NBER WORKING PAPER SERIES

QUANTIFYING THE UNCERTAINTY OF IMPUTED DEMOGRAPHIC DISPARITY ESTIMATES: THE DUAL-BOOTSTRAP

Benjamin Lu
Jia Wan
Derek Ouyang
Jacob Goldin
Daniel E. Ho

Working Paper 32312
http://www.nber.org/papers/w32312

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 2024

We thank Stanford's Center for Population Health Sciences for data, and Isabel Gallegos, Cameron Guage, Dan Soriano, Melody Huang, Peng Ding, Thomas Hertz, Peter Hull, Qiwei Lin, Mark Loewenstein, and participants in the NBER CRIW Race, Ethnicity, and Economic Statistics for the 21st Century conference for helpful comments. This material is based upon work supported by the National Science Foundation under Grant No. 1745640. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Quantifying the Uncertainty of Imputed Demographic Disparity Estimates: The Dual-Bootstrap
Benjamin Lu, Jia Wan, Derek Ouyang, Jacob Goldin, and Daniel E. Ho
NBER Working Paper No. 32312
April 2024
JEL No. C10,J10,J15

ABSTRACT

Measuring average differences in an outcome across racial or ethnic groups is a crucial first step for equity assessments, but researchers often lack access to data on individuals' races and ethnicities to calculate them. A common solution is to impute the missing race or ethnicity labels using proxies, then use those imputations to estimate the disparity. Conventional standard errors mischaracterize the resulting estimate's uncertainty because they treat the imputation model as given and fixed, instead of as an unknown object that must be estimated with uncertainty. We propose a dual-bootstrap approach that explicitly accounts for measurement uncertainty and thus enables more accurate statistical inference, which we demonstrate via simulation. In addition, we adapt our approach to the commonly used Bayesian Improved Surname Geocoding (BISG) imputation algorithm, where direct bootstrapping is infeasible because the underlying Census Bureau data are unavailable. In simulations, we find that measurement uncertainty is generally insignificant for BISG except in particular circumstances; bias, not variance, is likely the predominant source of error. We apply our method to quantify the uncertainty of prevalence estimates of common health conditions by race using data from the American Family Cohort.

Benjamin Lu
University of California, Berkeley
b.lu@berkeley.edu

Jia Wan
Massachusetts Institute of Technology
jiawan@mit.edu

Derek Ouyang
Stanford Law School
Room 313, Crown Quadrangle 559
Nathan Abbott Way
Stanford, CA 94305-8610
douyang1@law.stanford.edu

Jacob Goldin
University of Chicago Law School
1111 E 60th Street
Chicago, IL 60637
and NBER
jsgoldin@gmail.com

Daniel E. Ho
Stanford Law School
dho@law.stanford.edu

An appendix is available at http://www.nber.org/data-appendix/w32312

Quantifying the Uncertainty of Imputed Demographic Disparity Estimates: The Dual-Bootstrap

Benjamin Lu[*], Jia Wan[†], Derek Ouyang[‡],
Jacob Goldin[§], Daniel E. Ho[¶]

March 12, 2024

Abstract

Measuring average differences in an outcome across racial or ethnic groups is a crucial first step for equity assessments, but researchers often lack access to data on individuals' races and ethnicities to calculate them. A common solution is to impute the missing race or ethnicity labels using proxies, then use those imputations to estimate the disparity. Conventional standard errors mischaracterize the resulting estimate's uncertainty because they treat the imputation model as given and fixed, instead of as an unknown object that must be estimated with uncertainty. We propose a dual-bootstrap approach that explicitly accounts for measurement uncertainty and thus enables more accurate statistical inference, which we demonstrate via simulation. In addition, we adapt our approach to the commonly used Bayesian Improved Surname Geocoding (BISG) imputation algorithm, where direct bootstrapping is infeasible because the underlying Census Bureau data are unavailable. In simulations, we find that measurement uncertainty is generally insignificant for BISG except in particular circumstances; bias, not variance, is likely the predominant source of error. We apply our method to quantify the uncertainty of prevalence estimates of common health conditions by race using data from the American Family Cohort.

1 Introduction

Racial and ethnic disparities are a common focus of academic study, policymaking, and advocacy efforts across many domains, including criminal justice (Gelman et al., 2007; Berdejó, 2018), health care (Azin et al., 2020; Mackey et al., 2021), technology (Buolamwini and Gebru, 2018; Koenecke et al., 2020), and taxation (Brown, 2022; Avenancio-León and Howard, 2022).[1]

Such disparities are straightforward to compute—if individual-level demographic and outcome data are available. In many settings where the measurement of racial disparities is of interest, however, race data are missing or otherwise inaccessible. For example, Regulation B of the Equal

[*]University of California, Berkeley
[†]Massachusetts Institute of Technology
[‡]Stanford University
[§]University of Chicago and NBER
[¶]Stanford University

[1]For brevity, we use "race" to refer to both race and ethnicity throughout the remainder of this paper.

Credit Opportunity Act prohibits creditors from discriminating against an applicant on the basis of race. But monitoring and enforcing this prohibition is complicated by the fact that the very same laws also prohibit creditors from inquiring about an applicant's race at all.

Some researchers work around this problem by imputing individuals' races based on observable proxy features, then using those imputations to estimate the racial disparity. One of the most common imputation methods is Bayesian Improved Surname Geocoding (BISG), which imputes an individual's race based on their surname and geolocation (Elliott et al., 2009). Recent work has also investigated the potential power of machine learning for this task (Cheng et al., 2023; Xue et al., 2019; Kim et al., 2018).

But imputations are estimates, not oracles. Like any other statistic, each imputation is the output of an estimator fit on data, with its own bias and variance. This statistical uncertainty could affect the quality of the downstream racial disparity estimate. Many studies that use imputations ignore this potential error propagation, instead treating the imputation model as known with certainty (e.g., Brown et al., 2016; Zhang, 2018; Yee et al., 2022). Doing so can imperil the reliability of the final estimate in different ways (e.g., Labgold et al., 2021, adjusting for imputation bias).

This paper examines one aspect of the problem: the effect of measurement uncertainty on statistical inference. It is standard practice in academic research to report the confidence interval or standard error associated with a racial disparity estimate. But typical confidence intervals and standard errors reflect only classical *sampling uncertainty*—i.e., uncertainty arising from the fact that the disparity estimate is based on only a sample of the broader population of interest. They do not reflect the *measurement uncertainty* that arises from estimating the race probability model and thus risk mischaracterizing the degree of confidence in the disparity estimate.

We make three contributions to the study of this issue. First, we offer a "dual-bootstrap" procedure that incorporates both sampling and measurement uncertainty and thus offers more accurate statistical inference. We prove that our procedure is consistent for some race probability models under standard regularity conditions.

Second, we adapt our procedure to the special case of BISG, where Census Bureau-imposed constraints on data availability raise particular challenges. The Census Bureau does not disclose the individual-level survey responses on which its popularly used American Community Survey race-by-geolocation estimates are based. This prevents researchers from directly applying the general dual-bootstrap algorithm. We propose one way to nonetheless approximate the measurement uncertainty of BISG race probability estimates using other information provided by the Census Bureau.

Third, we apply our approach to simulated and real data to investigate how much measurement uncertainty contributes to the final disparity estimate's standard error. Our findings suggest that, in general, the uncertainty of BISG imputations only negligibly increases standard errors because BISG is a relatively inflexible model based on large-scale data (i.e., full-scale census records); bias, not variance, is likely the predominant type of error in BISG. But we do find some exceptions: BISG measurement uncertainty, and the way it is estimated, can substantially affect the final inference when studying particular demographic or geographic groups. We also show that when race probability models more flexible than BISG are employed, properly accounting for measurement uncertainty can substantially affect the widths of resulting confidence intervals. We illustrate these findings through an analysis of racial disparities in common health outcomes in the American Family Cohort, a dataset containing the electronic health records of primary care visits by patients in the United States. Our method has also been applied in a recent working paper studying racial disparities in tax audit rates (Elzayn et al., 2023).

2 Related Work

To our knowledge, the role of measurement uncertainty in the specific context of race imputation has not been thoroughly studied. As mentioned in Section 1, many studies where race is imputed simply ignore it. One exception is a concurrent working paper by Derby et al. (2024), who propose a fully Bayesian approach where a prior distribution for the conditional race probabilities is assumed, then updated based on reported Census Bureau estimates to obtain a posterior distribution from which conditional race probabilities are sampled.[2] Our proposal for BISG is similar in spirit. It can be viewed as the frequentist analog—but with the distinct advantage of constructing a sampling distribution of the conditional race probability estimates based on the uncertainty that the Census Bureau actually reports for those estimates, instead of a purely assumed prior model. This fidelity comes at some cost to flexibility; see Section 5 for discussion.

This paper draws from a rich body of research on missing data, survey design, and causal inference. Especially relevant are two strands of work, on inverse propensity weighting (IPW) and Z-estimation. The disparity estimator we consider, analyzed by Chen et al. (2019) and described in Section 3 below, weights individuals by their estimated probability of being of a given race. It is thus very similar in form to Hájek IPW estimators, which have been extensively studied by, for example, Miratrix et al. (2018) and Matsouaka et al. (2024). And, following some prior work on the properties of IPW estimators in the context of causal inference (Reifeis and Hudgens, 2022; Shu et al., 2021), we rely on Z-estimator theory to establish the asymptotic properties of our proposed method (Kosorok, 2008; Stefanski and Boos, 2002).

We distinguish our subject of investigation from several other important but distinct areas of study. First, we focus solely on the variability of the imputed disparity estimator, as typically reflected in metrics like the standard error or confidence intervals. Prior work has examined identification and bias properties of the specific imputed disparity estimator on which we focus (Chen et al., 2019; Kallus et al., 2022; Elzayn et al., 2023). Others have examined the accuracy and bias of specific race imputation models that often underlie imputed disparity estimators. For example, as mentioned above, Imai et al. (2022) propose ways of improving the accuracy of BISG by accounting for the possible migration of racial minorities to geographic areas where none resided prior to the latest census count. These issues are largely orthogonal to the challenge of accurately characterizing an imputation-based estimator's variability. In our theory and simulations, we assume that these issues have been favorably resolved.

Second, our work is distinct from multiple imputation, at least in its classical formulation. In the typical setting amenable to multiple imputation, race is observed as a categorical variable for a subset of the data to be analyzed, and the researcher seeks to impute the categorical race variable for the remaining subset of the data where it is missing; Fong and Tyler (2021) discuss some challenges of proper statistical inference in that setting that are similar to the ones we address here. But the setting we consider (described in Section 3) is one where race is completely unobserved for the data to be analyzed, and the researcher seeks to estimate each unit's real-valued probability of being a given race, not the unit's actually realized race. Our problem setting is thus more closely related to that of measurement error models (e.g., Fuller, 2009) and two- or split-sample instrumental variables (e.g., Angrist and Krueger, 1992, 1995), with particular focus on uncertainty quantification for general, nonlinear measurement models. Nonetheless, some conceptual similarities to multiple imputation can be drawn. Perhaps the most salient connection is to the concept of

[2] Imai et al. (2022) similarly impose a prior, but they focus on how doing so improves the accuracy of race predictions, not on how it can more accurately quantify the uncertainty of downstream estimates.

"proper" multiple imputation, defined by Rubin (1987). As Murray (2018) summarizes it, multiple imputation generally yields valid inference only if, among other conditions, the uncertainty of the imputation model itself is accounted for. This same concept underpins our work.

3 Setup and Notation

Consider two datasets: a training dataset $\mathcal{T} \equiv \{Z_i, A_i\}_{i=1}^{n_\mathcal{T}}$ and a primary dataset $\mathcal{P} \equiv \{Z_j, Y_j\}_{j=1}^{n_\mathcal{P}}$, where Y denotes the outcome, A is a binary indicator of race, Z denotes observable proxies of race, and $n_\mathcal{T}$ and $n_\mathcal{P}$ are the number of units in the training and primary datasets, respectively. The training dataset is drawn i.i.d. from some population $\mathbb{T}$, and the primary dataset is drawn i.i.d. from a potentially different population $\mathbb{P}$. Our estimand is the racial disparity in outcomes in the primary population $\mathbb{P}$:

$$\delta \equiv \mathbb{E}_\mathbb{P}\left[Y \mid A = 1\right] - \mathbb{E}_\mathbb{P}\left[Y \mid A = 0\right].$$

If A were observed in the primary dataset, estimation and inference of δ would be straightforward. But it is not—so we impute A using Z based on some model class $\mathcal{F}_A$ instead. Specifically, we fit a model $f \in \mathcal{F}_A$ of A on Z using the training dataset, where (Z, A) is jointly observed. We then use that model to estimate the race probability of each unit in the primary dataset: $\widehat{\Pr}_\mathbb{P}\left(A = 1 \mid Z = Z_j\right) \equiv f(Z_j)$. Finally, we estimate the racial disparity by the probabilistic weighting estimator that Chen et al. (2019) propose:

$$\hat{\delta} \equiv \frac{\sum_{j=1}^{n_\mathcal{P}} \widehat{\Pr}_\mathbb{P}\left(A = 1 \mid Z = Z_j\right) Y_j}{\sum_{j=1}^{n_\mathcal{P}} \widehat{\Pr}_\mathbb{P}\left(A = 1 \mid Z = Z_j\right)} - \frac{\sum_{j=1}^{n_\mathcal{P}} \widehat{\Pr}_\mathbb{P}\left(A = 0 \mid Z = Z_j\right) Y_j}{\sum_{j=1}^{n_\mathcal{P}} \widehat{\Pr}_\mathbb{P}\left(A = 0 \mid Z = Z_j\right)}.$$

Other imputation-based disparity estimators have been used or analyzed in past work. For example, Chen et al. (2019) discuss a thresholding estimator that estimates the mean outcome in each race by classifying individuals' races instead of using soft probabilities. And Elzayn et al. (2023) consider the slope coefficient in a linear regression of the outcomes on the estimated race probabilities, which they show can in conjunction with $\hat{\delta}$ bound the true disparity. But we focus on $\hat{\delta}$ because it is a commonly used estimator with favorable statistical properties (Chen et al., 2019; McCaffrey and Elliott, 2008). We briefly outline how our framework might extend to the linear disparity estimator of Elzayn et al. (2023) in Appendix C.2, but we defer a detailed examination of this and other extensions to future work.

We invoke standard assumptions so that the estimator $\hat{\delta}$ is consistent for δ. First, we assume that the probability model $\widehat{\Pr}_\mathbb{P}\left(A = 1 \mid Z\right)$ is correctly specified. Since the model is fit on the training dataset $\mathcal{T}$ but used to characterize the primary population $\mathbb{P}$, this assumption typically also implies that the conditional distribution of A given Z is the same in $\mathbb{T}$ and $\mathbb{P}$ everywhere Z has positive density in $\mathbb{P}$. Second, we assume that $\mathbb{E}_\mathbb{P}\left[\mathrm{Cov}_\mathbb{P}\left(A, Y \mid Z\right)\right] = 0$; Chen et al. (2019) show that this condition is sufficient for $\hat{\delta}$ to be consistent when the true probabilities are given. Since our focus is inference, not estimation, we take these assumptions for granted and refer interested readers to past work on the consistency of $\hat{\delta}$.

4 The Dual-Bootstrap

We propose a "dual-bootstrap" procedure to enable proper inference of $\hat{\delta}$ that accounts for both sampling and measurement uncertainty, as Figure 1 illustrates. We first state the procedure in

general terms, then investigate via simulation the effects of measurement uncertainty and the dual-bootstrap's ability to account for it.

4.1 General Procedure

Algorithm 1 states the dual-bootstrap procedure. It frames the desired output as a confidence interval estimated via the percentile bootstrap, but other uncertainty metrics can be estimated too. As Algorithm 1 shows, the dual-bootstrap is straightforward: We simply resample with replacement both the training and the primary datasets, then refit the race probability model on the resampled training dataset and apply it to estimate the racial disparity in the resampled primary dataset. The algorithm here calls for simple resampling with replacement, but other, more complex forms of resampling may be appropriate—for example, if the data are clustered (Owen, 2007; Derby et al., 2024).

The key contribution of the dual-bootstrap stems from its resampling of the training dataset and refitting of the race probability model. Doing so accounts for the uncertainty of the race probability estimates themselves. This uncertainty is then propagated downstream to the bootstrap statistic $\hat{\delta}^{*b}$. As discussed above, some prior work has ignored this measurement uncertainty entirely, instead treating $\widehat{\Pr}_{\mathbb{P}}(A = 1 \mid Z = Z_j)$ as true. This corresponds to skipping the first two lines of the for-loop in Algorithm 1.

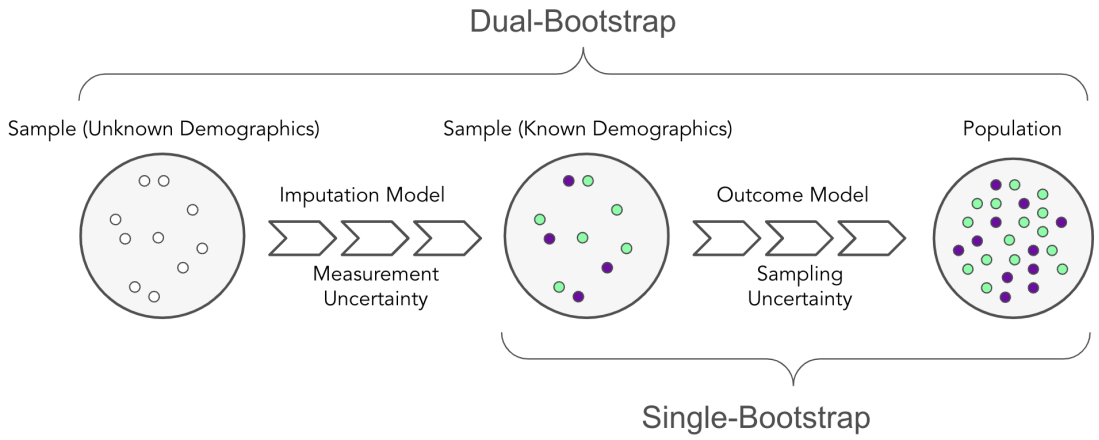

Figure 1: Illustration of the uncertainties captured by the dual-bootstrap, compared to those captured by the single-bootstrap.

We prove in Appendix C.1 that $\hat{\delta}$ and its dual-bootstrap analogs $\hat{\delta}^{*b}$ are asymptotically normal when the race probabilities are estimated via logistic regression and general regularity conditions hold. We limit the proof to logistic regression because doing so allows us to frame $\hat{\delta}$ as a Z-estimator—broadly, any estimator that can be expressed as the approximate zero of a data-dependent function—to which standard theoretical results can apply. The proof strategy likely applies readily to other race probability models that fall within the Z-estimation framework, albeit possibly with slight modifications to the regularity conditions. We leave such extensions to future work. It is less clear what theoretical properties the dual-bootstrap of $\hat{\delta}$ has when the race prob-

ability model does not fall within the Z-estimation framework. A closer examination of this issue might prove fruitful.

The Z-estimation theory that we apply to prove asymptotic normality also provides a closed-form expression for the variance of the limiting distribution, but this is mostly of theoretical interest. In practice, deriving the closed-form expression usually has little utility. Statistical software like the `geex` package in `R` can compute the empirical variance estimator using numerical routines, without requiring analytic derivations (Saul and Hudgens, 2020). When the race probability model falls within the Z-estimation framework, using such numerical solvers can often require much less computational power than the dual-bootstrap.

Algorithm 1: Dual-Bootstrap

Data: Training Dataset $\mathcal{T} = \{Z_i, A_i\}_{i=1}^{n_\mathcal{T}}$, Primary Dataset $\mathcal{P} = \{Z_j, Y_j\}_{j=1}^{n_\mathcal{P}}$, Model Class $\mathcal{F}_A$, Number of Bootstrap Draws $B \in \mathbb{N}$, Level $\alpha \in [0, 1]$

Result: Confidence interval for the demographic disparity estimate $\hat{\delta}$

for b *in range* B **do**

Resample $\mathcal{T}^{*b}$ by sampling with replacement from $\mathcal{T}$

Fit $\widehat{\Pr}_\mathbb{P}^{*b}(A = 1 \mid Z) \in \mathcal{F}_A$ on $\mathcal{T}^{*b}$

Resample $\mathcal{P}^{*b}$ by sampling with replacement from $\mathcal{P}$

Compute

$$\hat{\delta}^{*b} \equiv \frac{\sum_{j=1}^{n_\mathcal{P}} \widehat{\Pr}_\mathbb{P}^{*b}\left(A = 1 \mid Z = Z_j^{*b}\right) Y_j^{*b}}{\sum_{j=1}^{n_\mathcal{P}} \widehat{\Pr}_\mathbb{P}^{*b}\left(A = 1 \mid Z = Z_j^{*b}\right)} - \frac{\sum_{j=1}^{n_\mathcal{P}} \widehat{\Pr}_\mathbb{P}^{*b}\left(A = 0 \mid Z = Z_j^{*b}\right) Y_j^{*b}}{\sum_{j=1}^{n_\mathcal{P}} \widehat{\Pr}_\mathbb{P}^{*b}\left(A = 0 \mid Z = Z_j^{*b}\right)}$$

Output $(1 - \alpha)$-level percentile bootstrap confidence interval

$$\left(\hat{\delta}_B^{(\alpha/2)}, \hat{\delta}_B^{(1-\alpha/2)}\right)$$

where $\hat{\delta}_B^{(\alpha)}$ is the empirical α-percentile of the $\hat{\delta}^{*b}$

4.2 Simulations

We demonstrate empirically that the dual-bootstrap more accurately accounts for the overall uncertainty of imputed disparity estimates. We do so through a simple simulation in which both the training and primary populations follow the same data-generating process:

- A single proxy is drawn i.i.d. from a standard normal: $Z \sim \mathcal{N}(0, 1)$.
- Race is drawn i.i.d. from a Bernoulli distribution with probability logistic in Z: $A \mid Z \sim \text{Bern}\left[\exp(Z)/\{\exp(Z) + 1\}\right]$.
- The outcome Y is i.i.d. normal and linear in Z: $Y \mid Z \sim \mathcal{N}(5Z, 9)$.

In each simulation repetition, we draw (Z, A) tuples as the training dataset $\mathcal{T}$ and (Z, Y) tuples as the primary dataset $\mathcal{P}$. We fit a logistic regression of A on Z with $\mathcal{T}$, then apply it to $\mathcal{P}$ to

		Coverage Rate			Interval Width		
$\lvert\mathcal{T}\rvert$	$\lvert\mathcal{P}\rvert$	Dual-Bootstrap	Single-Bootstrap	Empirical	Dual-Bootstrap	Single-Bootstrap	Empirical
100	100	0.91	0.81	0.92	3.8	2.2	3.7
100	1000	0.94	0.66	0.95	3.1	0.7	3.0
100	5000	0.94	0.54	0.94	3.0	0.3	3.0
1000	100	0.89	0.87	0.90	2.4	2.2	2.4
1000	1000	0.93	0.80	0.93	1.2	0.7	1.2
1000	5000	0.97	0.67	0.97	1.0	0.3	1.0
5000	100	0.87	0.86	0.87	2.2	2.2	2.2
5000	1000	0.88	0.84	0.88	0.8	0.7	0.8
5000	5000	0.94	0.83	0.94	0.5	0.3	0.5

Table 1: Coverage rates and widths of 95% confidence intervals estimated using the dual-bootstrap, the single-bootstrap, and the empirical variance estimator for varying sample sizes of $\mathcal{T}$ and $\mathcal{P}$.

obtain our point estimate $\hat{\delta}$. We then apply the dual-bootstrap with 2,000 bootstrap iterations to estimate a 95% confidence interval for δ. For comparison, we also estimate what we call the "single-bootstrap" standard 95% confidence interval, in which the race probability estimates are treated as given and only $\mathcal{P}$ is resampled. We also estimate a 95% confidence interval based on the empirical variance estimator implied by Z-estimation theory using the `geex` package.

Table 1 reports the resulting coverage rates of the three types of confidence intervals over 500 simulation repetitions for various sample sizes of $\mathcal{T}$ and $\mathcal{P}$. We note four trends. First, the coverage rate of the single-bootstrap is worst when $\mathcal{T}$ is small and best, though still inadequate, when it is large. This reflects the effect of measurement uncertainty on the variance of the ultimate disparity estimator in this specific simulation setup; as $\mathcal{T}$ increases, the variability of the imputations decreases and becomes less influential. Second, for any size of $\mathcal{T}$, the single-bootstrap's coverage rate decreases as $\mathcal{P}$ increases. This suggests that the importance of measurement uncertainty can amplify with the number of imputations required. Third, the dual-bootstrap provides better coverage than the single-bootstrap, but it requires sufficiently large sample sizes to achieve the desired 95% rate. This reflects the fact that the theoretical properties of the bootstrap take effect asymptotically, as we discuss in Appendix C.1. Finally, the empirical variance estimator and the dual-bootstrap behave nearly identically; this is consistent with the theoretical results of Appendix C.1.

5 Special Case: BISG

In this section, we adapt the dual-bootstrap to the BISG algorithm, which is commonly used in empirical applications as the race probability model. The key challenge to applying our dual-bootstrap approach in this setting is that the Census Bureau does not generally make publicly available the training dataset $\mathcal{T}$ on which the BISG prior probabilities are based. We suggest one way to overcome this constraint while still upholding the fundamental principle animating the generic dual-bootstrap procedure of Section 4.1. We then apply our method to assess how prominent the uncertainty of BISG imputations are in practice. We conclude that, with some notable exceptions, the variability of BISG imputations is generally negligible in practice; bias, not

variance, is likely the primary source of error in BISG.

5.1 BISG-Specific Procedure

BISG imputes race by naively applying Bayes' Theorem to Census Bureau estimates of the racial composition of people by surname and geolocation. In this context, A is typically categorical instead of binary, containing all race categories defined by the Census Bureau; $Z = (S, G)$, where S is a categorical variable denoting the individual's surname and G is a categorical variable denoting the individual's geolocation; and the race probability model is

$$\widehat{\Pr}_{\mathbb{P}}(A = a' \mid S = s, G = g) \equiv \frac{\widehat{\Pr}_{\mathbb{P}}(A = a' \mid G = g)\,\widehat{\Pr}_{\mathbb{P}}(S = s \mid A = a')}{\sum_a \widehat{\Pr}_{\mathbb{P}}(A = a \mid G = g)\,\widehat{\Pr}_{\mathbb{P}}(S = s \mid A = a)},$$

where the prior probabilities on the right are parameter estimates based on Census Bureau surveys. Specifically, researchers commonly use as $\widehat{\Pr}_{\mathbb{P}}(A \mid G)$ the Census Bureau's American Community Survey (ACS) estimates of the number of people of each race residing in each geolocation. And they compute $\widehat{\Pr}_{\mathbb{P}}(S \mid A)$ based on the Census Bureau's 2010 table of frequently occurring surnames. For BISG, then, the training dataset $\mathcal{T}$ is the microdata—i.e., individual-level survey responses—that the Census Bureau collects to generate the ACS and surname estimates. The standard dual-bootstrap procedure outlined in Algorithm 1 thus calls for the analyst to resample the microdata with replacement and recompute the racial composition of each geolocation and surname.[3]

The challenge, however, is that $\mathcal{T}$ is inaccessible. The Census Bureau generally does not publish microdata for privacy reasons. This is not an issue for the microdata on which the surname-race probabilities are based: Since the surname table is just a raw tabulation, we can still essentially reconstruct the microdata that produced it and resample from it.[4] But the same is not true of the microdata on which the ACS race-by-geolocation estimates are based. The ACS estimates of the racial composition of geographic areas are not raw tabulations of survey microdata; rather, they are produced by re-weighting the microdata to adjust for factors like probability of selection in an unknown and presumably complex way.[5] Thus, a solution for the race-by-geolocation probabilities is required.

The key intuition behind the solution we propose is that we seek to resample the microdata $\mathcal{T}$ and recompute the prior $\widehat{\Pr}_{\mathbb{P}}(A \mid G)$ only as a means of, essentially, drawing from the approximate sampling distribution of $\widehat{\Pr}_{\mathbb{P}}(A \mid G)$. If we can approximate the sampling distribution of $\widehat{\Pr}_{\mathbb{P}}(A \mid G)$ in some other way, then we can just draw from it directly—no resampling or model-refitting needed. Fortunately, the Census Bureau suggests and endorses a way of estimating key parameters of the

[3] The surname table is a raw tabulation of data from the decennial census, which covers the entire population of the United States. Thus, depending on the population for which race-specific outcomes are to be estimated, resampling the data that produced the surname table might be unnecessary. Nonetheless, we outline our procedure to include resampling for two reasons. First, it could be appropriate to do so, depending on the estimand. Second, race-by-name probabilities are sometimes sourced from data that are properly characterized as a sample, rather than the entire population. For example, Imai et al. (2022) use voting records from a handful of states to estimate the race-by-name probabilities. In such cases, resampling would likely be appropriate.

[4] Such a reconstruction is necessarily imprecise since the surname table aggregates all surnames held by fewer than 100 individuals into a single "Other" category. Our reconstruction of the microdata can never recover these surnames. We leave a detailed examination of the significance of this issue to future work.

[5] Some ACS microdata are available, but only at levels of geographic granularity that are too low to be useful in most applications. Such microdata are also incomplete—they consist of only about two-thirds of the records used to produce the ACS estimates—and thus might not be any more amenable to direct resampling.

sampling distribution even without microdata. We outline the approach below, with further details available in Appendix A.

As an initial matter, we center the sampling distribution of $\widehat{\Pr}_{\mathbb{P}}(A \mid G)$ at the published ACS estimate, which we denote by $\hat{\mu}_G$. Then, to estimate the covariance of this sampling distribution, which we denote by $\hat{\Sigma}_G$, we use the publicly available ACS variance replicates. These are 80 "pseudo-estimates" of the racial composition of each geolocation, which we denote by $\hat{\mu}_G^{\dagger r}$ for $r = 1, \ldots, 80$. The Census Bureau uses them to estimate variances via the successive differences replication (SDR) method. The variance replicates are not bootstrap statistics; they have "no other use [beyond calculating SDR variances] and no independent meaning" (Census Bureau, 2022). So we cannot directly apply the dual-bootstrap algorithm to them. Instead, we use the variance replicates to estimate the covariance of the race-by-geolocation probability estimates based on the formula prescribed by the Census Bureau:

$$\hat{\Sigma}_G \equiv \frac{4}{80} \sum_{r=1}^{80} \left(\hat{\mu}_G^{\dagger r} - \hat{\mu}_G\right) \left(\hat{\mu}_G^{\dagger r} - \hat{\mu}_G\right)^{\intercal}.$$

When ACS estimates that there are zero people of a given race in a geographic area, all associated variance replicates are zero. In such "zero-count" cases, we follow the Census Bureau's recommendation not to use the above formula; instead, we assume the estimate has zero covariance with the other estimates and essentially derive the variance from the Census Bureau's estimated margin of error (Census Bureau, 2022). Appendix A.1 describes the procedure in more detail. As we discuss in Section 5.2, the choice to account for uncertainty in zero-count geolocations can be influential in specific circumstances.

Finally, we assume that the sampling distribution of $\widehat{\Pr}_{\mathbb{P}}(A \mid G)$ is normal, so the parameter estimates $(\hat{\mu}_G, \hat{\Sigma}_G)$ fully specify the sampling distribution as $\mathcal{N}(\hat{\mu}_G, \hat{\Sigma}_G)$. We leave generalizations of the form of the sampling distribution to future work. See Appendix A.2 for more discussion.

With the sampling distribution of $\widehat{\Pr}_{\mathbb{P}}(A \mid G)$ in hand, our modified dual-bootstrap routine can be executed. Algorithm 2 states the modified implementation. The key distinction from Algorithm 1 is that the bootstrap race-by-geolocation probability estimate $\widehat{\Pr}_{\mathbb{P}}^{*b}(A \mid G)$ is computed by drawing directly from the sampling distribution $\mathcal{N}(\hat{\mu}_G, \hat{\Sigma}_G)$ instead of by refitting on resampled microdata. As with Algorithm 1, the resampling steps in this algorithm use simple resampling with replacement, but more complex forms of resampling may be appropriate (Owen, 2007; Derby et al., 2024).

As mentioned in Section 2, recent work has proposed an alternative approach: setting the sampling distribution to the posterior distribution obtained by updating an assumed prior with the ACS race-by-geolocation estimates (Derby et al., 2024). We do not adopt this approach because, as discussed above, the Census Bureau offers its own account of the uncertainty of its estimates. This uncertainty is multifaceted—it includes considerations like the Census Bureau's sampling scheme and survey nonresponse, as well as adjustments the Census Bureau has made to account for them—and inscrutable to the general public. So, rather than impose our own model for this uncertainty by specifying an ultimately arbitrary prior, we prefer to use the model offered by the Census Bureau, which is best-positioned to develop one.

One potential advantage to assuming a prior distribution instead of using the Census Bureau's uncertainty model is that it can accommodate a superpopulation framework for the race probability model. For example, a researcher can assume an abstract superpopulation of which each year's demographic composition is a sample. The prior distribution characterizes the superpopulation of race probabilities and is updated by a given year's observed demographic composition.

Algorithm 2: BISG Dual-Bootstrap

Data: ACS Estimate $\hat{\mu}$, ACS Covariance Matrix Estimate $\hat{\Sigma}$, Surname Table $\mathcal{S}$, Primary Dataset $\mathcal{P} = \{S_j, G_j, Y_j\}_{j=1}^{n_\mathcal{P}}$, Number of Bootstrap Draws $B \in \mathbb{N}$, Level $\alpha \in [0, 1]$, Race Groups a' and a''

Result: Confidence interval for the demographic disparity estimate $\hat{\delta}$

for *b in range B* **do**

Resample $\mathcal{S}^{*b}$ by sampling with replacement from $\mathcal{S}$ `// optional; see Footnote 3`

for *s in* $\{S_j\}_{j=1}^{n_\mathcal{P}}$*, a in supp(A)* **do**

Compute $\widehat{\Pr}_\mathbb{P}^{*b}(S = s \mid A = a)$ from $\mathcal{S}^{*b}$ `// optional; see Footnote 3`

for *g in* $\{G_j\}_{j=1}^{n_\mathcal{P}}$ **do**

Sample $\widehat{\Pr}_\mathbb{P}^{*b}(A \mid G = g) \sim \mathcal{N}\left(\hat{\mu}_g, \hat{\Sigma}_g\right)$

Resample $\mathcal{P}^{*b}$ by sampling with replacement from $\mathcal{P}$

for *j in range* $n_\mathcal{P}$ **do**

Compute

$$\widehat{\Pr}_\mathbb{P}^{*b}\left(A = a' \mid S = S_j^{*b}, G = G_j^{*b}\right) \equiv \frac{\widehat{\Pr}_\mathbb{P}^{*b}\left(A = a' \mid G = G_j^{*b}\right) \widehat{\Pr}_\mathbb{P}^{*b}\left(S = S_j^{*b} \mid A = a'\right)}{\sum_a \widehat{\Pr}_\mathbb{P}^{*b}\left(A = a \mid G = G_j^{*b}\right) \widehat{\Pr}_\mathbb{P}^{*b}\left(S = S_j^{*b} \mid A = a\right)},$$

$$\widehat{\Pr}_\mathbb{P}^{*b}\left(A = a'' \mid S = S_j^{*b}, G = G_j^{*b}\right) \equiv \frac{\widehat{\Pr}_\mathbb{P}^{*b}\left(A = a'' \mid G = G_j^{*b}\right) \widehat{\Pr}_\mathbb{P}^{*b}\left(S = S_j^{*b} \mid A = a''\right)}{\sum_a \widehat{\Pr}_\mathbb{P}^{*b}\left(A = a \mid G = G_j^{*b}\right) \widehat{\Pr}_\mathbb{P}^{*b}\left(S = S_j^{*b} \mid A = a\right)}$$

Compute

$$\hat{\delta}^{*b} \equiv \frac{\sum_{j=1}^{n_\mathcal{P}} \widehat{\Pr}_\mathbb{P}^{*b}\left(A = a' \mid S = S_j^{*b}, G = G_j^{*b}\right) Y_j^{*b}}{\sum_{j=1}^{n_\mathcal{P}} \widehat{\Pr}_\mathbb{P}^{*b}\left(A = a' \mid S = S_j^{*b}, G = G_j^{*b}\right)} - \frac{\sum_{j=1}^{n_\mathcal{P}} \widehat{\Pr}_\mathbb{P}^{*b}\left(A = a'' \mid S = S_j^{*b}, G = G_j^{*b}\right) Y_j^{*b}}{\sum_{j=1}^{n_\mathcal{P}} \widehat{\Pr}_\mathbb{P}^{*b}\left(A = a'' \mid S = S_j^{*b}, G = G_j^{*b}\right)}$$

Output $(1 - \alpha)$-level percentile bootstrap confidence interval

$$\left(\hat{\delta}_B^{(\alpha/2)}, \hat{\delta}_B^{(1-\alpha/2)}\right)$$

where $\hat{\delta}_B^{(\alpha)}$ is the empirical α-percentile of the $\hat{\delta}^{*b}$

To our knowledge, the Census Bureau's uncertainty model cannot accommodate such a superpopulation framework because it only models uncertainty arising from its survey sampling procedure. In our view, however, this limitation is significant only if the superpopulation parameters of the race probability model are of independent interest. In most applications—like the study of racial disparities—the race probability model is merely nuisance.

Consider, for example, the task of estimating racial disparities in tax audit rates.[6] It is true

[6] Elzayn et al. (2023) apply our proposed approach to this setting to quantify the uncertainty associated with their tax audit disparity estimates.

that researchers might be interested in the racial disparity at the superpopulation level (e.g., as a parameter of an abstract data-generating process that produces the observed tax audit rates by race each year). But, to estimate such a disparity, they necessarily use observed tax audit data from certain, well-defined years. Suppose that the race of the taxpayer in each audit decision is unavailable, so the researchers impute it using BISG with ACS data from the relevant years. Quantifying the uncertainty of any individual race imputation and how it affects the uncertainty of the final disparity estimate requires only an understanding of the error of the year-specific BISG model used for that imputation; it does not require reference to any BISG (or other race probability) model at the superpopulation level.

5.2 BISG Simulations

We show via simulation that the uncertainty of BISG imputations generally has little effect on the variance of the resulting racial disparity estimate. For this simulation, we use the 2017-2021 ACS 5-year estimates of the racial composition of each census block group and the 2010 Census Bureau surname table. We use the following data-generating process for both the training and primary populations:

- The proxy tuple (G, S) is sampled i.i.d. from the marginal census block group and surname frequencies given by the ACS estimates and the Census Bureau surname table, excluding tuples where the racial composition of the surname is withheld by the Census Bureau and tuples where the surname and census block group have mutually exclusive racial compositions.
- The outcome Y is i.i.d. standard normal, independent of G, S, and A: $Y \sim \mathcal{N}(0, 1)$.

Because the race probability model is estimated using BISG and the outcome is independent of race, no concrete race indicators need to be generated for this simulation. In each of 100 simulation repetitions, we draw 1,000 (G, S, Y) tuples as the primary dataset $\mathcal{P}$. On this dataset, we estimate the average outcome $\hat{\delta}$ for each race using BISG-estimated probabilities. We then estimate the standard error using both the dual-bootstrap and the single-bootstrap.

Table 2 reports the results. Overall, accounting for measurement uncertainty in this setting barely affects the resulting standard errors.[7] Figure 2 offers one explanation: For most units, the bootstrap standard error of the posterior race probability is low, close to 0.05 on average. For comparison, the bootstrap standard errors of the race probabilities in the machine learning simulation of Section 4.2 are about 0.23, nearly five times larger. The standard errors are much smaller here likely because the BISG model is fairly rigid, and the ACS and surname prior probabilities that parameterize it are based on millions of individual-level training points. The notable exception in Table 2 is the American Indian and Alaska Native group, for which the dual-bootstrap standard error is substantially less than the single-bootstrap standard error. As a theoretical matter, it might generally be possible for standard errors to decrease after properly accounting for measurement error; we do not prove so in our specific setting, but Reifeis and Hudgens (2022) show that this can occur in the closely related setting of IPW estimation of the average treatment effect on the treated. However, we believe that the specific reduction observed here is due to our handling of zero counts and our specific data-generating process, as described in more detail below.

[7] Consistent with this finding, Elzayn et al. (2023) obtain only slightly larger standard errors when they apply our method to estimate the uncertainty of their BISG-based estimates of tax audit disparities by race.

	Average Standard Error	
Race Group	Dual-Bootstrap	Single-Bootstrap
American Indian and Alaska Native	0.11 (0.02)	0.18 (0.06)
Asian and Pacific Islander	0.11 (0.01)	0.11 (0.01)
Black	0.07 (0.00)	0.07 (0.00)
Hispanic	0.06 (0.00)	0.06 (0.00)
Multiracial	0.07 (0.01)	0.06 (0.01)
White	0.04 (0.00)	0.04 (0.00)

Table 2: Average standard error of the estimated average outcome of each race, as estimated by the dual-bootstrap and the single-bootstrap when BISG is used for imputation. Standard deviations over the simulation repetitions are in parentheses. The dual-bootstrap's average estimate of the standard error is the same as that of the single-bootstrap except for the American Indian and Alaska Native group, for which it is lower. We offer one explanation for this in Figures 3-4 and the associated discussion below.

Although measurement uncertainty appears to be of nominal significance marginally over the entire population of the United States, we find evidence that it, and the way it is modeled, can be influential in certain situations. To illustrate, we rerun the above simulations for each state—that is, we sample census block groups from the marginal frequencies given by the ACS estimates within each state. Figure 3 shows the state-by-state results for three racial groups that we highlight here because of their particularly prominent trends; Appendix B.1 contains corresponding figures for other racial groups. In most states, accounting for measurement uncertainty has essentially no effect on the uncertainty of the average outcome estimate for White people; it increases the uncertainty of the average outcome estimate for multiracial people; and it decreases the uncertainty of the average outcome estimate for American Indians and Alaska Natives.

We believe that the key to understanding these seemingly incompatible phenomena lies primarily in (1) the prevalence of each race group overall, (2) the geographic concentration of certain race groups, and (3) the distribution of the outcome among race groups. White people are much more prevalent than multiracial people overall: The 2017-2021 ACS data we use estimates that 59% of the population is White, while 3% are multiracial. Thus, ACS estimates of the proportion of White people in each census block group are more precise—in other words, have less measurement uncertainty—than estimates of the proportion of multiracial people. This explains why measurement uncertainty increases the uncertainty of the average outcome estimate for multiracial people more than it does for White people.

American Indians and Alaska Natives are even less prevalent than multiracial people overall: The 2017-2021 ACS data we use estimates that about 0.6% of the population is American Indian or Alaska Native. But accounting for measurement uncertainty generally *decreases* the uncertainty of the average outcome estimate in these simulations because they are a geographically concentrated minority: As Figure 3 shows, the ACS estimates that there are zero American Indians and Alaska Natives in most census block groups in most states. As discussed in Section 5.1, we use the ACS estimated margin of error instead of the ACS variance replicates to approximate the sampling distribution of such estimates since it is possible that American Indians and Alaska Natives in fact reside in those block groups and were simply not sampled by ACS. Accounting for measurement uncertainty in this way gives nonzero weight to people who otherwise would have none. When the

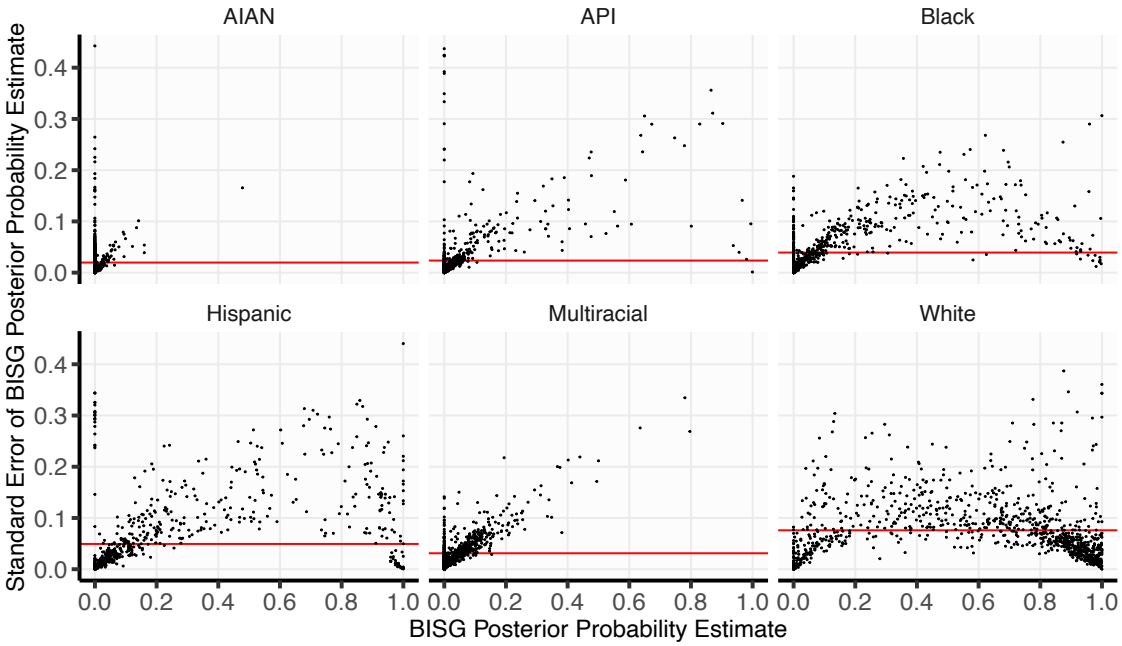

Figure 2: Bootstrap standard error of the BISG-estimated probability of being a given race plotted against the BISG-estimated probability in one simulation repetition. The horizontal red line indicates the average bootstrap standard error. "AIAN" is the abbreviation for American Indian and Alaska Native, and "API" is the abbreviation for Asian and Pacific Islander.

outcomes of these people are informative of the average outcome of American Indians and Alaska Natives—as they are in this simulation, since all units have outcomes drawn from a standard normal distribution—giving them nonzero weight increases the effective sample size and thus decreases the standard error of the average outcome estimate. This phenomenon likely explains the overall decrease in the standard error for American Indians and Alaska Natives reported in Table 2.

We illustrate some of these dynamics through an additional simulation focused on the American Indian and Alaska Native population in New Mexico. In this simulation, we generate synthetic states by taking ACS estimates from New Mexico and altering (1) the total prevalence of American Indian and Alaska Native people in the state and (2) the percentage of census block groups in the state in which zero American Indian and Alaska Native people are estimated to reside. Appendix B.2 describes this process in detail. We then rerun the previous simulation on each synthetic state. As Figure 4 shows, accounting for measurement uncertainty decreases standard errors the less prevalent and the more concentrated the race group is.

These results for the American Indian and Alaska Native group demonstrate that a proper accounting of uncertainty in zero-count geolocations can, in specific circumstances, be influential. Our choice to take at face value the Census Bureau's margins of error when quantifying the uncertainty of the estimated average outcomes of a race group gives influence to people in geolocations where ACS estimates no people of that race group reside. This motivation to properly leverage data from zero-count geolocations also underlies some of the work of Imai et al. (2022)—though they focus on

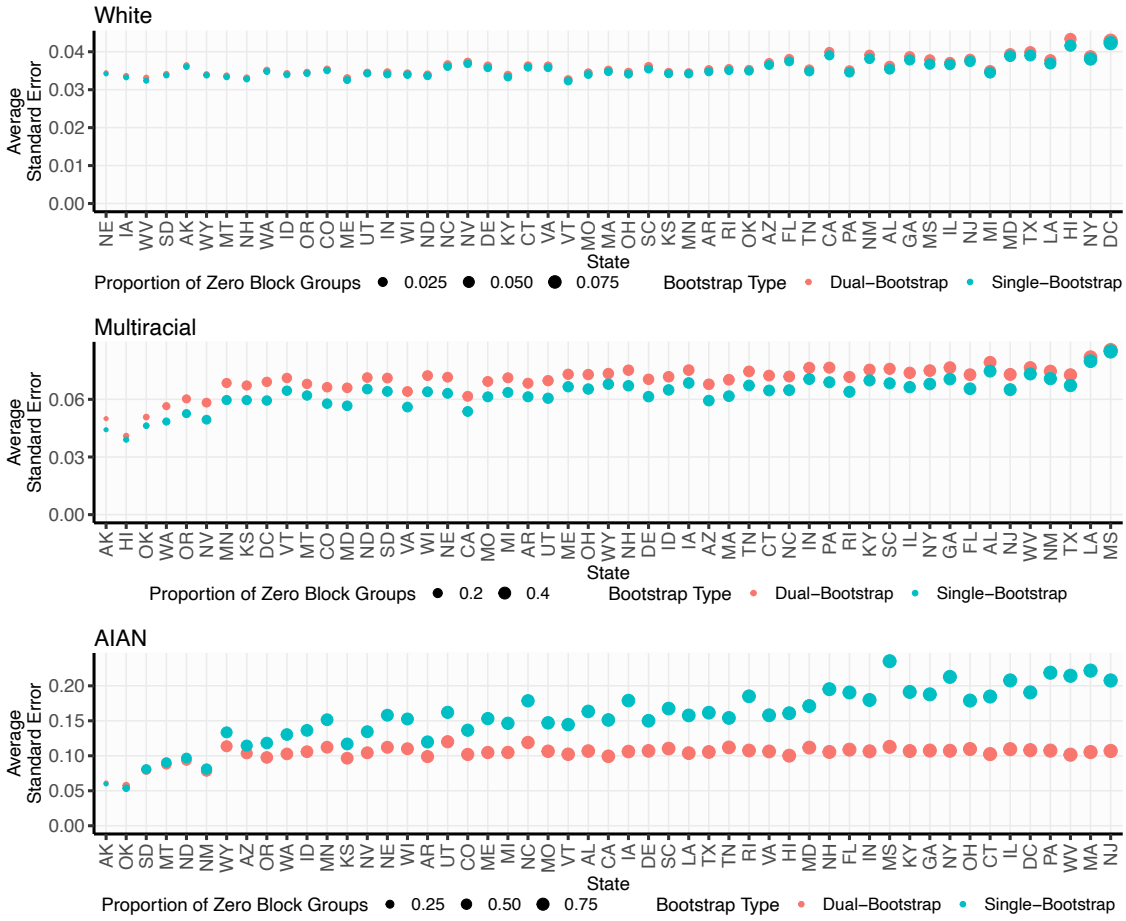

Figure 3: Dual-bootstrap and single-bootstrap standard errors of the estimated average outcome for the White, Multiracial, and American Indian and Alaska Native (AIAN) race groups in each state. The states are ordered by the proportion of census block groups in which the American Community Survey estimates there are zero people of the given race.

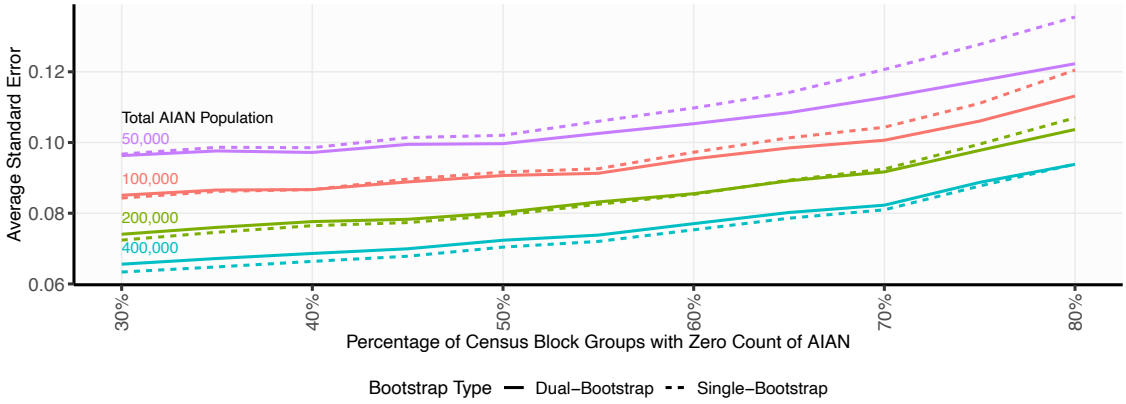

Figure 4: Dual-bootstrap and single-bootstrap standard errors of the estimated average outcome for the American Indian and Alaska Native (AIAN) race group in New Mexico when American Community Survey estimates of the total counts and geographic concentrations of AIAN people in the state are artificially altered.

how accounting for the migration of racial minorities since the last decennial census can improve imputation accuracy, whereas we focus on how accounting for the possible nonselection of racial minorities in ACS sampling can improve uncertainty quantification of downstream estimates. We emphasize that in practice, however, properly accounting for uncertainty in zero-count geolocations might not have as drastic or counterintuitive an effect as shown here, where the outcomes of all simulated units are equally informative for all race groups. On the contrary, it might in some cases increase standard errors if the outcomes of people in zero-count geolocations are substantially different from those of the target race group. In other cases, it might have no effect on balance.

Taken together, the phenomena identified in the simulations above highlight that, although the uncertainty of BISG imputations might not be substantial in studies of the general U.S. population, properly accounting for it in studies of particular geographic areas or demographic groups can be important to ensuring that the resulting inference is neither conservative nor anti-conservative. More generally, our finding is consistent with a broader literature on the challenges of race imputation for certain demographic groups (e.g., Imai et al., 2022).

6 Application

We apply the dual-bootstrap to study racial disparities in health outcomes using the American Family Cohort (AFC) dataset (Vala et al., 2023). The dataset contains electronic health records from the primary care visits of patients in the United States. Relevant features for our purposes include patient geolocation; first name; surname; self-reported race, which are provided as mapped to White, Black, Hispanic, Asian and Pacific Islander, American Indian and Alaska Native, Multiracial, and Other; and indicators for the diagnosis of asthma, obesity, and diabetes at any point during the time period covered by the dataset. We downsample the data due to computational constraints by taking a stratified random sample of 100,000 patients with the same race proportions as the full dataset. Although we do not adopt them here, general steps can be taken in practice to

improve the computational efficiency of the bootstrap (e.g., Kleiner et al., 2014).

We preprocess the dataset as follows. First, we produce race proxies by converting categorical geolocation, first name, and surname data into numerical race probability estimates. Specifically, we convert the surnames into "prior probability" features by computing the probability of each of the six race categories (excluding Other) given surname based on the Census Bureau's 2010 surname table. And we convert the first names and geolocations into "update" features by computing the probability of the first name or geolocation given each of the six race categories. We use mean imputation for any missing geolocation, first name, or surname probabilities and include a binary missingness indicator for each as a separate feature. Second, we randomly split the data into a primary dataset of size 20,000 and a training dataset of size 80,000. We mask self-reported race in the primary dataset and health outcomes in the training dataset.

On the training dataset, we fit a random forest of patients' races on the processed features defined above. Although these same features could be run through BISG to output race probabilities, we choose to use a random forest because Cheng et al. (2023) find that it produces more accurate estimates in this dataset. We allow for slight tuning of the random forest hyperparameters. Specifically, we perform a grid search of the following hyperparameters using 5-fold cross-validation.

Number of Trees: 100.

Maximum Tree Splits: 20, 50, 100.

Proportion of Features Per Split ($p = 21$)**:** $\sqrt{p}$, $0.5p$, $0.75p$.

Minimum Number of Units to Initiate Split: 10, 25, 100.

We then apply the random forest to the primary dataset to estimate patients' race probabilities. With those estimates in hand, we estimate the incidence rates of asthma, obesity, and diabetes for each race in the primary dataset. To compute confidence intervals for each estimate, we use both the single-bootstrap, which retains the original random forest model, and the dual-bootstrap, which refits a new random forest model on bootstrapped draws of the training dataset. We run 100 bootstrap iterations.

Figure 5 shows the results. In general, the single-bootstrap understates the uncertainties of the prevalence estimates compared to the dual-bootstrap. But this trend is not uniform across races. For Asian and Pacific Islander, Black, Hispanic, and White patients, the widths of the dual-bootstrap and single-bootstrap confidence intervals are essentially the same. On the other hand, the dual-bootstrap confidence intervals are substantially wider than the single-bootstrap confidence intervals for American Indian and Alaska Native, Multiracial, and Other patients—in some cases doubly so. This appears largely to be because those patients appear infrequently in the data on which the random forest model was trained, so their race probability estimates are more variable. As Figure 6 shows, the dual-bootstrap confidence interval widths are nearly identical to the single-bootstrap ones when we alter our downsample so that all races are equally represented in the training dataset.

7 Discussion

We propose a dual-bootstrap procedure to more accurately account for the uncertainty of race imputations that are subsequently used to estimate racial disparities and other race-specific outcomes.

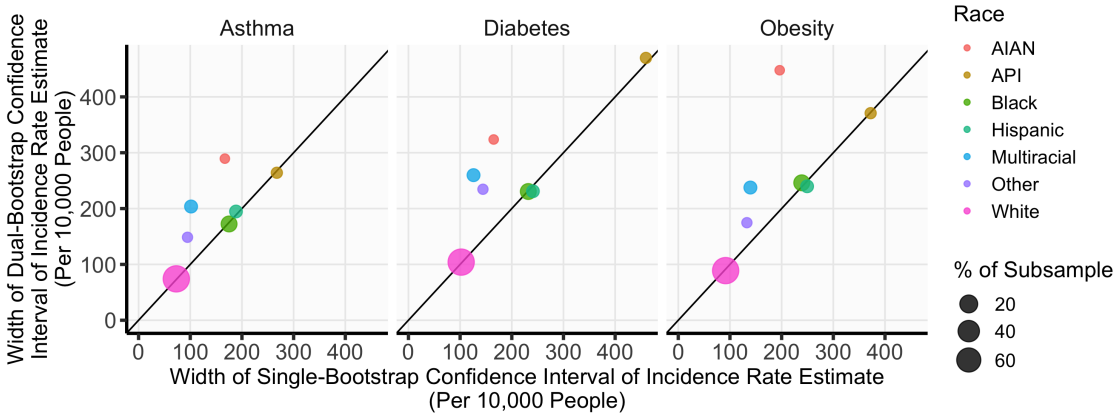

Figure 5: Widths of dual-bootstrap vs. single-bootstrap confidence intervals of the estimated prevalence of certain health conditions by race. This analysis was conducted on a 100,000-unit subsample of the American Family Cohort population with the same racial composition as the full population. Points are sized by the proportion of units in the subsample that are of the given race group. "AIAN" is the abbreviation for American Indian and Alaska Native, and "API" is the abbreviation for Asian and Pacific Islander.

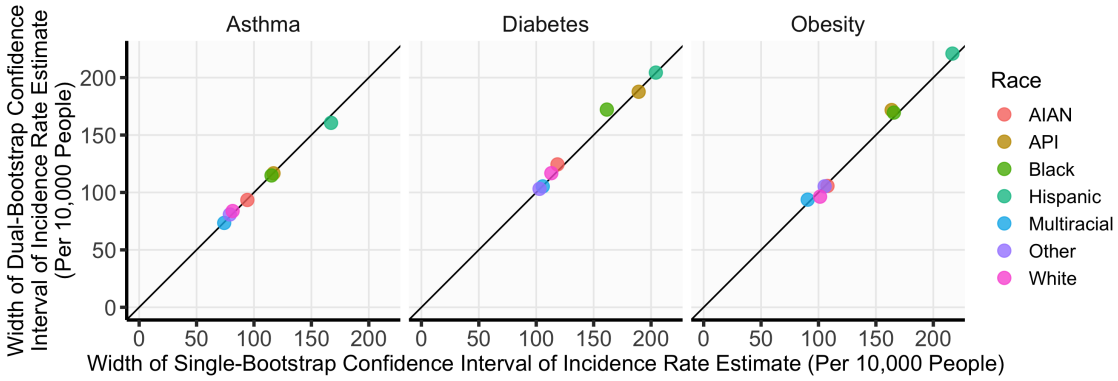

Figure 6: Widths of dual-bootstrap vs. single-bootstrap confidence intervals of the estimated prevalence of certain health conditions by race. This analysis was conducted on a 100,000-unit subsample of the American Family Cohort population where each race was equally represented in the training dataset. "AIAN" is the abbreviation for American Indian and Alaska Native, and "API" is the abbreviation for Asian and Pacific Islander.

Our method is straightforward to implement, although complications can arise when the underlying data used to train the race probability model are unavailable. We offer one way of overcoming such difficulties in the specific case of BISG, an imputation model that is often parameterized by ACS race-by-geolocation estimates that are based on undisclosed microdata. Our simulation results suggest that the measurement uncertainty of BISG generally does not impact the uncertainty of downstream estimates, likely because it is a fairly rigid model with a relatively large sample size underpinning its parameter estimates. But it can be significant for specific race groups in specific geographies, with the potential to increase or decrease the standard error of downstream estimates, as our state-by-state results show. And we emphasize that despite its overall low variability—or perhaps because of it—BISG still suffers from bias, as others have shown and sought to improve (e.g., Imai et al., 2022).

We see several opportunities for future work in this direction. Most immediately, an investigation of the theoretical properties of the dual-bootstrap when the race probability model falls outside the Z-estimator framework could be informative. On the practical side, a closer examination and improvement of some of the design choices made in our adaptation of the dual-bootstrap to BISG—such as our choice to use a normal distribution and other choices outlined in Appendix A—could produce more accurate inference. Any changes or additions by the Census Bureau to the data products it publishes could help or hinder these efforts. We also see broader opportunities in this space. For example, the development of prospective heuristics for study design akin to a power analysis might prove useful to applied researchers. In some settings, researchers have a choice between analyzing a small dataset where race is observed and analyzing a larger dataset where race must be imputed. The need to account for measurement uncertainty—which, as shown in this paper, can be substantial or not—only complicates this choice. A set of heuristics that allows researchers to prospectively approximate the standard error that would result from each choice given certain parameters like the sample sizes of the datasets, the accuracy of the imputations, and the variability of the outcome might help with the decision.

Acknowledgements

We thank Stanford's Center for Population Health Sciences for data, and Isabel Gallegos, Cameron Guage, Dan Soriano, Melody Huang, Peng Ding, Thomas Hertz, Peter Hull, Qiwei Lin, Mark Loewenstein, and participants in the NBER CRIW Race, Ethnicity, and Economic Statistics for the 21st Century conference for helpful comments. This material is based upon work supported by the National Science Foundation under Grant No. 1745640.

References

Angrist, J. D. and Krueger, A. B. (1992). The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples. *Journal of the American statistical Association*, 87(418):328–336.

Angrist, J. D. and Krueger, A. B. (1995). Split-Sample Instrumental Variables Estimates of the Return to Schooling. *Journal of Business & Economic Statistics*, 13(2):225–235.

Avenancio-León, C. F. and Howard, T. (2022). The Assessment Gap: Racial Inequalities in Property Taxation. *The Quarterly Journal of Economics*, 137(3):1383–1434.

Azin, A., Hirpara, D. H., Doshi, S., Chesney, T. R., Quereshy, F. A., and Chadi, S. A. (2020). Racial Disparities in Surgery: A Cross-Specialty Matched Comparison Between Black and White Patients. *Annals of Surgery Open*, 1(2):e023.

Berdejó, C. (2018). Criminalizing Race: Racial Disparities in Plea-Bargaining. *Boston College Law Review*, 59:1187–1249.

Brown, D. A. (2022). *The Whiteness of Wealth: How the Tax System Impoverishes Black Americans—And How We Can Fix It.* Crown.

Brown, D. P., Knapp, C., Baker, K., and Kaufmann, M. (2016). Using Bayesian Imputation to Assess Racial and Ethnic Disparities in Pediatric Performance Measures. *Health Services Research*, 51(3):1095–1108.

Buolamwini, J. and Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR.

Census Bureau (2022). Documentation for the 2017-2021 Variance Replicate Estimates Tables. Technical report.

Chen, J., Kallus, N., Mao, X., Svacha, G., and Udell, M. (2019). Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 339–348, New York, NY, USA. Association for Computing Machinery.

Cheng, L., Gallegos, I. O., Ouyang, D., Goldin, J., and Ho, D. E. (2023). How Redundant are Redundant Encodings? Blindness in the Wild and Racial Disparity when Race Is Unobserved. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 667–686, New York, NY, USA. Association for Computing Machinery.

Derby, E., Dowd, C., and Mortenson, J. (2024). Constructing Confidence Intervals for BIFSG Disparity Estimates.

Elliott, M. N., Morrison, P. A., Fremont, A., McCaffrey, D. F., Pantoja, P., and Lurie, N. (2009). Using the Census Bureau's Surname List to Improve Estimates of Race/Ethnicity and Associated Disparities. *Health Services and Outcomes Research Methodology*, 9:69–83.

Elzayn, H., Smith, E., Hertz, T., Ramesh, A., Fisher, R., Ho, D. E., and Goldin, J. (2023). Measuring and Mitigating Racial Disparities in Tax Audits.

Fong, C. and Tyler, M. (2021). Machine Learning Predictions as Regression Covariates. *Political Analysis*, 29(4):467–484.

Fuller, W. A. (2009). *Measurement Error Models*. John Wiley & Sons.

Gelman, A., Fagan, J., and Kiss, A. (2007). An Analysis of the New York City Police Department's "Stop-and-Frisk" Policy in the Context of Claims of Racial Bias. *Journal of the American Statistical Association*, 102(479):813–823.

Imai, K., Olivella, S., and Rosenman, E. T. (2022). Addressing Census Data Problems in Race Imputation via Fully Bayesian Improved Surname Geocoding and Name Supplements. *Science Advances*, 8(49):eadc9824.

Kallus, N., Mao, X., and Zhou, A. (2022). Assessing Algorithmic Fairness with Unobserved Protected Class Using Data Combination. *Management Science*, 68(3):1959–1981.

Kim, J.-S., Gao, X., and Rzhetsky, A. (2018). RIDDLE: Race and Ethnicity Imputation from Disease History with Deep Learning. *PLOS Computational Biology*, 14(4):e1006106.

Kleiner, A., Talwalkar, A., Sarkar, P., and Jordan, M. I. (2014). A Scalable Bootstrap for Massive Data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(4):795–816.

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., and Goel, S. (2020). Racial Disparities in Automated Speech Recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer.

Labgold, K., Hamid, S., Shah, S., Gandhi, N. R., Chamberlain, A., Khan, F., Khan, S., Smith, S., Williams, S., Lash, T. L., and Collin, L. J. (2021). Estimating the Unknown: Greater Racial and Ethnic Disparities in COVID-19 Burden After Accounting for Missing Race/Ethnicity Data. *Epidemiology*, 32(2):157–161.

Mackey, K., Ayers, C. K., Kondo, K. K., Saha, S., Advani, S. M., Young, S., Spencer, H., Rusek, M., Anderson, J., Veazie, S., Smith, M., and Kansagara, D. (2021). Racial and Ethnic Disparities in COVID-19-Related Infections, Hospitalizations, and Deaths: A Systematic Review. *Annals of Internal Medicine*, 174(3):362–373.

Matsouaka, R. A., Liu, Y., and Zhou, Y. (2024). Overlap, Matching, or Entropy Weights: What Are We Weighting For? *Communications in Statistics – Simulation and Computation*.

McCaffrey, D. F. and Elliott, M. N. (2008). Power of Tests for a Dichotomous Independent Variable Measured with Error. *Health Services Research*, 43(3):1085–1101.

Miratrix, L. W., Sekhon, J. S., Theodoridis, A. G., and Campos, L. F. (2018). Worth Weighting? How to Think About and Use Weights in Survey Experiments. *Political Analysis*, 26(3):275–291.

Murray, J. S. (2018). Multiple Imputation: A Review of Practical and Theoretical Findings. *Statistical Science*, 33(2):142–159.

Ouimet, F. (2022). A Multivariate Normal Approximation for the Dirichlet Density and Some Applications. *Stat*, 11(1):e410.

Owen, A. B. (2007). The Pigeonhole Bootstrap. *Annals of Applied Statistics*, 1(2):386–411.

Reifeis, S. A. and Hudgens, M. G. (2022). On Variance of the Treatment Effect in the Treated When Estimated by Inverse Probability Weighting. *American Journal of Epidemiology*, 191(6):1092–1097.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.

Saul, B. C. and Hudgens, M. G. (2020). The Calculus of M-Estimation in R with geex. *Journal of Statistical Software*, 92(2):1–15.

Shu, D., Young, J. G., Toh, S., and Wang, R. (2021). Variance Estimation in Inverse Probability Weighted Cox Models. *Biometrics*, 77(3):1101–1117.

Stefanski, L. A. and Boos, D. D. (2002). The Calculus of M-Estimation. *The American Statistician*, 56(1):29–38.

Vala, A., Hao, S., Chu, I., Phillips, R. L., and Rehkopf, D. (2023). *The American Family Cohort (v12.2)*. Redivis, Stanford, CA.

Van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press.

Xue, Y., Harel, O., and Aseltine, R. (2019). Comparison of Imputation Methods for Race and Ethnic Information in Administrative Health Data. In *13th International Conference on Sampling Theory and Applications*, pages 1–4, Bordeaux, France. IEEE.

Yee, K., Hoopes, M., Giebultowicz, S., Elliott, M. N., and McConnell, K. J. (2022). Implications of Missingness in Self-Reported Data for Estimating Racial and Ethnic Disparities in Medicaid Quality Measures. *Health Services Research*, 57(6):1370–1378.

Zhang, Y. (2018). Assessing Fair Lending Risks Using Race/Ethnicity Proxies. *Management Science*, 64(1):178–197.