Prop'n of Trips Starting or Ending in Historically Redlined Neighborhoods	0.10 (0.001)	0.11*** (0.003)	0.09 (0.003)	0.09 (0.003)	0.09 (0.003)
Average PM 2.5 Levels in Pickup Census Tracts	9.18 (0.026)	9.13 (0.065)	9.04*** (0.045)	9.71*** (0.055)	8.88*** (0.052)
Trip requested from iOS	0.75 (0.004)	0.71*** (0.008)	0.75 (0.008)	0.76 (0.008)	0.80*** (0.008)
Prop'n of Trips on Weekend Nights	0.14 (0.002)	0.12*** (0.003)	0.12*** (0.003)	0.15*** (0.004)	0.16*** (0.004)
Prop'n of Trips During Commuting Hours	0.22 (0.002)	0.23 (0.004)	0.24*** (0.004)	0.22 (0.004)	0.21*** (0.004)
Prop'n of Trips in City Core	0.29 (0.003)	0.28 (0.006)	0.33*** (0.006)	0.27 (0.006)	0.28 (0.006)
Prop'n of Trips that are Airport Trips	0.17 (0.002)	0.13*** (0.004)	0.22*** (0.005)	0.12*** (0.004)	0.21*** (0.005)
Prop'n of Trips in Area above the 90th percentile of People Age 18 - 24	0.09 (0.002)	0.08 (0.003)	0.11*** (0.004)	0.06*** (0.003)	0.09 (0.004)
Prop'n of Trips in Areas with Above-Median Level of Uber Eats Businesses	0.06 (0.001)	0.05*** (0.002)	0.08*** (0.003)	0.05*** (0.002)	0.06 (0.003)

Note: The four self-reported races shown – African American, Asian, Hispanic, and White – are the most common in the data. The category "All Users" includes respondents who did not answer the question, answered "Other", or answered "Native American or Alaskan Native". All data looks at a user's Uber trips in 2019. *** indicates p < 0.01 for a comparison between the group in question versus all users excluding that group. Standard errors are given in parentheses alongside means. The Appendix contains a data dictionary with variable definitions.

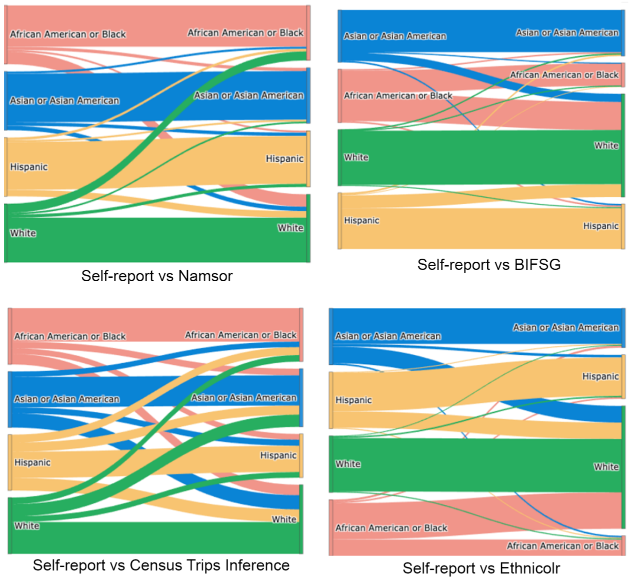


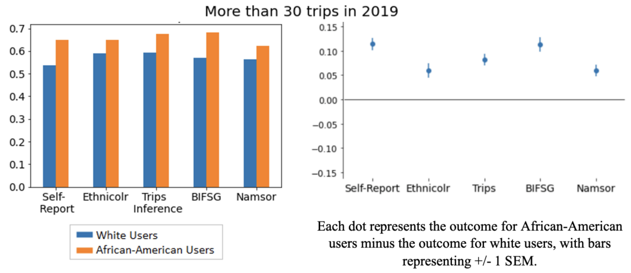
Figure 1: Each Sankey graph depicts the accuracy of a demographic inference method versus the self-reported race data among four groups: African American or Black, Asian or Asian American, Hispanic, and White. From top left, moving clockwise: Self-report vs Namsor; Self-report vs BIFSG; Self-report vs Census Trips Inference; Self-report vs Ethnicolr.

these imperfect inferences risk fooling us into thinking there is no disparity when one does exist (or show a disparity that doesn't exist)?

To answer this question, we compare the self-reported and inferred racial labels to measure differences among five outcomes: whether the user took more than 30 trips in 2019; the proportion of trips where the driver canceled; the proportion of trips that started or ended in a historically redlined neighborhood; the proportion of trips in areas with an above-median number of Uber Eats restaurants; and the proportion of trips taken during weekend nights. Figures 2-6 show the results. For each outcome, we depict two graphs. The left panels contain bar graphs showing outcomes by race (focusing on white users and African-American users), using different measures of race: the self-reports and each of the four different inference methods. The panels on the right contain graphs depicting the difference in each outcome – African-American minus white users – using the five different measures of race. While we focus on these five variables and these two race categories for simplicity, our Appendix shows that the same story holds when looking at a host of other outcomes, and the same story holds when we compare different racial groups.

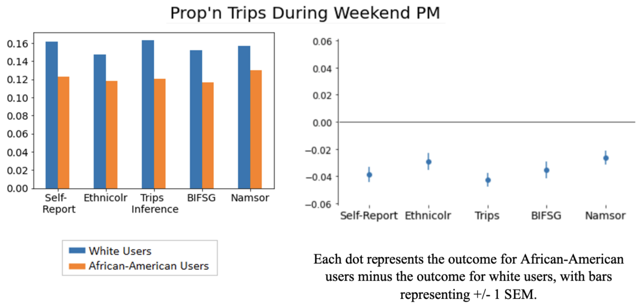
In short, the inferences "work," at least as applied to these measurement tasks. In other words, despite the inaccuracies described in 5.2 (and other literature), the observer's ability to detect a disparity when one exists among self-reported racial groups generally remains consistent. When there exists a racial disparity in the underlying data, we can reliably detect it with inferred race labels, at least directionally (if not always with the same magnitude).

Going further, there is even some evidence that inferred categories might, in some sense, work better than self-reported ones. Consider trips in historically red-lined areas. Even after nearly a century, areas that were redlined continue to exhibit the impact



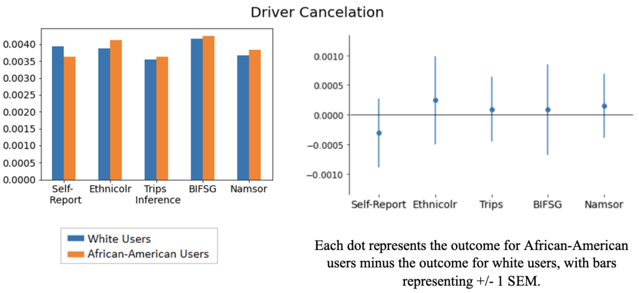
Each dot represents the outcome for African-American users minus the outcome for white users, with bars representing +/- 1 SEM.

Figure 2: Demographic comparison of users with more than 30 trips in 2019.



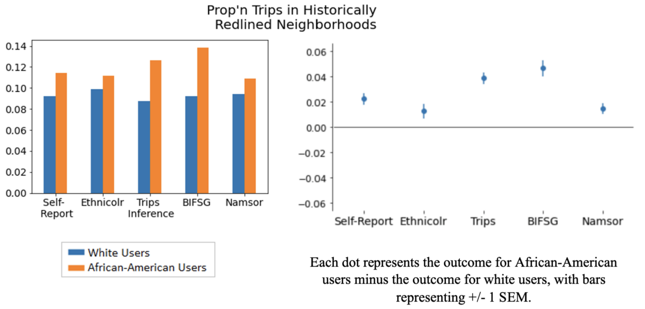
Each dot represents the outcome for African-American users minus the outcome for white users, with bars representing +/- 1 SEM.

Figure 3: Demographic comparison of users based on proportion of trips during weekend evenings.



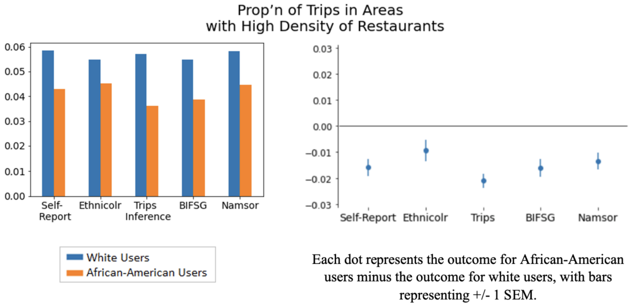
Each dot represents the outcome for African-American users minus the outcome for white users, with bars representing +/- 1 SEM.

Figure 4: Demographic comparison of users based on driver cancellation.



Each dot represents the outcome for African-American users minus the outcome for white users, with bars representing +/- 1 SEM.

Figure 5: Demographic comparisons for users based on proportion of trips in historically redlined neighborhoods.



Each dot represents the outcome for African-American users minus the outcome for white users, with bars representing +/- 1 SEM.

Figure 6: Demographic comparisons for users based on proportion of trips in areas with high density of restaraunts.

of the policy, with higher pollution levels [23] and lower property values [37]. As such, people who live and move in these areas could be considered as being more exposed to historically racist policies. In our data, we indeed find that African-American users are more likely to take trips in these areas, but this difference is higher among the two inference methods that rely on geography to infer race, while lower among the inference methods that only rely on name. Similarly, consider driver cancellations. Here, we see slightly lower driver cancel rates against African-American riders by self-reported race. But cancel rates are slightly higher – though still not statistically significantly higher – among riders classified as African-American by the inferences. Why? One possibility is that drivers cancel trips using a rider's name (which can be seen after accepting a trip) or pickup location, which might correlate with race. It's possible that some drivers might rely on the same signals to infer a rider's race as these inferences do. If these assumptions are true, then the race inferences would be more accurate than self-reported race for identifying this kind of outcome disparity, by virtue of relying on name or location as a proxy.

This suggests that, when measuring disparities, the choice of inference or racial proxy might be an important part of the analysis design. If one wants to measure disparities, one should have in mind what might be driving those disparities, or at least be willing to speculate and explore a range of potential causes. If the discrimination is spatial in nature, a geography-based inference might be better suited. If the discrimination is based on a user's photograph, then it might make more sense to infer a user's demographic category based on their photograph, as Airbnb's discrimination research team does. In short, the goal is measuring disparities as they occur, and by the phenomena that cause them, in the real world.

6 DISCUSSION

Recent literature has emphasized the fraught nature of demographic measurement along legal, social, and practical dimensions. We acknowledge these challenges. However, our analyses suggest that practitioners need not throw up their hands. To borrow a phrase from James Baldwin, "[n]ot everything that is faced can be changed;

but nothing can be changed until it is faced. [10]" The central finding of our research is that racial inference methods may be up to the task of aiding this important work.

In closing, we offer the following reflections:

First, as we discussed above, racial labels — uncertain, outdated, and crude as they might be — still undergird the civil rights laws and policies that practitioners leverage to advance racial justice. American courts and other institutions are actively relying on racial inferences and labels to make decisions that affect people's lives. While it is critical that we do not become limited or trapped by socially constructed labels, the enduring power of racism demands work to address its implications. Given this reality, we believe demographic measurement still has an important and practical place.

Second, we believe that the distinction between measurement and intervention is critical. Much of the FAccT community's work has focused on fairness interventions. However, we purposefully address only measurement in this paper, because (1) it is a necessary condition for any reliable intervention, and (2) we believe measurement and interventions should be informed by separate standards and debates. Measurement can be useful even if it is imperfect. It can provide organizations with valuable and reliable information about directional disparities. It should prompt "an interrogation of the full decision-making pipeline. [27]" It can, and should, be the basis of a multidisciplinary and inclusionary analysis. Intervention, on the other hand, demands more empirical and normative justification. We hope that by distinguishing these two concepts, organizations are encouraged to "start the work" of recognizing racial disparities, without having to first decide about what automated fairness interventions might ultimately be appropriate.

Third, practitioners should not attempt to find a single, ideal inference method. There is no such thing. Our research suggests that different proxies are likely to be more or less suited to different kinds of measurement tasks. And, as we've demonstrated, it may be valuable to assess several inference methods simultaneously and learn about their strengths and weaknesses for a particular scenario.

Finally, companies should be held to the highest standards when undertaking demographic measurement. At minimum, they should engage directly with affected individuals and other stakeholders, including civil rights organizations, social scientists, and experts on demographic characteristics, like race. The questions of what to measure, under what circumstances, and how those results will be shared with the public, are just as important as the questions of theoretical or technical feasibility that we address here.

7 FUTURE RESEARCH

We see many useful directions for future research and hope to see more work on these issues, as they will not go away by themselves.

First, we hope for more work that challenges and analyzes the categories themselves. We focused on simple categories, but future work could look at intersections between ethnicity and race, or gender and race. A Hispanic woman with dark skin is likely to face a different type of discrimination than a Hispanic man with light skin. More research needs to be done to understand how to confront this challenge.

Second, recent Census demographics show another shortcoming with these categories: the meaning of them changes over time. Future research can look at how these categories change on a societal level, as well as how people's self-definitions might change in different contexts. This is especially important as more people identify as being in multiple racial and ethnic categories.

Finally, this project largely leaves untouched other groups that face significant discrimination. Here, we conclude that inference can be a useful way to measure racial disparities. There are significant practical and normative questions about other kinds of categories, such as disability and sexual orientation, which we do not consider here, but which do merit further study.

REFERENCES

[1] 2020. Facebook's Civil Rights Audit - Final Report. https://about.fb.com/wp-content/uploads/2020/07/Civil-Rights-Audit-Final-Report.pdf

[2] Janet Adamy and Paul Overberg. 2020. The Census Predicament: Counting Americans by Race. *The Wall Street Journal* (2020). https://www.wsj.com/articles/the-predicament-of-counting-americans-by-race-11606492632

[3] Airbnb. 2020. A new way we're fighting discrimination on Airbnb. https://www.airbnb.com/resources/hosting-homes/a/a-new-way-were-fighting-discrimination-on-airbnb-201

[4] Andrew Smart Alex Hanna, Emily Denton and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. *In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (2020).

[5] Sapiezynski P. Bogen M. Korolova A. Mislove A. Ali, M. and A. Rieke. 2019. Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes. *Proceedings of the ACM on Human-Computer Interaction* (2019).

[6] Airbnb anti-discrimination team. 2020. Measuring discrepancies in Airbnb guest acceptance rates using anonymized demographic data. https://news.airbnb.com/wp-content/uploads/sites/4/2020/06/Project-Lighthouse-Airbnb-2020-06-12.pdf

[7] Ritam Dutt Avijit Ghosh and Christo Wilson. 2021. When Fair Ranking Meets Uncertain Inference. *44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2021). https://ai.facebook.com/research/publications/how-meta-is-working-to-assess-fairness-in-relation-to-race-in-the-us-across-its-products-and-systems

[8] J. A. List I. Muir B. Chandar, U. Gneezy. 2019. The Drivers of Social Preferences: Evidence from a Nationwide Tipping Field Experiment. *National Bureau of Economic Research, Working Paper 26380* (2019).

[9] Arthur P. Baines and Dr. Marsha J. Courchane. 2014. Fair Lending: Implications for the Indirect Auto Finance Market. *American Financial Services Association* (2014). https://media.crai.com/sites/default/files/publications/Fair-Lending-Implications-for-the-Indirect-Auto-Finance-Market.pdf

[10] James Baldwin. 1962. As Much Truth as One Can Bear. (1962).

[11] Derrick Bell. 1992. Faces at the Bottom of the Well: The Permanence Of Racism. (1992).

[12] Cynthia L. Bennett and Os Keyes. 2019. What Is the Point of Fairness?: Disability, AI and the Complexity of Justice. (2019). https://arxiv.org/abs/1908.01024

[13] Sebastian Benthall and Bruce D. Haynes. 2019. Racial Categories in Machine Learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19* (2019). https://doi.org/10.1145/3287560.3287575

[14] Consumer Financial Protection Bureau. 2014. Using publicly available information to proxy for unidentified race and ethnicity. (2014). https://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf

[15] Devon W. Carbado. 2011. Critical What What? *43 Conn. L. Rev. 1593* (2011).

[16] Hannak A. Ma R. Chen, L. and C. Wilson. 2018. Investigating the Impact of Gender on Rank in Resume Search Engines. *Annual Conference of the ACM Special Interest Group on Computer Human Interaction* (2018).

[17] R. L Davis S. B Omer D. Adjaye-Gbewonyo, R. A Bednarczyk. 2014. Using the Bayesian Improved Surname Geocoding Method (BISG) to Create a Working Classification of Race and Ethnicity in a Diverse Managed Care Population: A Validation Study. *Health Serv Res* (2014).

[18] Morgan Klaus Scheuerman Foad Hamidi and Stacy M. Branham. 2018. Gender Recognition or Gender Reductionism?: The Social Implications of Embedded Gender Recognition Systems. *In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18* (2018).

[19] Centers for Disease Control and Prevention. 2021. National Environmental Public Health Tracking Network. www.cdc.gov/ephtracking

[20] S. Laohaprapanon G. Sood. 2018. Predicting Race and Ethnicity From the Sequence of Characters in a Name. (2018). https://arxiv.org/abs/1805.02109

[21] J. Chevalier Peter E. Rossi L. Currier G. Work, M. K. Chen. 2020. Suppliers and Demanders of Flexibility: The Demographics of Gig Work. (2020).

[22] Lani Guinier. 2021. Enlisting Race, Resisting Power, Transforming Democracy, in Race Rights, and Redemption: The Derrick Bell Lectures on the Law and Critical Race Theory. (2021).

[23] Pendleton N. Hoffman JS, Shandas V. 2020. The Effects of Historical Housing Policies on Resident Exposure to Intra-Urban Heat: A Study of 108 US Urban Areas. *Climate* (2020). https://doi.org/10.3390/cli8010012

[24] Lily Hu and Issa Kohler-Hausmann. 2020. What's sex got to do with machine learning? *In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (2020).

[25] Xiaojie Mao Geoffry Svacha Jiahao Chen, Nathan Kallus and Madeleine Udell. 2019. Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved. *Conference on Fairness, Accountability, and Transparency* (2019).

[26] O. H. Gandy Jr. 2010. Engaging rational discrimination: exploring reasons for placing regulatory constraints on decision support systems. *Ethics and Information Technology* 12 (2010), 29–42.

[27] J. Brown A. Xiang M. Andrus, E. Spitzer. 2021. What We Can't Measure, We Can't Understand: Challenges to Demographic Data Procurement in the Pursuit of Fairness. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2021).

[28] S. Ahmed M. Bogen, A. Rieke. 2020. Awareness in practice: tensions in access to sensitive attribute data for antidiscrimination. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (2020).

[29] Caitlin Lustig Morgan Klaus Scheuerman, Kandrea Wade and Jed R. Brubaker. 2020. How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis. *Proceedings of the ACM on Human Computer Interaction 4, CSCW1* (2020).

[30] A. Zhou N. Kallus, X. Mao. 2020. Assessing Algorithmic Fairness with Unobserved Protected Class Using Data Combination. (2020). https://arxiv.org/pdf/1906.00285.pdf

[31] Namsor. 2021. API v2.0.16 Documentation. https://namsor.app/api-documentation.html#namsor-api

[32] J. Novet. 2018. LinkedIn wants to help recruiters do a better job with gender diversity. https://www.cnbc.com/2018/10/10/linkedin-recruiter-starts-reflecting-gender-mix-in-search-results.html

[33] Powers B. Vogeli C. Obermeyer, Z. and S. Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 266 (2019), 6464.

[34] Color of Change. 2020. Beyond the Statement: Tech Framework. https://colorofchange.org/btstech/

[35] Partnership on AI. 2021. Fairer Algorithmic Decision-Making and Its Consequences: Interrogating the Risks and Benefits of Demographic Data Collection, Use, and Non-Use. https://partnershiponai.org/paper/fairer-algorithmic-decision-making-and-its-consequences/

[36] L. Kaplan A. Mislove A. Rieke P. Sapiezynski, A. Ghosh. 2019. Algorithms that 'Don't See Color': Comparing Biases in Lookalike and Special Ad Audiences. (2019). https://arxiv.org/abs/1912.07579

[37] J. Passy. 2018. How 'redlining' still hurts home values. *MarketWatch* (2018). https://www.marketwatch.com/story/how-redlining-still-hurts-home-values-2018-04-26

[38] Relman Colfax PLLC. 2020. Fair Lending Monitorship of Upstart Network's Lending Model.

[39] J. Miao I. Mironov J. Tannen R. Alao, M. Bogen. 2021. How Meta is working to assess fairness in relation to race in the U.S. across its products and systems. (2021). https://ai.facebook.com/research/publications/how-meta-is-working-to-assess-fairness-in-relation-to-race-in-the-us-across-its-products-and-systems

[40] G. B. Canner R. B. Avery, K. P. Brevoort. 2009. Credit Scoring and Its Effects on the Availability and Affordability of Credit. *The Journal of Consumer Affairs* 43, 3 (2009).

[41] Jr. R. L. Austin. 2021. Race Data Measurement and Meta's Commitment to Fair and Inclusive Products. https://about.fb.com/news/2021/11/inclusive-products-through-race-data-measurement/

[42] Richard Marciano Nathan Connolly et al. Robert K. Nelson, LaDale Winling. 2021. Mapping Inequality. https://dsl.richmond.edu/panorama/redlining

[43] Vincent Southerland. 2021. The Intersection of Race and Algorithmic Tools in the Criminal System. *80 Md. L. Rev. 487* (2021).

[44] McKee K. R. Kay J. & Mohamed S. Tomasev, N. 2021. Fairness for Unobserved Characteristics: Insights from Technological Impacts on Queer Communities. (2021). https://doi.org/10.1145/3461702.3462540

[45] I. Voicu. 2016. Using First Name Information to Improve Race and Ethnicity Classification. (2016). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2763826

[46] Isabel Wilkerson. 2020. Caste. (2020).

[47] Michel Verleysen & Vincent D. Blondel Yves-Alexandre de Montjoye, César A. Hidalgo. 2013. Unique in the Crowd: The privacy bounds of human mobility. *Scientific Reports* (2013).

A APPENDICES

A.1 Data Dictionary

These are definitions for the terms used in Table 1 and Figures 2-18:

- **More than 30 trips / year**: Binary variable indicating whether the user took more than 30 trips in 2019.
- **Prop'n of Trips with a Driver Cancel**: Proportion of all trips in 2019 a user attempted to take where the driver canceled the trip. If a trip request was never matched with a driver, this was not included in the denominator of all trips in 2019.
- **Proportion of Trips Starting or Ending in Historically Redlined Neighborhoods**: Proportion of trips that began or ended in a census tract that overlapped with a HOLC area graded D, as measured in the University of Richmond Mapping Inequality Project.
- **Average PM 2.5 Levels in Pickup Census Tracts**: Average PM 2.5 Level in the Pickup Census Tract across all 2019 trips for a user, as measured by the CDC.
- **Trip requested from iOS**: Whether the trip request was made from an iOS operating system.
- **Prop'n of Trips on Weekend Nights**: Proportion of all 2019 trips that began between Friday 10 pm and Saturday 2 am or Saturday 10 pm and Sunday 2 am.
- **Prop'n of Trips During Commuting Hours**: Proportion of all 2019 trips that began on a weekday between 8 am and 10 am or 4 pm and 6 pm.
- **Prop'n of Trips in City Core**: Proportion of all 2019 trips that begin in a "City Core." "City Core" is defined as the contiguous area of a city, excluding airports, with the highest density of trips.
- **Prop'n of Trips that are Airport Trips**: Proportion of all 2019 trips that started or began at an airport.
- **Prop'n of Trips in Area with above the 90th percentile of People Age 18 - 24**: Proportion of all 2019 trips that started or ended in a census tract with above the 90th percentile for number of people aged 18 to 24.
- **Prop'n of Trips in Areas with Above-Median Level of Uber Eats Businesses**: Proportion of all 2019 trips that started or ended in a census tract with an above-median level of Uber Eats businesses.

A.2 Outcomes by Self-Reported and Inferred Race

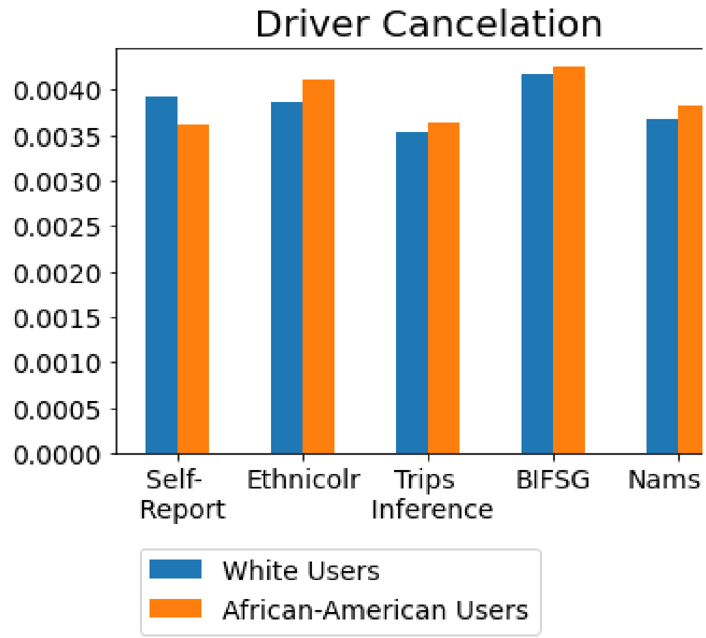


Figure 9: Demographic outcomes for driver cancellation.

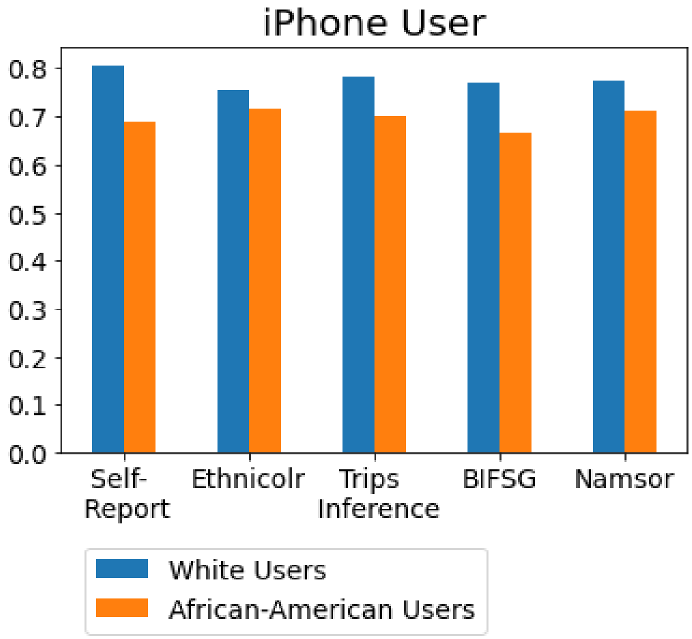


Figure 10: Demographic outcomes for iPhone users.

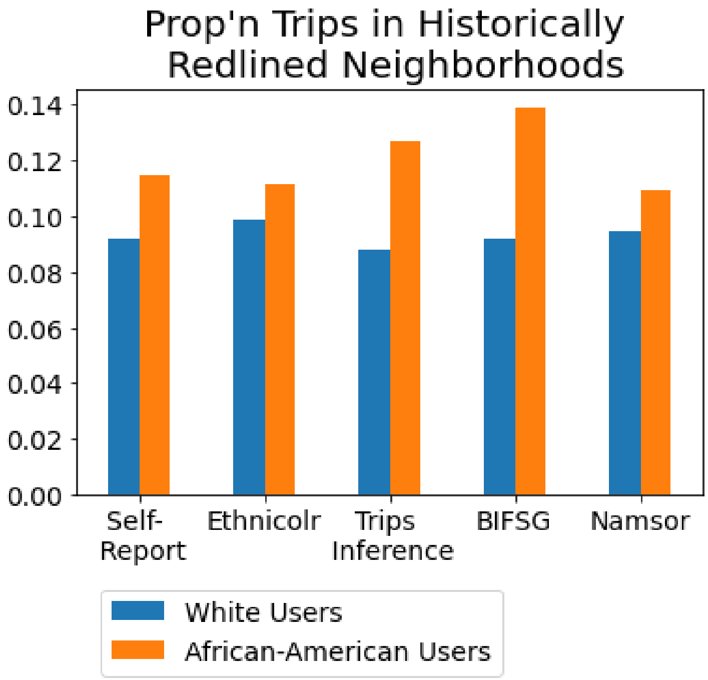


Figure 11: Demographic outcomes for proportion of trips in historically redlined neighborhoods.

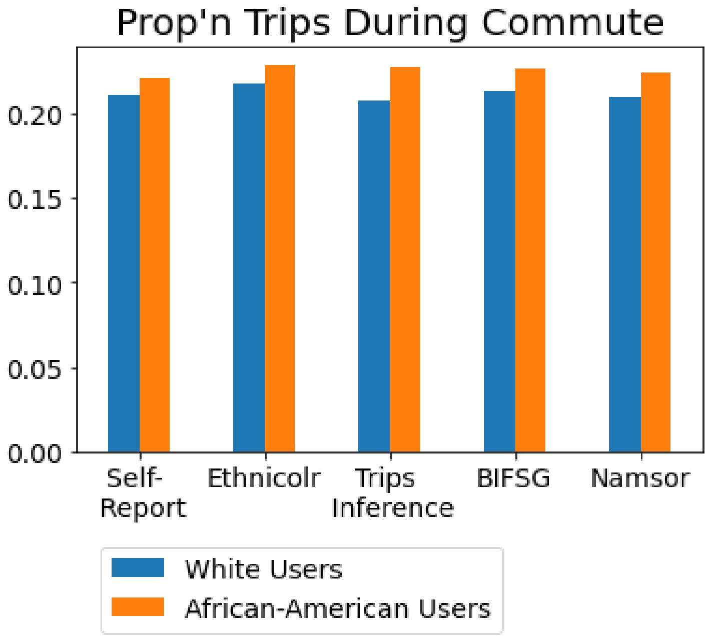


Figure 12: Demographic outcomes for trips during commute.

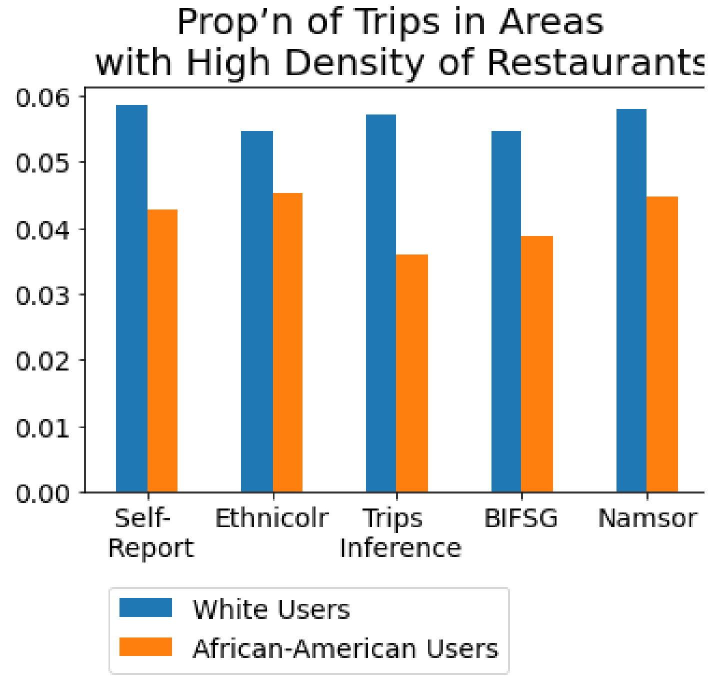


Figure 15: Demographic outcomes for trips in areas with a high density of restaurants.

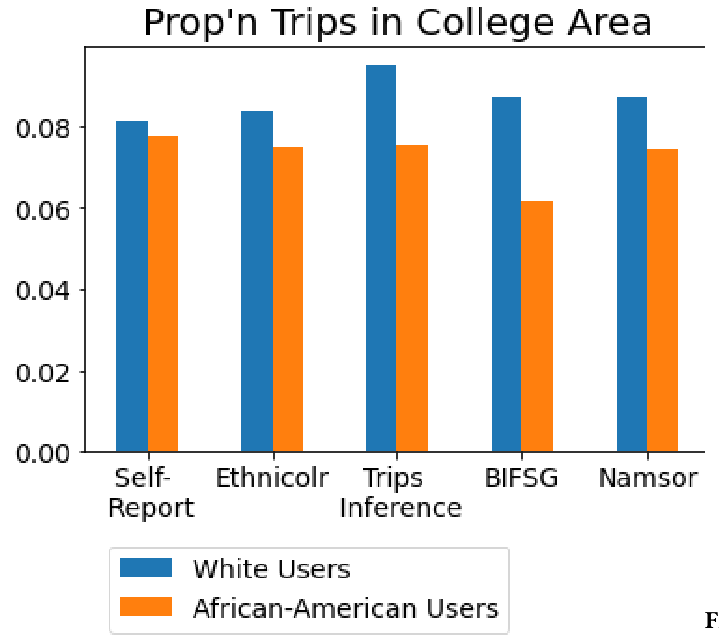


Figure 16: Demographic outcomes for trips in a college area.

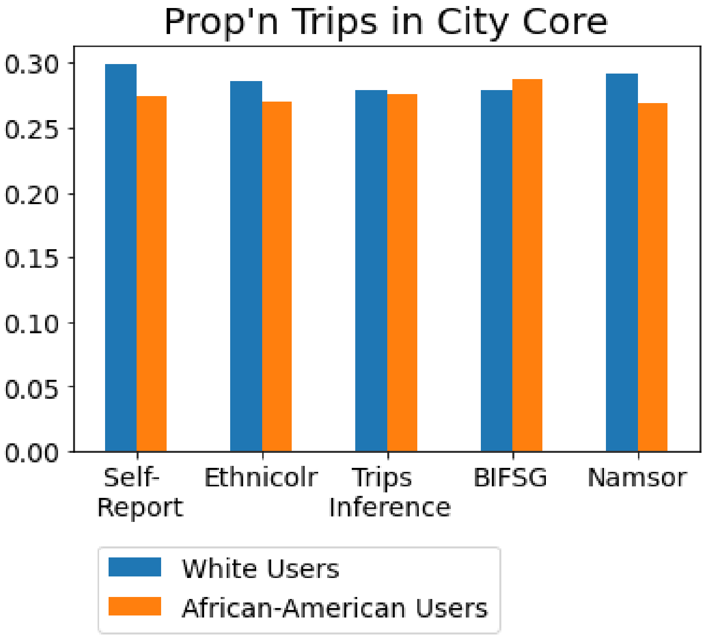


Figure 17: Demographic outcomes for trips in a city core.

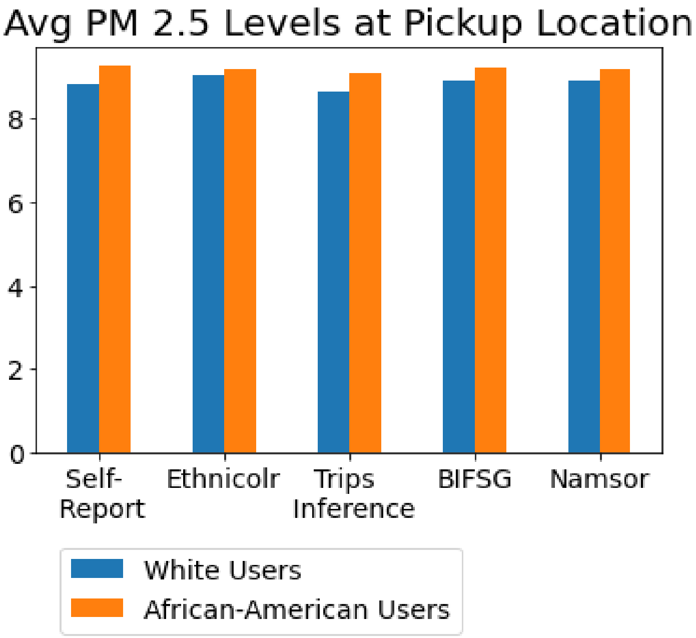


Figure 18: Demographic outcomes for users with more than 30 trips in 2019.