Leveraging Administrative Data for Bias Audits: Assessing Disparate Coverage with Mobility Data for COVID-19 Policy

Amanda Coston
acoston@cs.cmu.edu
Carnegie Mellon University

Neel Guha
nguha@stanford.edu
Stanford University

Derek Ouyang
douyang1@stanford.edu
Stanford University

Lisa Lu
lcl@law.stanford.edu
Stanford University

Alexandra Chouldechova
achould@cmu.edu
Carnegie Mellon University

Daniel E. Ho
dho@law.stanford.edu
Stanford University

ABSTRACT

Anonymized smartphone-based mobility data has been widely adopted in devising and evaluating COVID-19 response strategies such as the targeting of public health resources. Yet little attention has been paid to measurement validity and demographic bias, due in part to the lack of documentation about which users are represented as well as the challenge of obtaining ground truth data on unique visits and demographics. We illustrate how linking large-scale administrative data can enable auditing mobility data for bias in the absence of demographic information and ground truth labels. More precisely, we show that linking voter roll data—containing individual-level voter turnout for specific voting locations along with race and age—can facilitate the construction of rigorous bias and reliability tests. Using data from North Carolina's 2018 general election, these tests illuminate a sampling bias that is particularly noteworthy in the pandemic context: older and non-white voters are less likely to be captured by mobility data. We show that allocating public health resources based on such mobility data could disproportionately harm high-risk elderly and minority groups.

1 INTRODUCTION

Mobility data has played a central role in the response to COVID-19. Describing the movement of millions of people, smartphone-based mobility data has been used to analyze the effectiveness of social distancing polices (non-pharmaceutical interventions), illustrate how movement impacts the transmission of COVID-19, and probe how different sectors of the economy have been affected by social distancing policies [1, 6, 9, 12, 22, 24, 35, 50]. Despite the high-stakes settings in which this data is deployed, there has been no independent assessment of the reliability of this data. In this paper we show how administrative data (i.e., data from government agencies kept for administrative purposes) can be used to perform such an assessment.

Data reliability should be a foremost concern in all policy-making and policy evaluation settings, and is especially important for mobility data due to the lack of transparency surrounding data provenance. Mobility data providers obtain their data from opt-in location-sharing mobile apps, such as navigation, weather, or social media apps, but do not disclose which specific apps feed into their data [33]. This opacity prevents data consumers such as policymakers and researchers from understanding who is represented in the mobility data, a key question for enabling effective and equitable policies in high-stakes settings such as the COVID-19 pandemic. Grantz et al. describe "a critical need to understand where and to what extent these biases may exist" in their discussion on the use of mobility data for COVID-19 response.

Of particular interest is potential sampling bias with respect to important demographic variables in the context of the pandemic: age and race. Older age has been established as an increased risk factor for COVID-19-related mortality [56]. African-American, Native-American and Latinx communities have seen disproportionately high case and death counts from COVID-19 [49] and the pandemic has reinforced existing health inequities that affect vulnerable communities [26]. If certain races or age groups are not well-represented in data used to inform policy-making, we risk enacting policies that fail to help those at greatest risk and serve to further exacerbate disparities.

In this paper we assess SafeGraph, a widely-used point-of-interest (POI)-based mobility dataset[1] for disparate coverage by age and race. We define **coverage** with respect to a POI: coverage is the proportion of traffic at a POI that is recorded in the mobility data. For privacy reasons, many mobility datasets are aggregated up from the individual level to the physical POI level. Due to this aggregation, we lack the resolution to assess individual-level coverage quantities like the fraction of members of a demographic subgroup of interest who are represented in the data. Nonetheless, our POI-based notion of coverage is relevant for many COVID-19 policies that are made based on traffic to POIs, such as deciding to close certain business sectors, allocating resources like pop-up testing sites to high-risk areas, and determining where to target investigations of public health

---

[1]POIs refer to anywhere people spend money or time, including schools, brick-and-mortar stores, parks, places of worship, and airports. See https://www.safegraph.com/.

Amanda Coston, Neel Guha, Derek Ouyang, Lisa Lu, Alexandra Chouldechova, and Daniel E. Ho

order violations. We use differences in the distributions of age and race across POIs to assess demographic disparities in coverage.

While we focus here on a specific dataset and implications for COVID-19 policy, the question of how one can assess disparate coverage is a more general one in algorithmic governance. Ground truth is often lacking, which is precisely why policymakers and academics have flocked toward big data, on the implicit assumption that scale can overcome more conventional questions of data reliability, sampling bias, and the like [2, 37]. Government agencies may not always have access to protected attributes, making fairness and bias assessments challenging [34].

The main contributions of our paper are as follows:

(1) We show how administrative data can enable audits for bias and reliability (§ 4)
(2) We characterize the measurement validity of a smartphone-based mobility dataset that is widely used for COVID-19 research, SafeGraph (§ 4.2, 5.1)
(3) We illuminate significant demographic disparities in the coverage of SafeGraph (§ 5.2)
(4) We illustrate how this disparate coverage may distort policy decisions to the detriment of vulnerable populations (§ 5.3)

Our paper proceeds as follows. Sections 2 and 3 discuss related work and background on the uses of mobility data in the pandemic. Section 4 provides an overview of our auditing framework, formalizes the assumptions to construct bias and reliability tests, and discusses the estimation approach using voter roll data from North Carolina's 2018 general election. Section 5 presents results that while SafeGraph can be used to estimate voter turnout, the mobility data systematically undersamples older individuals and minorities. Section 6 discusses interpretation and limitations.

2 RELATED WORK

Our assessment of disparate coverage is related to several strands in the literature. First, the most closely related work to ours is SafeGraph's own analysis of sampling bias discussed below (§ 3.3). SafeGraph's analysis examines demographic bias only at the national aggregated level and does not address the question of demographic bias for POI-specific inferences. Ours is the first independent assessment of demographic bias to the extent we are aware.

Second, our work relates to existing work on demographic bias in smartphone-based estimates [55]. A notable line of survey research has examined the distinct demographics of smartphone users [20, 38]. [53] and [54] document significant concerns about mobility-based estimates from mobile phone data, including particularly low coverage for elderly. The literature further finds that smartphone ownership in the United States varies significantly with demographic attributes [8]. In 2019 an estimated 81% of Americans owned smartphones with ownership rates of 96% for those aged 18-29 and ownership rates of 53% for those aged over 65 [44]. Racial disparities in smartphone ownership are less pronounced, with an ownership rate of 82%, 80%, and 79% for White, Latinx, and African-American individuals, respectively. Even conditional on mobile phone ownership, however, demographic disparities may still exist. App usage may differ by demographic group. According to one report, 69% of U.S. teenagers, for instance, use Snapchat,

compared to 24% of U.S. adults [4]. Of particular relevance to mobility datasets, the rate at which users *opt in* to location sharing may vary by demographic subgroup. Hoy and Milne, for instance, reported that college-aged women exhibit greater concerns with third party data usage. And even among users who who opt in to a specific app, usage behavior may differ according to demographics. Older users, for instance, may be more likely to use a smartphone as a "classic phone" [3].

Our work responds to a recent call to characterize the biases in mobility data used for COVID-19 policies [25]. Grantz et al. highlight the potential for demographic bias, citing "clear sociodemographic and age biases of mobile phone ownership." They note, "Identifying and quantifying these biases is particularly challenging, though, when there is no clear gold standard against which to validate mobile phone data." We provide the first rigorous test for demographic bias using auxiliary estimates of ground truth.

Third, our work bears similarity to the literature on demographic bias in medical data and decision-making. A long line of research has demonstrated that medical research is disproportionately conducted on white males [19, 40, 43]. This literature has cataloged the harmful effects of making treatment decisions for subgroups that were underrepresented in the data [7, 51, 52]. In much the same vein, our work calls into question research conclusions based on SafeGraph data that may not be relevant for older or minority subgroups.

Last, our work relates more broadly to the sustained efforts within machine learning to understand sources of demographic bias in algorithmic decision making [14, 15, 23, 27, 36]. Important work has audited demographic bias of facial recognition technology [10], child welfare screening tools [13], criminal risk assessment scores [45], and health care allocation tools [2, 41]. Often the underlying data is identified as a major source of bias that propagates through the algorithm and leads to disparate impacts in the decision-making stage. Similarly, our study illustrates how disparate coverage in smartphone-based data can misallocate COVID-19 resources.

3 BACKGROUND ON SAFEGRAPH MOBILITY DATA

We now discuss the SafeGraph mobility dataset, illustrate how this data has been widely deployed to study and provide policy recommendations for the public health response to COVID-19, and discuss SafeGraph's own assessment of sampling bias.

3.1 SafeGraph Mobility Data

SafeGraph contains mobility data from roughly 47M mobile devices in the United States. The company sources this data from mobile applications, such as navigation, weather, or social media apps, where users have opted in to location tracking. It aggregates this information by points-of-interest (POIs) such as schools, restaurants, parks, airports, and brick-and-mortar stores. Hourly visit counts are available for each of over 6M POIs in their database.[2] Individual device pattern data is not distributed for researchers due to privacy concerns. Our analysis relies on SafeGraph's 'research release' data which aggregates visits at the POI level.

---

[2]See https://docs.safegraph.com/docs/places-summary-statistics.

3.2 Use of SafeGraph Data in COVID-19 Response

When the pandemic hit, SafeGraph released much of its data for free as part of the "COVID-19 Data Consortium" to enable researchers, non-profits, and governments to leverage insights from mobility data. As a result, SafeGraph's mobility data has become the dataset de rigueur in pandemic research. The Centers for Disease Control and Prevention (CDC) employs SafeGraph data to examine the effectiveness of social distancing measures [39]. According to SafeGraph, the CDC also uses SafeGraph to identify healthcare sites that are reaching capacity limits and to tailor health communications. The California Governor's Office, and the cities of Los Angeles [21], San Francisco, San Jose, San Antonio, Memphis, and Louisville, have each relied on SafeGraph data to formulate COVID-19 policy, including evaluation of transmission risk in specific areas and facilities and enforcement of social distancing measures. Academics, too, have employed the data widely to understand the pandemic: [12] used SafeGraph data to examine how social distancing compliance varied by demographic group and recommend occupancy limits for business types; [17, 18] used SafeGraph to infer the effect of "superspreader" events such as the Sturgis Motorcycle Rally and campaign events; [42] examined whether social distancing was more prevalent in in areas with higher xenophobia; and [1] examined whether social distancing compliance was driven by political partisanship, to name a few. What is common across all of these works is that they assume that SafeGraph data is representative of the target population.

3.3 SafeGraph Analysis of Sampling Bias

SafeGraph has issued a public report about the representativeness of its data [46, 47]. While SafeGraph does not have individual user attributes (e.g., race, education, income), it merged census data based on census block group (CBG) to assess bias along demographic characteristics.[3] SafeGraph assigns each device an estimated home CBG based on where the device spends most of its nights and uses the demographics of the estimated home CBG for the bias assessment. The racial breakdown of device holders, for instance, was allocated proportionally based on the racial breakdown of the devices' estimated home CBGs. SafeGraph then compared the total SafeGraph imputed demographics against census population demographics at the national level. According to SafeGraph, the results suggest that their data is "well-sampled across demographic categories" [46].

SafeGraph's examination for sampling bias should be applauded. Companies may not always have the incentive to address these questions directly, and SafeGraph's analysis is transparent, with data and replication code provided. As far as we are aware, it remains the only analysis of SafeGraph sampling bias.

Nevertheless, their analysis suffers from several key limitations. Most notably, this analysis does not use ground-truth demographic information and instead relies on imputed demographics using a method which suffers systematic biases. For instance, home CBG estimation is inaccurate for certain segments of the population, such as nighttime workers. Even when the estimated home CBG itself is correct, their imputation of *demographics* from the CBG

imposes a strong homogeneity assumption: The mere fact that 52% of Atlanta's population is African American does not guarantee that five out of ten SafeGraph devices in Atlanta belong to African-Americans.

Additionally, the analysis uses an aggregation scheme which introduces two methodological limitations. First, because their analysis aggregates CBGs nationally, the results are susceptible to undue influence from outliers, such as those resulting from errors in home CBG estimation. We anticipate these errors to be substantial since SafeGraph reports highly unrepresentative sampling rates at the CBG level, including CBGs with four times as many devices as residents.[4] Second, the results may also miss significant differences in the joint distribution of features because the analysis aggregates CBGs for a single attribute at a time. For example, if coverage is better for younger populations and for whiter populations, but whiter populations are on average older than non-white populations, then evaluating coverage marginally against either race or age will underestimate disparities. Indeed we present evidence for such an effect in § 5.

Lastly, this analysis uses CBGs as the unit of analysis which may miss disparities that exist at finer geographic units, such as POIs. This distinction is noteworthy since many of the COVID-19 analyses referenced above leverage SafeGraph data at finer geographic units than CBGs (e.g. POIs). This risks drawing conclusions from data at a level of resolution that SafeGraph has not established to be free from coverage disparities. SafeGraph warns that "local analyses examining only a few CBGs" should proceed with caution. Because SafeGraph's analysis examines demographic bias only at census aggregated levels and does not address the question of demographic bias for POI-specific inferences, an independent coverage audit remains critical. We provide such an audit using a method that uses POIs as the unit of analysis and avoids the noted methodological limitations.

4 AUDITING FRAMEWORK

In this section we outline our proposed auditing methodology and state the conditions under which the proposed method allows us to detect demographic disparities in coverage. We motivate our approach by first describing the idealized audit we would perform if we had access to ground truth data. We then introduce our administrative data and subsequently modify this framework to account for the limitations of the available data.

4.1 Notation

Let $\mathcal{I} = \{1, ..., n\}$ denote a set of SafeGraph POIs. Let $S^j \in \mathbb{R}^n$ denote a vector of the SafeGraph traffic count (i.e. number of visits) for day $j \in \mathcal{J}$ where each element S_i^j indicates the traffic to POI i on day j. Similarly let T_i^j denote the ground truth traffic (visits) to POI i during day j. When the context is clear, we omit the superscript j when referring to vectors $S \in \mathbb{R}^n$ and $T \in \mathbb{R}^n$. We use $\oslash$ to denote Hadamard division (the element-wise division of two matrices). With this, we define our coverage function $C(S, T)$.

---

[3]CBGs are geographic regions that contain typically between 600 and 3000 residents. CBGs are the smallest geographic unit for which the census publishes data.

[4]See Fig. 3 of [48].

DEFINITION 1 (COVERAGE FUNCTION). *Let* $C(S, T) : \mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbb{R}^n$ *denote the following coverage function:*

$$C(S, T) = S \oslash T$$

The coverage function yields a vector where the ith element equals $\frac{S_i}{T_i}$ *and describes the coverage of POI i.*

Let D_i^j denote a numeric measure of the demographics of visitors to POI i on day j; for instance D_i^j may be the percentage of visitors to a location on a specific day that are over the age of 65. Let $\text{cor}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(y)}}$ denote the Pearson correlation between vectors X and Y and let $r(X)$ be the rank function that returns the rank of vector X.[5] Our audit will consider the (Spearman) rank correlation $\text{cor}(r(X), r(Y))$, which provides a more flexible and robust measure of association than the Pearson correlation.

4.2 Idealized Audit

Our audit assesses how well SafeGraph measures ground truth visits and whether this coverage varies with demographics. We operationalize these two targets as follows:

DEFINITION 2 (MEASUREMENT SIGNAL AND VALIDITY). *Define the* ***strength of measurement signal*** *as*

$$\text{cor}(r(S), r(T)).$$

A positive signal indicates facial measurement validity, and a signal close to one indicates high measurement validity.

DEFINITION 3 (DISPARATE COVERAGE). *We will say that* ***disparate coverage*** *exists when the rank correlation between coverage and the demographic measure is statistically different from zero:*

$$\text{cor}(r(C(S, T), r(D)) \neq 0.$$

We are interested in identifying an association of any kind; we are not concerned with identifying a causal effect per se. Age might have a causal effect on smartphone usage, setting aside the question of manipulability [30], as depicted in the top panel (a) of Fig. 1. But as the bottom panel (b) depicts, age may not directly affect SafeGraph coverage but be directly correlated with a factor like urban/rural residence, which in turn does affect SafeGraph coverage. For either mechanism, the policy-relevant conclusion remains that SafeGraph is underrepresenting certain age groups.

In reality, there is no ground truth source of information about foot traffic and the corresponding demographics for all 6 million POIs. Instead, we must make do with estimates of T and D based on auxiliary data sources about some subset of visits to a subset of POIs. In order to identify the relationship of interest (Def. 3) between coverage and demographics, we need the following to hold:[6]

DEFINITION 4 (NO INDUCED CONFOUNDING). *The estimation procedure does not induce a confounding factor that affects both the estimate of demographics and the estimate of coverage.*

(a) Causal association
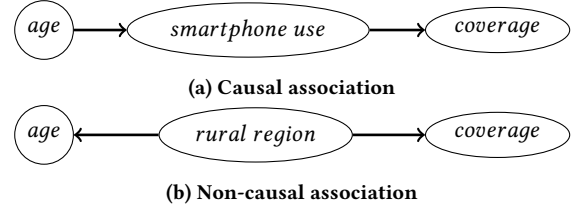

(b) Non-causal association

Figure 1: Possible mechanisms under which disparate coverage arises. Disparate coverage may be a result of a causal associations such as (a) whereby older people are less likely to own or use smartphones and therefore places frequented by older people have lower coverage. Disparate coverage may also arise due to a non-causal associations such as (b) whereby rural regions have higher percentages of older residents and worse cell reception which reduces coverage. Both types of associations are policy-relevant because in both cases, certain age groups are underrepresented.

DEFINITION 5 (NO SELECTION BIAS). *The selection is not based on an interaction between factors that affect coverage and demographics.*

We emphasize the difficulty in obtaining this information. It is challenging to obtain estimates of foot traffic to POIs. In fact, researchers typically treat smartphone-based mobility data as if it were ground truth (e.g. [5]). It is even more challenging to identify data sources for ground truth visits to POIs *with* corresponding demographic information [25]. Consider for instance large sporting events where stadium attendance is closely tracked. Can we leverage differences in audience demographics based on the event (e.g., international soccer game between two countries) in order to assess disparate coverage? Two major impediments are lack of access to precise demographic estimates as well as confounding factors such as tailgating that may vary with demographics.

4.3 Administrative data on voter turnout

We propose a solution using large-scale administrative data that records individual-level visits along with demographic information: voter turnout data in North Carolina's 2018 general election from L2, a private voter file vendor which aggregates publicly available voter records from jurisdictions nationwide.[7] Our analysis relies primarily on four fields in the L2 voter files: age, race, precinct, and turnout. The L2 data is missing one key piece of information: the poll location. We use a crosswalk of voting precinct to poll location obtained from the North Carolina Secretary of State to map each voter via their voting precinct to a SafeGraph POI. Overall, our data includes 539K voters who turned out to vote at 558 voting locations that could be matched. Table 1 presents summary statistics on voters associated with polling locations that could be matched, showing that our data is highly representative of all voting locations. (Details on the data and preprocessing are provided in Appendices A and B.)

Derived from official certified records by election authorities, voter turnout information is of uniquely high fidelity. In an analysis of five voter file vendors, Pew Research, for instance, found that the vendors had 85% agreement about turnout in the 2018 election [32].

---

[5]The rank assigns each element of the vector the value of its rank in an increasing ranking of all elements in the vector. For example the rank of vector "(5, 1, 3)" would be "(3, 1, 2)".

[6]Appendix C discusses the analogous assumptions required to identify the target for measurement validity (Def. 2).

[7]See https://l2political.com/.

	Matched Voters	All Voters
Voters	539,607	1,581,937
Mean Age	52.57	52.78
Std Age	16.67	16.59
Proportion over 65	0.25	0.26
Proportion Hispanic	0.04	0.04
Proportion Black	0.20	0.19
Proportion White	0.70	0.71

Table 1: Demographics of all voters in North Carolina's 2018 general election compared to voters included in our analysis ("matched voters"). The matched voters are representative of the full voting population. Details of the matching procedure are given in Appendix B.

Voter registration forms typically include fields for date of birth, gender, and often race.[8] When race is not provided, data vendors estimate race. The Pew study found race to be 79% accurate across the five vendors, with accuracies varying from 67% for African-Americans to 72% for Hispanics to 93% for non-Hispanics.[9] We can identify individuals visiting a specific voting location on election day because North Carolina differentiates *in person, election day* voters from absentee, mail, and early voters. We note that poll locations are often schools, community centers, religious institutions, and fire stations. These POIs may hence also have non-voter traffic on election day. We address this possible source of confounding by adjusting the SafeGraph traffic using an estimate of non-voter traffic.

4.4 Adjustment for non-voter traffic

Non-voter traffic may be incorporated into SafeGraph measures and may confound our analysis if the magnitude of that non-voter traffic varies with the demographic attributes of the voters. For instance, if younger voting populations are more likely to vote at community centers which have large non-voter traffic and older voting populations are more likely to vote at fire stations which have small non-voter traffic, then even if SafeGraph has no disparate coverage, we would observe a negative relationship between coverage and age.[10] We control for this confounding by estimating non-voter traffic using mean imputation. In Appendix D, we provide similar results using a linear regression imputation procedure.

4.4.1 Additional notation. Letting j^* denote election day, we estimate the non-voter traffic at poll location i on election day, $Z_i^{j^*}$, by

averaging SafeGraph traffic to i on adjacent days:

$$Z_i^{j^*} = \frac{S_i^{j^*-1} + S_i^{j^*+1}}{2}$$

This adjustment enables us to compute the marginal traffic over the estimated baseline, which we term SafeGraph marginal traffic.[11]

DEFINITION 6 (MARGINAL TRAFFIC). *SafeGraph marginal traffic denotes device counts above estimated baseline: $S_i^{j^*} - Z_i^{j^*}$.*

Let $V_i^{j^*}$ denote the number of voters at poll location i as recorded by L2. With this, we refine our definition of coverage using the coverage function from Def. 1:

DEFINITION 7 (SAFEGRAPH COVERAGE). *SafeGraph coverage is $C(S^{j^*} - Z^{j^*}, V^{j^*})$. Each element i of this vector refers to the ratio of marginal traffic at POI i to voter turnout at i.*

4.5 Audit via voter turnout

The disparate impact question in this setting is *does SafeGraph coverage of voters at different poll locations vary with voter demographics?* We focus on two key demographic risk factors for COVID-19: age and race. We summarize the age distribution at a polling location i by computing the proportion of voters over age 65. For race, we consider the proportion of voters who are an ethnic group besides white.[12]

Def. 3 formalizes this question as testing whether there is a rank correlation between $C(S^{j^*} - Z^{j^*}, V^{j^*})$ and demographic measure D. However such a test may be misleading if we have induced confounding by our estimation procedure (Def. 4). We can incorporate a test of confounding into our audit. Specifically, we can test for *time-invariant* confounding.

DEFINITION 8 (TIME-INVARIANT CONFOUNDING). *A time-invariant confounder affects our demographic estimate as well as traffic on election day and on non-election days.*

This contrasts to a **time-varying confounding**:

DEFINITION 9 (TIME-VARYING CONFOUNDING). *A time-varying confounder affects our demographic estimate and traffic on election day only. It does not affect traffic on non-election days.*

Examples of time-invariant and time-varying confounding are given in Figure 2. The assumption of no time-varying confounding is untestable but it is reasonable to believe this holds in our setting. Most voting places, for instance, are public places making it unlikely that the non-voter traffic is affected differentially on election and non-election days. Another possible time-varying confounder would be if voting locations with older (or largely non-white) voters are more likely to be placed outside of the SafeGraph geometry for device attribution (e.g., parking lot). We do not believe this is likely because voting locations are typically indoors for security and climate reasons during a November election. We can accommodate time-invariant confounding in our audit by modifying the definition of disparate coverage.

---

[8]North Carolina, for instance requests both race and ethnicity (https://s3.amazonaws.com/dl.ncsbe.gov/Voter_Registration/NCVoterRegForm_06W.pdf).

[9]The study did not name which voter file vendors were analyzed.

[10]Non-voter traffic may be affected by device attribution errors, in which device GPS locations are incorrectly assigned to one of two adjacent POIs. SafeGraph reports in its user documentation that "[it] is more difficult to measure visits to a midtown Manhattan Starbucks than a visit to a suburban standalone Starbucks." If younger voting populations are more likely to vote in dense urban polling locations, then even if there isn't large non-voter traffic in the same facility, large traffic in an adjacent facility could still be incorrectly attributed to the polling location with greater likelihood than to a suburban polling location. However, this source of confounding can be controlled for using the same technique described.

[11]The adjustment resulted in negative estimates of voter traffic for poll locations at schools. In the Appendix B, we show that baseline traffic estimation is generally much worse for school, due in part to school holidays or large-scale events such as sports games. As a result, we exclude schools from our analysis.

[12]In what follows we use the generic variable D to indicate either measure of demographics.

Time-invariant confounding	Time-varying confounding
Example: Younger voting populations vote at places like community centers with large non-voter traffic whereas older populations vote at places like fire stations with little non-voter traffic.	**Example:** Younger voting populations vote at places that are open to non-voter traffic on election day whereas older populations vote at places that are closed to non-voter traffic on election day
Testable (see § 4.5)	Untestable assumption

Figure 2: We distinguish between two types of confounding: time-invariant versus time-varying confounding. We test for time-invariant confounding (§ 4.5) but we cannot test for time-varying confounding. Our results assume no time-varying confounding.

DEFINITION 10 (DISPARATE COVERAGE). *We will say that **disparate coverage** exists when the rank correlation between coverage on **election** day and voter demographics is statistically different from the rank correlation between coverage on **non-election** day and voter demographics: For $j \neq j*$,*

$$\mathrm{cor}\big(r(C(S^{j*} - Z^{j*}, V^{j*}), r(D^{j*})\big) \neq$$
$$\mathrm{cor}\big(r(C(S^{j} - Z^{j}, V^{j*}), r(D^{j*})\big)$$

We evaluate this more robust notion of disparate coverage using 40 weekdays in October and November of 2018 to generate a placebo distribution of the estimated correlation coefficients against which we compare the election-day estimate.[13] Algorithm 1 provides details (note that $\mathbb{I}$ denotes the indicator function). This procedure is similar to methods of randomization inference in the literature on treatment effects [29]. If the election-day correlation is unlikely under placebo distribution (i.e. small p-value), and we additionally believe there is no time-varying confounding, then we can conclude that SafeGraph has disparate coverage of voters on election day.

Algorithm 1: Assessing Disparate Coverage (Def. 10)

Input: Voter data (V^{j*}, D^{j*}) SafeGraph data $\{(S^j, Z^j)\}_{j=1}^{n}$
Result: p-value for the election-day correlation under the placebo distribution
for $j = 1, 2, \dots n$ **do**
　　Compute $\rho_j = \mathrm{cor}(r(C(S^j - Z^j, V^{j*})), r(D^{j*}))$.
end
return $p = \frac{1}{n} \sum_{j=1}^{n} \mathbb{I}\{(\rho_j \leq \rho_{j*})\}$

In order to generalize these findings to the broader population on non-election day, the selection cannot be based on factors that affect both coverage and demographics (See Def. 5). Example violations might include: (i) The older (or non-white) population that doesn't vote is more likely to use smartphones than the older (or non-white) population that does vote; and (ii) Older (or non-white) voters leave

---

[13]We use weekdays in October and November except October 1 and November 5, 7, and 30. November 5 and 7 are adjacent to election day, so the baseline adjustment (§ 4.4) would be biased. Out of convenience, we drop October 1 and November 30 to avoid having to pull September and December data to respectively compute their baseline adjustments.

their smartphones at home when they go vote but always carry their smartphones otherwise, whereas younger (or white) voters bring their smartphones to the polls and elsewhere. We believe such mechanisms are unlikely. Testing this assumption would require the use of an additional auxiliary dataset which is outside the scope of this paper. We emphasize that this assumption of no selection bias can still hold even though the voting population is not a random sample of the population with respect to demographics. In fact, since the voting population is older and more white than the general population [32], the association between coverage and age/race among voters could very well underestimate the magnitude of the population association.

We should also consider the association between demographics age and race. It is well known that younger populations have a larger proportion non-white relative to older populations, and this holds in our sample. Polling locations with younger voters are also more likely to have higher proportions of minority voters (Appendix A). Additionally, there are widespread concerns that disparate impact can be more pronounced at the "intersection" of protected groups [10, 11, 16]. We can jointly test for disparate coverage by modifying Alg. 1 for the multiple regression setting. We perform $n = 41$ linear regressions to model coverage as a function of the percentage over 65 and the percentage non-white for each weekday j in October and November 2018. We test whether the election-day coefficients on age/race are different from the 40 non-election day coefficients on age/race. Alg. 2 provides details using the notation that A^{j*} denotes the proportion of voters over age 65 and R^{j*} denotes the proportion of voters who are non-white.

5 RESULTS

5.1 Measurement Validity

Election day brings a dramatic increase in traffic to polling locations relative to non-election days, and any valid measure of visits should detect this outlier. Figure 3 shows the daily aggregate traffic across poll locations for October and November of 2018, and as expected, we see a significant increase in both total traffic (top panel) and marginal traffic (bottom panel) on election day. To assess the strength of this signal using the framework described above (Def. 2), we present the correlation between marginal SafeGraph traffic on election day and actual voter turnout. The rank correlation test yields a positive correlation: $\mathrm{cor}\big(r(S^{j*} - Z^{j*}), r(V^{j*})\big) = 0.383$ with

Algorithm 2: Assessing Joint Disparate Coverage

Input: Voter data (V^{j*}, D^{j*}) SafeGraph data $\{(S^j, Z^j)\}_{j=1}^n$

Result: p-values for the election-day coefficients on race and age under the placebo distribution

for $j = 1, 2, \ldots n$ **do**

    Fit a linear regression:
$$C(S^j - Z^j, V^{j*}) = \alpha_j A^{j*}, +\beta_j R^{j*} + \gamma_j$$

end

return $p_A = \frac{1}{n} \sum_{j=1}^n \mathbb{I}\{(\alpha_j \leq \alpha_{j*})\}$ *and*

$$p_R = \frac{1}{n} \sum_{j=1}^n \mathbb{I}\{(\beta_j \leq \beta_{j*})\}$$

p-value < 0.001.[14] Figure 4 displays this relationship by comparing the marginal election traffic $S_i^{j*} - Z_i^{j*}$ on the x-axis against actual voter counts T_i^{j*} on the y-axis for each polling location.

This corroborates that SafeGraph data is able to detect broad patterns in movement and visits. That said, the estimates at the individual polling place location level are quite noisy: root mean-squared error is 1375 voters. For instance, amongst polling places that registered 20 marginal devices, roughly 300 to 2300 actual voters turned out. This significant noise is likely due to a combination of factors. First, SafeGraph may incorrectly attribute voters to nearby POIs because of incorrect building geometries. Second, we may not be able to perfectly adjust for non-election traffic. Third, SafeGraph may have disparate coverage of voters by demographic attributes. This last factor is the focus of our analysis.

5.2 Demographic Bias

We assess whether the demographic composition of voters who actually turned out to vote in person is correlated with coverage. We begin with preliminary results and then proceed to our main disparate coverage results (as defined in Def. 10).

5.2.1 Preliminary results. Polling locations with older votes have lower coverage rates. The top panel of Figure 5 shows how Safe-Graph coverage $C(S-Z, V)$ varies with A, the proportion of voters over age 65. The rank correlation test yields $\text{cor}\big(r(C(S-Z, V)), r(A)\big) = -0.14$ with p-value < 0.001. We also show how coverage decreases as the proportion of non-white voters increases (bottom panel). The rank correlation of race and coverage is $\text{cor}\big(r(C(S - Z, V)), r(R)\big) = -0.11$ with p-value $= 0.0067$. The top panel of Figure 6 presents a heat map of coverage with age bins (quartiles) on the x-axis and race bins (quartiles) on the y-axis. This bottom left cell, for instance, shows that precincts that are the most white and young have highest coverage rates. The lowest coverage is for older minority precincts. The lower panel of Figure 6 similarly plots race on the x-axis against coverage on the y-axis, separating older precincts (yellow) and younger precincts (blue). Older precincts on average have lower coverage rates than younger precincts, and coverage declines as the minority population increases.

---

[14]The correlation is similar but slightly lower for unadjusted SafeGraph traffic: $\text{cor}\big(r(S^{j*}), r(V^{j*})\big) = 0.373$ with p-value < 0.001.
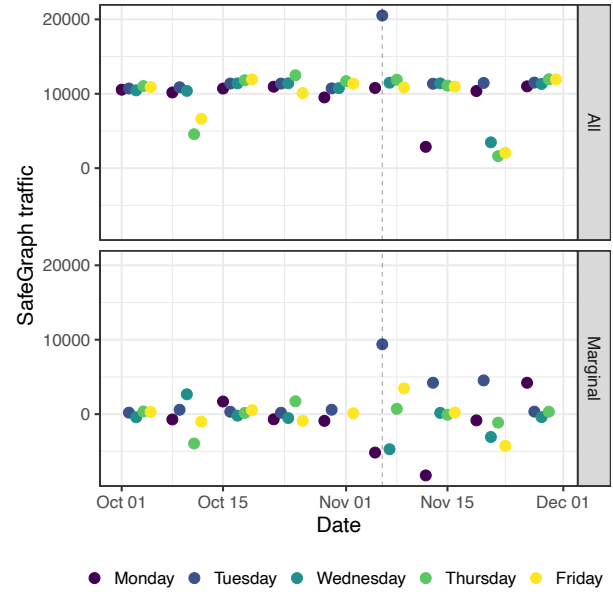
Figure 3: SafeGraph traffic by weekday over October and November 2018 for all polling locations in North Carolina. The top panel shows all SafeGraph traffic and the bottom panel shows the marginal traffic computed using the method in § 4.4. In both total and marginal traffic, the election day (dotted) line shows a significant boost in traffic.
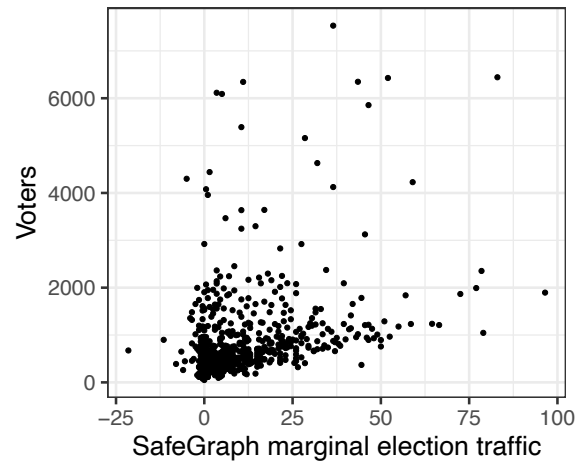

Figure 4: Election day traffic as observed by SafeGraph (x-axis) and actual voter turnout across polling locations (y-axis). Each dot represents a polling location in North Carolina in the 2018 general election.

5.2.2 Main results. Figure 7 shows that the negative election-day rank correlation between coverage and voter demographics is significantly outside the placebo distribution for non-election days

Amanda Coston, Neel Guha, Derek Ouyang, Lisa Lu, Alexandra Chouldechova, and Daniel E. Ho
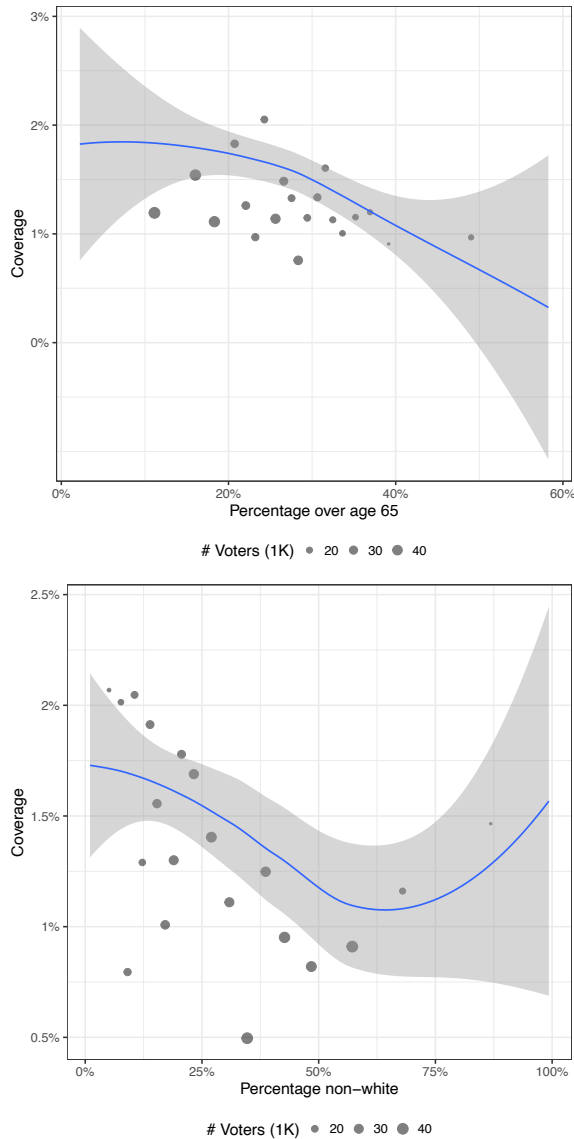

Figure 5: Estimated SafeGraph coverage rates against age and race for North Carolina 2018 general election. Each point displays a ventile of poll location by age (top) and race (bottom). The blue lines depict LOESS smoothing on the individual poll locations.
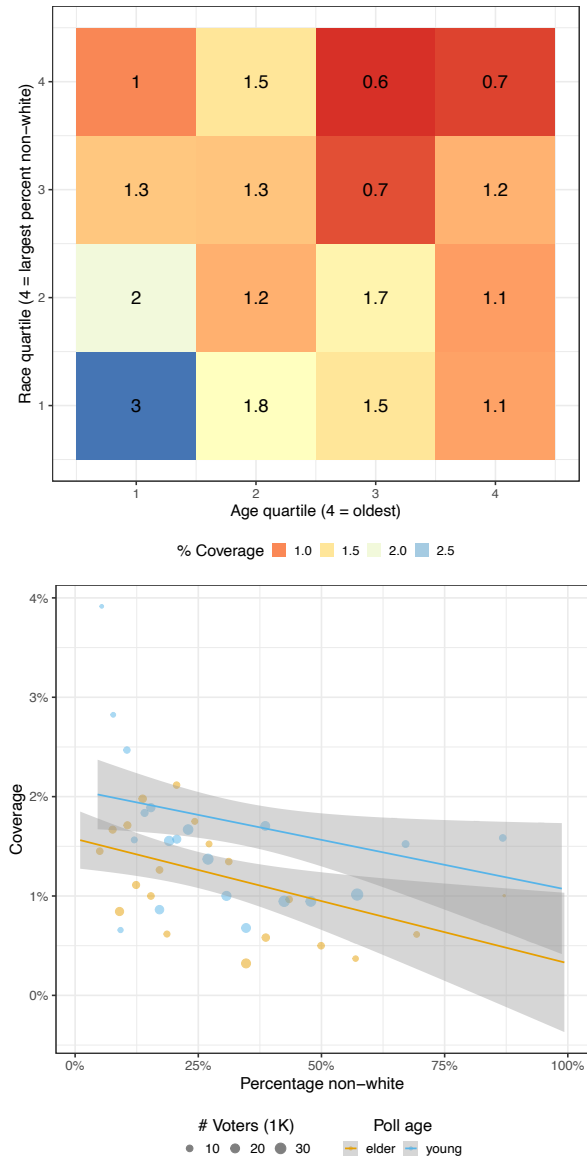

Figure 6: Intersectional coverage effects by race and age. The top panel presents the coverage rate by quartiles of age on the x-axis and race on the y-axis. The bottom panel plots the coverage rate on the y-axis against percentage of non-white voters at the polling location on the x-axis for older polling locations (yellow) versus younger polling locations (blue) for ventiles of poll location by race. (Lines display linear smoothing of the individual poll locations.) Coverage is lowest among older minority populations and highest among younger whiter populations.

(empirical one-sided p-values are ≈ 0.024 for both age and race, respectively). For our joint analysis of disparate coverage (Alg. 2), we find that the negative coefficients for age and race are statistically outside the placebo distribution (See Fig. 8; empirical one-sided p-values are ≈ 0.024 and 0.049 for age and race respectively).[15] Our findings are robust to time-invariant confounding. Assuming no selection bias (Def. 5) or time-varying confounding (Def. 9), we can

conclude that SafeGraph has disparate coverage by age and race, two demographic risk factors for COVID-19.

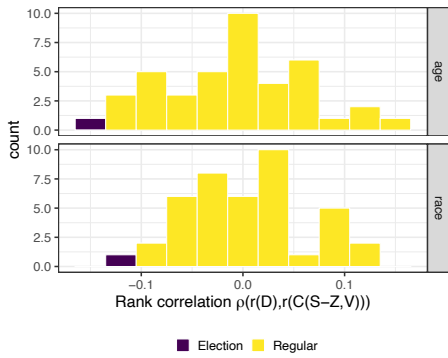[15]In Appendix C, we present similar placebo results for measurement validity.

Figure 7: Distribution of placebo rank correlations between election-day demographics and marginal SafeGraph traffic on non-election days. Under the empirical placebo distribution, the election-day coverage's negative correlations with age (top panel) and race (bottom panel) are very unlikely (p-value $<$ 0.05). Placebo correlations computed for 40 weekdays in October and November 2018.
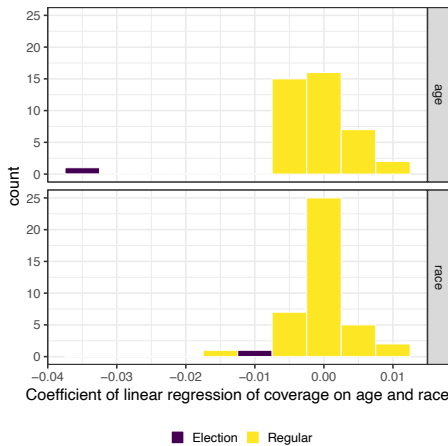

Figure 8: Placebo distribution of coefficients of the linear regression of marginal SafeGraph coverage on election-day age and race demographics. Under the empirical placebo distribution, the election-day's negative coefficients for age and (top panel) and race (bottom panel) are very unlikely (p-value $<$ 0.05). This suggests that SafeGraph data has disparate coverage by age and race. Regressions computed for 40 weekdays of October and November 2018.

5.3 Policy implications

We now examine the policy implications of disparate coverage in light of the widespread adoption of SafeGraph data in COVID-19 response. In particular, we show how disparate coverage may lead to under-allocation of important health resources to vulnerable populations. For instance, suppose the policy decision at hand is where to locate mobile pop-up COVID-19 testing sites, and suppose

the aim is to place these sites in the most trafficked areas to encourage asymptomatic individuals to get tested. One approach could use SafeGraph traffic estimates to rank order POIs. How would this ordering compare to the optimal ordering by ground truth traffic? Using voter turnout as an approximation to ground truth traffic, we perform linear regression of the rank of voter turnout against rank according to SafeGraph marginal traffic as well as age and race: $r(V) \sim r(S - Z) + A + R$. Table 2 presents results of this rank regression (where rank is in descending order), confirming that the SafeGraph rank is significantly correlated with ground truth rank. But the large coefficient on age indicates that each percentage point increase in voters over 65 is associated with a 4 point drop in rank relative to the optimal ranking. Similarly, the coefficient on race indicates that a 1.5 point increase in percent non-white is associated with a one point drop in rank relative to the optimal ranking. This demonstrates that ranking by SafeGraph traffic may disproportionately harm older and minority populations by, for instance, failing to locate pop-up testing sites where needed the most.

Table 2: To evaluate a potential rank-based policy allocation, we compare the rank of voter turnout against rank by Safe-Graph traffic, controlling for age and race in a linear regression. Although SafeGraph rank is correlated with the optimal rank by voter turnout, the coefficients on age and race indicate that each demographic percentage point increase is associated with a 4-point and 1-point drop in rank for age and race, respectively. This indicates that significant adjustments based on demographic composition should be made to a SafeGraph ranking. Failure to do so may direct resources away from older and more minority populations.

	Dependent variable:
	Voter turnout rank
SafeGraph rank	0.317***
	(0.040)
% over 65	4.716***
	(0.748)
% non-white	0.681**
	(0.295)
Constant	40.278
	(24.830)
Observations	558
R^2	0.203
Adjusted R^2	0.199
Residual Std. Error	144.264 (df = 554)
F Statistic	47.027*** (df = 3; 554)
Note:	*p<0.1; **p<0.05; ***p<0.01

We also consider the implications of using SafeGraph to inform proportional resource allocation decisions, such as the provision of masks. We compare the allocation based on SafeGraph traffic to the allocation based on voter turnout data. Table 3 presents results for polling locations binned into four age-race groups by partitioning at the median proportion over 65 and median proportion non-white. Each cell presents the proportion of resources that would be allocated to that age-race bin, demonstrating that strict reliance on SafeGraph would under-allocate resources by 37% to the oldest/most non-white category (p-value < 0.05) and over-allocate resources by 33% to the youngest/whitest category (p-value < 0.05).

	SafeGraph allocation	Optimal allocation	Percent difference
young white	0.33 (0.03)	0.25 (0.02)	+33%
young non-white	0.33 (0.03)	0.35 (0.03)	-5%
older white	0.19 (0.02)	0.18 (0.01)	+9%
older non-white	0.13 (0.02)	0.21 (0.02)	-37%

Table 3: Allocation of resources for age-race groups by Safe-Graph versus by true voter counts, with standard errors in parentheses. The SafeGraph allocation redirects over one-third of the optimal allocation from the oldest, most non-white group to the youngest, whitest group (p-value < 0.05).

The clear policy implication here is that while SafeGraph information may aid in a policy decision, auxiliary information (including prior knowledge) should likely be combined to make final resource allocation decisions.

6 DISCUSSION

We have provided the first independent audit of demographic bias of a smartphone-based mobility dataset that has been widely used in the policy response to COVID-19. Our audit indicates that the data underrepresents two high risk groups: older and more non-white populations. Our results suggest that policies made without adjustment for this sampling bias may disproportionately harm these high risk groups. However, we note a limitation to our analysis. Because SafeGraph information is aggregated for privacy reasons, we are not able to test coverage at the individual level. To avoid a potential ecological fallacy, our results should be interpreted as a statement about *POIs* rather than *individuals*. That is, POIs frequented by older (or minority) visitors have lower coverage than POIs frequented by younger (or whiter) populations. Of course, policy decisions are typically made at some level of aggregation, so the demographic bias we document at this level remains relevant for those decisions.

A key future research question is how to use the results of this audit to improve policy decisions. We suggest a few possible future directions. A bias correction approach would construct weights to adjust estimates based on race and age. Such an approach crucially requires knowledge about demographic composition. In policy settings where such information is not readily available, it may be fruitful to investigate whether mobility data companies like Safe-Graph can provide normalized visit counts based on the estimated demographic profile of the smartphone user. This could offer a significant improvement over current normalization approaches which, per SafeGraph's recommendation, use census block group (CBG)-based normalization factors [47]. While this bias correction might help to estimate population parameters (e.g., percentage of CBG population not abiding by social distancing), it is unlikely to capture the kind of demographic interaction effects we document here. Much more work should be done to study disparate coverage and ideally provide, for instance, a weighing correction to the normalization factors that properly accounts for the demographic disparities documented in this audit.

Another possible solution is increased transparency. Researchers do not know details about the source of SafeGraph's mobility data, namely which mobile apps feed into the SafeGraph ecosystem. Access to such information may make the bias correction approach more tractable. If, for instance, researchers could identify that a data point emanates from Snapchat, then they could use what is known about the Snapchat user base to make adjustments. Given its increasing importance for policy, SafeGraph should consider disclosing more details about which apps feed into their ecosystem.

7 CONCLUSION

Mobility data based on smartphones has been rapidly adopted in the COVID-19 response. As [25] note, one of the most profound challenges arising with such rapid adoption has been the need to assess the potential for demographic bias "when there is no clear gold standard against which to validate mobile phone data." Our paper illustrates one potential path forward, by linking smartphone-based data to high-fidelity ground truth administrative data. Voter turnout records, which record at the individual level whether a registered voter traveled to a polling location on a specific day and describe the voter's demographic information, enable us to develop a straightforward audit test for disparate coverage. We find that coverage is notably skewed along race and age demographics, both of which are significant risk factors for COVID-19 related mortality. Failure to address such disparities risks policy distortions based on mobility data that could exacerbate serious existing inequities in the health care response to the pandemic.

10

REFERENCES

[1] Hunt Allcott, Levi Boxell, Jacob Conway, Matthew Gentzkow, Michael Thaler, and David Y Yang. 2020. *Polarization and public health: Partisan differences in social distancing during the Coronavirus pandemic.* Working Paper w26946. National Bureau of Economic Research (NBER).

[2] Kristen M Altenburger, Daniel E Ho, et al. 2018. When Algorithms Import Private Bias into Public Enforcement: The Promise and Limitations of Statistical Debiasing Solutions. *Journal of Institutional and Theoretical Economics* 174, 1 (2018), 98–122.

[3] Ionut Andone, Konrad Błaszkiewicz, Mark Eibes, Boris Trendafilov, Christian Montag, and Alexander Markowetz. 2016. How Age and Gender Affect Smartphone Usage. In *UbiComp '16.* Association for Computing Machinery, New York, NY, USA, 9–12. https://doi.org/10.1145/2968219.2971451

[4] Salman Aslam. 2021. Snapchat by the Numbers: Stats, Demographics Fun Facts. https://www.omnicoreagency.com/snapchat-statistics/#:~:text=Snapchat%20Demographics

[5] Han Bao, Xun Zhou, Yingxue Zhang, Yanhua Li, and Yiqun Xie. 2020. COVID-GAN: Estimating Human Mobility Responses to COVID-19 Pandemic through Spatio-Temporal Conditional Generative Adversarial Networks. In *Proceedings of the 28th International Conference on Advances in Geographic Information Systems.* 273–282.

[6] Seth G. Benzell, Avinash Collis, and Christos Nicolaides. 2020. Rationing social contact during the COVID-19 pandemic: Transmission risk and social benefits of US locations. *Proceedings of the National Academy of Sciences* 117, 26 (2020), 14642–14644. https://doi.org/10.1073/pnas.2008025117 arXiv:https://www.pnas.org/content/117/26/14642.full.pdf

[7] Guillermo Bernal and María R Scharró-del Río. 2001. Are empirically supported treatments valid for ethnic minorities? Toward an alternative approach for treatment research. *Cultural Diversity and Ethnic Minority Psychology* 7, 4 (2001), 328.

[8] Krishna K Bommakanti, Laramie L Smith, Lin Liu, Diana Do, Jazmine Cuevas-Mota, Kelly Collins, Fatima Munoz, Timothy C Rodwell, and Richard S Garfein. 2020. Requiring smartphone ownership for mHealth interventions: who could be left out? *BMC public health* 20, 1 (2020), 81.

[9] Adam Brzezinski, Valentin Kecht, and David Van Dijcke. 2020. *The Cost of Staying Open: Voluntary Social Distancing and Lockdowns in the US.* Technical Report. Oxford University.

[10] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Conference on Fairness, Accountability, and Transparency (Proceedings of Machine Learning Research),* Sorelle A. Friedler and Christo Wilson (Eds.), Vol. 81. PMLR, New York, NY, USA, 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html

[11] Ángel Alexander Cabrera, Will Epperson, Fred Hohman, Minsuk Kahng, Jamie Morgenstern, and Duen Horng Chau. 2019. FairVis: Visual analytics for discovering intersectional bias in machine learning. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST).* IEEE, Virtual, 46–56.

[12] Serina Chang, Emma Pierson, Pang Wei Koh, Jaline Gerardin, Beth Redbird, David Grusky, and Jure Leskovec. 2021. Mobility network models of COVID-19 explain inequities and inform reopening. *Nature* 589 (2021), 82–-87.

[13] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research),* Sorelle A. Friedler and Christo Wilson (Eds.), Vol. 81. PMLR, New York, NY, USA, 134–148. http://proceedings.mlr.press/v81/chouldechova18a.html

[14] Alexandra Chouldechova and Aaron Roth. 2018. The Frontiers of Fairness in Machine Learning. arXiv:arXiv:1810.08810

[15] Sam Corbett-Davies and Sharad Goel. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. arXiv:arXiv:1808.00023

[16] Kimberlé Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum* 1, 8 (1989), 139. Issue 1.

[17] Dhaval M Dave, Andrew I Friedson, Kyutaro Matsuzawa, Drew McNichols, Connor Redpath, and Joseph J Sabia. 2020. *Did President Trump's Tulsa Rally Reignite COVID-19? Indoor Events and Offsetting Community Effects.* Technical Report. National Bureau of Economic Research.

[18] Dhaval M Dave, Andrew I Friedson, Drew McNichols, and Joseph J Sabia. 2020. *The Contagion Externality of a Superspreading Event: The Sturgis Motorcycle Rally and COVID-19.* Technical Report. National Bureau of Economic Research.

[19] Rebecca Dresser. 1992. Wanted single, white male for medical research. *The Hastings Center Report* 22, 1 (1992), 24–29.

[20] David Dutwin, Scott Keeter, and Courtney Kennedy. 2010. Bias from wireless substitution in surveys of Hispanics. *Hispanic journal of behavioral sciences* 32, 2 (2010), 309–328.

[21] Philip Mielke Eva Pereira, Bryan Bonack and Chelsea Lawson. 2020. Using Data to Govern Through a Crisis. https://www.safegraph.com/webinar-govern-through-a-crisis

[22] Maryam Farboodi, Gregor Jarosch, and Robert Shimer. 2020. *Internal and external effects of social distancing in a pandemic.* Technical Report. National Bureau of Economic Research.

[23] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. 2019. A Comparative Study of Fairness-Enhancing Interventions in Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency.* Association for Computing Machinery, New York, NY, USA, 329–338. https://doi.org/10.1145/3287560.3287589

[24] Song Gao, Jinmeng Rao, Yuhao Kang, Yunlei Liang, and Jake Kruse. 2020. Mapping county-level mobility pattern changes in the United States in response to COVID-19. *SIGSPATIAL Special* 12, 1 (2020), 16–26.

[25] Kyra H Grantz, Hannah R Meredith, Derek AT Cummings, C Jessica E Metcalf, Bryan T Grenfell, John R Giles, Shruti Mehta, Sunil Solomon, Alain Labrique, Nishant Kishore, et al. 2020. The use of mobile phone data to inform analysis of COVID-19 pandemic epidemiology. *Nature Communications* 11, 1 (2020), 1–8.

[26] Darrell M Gray, Adjoa Anyane-Yeboa, Sophie Balzora, Rachel B Issaka, and Folasade P May. 2020. COVID-19 and the other pandemic: populations made vulnerable by systemic inequity. *Nature Reviews Gastroenterology & Hepatology* 17, 9 (2020), 520–522.

[27] Moritz Hardt and Solon Barocas. 2017. Fairness in machine learning.

[28] M Hlavac. 2018. Stargazer: Well-formatted regression and summary statistics tables (R Package version 5.2)[Computer software].

[29] Daniel E Ho and Kosuke Imai. 2006. Randomization inference with natural experiments: An analysis of ballot effects in the 2003 California recall election. *Journal of the American statistical association* 101, 475 (2006), 888–900.

[30] Paul W Holland. 1986. Statistics and Causal Inference. *J. Amer. Statist. Assoc.* 81, 396 (1986), 945–960.

[31] Mariea Grubbs Hoy and George Milne. 2010. Gender differences in privacy-related measures for young adult Facebook users. *Journal of Interactive Advertising* 10, 2 (2010), 28–45.

[32] Ruth Igielnik, Scott Keeter, Courtney Kennedy, and Bradley Spahn. 2018. *Commercial voter files and the study of US politics.* Technical Report. Pew Research Center. www.pewresearch.org/2018/02/15/commercial-voter-files-and-the-study-of-us-politics

[33] Michael H. Keller Jennifer Valentino-DeVries, Natasha Singer and Aaron Krolik. 2018. Your Apps Know Where You Were Last Night, and They're Not Keeping It Secret. https://www.washingtonpost.com/nation/2020/06/01/americans-are-delaying-medical-care-its-devastating-health-care-providers/?arc404=true

[34] Nathan Kallus, Xiaojie Mao, and Angela Zhou. 2020. Assessing Algorithmic Fairness with Unobserved Protected Class Using Data Combination. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20).* Association for Computing Machinery, New York, NY, USA, 110. https://doi.org/10.1145/3351095.3373154

[35] Benjamin D. Killeen, Jie Ying Wu, Kinjal Shah, Anna Zapaishchykova, Philipp Nikutta, Aniruddha Tamhane, Shreya Chakraborty, Jinchi Wei, Tiger Gao, Mareike Thies, and Mathias Unberath. 2020. A County-level Dataset for Informing the United States' Response to COVID-19. arXiv:arXiv:2004.00756

[36] Pauline T Kim. 2017. Auditing algorithms for discrimination. *University of Pennsylvania Law Review Online* 166 (2017), 189.

[37] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. The parable of Google Flu: traps in big data analysis. *Science* 343, 6176 (2014), 1203–1205.

[38] Sunghee Lee, J Michael Brick, E Richard Brown, and David Grant. 2010. Growing cell-phone population and noncoverage bias in traditional random digit dial telephone health surveys. *Health services research* 45, 4 (2010), 1121–1139.

[39] Amanda Moreland. 2020. Timing of State and Territorial COVID-19 Stay-at-Home Orders and Changes in Population Movement—United States, March 1–May 31, 2020. *MMWR. Morbidity and Mortality Weekly Report* 69 (2020), 1198–1203.

[40] Gina Moreno-John, Anthony Gachie, Candace M Fleming, Anna Napoles-Springer, Elizabeth Mutran, Spero M Manson, and Eliseo J Pérez-Stable. 2004. Ethnic minority older adults participating in clinical research. *Journal of Aging and Health* 16, 5_suppl (2004), 93S–123S.

[41] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.

[42] Maria Petrova, Ruben Enikolopov, Georgy Egorov, and Alexey Makarin. 2020. *Divided We Stay Home: Social Distancing and Ethnic Diversity.* Technical Report. National Bureau of Economic Research.

[43] Vickie L Shavers-Hornaday, Charles F Lynch, Leon F Burmeister, and James C Torner. 1997. Why are African Americans under-represented in medical research studies? Impediments to participation. *Ethnicity & health* 2, 1-2 (1997), 31–45.

[44] Mobile Fact Sheet. 2019. Pew Research Center, Internet and Technology. June 12, 2019.

[45] Jennifer L Skeem and Christopher T Lowenkamp. 2016. Risk, race, and recidivism: Predictive bias and disparate impact. *Criminology* 54, 4 (2016), 680–712.

[46] RF Squire. 2019. An Interactive Guide To Analyze Demographic Profiles from SafeGraph Patterns Data. https://colab.research.google.com/

drive/1qqLRxehVZr1OBpnbHRRyXPWo1Q98dnxA?authuser=1#scrollTo=fEFiU4ny9LYx

[47] RF Squire. 2019. Measuring and Correcting Sampling Bias in Safegraph Patterns for More Accurate Demographic Analysis. https://www.safegraph.com/blog/measuring-and-correcting-sampling-bias-for-accurate-demographic-analysis/?utm_source=content&utm_medium=referral&utm_campaign=colabnotebook&utm_content=panel_bias

[48] RF Squire. 2019. "What about bias in your dataset?" Quantifying Sampling Bias in SafeGraph Patterns. https://colab.research.google.com/drive/1u15afRytJMsizySFqA2EPlXSh3KTmNTQ#offline=true&sandboxMode=true

[49] Don Bambino Geno Tai, Aditya Shah, Chyke A Doubeni, Irene G Sia, and Mark L Wieland. 2020. The Disproportionate Impact of COVID-19 on Racial and Ethnic Minorities in the United States. *Clinical Infectious Diseases* 2020 (06 2020), 1–4. https://doi.org/10.1093/cid/ciaa815

[50] Laris Karklis Ted Mellnik and Andrew Ba Tran. 2020. Americans are delaying medical care, and it's devastating health-care providers. https://www.washingtonpost.com/nation/2020/06/01/americans-are-delaying-medical-care-its-devastating-health-care-providers/?arc404=true

[51] Sandra Millon Underwood. 2000. Minorities, women, and clinical cancer research: the charge, promise, and challenge. *Annals of Epidemiology* 10, 8 (2000), S3–S12.

[52] Darshali A Vyas, Leo G Eisenstein, and David S Jones. 2020. Hidden in plain sight—reconsidering the use of race correction in clinical algorithms.

[53] Amy Wesolowski, Caroline O Buckee, Kenth Engø-Monsen, and Charlotte Jessica Eland Metcalf. 2016. Connecting mobility to infectious diseases: the promise and limits of mobile phone data. *The Journal of infectious diseases* 214, suppl_4 (2016), S414–S420.

[54] Amy Wesolowski, Nathan Eagle, Abdisalan M Noor, Robert W Snow, and Caroline O Buckee. 2012. Heterogeneous mobile phone ownership and usage patterns in Kenya. *PloS one* 7, 4 (2012), e35319.

[55] Nathalie E Williams, Timothy A Thomas, Matthew Dunbar, Nathan Eagle, and Adrian Dobra. 2015. Measures of human mobility using mobile phone records enhanced with GIS data. *PloS one* 10, 7 (2015), e0133630.

[56] Fei Zhou, Ting Yu, Ronghui Du, Guohui Fan, Ying Liu, Zhibo Liu, Jie Xiang, Yeming Wang, Bin Song, Xiaoying Gu, et al. 2020. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The lancet* 395, 10229 (2020), 1054–1062.