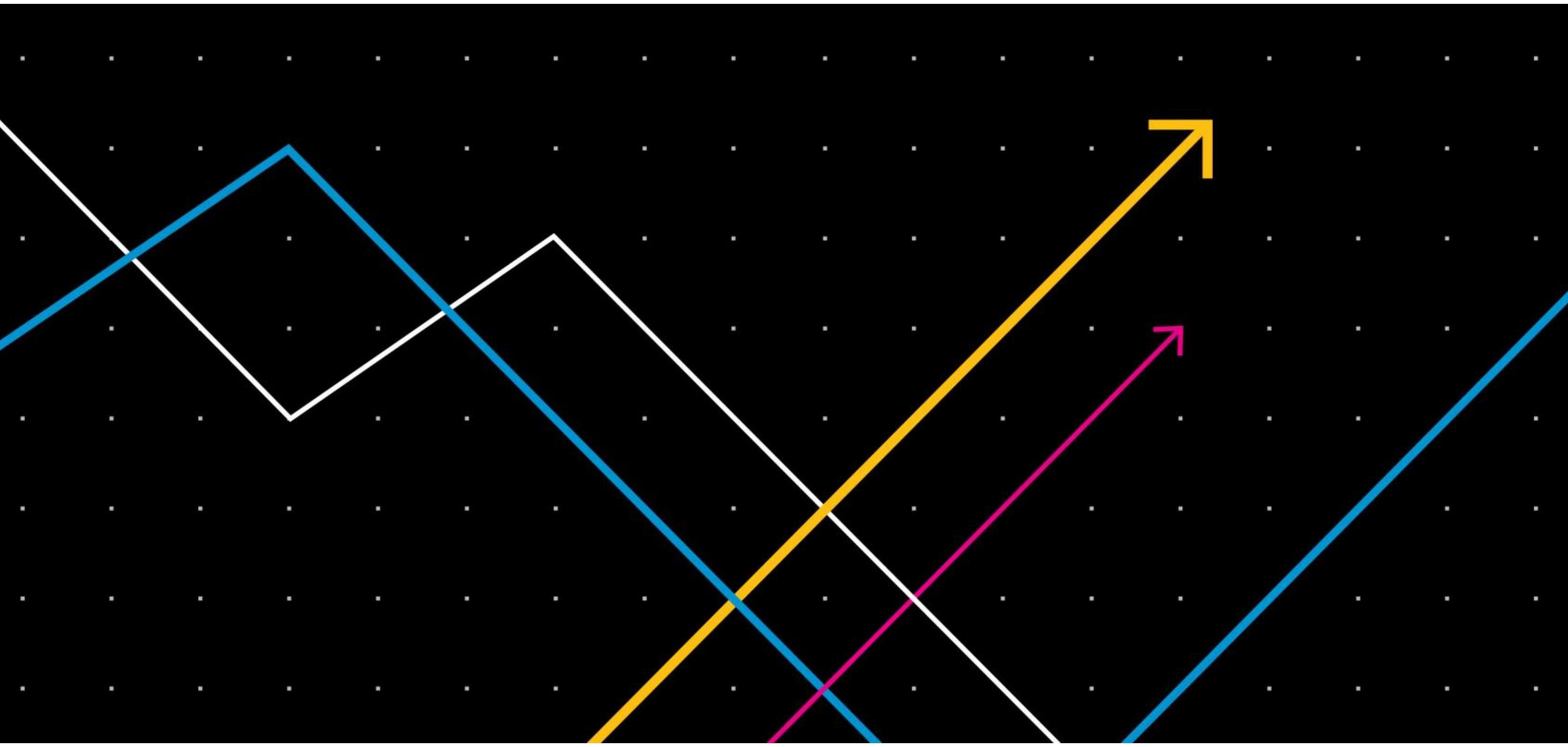


RESEARCH REPORT

Ethics and Empathy in Using Imputation to Disaggregate Data for Racial Equity

A Case Study Imputing Credit Bureau Data

Alena Stern *Ajjit Narayanan*

with Steven Brown, Graham MacDonald, LesLeigh Ford, and Shena Ashley

July 2021

URBAN INSTITUTE · ELEVATE · THE · DEBATE

ABOUT THE URBAN INSTITUTE

The nonprofit Urban Institute is a leading research organization dedicated to developing evidence-based insights that improve people's lives and strengthen communities. For 50 years, Urban has been the trusted source for rigorous analysis of complex social and economic issues; strategic advice to policymakers, philanthropists, and practitioners; and new, promising ideas that expand opportunities for all. Our work inspires effective decisions that advance fairness and enhance the well-being of people and places.

Contents

Acknowledgments

This report was funded by the Robert Wood Johnson Foundation. We are grateful to them and to all our funders, who make it possible for Urban to advance its mission.

The views expressed are those of the authors and should not be attributed to the Urban Institute, its trustees, or its funders. Funders do not determine research findings or the insights and recommendations of Urban experts. Further information on the Urban Institute's funding principles is available at urban.org/fundingprinciples.

The authors would like to thank Megan Randall for her excellent project management; Jahnavi Jagannath, Jacqueline Rayfield, and Yipeng Su for their contributions to the Ethics and Empathy project; Graham MacDonald and Aaron Williams for their significant contributions to developing the imputation methodology; Graham MacDonald, LesLeigh Ford, and Steven Brown for providing valuable feedback on drafts of this report; Aaron Williams and Chris Davis for providing excellent code review; Claire McKay Bowen, Rob Santos, and Breno Braga for providing feedback on the imputation methodology; Allison Feldman for creating the methodology visualization; and Wes Jenkins and Zach VeShancey for providing excellent editing.

Ethics and Empathy in Using Imputation to Disaggregate Data for Racial Equity

Disaggregating data by race and ethnicity is a critical method for shining light on racialized systems of privilege and oppression.[1] As City of Austin chief equity officer Brion Oaks told the Urban Institute, "Only when the city can segment data can we see what is truly happening. Aggregates can conceal reality."[2] But many high-value datasets do not collect or report information on race and ethnicity. For example, such information is missing in credit bureau data, which has inhibited efforts to examine how credit scores affect racial homeownership gaps and to challenge the use of credit screens in hiring.[3]

Imputation is a powerful tool for disaggregating data by appending racial and ethnic identifiers onto datasets lacking that information. Although failing to disaggregate data by race and ethnicity can pose considerable harm to Black people, Indigenous people, and other people of color, efforts to fill data gaps using imputation can risk the same or even greater harm, particularly if done without a proactive focus on equity.

This report describes lessons we learned from a case study in which we proactively incorporated equity in imputing race and ethnicity onto a nationally representative sample of credit bureau data. We organize these lessons around three "ethics checkpoints" where we examine our source datasets, our imputation methodology, and the resulting race and ethnicity imputations for potential racial bias and inaccuracy. At each checkpoint, we share how we approached mitigating bias where possible and transparently communicating any bias that could not be mitigated; we also discuss how to determine when the unmitigated risk is unacceptably high and therefore warrants terminating the production or use of the imputed data.

Although this report focuses on how to implement these ethics checkpoints, just as important are the researchers involved in the process and the institutional structures that hold teams accountable to the checkpoint outcomes. It is vital for researchers to engage impacted communities from the very beginning of the research process and collaborate with them at each checkpoint to identify potential risks and weigh those risks against the potential benefits of disaggregated data for their communities. Moreover, at the outset, researchers should create institutional structures, such as community advisory boards, that give community members power to affect the imputation process and hold researchers accountable for following the outcomes of the ethics checkpoints. Before beginning any imputation process, we recommend that

researchers consult our ethics and empathy standards guide for guidance on creating diverse teams and accountability structures to ensure the ethics checkpoints outlined in this report yield equitable results.

Background

Although imputation can be used in a variety of contexts—for instance, imputing missing race and ethnicity values in datasets that already provide this information—we focus on using imputation to generate an entirely new race and ethnicity variable onto datasets that lack information on race and ethnicity.

The most widely used method for imputing race and ethnicity on administrative data is Bayesian Improved Surname Geocoding, which the RAND Corporation developed for the US Department of Health and Human Services and which is also used by the Equal Opportunity Employment Commission and the Consumer Financial Protection Bureau (CFPB). The latest method involving this tool, Medicare Bayesian Improved Surname Geocoding 2.0, combines name, administrative data, and census data based on address in a calibrated Bayesian framework (a multinomial logistic regression model) to estimate probabilities by race and ethnicity for each record in a dataset.

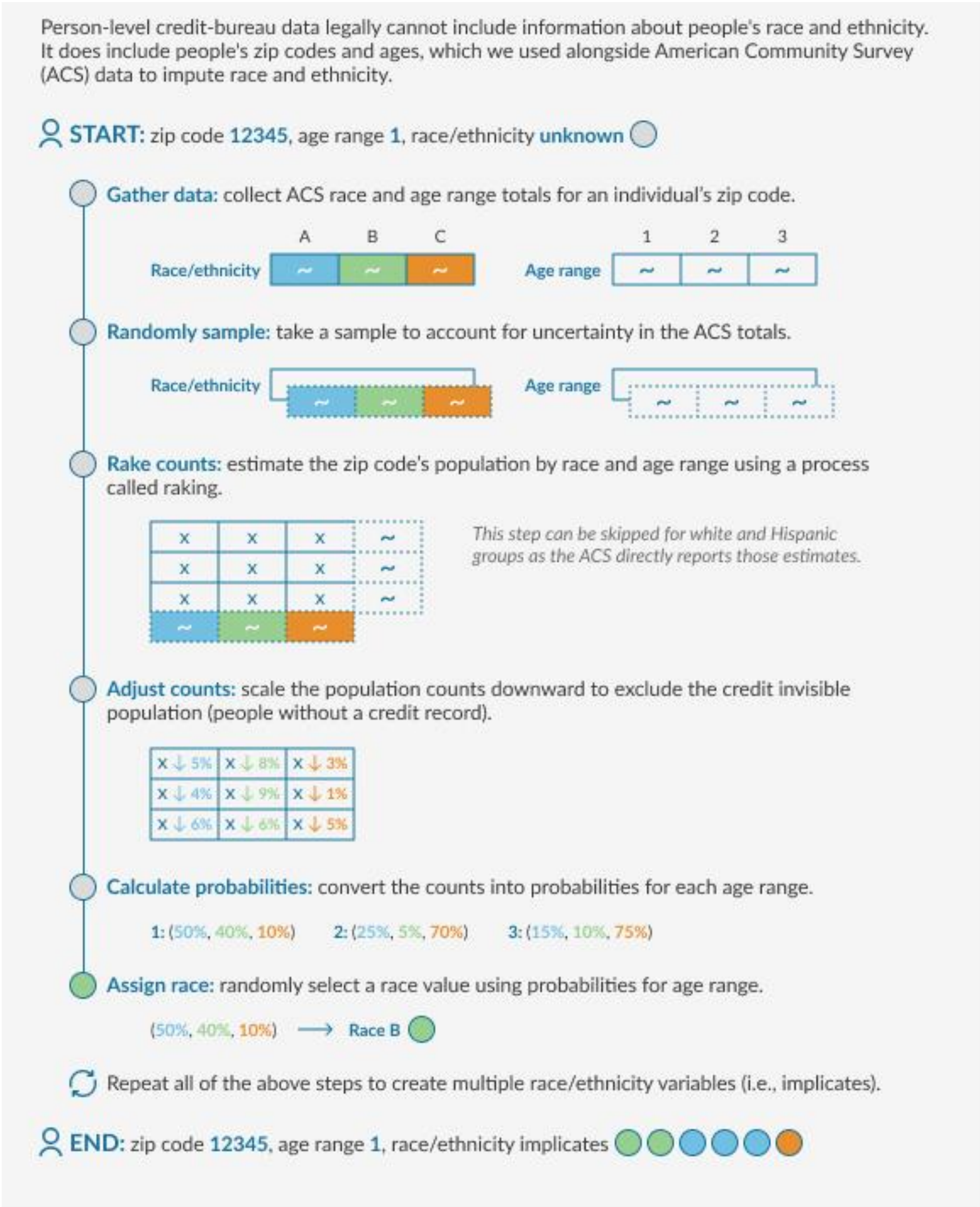
Multiple imputation, which involves creating multiple copies, or implicates, of an imputed race and ethnicity variable, is another standard procedure used in many public data products, such as the SIPP Synthetic Beta, the National Survey of Children's Health, and the Survey of Consumer Finances (SCF). Multiple imputation allows researchers to analyze variation resulting from the uncertainty in input data sources and from the imputation process when assessing the robustness of results.

Methodology

We used multiple imputation to add a combined race and ethnicity variable onto a 2013 dataset from a major credit bureau that represents a 2 percent random sample of adults with credit records in the United States. First, we used the Census Bureau's 2011–2015 five-year American Community Survey (ACS) estimates to calculate the probabilities of belonging to each racial/ethnic group for every person in the credit data based on their reported zip code and age, drawing on the geospatial imputation component of the Bayesian Improved Surname Geocoding method. With those probabilities, we randomly assigned a race/ethnicity group value. We then repeated the entire imputation process multiple times to produce multiple copies, or implicates, of the assigned race/ethnicity variable. Our approach accounts for the uncertainty in the ACS population count estimates and the inherent uncertainty of randomly assigning a race/ethnicity value based on a set of probabilities. The specific steps of this process are outlined in figure 1, and a detailed account of our methodology can be found in appendix A.

FIGURE 1

How We Imputed Race and Ethnicity onto Credit Bureau Data



URBAN INSTITUTE

Source: Urban research team; visual designed by Allison Feldman.

Ethical Imputation Checkpoints

We outline three checkpoints that we used before, during, and after the data imputation process. We used these checkpoints to identify and address areas where our methodology could introduce bias and assess whether the resulting imputed race/ethnicity variables were fit to use for equity analyses. These checkpoints and the risks we discuss are not unique to imputation; some amount of bias is unavoidable in any data analysis given the realities of imperfect data and constrained analytic choices.

Policy decisions made without disaggregated data are also prone to considerable bias and therefore often harm communities of color and other historically marginalized groups whose realities are concealed by aggregates. When deciding whether to proceed with or terminate imputation processes, researchers should weigh the potential harms of any unmitigated bias against the potential harms of not having imputed disaggregated data.

Our aim in this report is not to deter potential producers and users of imputed race and ethnicity data, but to empower them with tools to identify potential sources of bias, mitigate the bias where possible, and transparently communicate how any remaining bias limits ethical uses of the resulting data.

Checkpoint 1: Before Imputation, Audit Input Data for Bias

Data ethics advocates have produced considerable evidence of the potential for data analysis to encode racial biases present in input data sources. This risk can be harder to detect in complex methodologies, like machine learning or imputation, because of the layers of transformation between the input data and the analytic output—in our case, the imputed race/ethnicity variable. Accordingly, when **gathering datasets** before imputation, we audited each of the following input data sources for bias:

- **Credit bureau data:** a 2 percent random sample of all people with credit records in the United States in 2013, provided by a major credit bureau. We used a person's zip code (or county when zip code was missing) and age variables for imputation.
- **American Community Survey data:**[4] 2011–2015 five-year ACS estimates of population counts by race/ethnicity and age by geography.[5]
- **Consumer Financial Protection Bureau credit-invisibility data:** data from the CFPB on the percentage of people without a credit record (i.e., people who are "credit invisible") in the United States by race and age (Brevoort, Grimm, and Kambara 2015).

We examined each of these datasets for potential bias using the three questions that follow.

Does the dataset accurately represent the underlying population that it aims to measure? How might structural racism drive unrepresentativeness? One source of potential bias is that certain subgroups of a population of interest may be over- or underrepresented in a given dataset. And, in many cases, structural racism may result in data that are systematically unrepresentative for communities of color. For example, overpolicing of communities of color can result in arrest data that overrepresent these communities rather than accurately representing the true prevalence of crime committed in the underlying population.

We reviewed each dataset's documentation and methodology and discussed the potential for bias with expert users of each dataset. We concluded that bias involving over- and underrepresentation was unlikely to be a significant issue in each of our datasets. Although undercounting communities of color is a known issue in the decennial census, the Census Bureau takes extensive measures to mitigate the potential effects of nonresponse and maintain representativeness in the ACS, making it the data source of record for demographic population estimates. The credit bureau data we used are a nationally representative 2 percent random sample of adults with credit records provided by a major nationwide credit bureau. The CFPB data compare population counts in the ACS and the CFPB's representative sample from a major credit bureau to calculate the percentage of the national population that is credit invisible, broken down by age group and racial/ethnic group. Although we determined that these datasets are all adequately representative to use, it is important to acknowledge that structural racism affects people's access to credit and drives higher rates of credit invisibility among people of color,[6] thereby inherently limiting the usefulness of credit records for understanding financial well-being in communities of color.

Another dimension of unrepresentativeness is whether datasets used in imputation accurately reflect the lived experience of the people represented in the data. For example, we acknowledge that the race/ethnicity categories available in the ACS data—which we accordingly used as our race/ethnicity categories for imputation—may not accurately reflect the way people represented in the credit bureau data would self-identify and may conceal important differences within groups.[7] Although we cannot **mitigate** this unrepresentativeness because of limitations in the ACS data, we can document and **communicate** this concern to potential data users.

Do all the datasets being used in imputation represent the same population? Even if all the input datasets are perfectly representative of their target populations, if the target populations differ across datasets, imputation will likely be less accurate. In our case study, we took several steps to align the

populations of our input datasets. First, in the credit bureau data, we had to exclude 4,302 records from US territories that are not included in the ACS data.[8] Second, we identified that the ACS data measure the total population of adults in the United States, whereas the credit bureau data measure the population of adults *with a credit record*. The disparities between these populations vary by race and ethnicity: the CFPB data on credit invisibility show that 9.4 percent of white adults are credit invisible, compared with 14.8 percent of Black adults and 15.8 percent of Hispanic adults. We used the CFPB data by race/ethnicity and age to adjust the ACS population counts to reflect the population with a credit record (we discuss this further in checkpoint 2 below). Without this mitigation step, we would have likely overrepresented Black and Hispanic groups in the imputed race/ethnicity variable, though this step has limitations that we endeavor to transparently **communicate** below.

We also considered using estimates from the 2013 Survey of Consumer Finances of the proportion of households with student debt, vehicle debt, and other installment debt by race and ethnicity at the national level to benchmark the accuracy of our imputations by comparing statistics calculated using our imputed data against published statistics in the SCF.[9] Because the SCF publishes estimates at the household level based on respondents' race/ethnicity, whereas the credit bureau data are at the individual level, we concluded that these datasets were not readily comparable and decided to **terminate** use of the SCF.[10] Subsequent analyses may want to explore the use of the SCF and other datasets for benchmarking.

Are there missing data? If so, does that missingness disproportionately affect certain racial/ethnic groups? Is it correlated with other variables of interest? Missing values in the variables used for imputation can reduce overall accuracy and inject bias if that missingness disproportionately affects some racial/ethnics groups. We used the age and zip code fields of the credit bureau data in our imputation methodology, because considerable research has shown that location and age are important predictors of race.[11] Because the ACS reports data at the level of Zip Code Tabulation Area (ZCTA), we used a zip-code-to-ZCTA crosswalk file to translate between the credit bureau and ACS data.[12] After completing this crosswalk, the location and age fields in the credit bureau data were missing at rates of 0.04 percent for the ZCTA variable, 4.45 percent for the county variable, and 15.44 percent for the age variable.[13]

In the cases where an individual's ZCTA was missing, we used the race/ethnicity data of that individual's age group (if present) for the county (or nation, if county was also missing). If an individual's age group was missing, we used the race/ethnicity data for all adults at the smallest available geography. If the most precise age and location data are disproportionately missing for people of color, this missingness could inject racial bias into the imputations. We used a logistic regression to assess

whether the missingness of the age variable (which we focused on, given how rarely zip code was missing) was correlated with the racial composition of the individual's ZCTA and other credit variables that may be correlated with race.[14]

We find that, holding all the variables equal, people living in majority-nonwhite ZCTAs are 37 percent more likely to have missing age data than people *not* living in majority-nonwhite ZCTAs, while subprime consumers and people with auto debt are 162 percent and 59 percent more likely to have missing age data than their counterparts, respectively.[15] Although this raises concern about the potential for bias, we know, given high levels of segregation across neighborhoods, that location (in our case, zip code or county) is likely to be a much better predictor of race than age. We therefore felt comfortable that our mitigation strategy of using the race/ethnicity proportions of all adults in the geography where an individual's age was missing was sufficient to continue the imputation while **communicating** this concern. With additional time, we could confirm this by testing the impact of the missing age data in checkpoint 2 and communicating our findings to users.[16] Table 1 summarizes when to take steps at checkpoint 1 to mitigate bias and risk, when to communicate with other analysts or end users about any risks, and when to terminate the imputation process.

TABLE 1

Recommendations for Mitigating Bias and Risk, Communicating with Analysts and End Users, and Terminating Projects at Checkpoint 1 of the Imputation Process

Mitigate	Communicate	Terminate
▪ Mitigate bias by using datasets that most accurately represent the target population of interest. ▪ Mitigate data population mismatches (e.g., the credit visible population versus total population) by adjusting estimates to match. ▪ Mitigate missingness by using consistent rules to impute with the most accurate available data.	▪ Communicate cases where datasets may not fully represent the target population, including in race and ethnicity categories. ▪ Communicate adjustments performed and the limitations of those adjustments (e.g., dropping geographies not included in all data). ▪ Communicate the prevalence of missing data and the imputation strategy used. ▪ Communicate the magnitude and direction of potential unmitigated bias where possible.	▪ Terminate if the focus population (e.g., people in US territories) is not adequately represented in the data. ▪ Terminate if populations cannot be adequately adjusted, especially if using unadjusted populations adversely impacts estimates for communities of color.

Source: Urban research team.

Checkpoint 2: During Imputation, Examine Where Bias Could Be Introduced at Each Step

Input data are not the only potential source of bias in data analysis; every methodological decision can bias the outcome. When imputing race and ethnicity, it is critical to think about bias in a statistical sense (i.e., divergence from the "true" outcome) as well as racial bias, especially in cases where a decision may not harm—or may even improve—overall accuracy (reducing overall statistical bias), while differentially *reducing* the accuracy of imputations for smaller racial and ethnic groups. For each step of the imputation process, we discuss how we incorporated equity into our methodological decisions.

SAMPLE DATA

We **randomly sampled** the ACS population estimates from a normal distribution based on the reported estimate and margin of error to account for uncertainty in the reported estimates. One concern we encountered was that for the larger racial/ethnic categories (non-Hispanic white and Hispanic), the ACS provides population counts broken down by race and age together (e.g., non-Hispanic white people ages 18–19, ages 20–25, etc., in a ZCTA), so we could randomly sample these estimates directly. For smaller racial/ethnic categories, those population counts were only available by race alone or age range alone (e.g., the total non-Hispanic Asian population in a ZCTA and the total population of people ages 18–19 in a ZCTA).[17] For these smaller racial/ethnic groups, we randomly sampled these race and age-range totals and used an additional raking step to "fill in the blanks" to generate counts broken down by race/ethnicity and age range. We recognize that the availability of more precise ACS data for the Hispanic and non-Hispanic white groups likely yields more precise imputations for these groups. By accounting for the uncertainty of the ACS estimates in our analysis at checkpoint 3, we can effectively capture and **communicate** this uncertainty to data users.

For some cases with small estimates and/or large margins of error, our samples returned negative values. We replaced these negative values with zero because we determined this was the best way to satisfy the requirement of positive counts while **mitigating** the potential bias of overrepresenting the groups with ACS population estimates of zero. We are also **communicating** the potential for this to slightly overestimate small population counts across implicates.

RAKING

Raking, or iterative proportional fitting, is a procedure that allows users to fill in table cells when the row and column totals are known (appendix B provides an in-depth explanation of this process). We knew the totals by race (column totals) and the totals by age range (row totals) and we used raking to

generate the counts by race and age range together. As mentioned above, we only performed raking for smaller racial/ethnic categories for which we needed to "fill in the blanks" (i.e., all categories besides non-Hispanic white and Hispanic).

We were concerned that this selective application of raking could lead to less accurate results for the smaller racial/ethnic groups, so we considered an alternate approach of performing raking for all racial/ethnic categories. But through early tests comparing the raking procedure at the national level against the Integrated Public Use Microdata Series ACS microdata (which can be used to directly calculate estimates of counts by age and race at the national level), we found that including all the racial/ethnic categories in raking yielded less accurate results overall and particularly underestimated counts for the smaller racial/ethnic groups. Therefore, we concluded that our approach of only applying raking to the smaller racial/ethnic groups was the most effective way to **mitigate** this bias.

Another challenge we encountered is that the raking procedure requires the sum of the row totals (i.e., age counts) to equal the sum of the column totals (i.e., race/ethnicity counts). But because we sampled the age and race/ethnicity counts independently from different ACS tables, this was almost never true. To satisfy this constraint, we treated the race/ethnicity totals (which ACS reports directly) as the "ground truth" and readjusted the age totals (which we calculated from multiple ACS estimates) to match.[18] This means that the actual counts by age range within a sample may often be incorrect. We also tested a "rolling sampling" procedure, in which each sample is constrained by the results of the previous sample(s) to ensure that the row and column totals sum to the same value. But we found that this forced many of the later-sampled totals to zero, especially in smaller geographies with higher margins of error for the age and race/ethnicity estimates. We therefore concluded that reweighting was the most effective way to satisfy the requirements of raking while **mitigating** potential bias. Though our testing concluded that our approach to each of these raking challenges was the best option, neither is a perfect solution, and we are therefore **communicating** the limitations to data users.

CREDIT-INVISIBILITY ADJUSTMENT

We adjusted the counts produced in the raking step downward to exclude the credit-invisible population so that the adjusted counts represented the population with a credit record, matching the population represented in the credit bureau data. We made these adjustments using estimates from a CFPB report (Brevoort, Grimm, and Kambara 2015) breaking down credit-invisibility rates by age group and race/ethnicity. These data have the following limitations that could introduce bias into our methodology:

- The CFPB estimates are at the national level. We therefore must assume that these rates hold within age and racial/ethnic groups across the United States, which is likely untrue and may be particularly so for some geographic subgroups, such as rural areas or areas with large immigrant populations.
- The CFPB estimates are from 2010, whereas the credit bureau data are from 2013. If credit invisibility changed differentially by race and ethnicity in that period, this could be a source of bias.
- The CFPB aggregates data for multiple smaller racial/ethnic groups reported in the ACS into a single "other" category,[19] which we apply to all the constituent groups. This masks potential heterogeneity in credit invisibility among these groups and is likely more accurate for the larger racial/ethnic groups in this "other" category.
- The age ranges in the CFPB report do not exactly match the age ranges present in the ACS data. For all groups older than 35, the CFPB data report smaller age ranges than the ACS data. For example, the ACS includes the range of 35–44, whereas the CFPB data report credit invisibility rates for 35–39 and 40–44. To mitigate this mismatch, we took a simple average across the two age ranges, but this means that older people, especially in geographies containing an uneven distribution within age ranges, will likely have less accurate credit-invisibility rates.

Unfortunately, we cannot **mitigate** these concerns without access to more timely and granular credit-invisibility figures, which do not exist as of this writing. We concluded that failing to adjust for credit invisibility would introduce more bias than was likely to be introduced by using the available CFPB data, even given the above issues. With respect to our ability to accurately **communicate** these concerns, one limitation is that the CFPB does not provide margins of error for the credit-invisibility estimates, which means we could not account for uncertainty in those estimates in our imputation process.

CALCULATING PROBABILITIES AND ASSIGNING RACE/ETHNICITY

From the adjusted population counts, we calculated probabilities (i.e., proportions) by racial/ethnic group for each age range in a given geography. We then used the probability vector corresponding to an individual's age and zip code to randomly assign the imputed race/ethnicity value. As discussed in checkpoint 1, the age and zip code variables were sometimes missing in the credit bureau data, in which case we used the proportions for the age group in a coarser geography (county or nation), if zip code was missing, and/or a coarser age group (all adults), if age was missing. If our raking and sampling procedure returned a total population of zero for an individual's age group and zip code, we also used

the coarser geography or age group probabilities to assign race/ethnicity for that individual. Table 2 summarizes when to take steps at checkpoint 2 to mitigate bias and risk, when to communicate with other analysts or end users about risks, and when to terminate the imputation process.

TABLE 2

Recommendations for Mitigating Bias and Risk, Communicating with Analysts and End Users, and Terminating Projects at Checkpoint 2 of the Imputation Process

Mitigate	Communicate	Terminate
▪ Mitigate differential accuracy/bias by testing multiple methods for impact across racial/ethnic groups and geography when unsure of the best approach. ▪ Mitigate variable mismatches by using the best information available to combine across datasets.	▪ Communicate the potential impact of data limitations and methodological decisions on the accuracy of imputation results. ▪ Communicate to data users how they can derive estimates from multiple implicates to accurately capture uncertainty in both the data and imputation process.	▪ Terminate if it becomes apparent at any point that the data present too many potential sources of bias and cannot support an accurate imputation process, such that the opportunity costs of completing imputation and proceeding to determine fitness for purpose in checkpoint 3 outweigh the potential benefits from imputation.

Source: Urban research team.

Checkpoint 3: After Imputation, Assess Whether Imputed Race/Ethnicity Data Are Accurate Enough to Be Used Ethically for Your Analytic Purpose

Even with efforts to mitigate bias before and during imputation, some amount of inaccuracy is inevitable, just as any prediction or estimation from data is unlikely to be perfect. The question when imputing race and ethnicity onto datasets is whether the resulting data are accurate *enough*, in terms of both overall accuracy and differential accuracy across racial/ethnic groups, to be responsibly used—whether the data have “fitness for purpose.” The answer to this question will vary based on the intended use. We assessed the fitness for purpose of our imputations through the two key questions that follow.

HOW ACCURATE IS THE IMPUTED RACE/ETHNICITY VARIABLE?

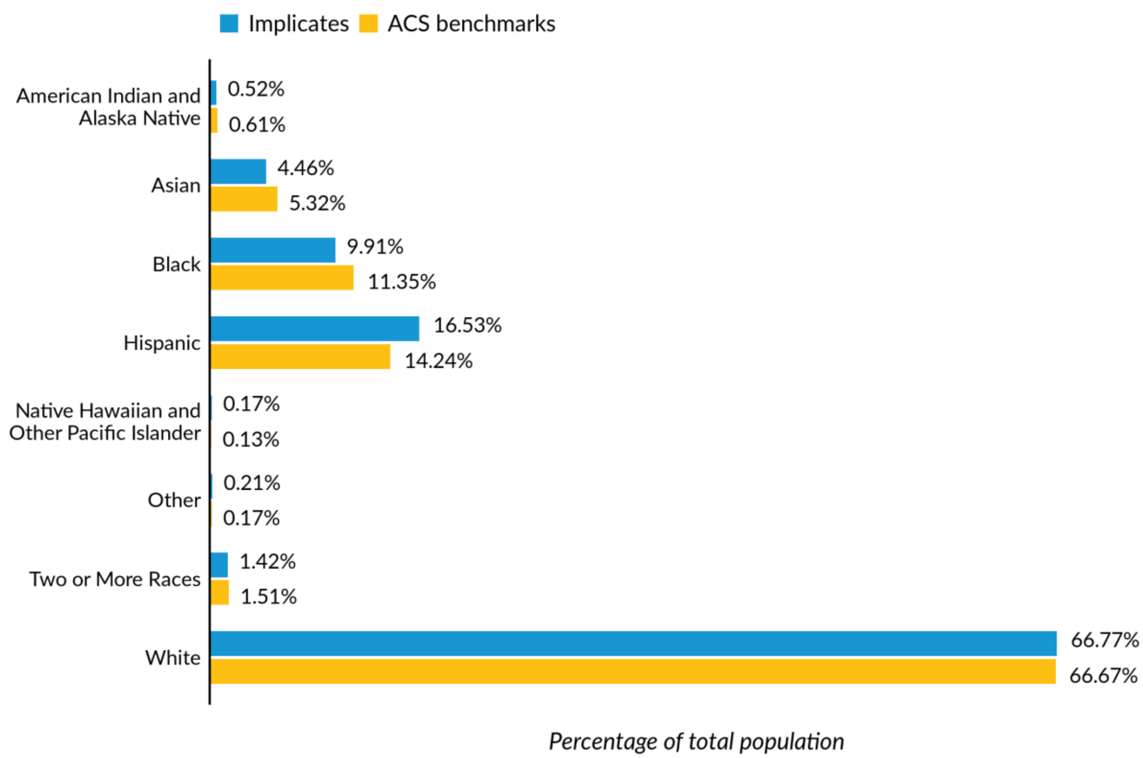
To measure the accuracy of our imputed race/ethnicity variable, we benchmarked our imputations against trusted aggregate statistics on population counts and proportions by race and ethnicity from the ACS. To create the ACS benchmarks, we adjusted the ACS data to exclude the credit-invisible population (as outlined above in the credit-invisibility adjustment step) and residents younger than 18 (using ACS microdata).[20] In both adjustments, the accuracy is limited by our only having national-level

data and coarser race/ethnicity categories. These adjustments ensured that the population represented by our imputations (credit-visible adults) matched the population we used for benchmarking.

Figure 2 shows that the population composition by race/ethnicity calculated from the imputed data reasonably follows the adjusted ACS benchmarks, with the key observation that the implicates underestimate Black and Asian populations by 1.4 percentage points and 0.8 percentage points, respectively, and overestimate Hispanic populations by 2.3 percentage points.

FIGURE 2

Population Composition by Race/Ethnicity from Implicates and the American Community Survey Benchmarks



URBAN INSTITUTE

Sources: Authors' analysis of 2013 data from a major credit bureau and 2011–2015 five-year Census Bureau American Community Survey estimates, adjusted using Consumer Financial Protection Bureau estimates of credit invisibility, and estimates of the adult population calculated from Integrated Public Use Microdata Series 2011–2015 American Community Survey microdata.

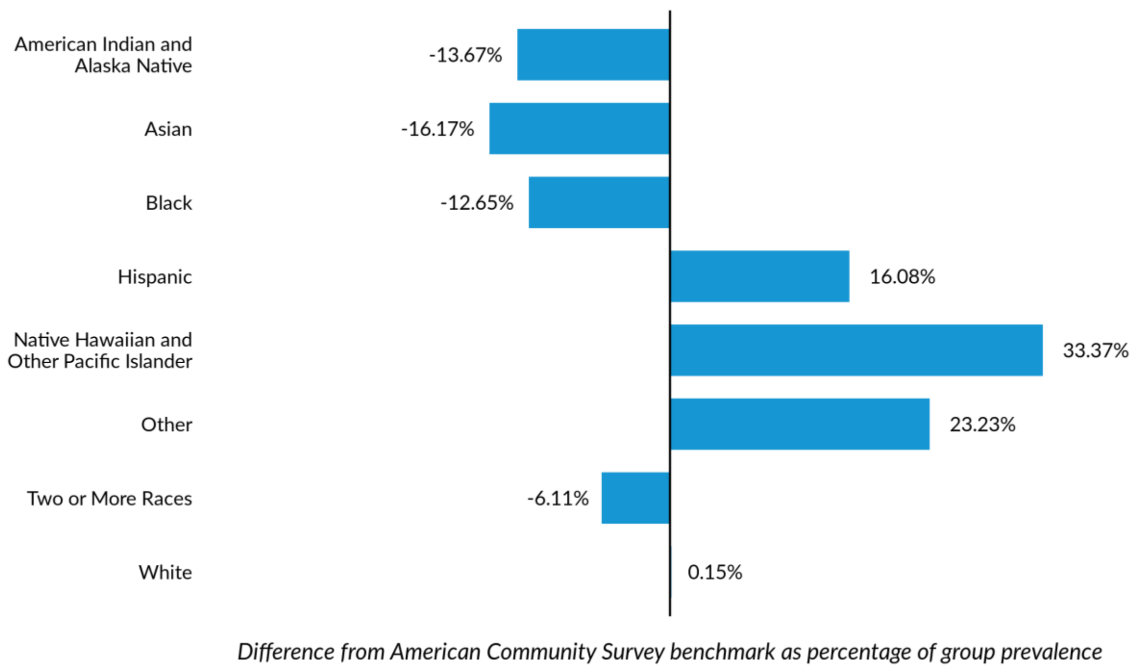
Notes: ACS = American Community Survey. The race and ethnicity values in the credit bureau implicates are estimated via multiple imputation. The major credit bureau does not collect this information.

Although these disparities are small in absolute terms, when evaluated as a share of each racial/ethnic group's original population proportion, the picture is quite different (figure 3). For

example, although the estimates of the Native Hawaiian and Other Pacific Islander population shares from the implicates and ACS benchmarks are only 0.04 percentage points apart, a very small absolute percentage-point difference, this corresponds to a 33.3 percent difference between the implicates and the benchmark, the largest discrepancy among all racial/ethnic groups. This example highlights the need to be careful and use multiple evaluation metrics so you can holistically evaluate accuracy, especially for smaller race and ethnicity categories.

FIGURE 3

Discrepancy in Population Composition by Race/Ethnicity between Implicates and American Community Survey Benchmarks as a Percentage of Race/Ethnicity Group Prevalence



URBAN INSTITUTE

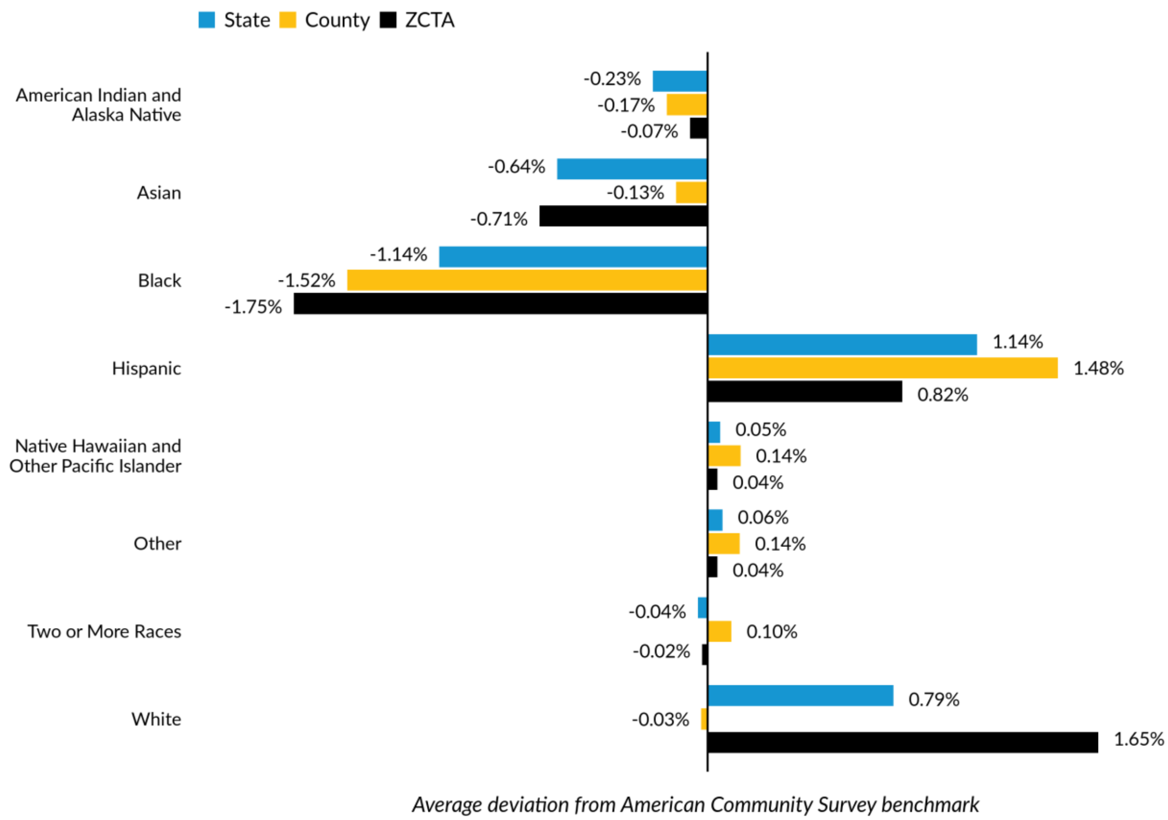
Sources: Authors' analysis of 2013 data from a major credit bureau and 2011–2015 five-year Census Bureau American Community Survey estimates, adjusted using Consumer Financial Protection Bureau estimates of credit invisibility, and estimates of the adult population calculated from the Integrated Public Use Microdata Series 2011–2015 American Community Survey microdata.

Notes: The race and ethnicity values in the credit bureau implicates are estimated via multiple imputation. The major credit bureau does not collect this information.

When we look at the average percentage-point discrepancy between our implicates and the ACS benchmarks at the state, county, and ZCTA levels (figure 4), we see a similar directional trend of overestimating Hispanic and underestimating Black and Asian population shares. As we move to the smaller ZCTA-level estimates the discrepancies generally get larger and we also overestimate the white population share.

FIGURE 4

Average Discrepancy in Population Composition by Race/Ethnicity between Implicates and American Community Survey Benchmarks by Geographic Level



Sources: Authors' analysis of 2013 data from a major credit bureau and 2011–2015 five-year Census Bureau American Community Survey estimates, adjusted using Consumer Financial Protection Bureau estimates of credit invisibility, and estimates of the adult population calculated from the Integrated Public Use Microdata Series 2011–2015 American Community Survey microdata.

Notes: ZCTA = zip code tabulation area. The race and ethnicity values in the credit bureau implicates are estimated via multiple imputation. The major credit bureau does not collect this information.

We do not know whether these disparities between the implicates and the "true" population proportions owe to inaccuracies in our imputation process or inaccuracies in the coarse adjustments we made to create the ACS benchmarks. Specifically, our adjusted ACS benchmarks rely on national estimates of the adult and credit-visible populations, so inaccuracies in the adjustments could be driving observed disparities at subnational levels. But the consistency of over- or underestimates for Hispanic, Black, and Asian populations are cause for concern.

To understand the potential drivers of this differential accuracy by racial/ethnic group, we examined the differences between the top and bottom 10 percent of ZCTAs where our imputations

most and least accurately matched the ACS benchmarks. We find that our imputations are least accurate in more racially diverse ZCTAs and most accurate in racially segregated ZCTAs. On average, the largest racial/ethnic group represented 93.2 percent of the population in the most accurate ZCTAs, whereas it represented 58.2 percent of the population in the least accurate ZCTAs. In the most accurate ZCTAs, the largest racial/ethnic group is always white or Hispanic. In the least accurate ZCTAs, the largest racial/ethnic group is Black eight times more frequently, and Asian nine times more frequently, than in all ZCTAs.

These discrepancies likely result from the greater accuracy of the ACS data for the white and Hispanic groups. By having smaller margins of error, the sampling step produced more consistent results for these groups. And, by directly reporting estimates by age for white and Hispanic groups, we were able to directly sample these groups instead of using raking to derive them. More accurate disaggregated race and ethnicity data from reliable providers like the Census Bureau is key for more reliable imputations in the future. Future research could explore whether not using age in imputation or using coarser race/ethnicity groups could improve accuracy, though this would come at the cost of granular disaggregation.[21]

All of the estimates in this report are calculated using 50 implicates. Because we account for the uncertainty of the ACS estimates and imputation process through random sampling, it stands to reason that with more implicates, we would converge toward a more consistent estimate. But that raises the question of how many implicates users need to create. Interestingly, we found that our estimates converged with relatively few implicates, though the number of implicates needed for convergence did increase for estimates at smaller geographies. The results of our testing can be found in appendix A. Users of multiply imputed race/ethnicity data can use similar tests to identify the appropriate number of implicates for their specific analytic use case. For more granular estimates, we recommend using more implicates.

IS THE IMPUTED RACE/ETHNICITY VARIABLE ACCURATE ENOUGH TO ETHICALLY USE FOR MY ANALYTIC PURPOSE?

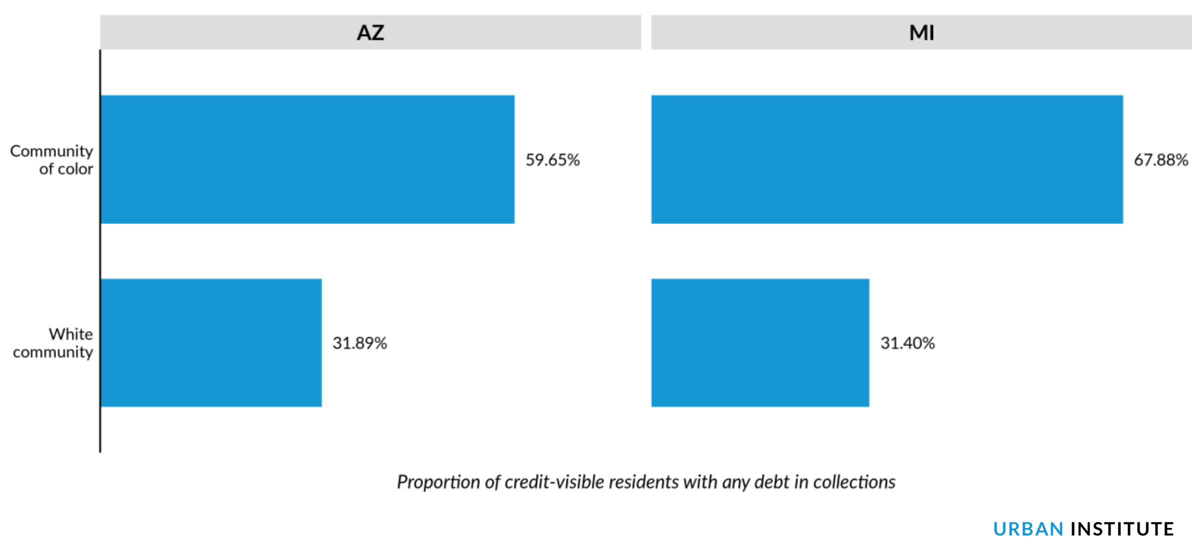
After confirming that our imputations reasonably tracked the ACS benchmarks (though we had concerns about inaccuracy, especially for Hispanic, Black, and Asian residents), we asked whether the imputations were accurate enough to use for equity analyses. In other words, we asked whether, given the inaccuracies we measured in our imputed data, we could ethically use the data to answer a given analytic question or whether we should terminate the process. Answering this question involves examining the ethical implications of using the imputed data relative to the potential next-best dataset

that could be used for the given analytic purpose. Calculating estimates using multiple implicates (we used 50) is key to determining fitness for purpose because it makes estimates more precise and, critically, allows the uncertainty of those estimates to be calculated accurately, which enables evidence-based conversations about the fitness of the estimates for ethical use.[22]

For example, imagine you are the director of research for a state advocacy group that is seeking to make the case for race-conscious programs to support people with debt in collections. Without access to disaggregated race and ethnicity data—which would likely only be available through imputation—you would likely take the common approach of analyzing disparities between communities of color and white communities, where people are identified as living in a white community if most residents (more than 60 percent) in their zip code are white and as living in a community of color if most residents (more than 60 percent) in their zip code are people of color.[23] As figure 5 shows, the research directors for organizations in Arizona and in Michigan would see very similar patterns of racial disparity using these data.

FIGURE 5

Proportions of People in Communities of Color and in White Communities with Any Debt in Collections in Arizona and Michigan



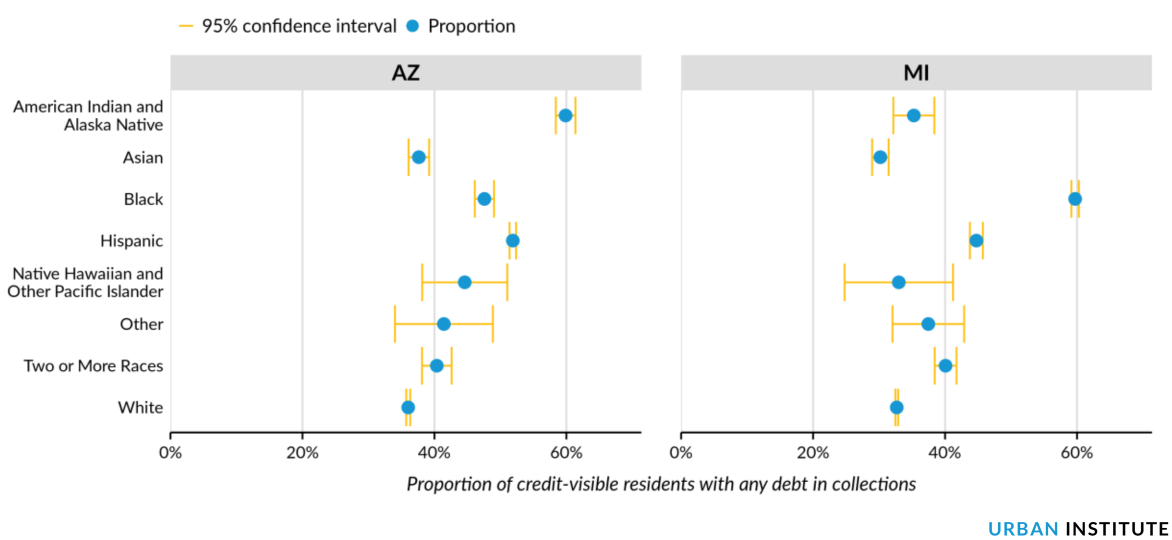
Sources: Authors' analysis of 2013 data from a major credit bureau and 2011–2015 five-year Census Bureau American Community Survey estimates.
Notes: Types of communities (i.e., white communities, communities of color) are estimated based on the American Community Survey data on population by race and ethnicity for the zip code tabulation area of each observation in the credit bureau data. The major credit bureau does not collect this information.

This approach would have at least three drawbacks for the research directors in Arizona and Michigan. First, this approach does not allow for disaggregation within the broad "communities of color" category to enable more precise targeting. Second, it excludes people in communities where neither the white population nor the population of color has a population share greater than 60 percent—or 22 and 6 percent of the credit bureau data observations in Arizona and Michigan, respectively. Third, although the disparity is large in both states, this approach does not enable the research directors to estimate the uncertainty of these estimates—that is, how confident they can be that the observed difference is the result of true differences in outcomes rather than sampling variability.

When we compare these results with the analysis of our imputed race and ethnicity data (figure 6), we see that using the broad category of "communities of color" conceals critical heterogeneity between racial/ethnic groups.

FIGURE 6

Proportions of Adults with Any Debt in Collections by Racial/Ethnic Group in Arizona and Michigan



Sources: Authors' analysis of 2013 data from a major credit bureau.
Notes: The race and ethnicity values for the credit bureau implicates are estimated via multiple imputation. The major credit bureau does not collect this information.

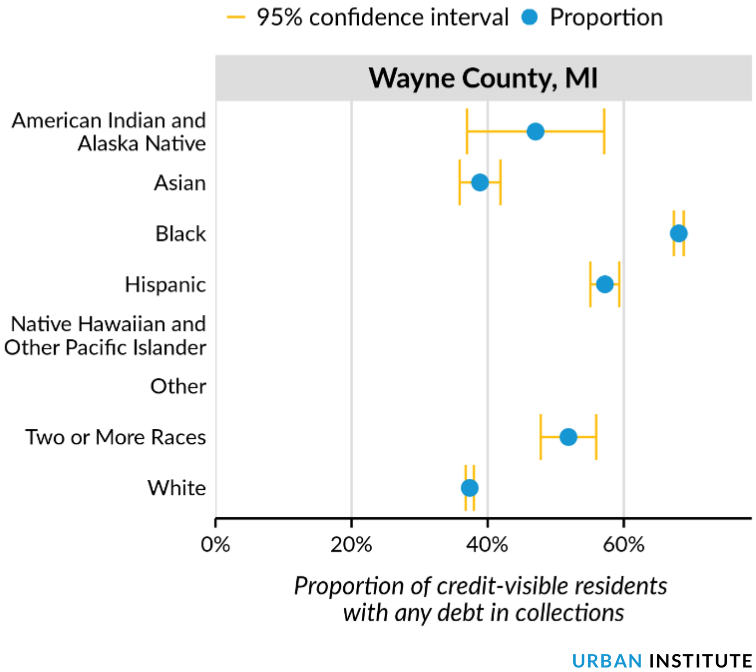
Increased visibility of racial/ethnic communities in data allows for greater awareness, for more effective targeting of programs, and for communities to organize and advocate for resources. For example, understanding the different rates at which people have any debt in collections by racial/ethnic group in each state would enable the hypothetical organization in Arizona to advocate for culturally sensitive programming focused in American Indian and Alaska Native communities, while the organization in Michigan could focus its work in Black communities. The small margins of error around the estimates for these communities in Arizona and Michigan, respectively, and the fact that the estimates include the full population unlike the coarser alternative approach shown in figure 5, can give the research directors and their target audiences confidence in the significance of the observed disparities. Moreover, the large magnitude of the disparity indicates that the results will be robust to the undercount of Black adults identified in our benchmarking.

At smaller geographies, small numbers of observations and greater uncertainty in estimates from imputed data might make the data difficult to use ethically. For example, users of county-level data in Michigan will quickly run into data limitations because only 15 of the 83 counties in Michigan have sufficient data to enable us to publish estimates for the Black population.[24]

Wayne County, where Detroit is located, has the largest Black credit-visible population (563,171 people) of all Michigan counties per our imputed data. Figure 7 shows that the imputed data can largely still support the identification of racial disparities in debt in collections at the county level in Wayne County, though notably there are insufficient data to estimate the shares of the Native Hawaiian and Other Pacific Islander population and the populations in the category “other” with debt in collections. Researchers using the imputed data to examine disparities in Wayne County may choose to proceed and **communicate** the limitations of the data or to **terminate** the analysis if they feel that not having sufficient data to examine debt in collections for all groups makes the data unfit for purpose.

FIGURE 7

Proportions of Adults with Any Debt in Collections by Racial/Ethnic Group in Wayne County, Michigan



Source: Authors' analysis of 2013 data from a major credit bureau.
Notes: The race and ethnicity values for the credit bureau implicates are estimated via multiple imputation. The major credit bureau does not collect this information.

Now imagine you are the Wayne County program director. Your task is to target grant funding to the 10 percent of neighborhoods with the highest proportions of Black adults with debt in collections. Using our imputed data, you could easily calculate the proportion of Black adults with debt in collections in each ZCTA and choose the 10 percent of ZCTAs (roughly 7 of Wayne County's 69) that have the highest values. But, you might ask yourself, how sure are you that these are the "right" neighborhoods to target grant funding to?

Given the high degree of uncertainty in the estimates at the ZCTA level, the answer is "not very." In fact, to be confident that you have reached all the neighborhoods that *could* be in the top 10 percent, you would have to provide funding to 21 neighborhoods, a threefold expansion in the intended size of the grant program.[25] And you know that the underrepresentation of Black adults in the imputed data revealed by benchmarking raises further concern about accuracy. For this geographic-targeting purpose, the data precision needed is high and the ethical consequences of getting it wrong are considerable. One way to **mitigate** that risk is to expand the program to account for uncertainty in targeting (this is not always feasible). Another mitigation approach could be to engage community

stakeholders to help gut check the estimates and provide additional information to complement the imputed data to inform targeting. Users should assess the risks associated with using the imputed data against the risks associated with using other data sources or taking different approaches to program design. If another approach involves lower risk (or even involves equal risk but is easier to **communicate** to communities), such as using alternative data, the right choice may be to **terminate** use of imputed data for that analytic purpose. Table 3 summarizes when to take steps in checkpoint 3 to mitigate bias and risk, when to communicate with other analysts or end users about the risks, and when to terminate the use of imputed race/ethnicity data.

TABLE 3

Recommendations for Mitigating Bias and Risk, Communicating with Analysts and End Users, and Terminating Projects at Checkpoint 3 of the Imputation Process

Mitigate	Communicate	Terminate
▪ Mitigate incomplete understanding of accuracy by racial/ethnic group by using multiple evaluation metrics. ▪ Mitigate inaccurate conclusions by using imputed data only for analyses where there are sufficient observations. ▪ Mitigate ethical harms where uncertainty is higher by expanding policy coverage to account for uncertainty.	▪ Communicate the uncertainty of estimates by accurately calculating margins of error and including them in publications. ▪ Communicate how imputed data were used to inform policy or programmatic decisions.	▪ Terminate the project when the uncertainty of estimates and the consequences of that uncertainty (i.e., getting it wrong) are high and there is another, lower-risk information source or decisionmaking approach.

Source: Urban research team.

Conclusion

Like all data analysis, data imputation involves risks. There are some that we did not engage with in this case study, such as the risk of violating the personhood of people represented by the data, the risk of reidentification through disaggregation, and the potential for bad actors to use imputed data. For additional guidance on how to address those risks, we encourage producers and users of imputed data to read our ethics and empathy standards guide.

Despite the risks involved, data imputation can fill critical gaps in disaggregated data to give researchers, policymakers, and community leaders a chance to make more ethical and more equitable decisions. To realize the wide-ranging benefits that better, more granular data offer, producers and users of imputed data must mitigate the risks by proactively centering equity in their methodological

decisions. This case study, in which we imputed race and ethnicity onto credit bureau data, surfaced several key lessons for ethical imputation, which are as follows.

Equity must be considered in every decision. Discussions of racial bias in advanced analytics often focus on the risk of replicating racial bias in the input data. Although the familiar adage "garbage in, garbage out" undoubtedly applies to imputation, it is important for analysts to check for bias in the input data and at every step of the imputation process. Establishing ethics checkpoints before, during, and after analysis can help analysts mitigate bias at each step. It is equally critical to ensure that the background and experience of the key decisionmakers at every step of the imputation process reflect diverse perspectives, including those from directly impacted communities, and to establish structures to hold researchers accountable to the checkpoint outcomes

Examine differential outcomes by race and ethnicity. Many of the potential sources of bias that we highlight in our checkpoints disproportionately impact smaller racial and ethnic groups and geographies because of their smaller sample sizes and greater estimate uncertainty. It is therefore critical to examine outcomes of each ethics checkpoint disaggregated by race and ethnicity. This includes methodological decisions (e.g., selecting a sampling method) and evaluating fitness for purpose (e.g., benchmarking against statistics disaggregated by race/ethnicity). Aggregate quality checks can conceal disparate accuracy across racial/ethnic groups that can serve to further reinforce racial disparities when those data are used.

Communication is key. Most of the risks that we highlight in this report are inherent to data analysis. The additional risk posed by imputation, like other advanced data analysis methods, is that the layers of complexity between the input data sources and output imputations make it harder for end users to identify potential bias. Being transparent about analytic decisions and the resulting limitations on the ethical use of the imputed data is critical. This also necessitates accurately calculating the margins of error for estimates derived from multiple implicates and clearly communicating them to users.

Fitness for purpose must be examined in context. Though producers of imputed data can broadly assess the accuracy of imputations by benchmarking against trusted statistics, users must determine fitness for purpose for each specific analytic case. They should evaluate whether the estimates are precise enough to draw accurate conclusions. Data producers can support these efforts by including as much information as possible in their data release, including using multiple imputation to provide measures of the uncertainty of imputed statistics. If concerns about imprecision are identified, users should engage impacted communities to weigh potential risks of making decisions using imprecise imputed estimates against risks of using the next-best data available.

In future work, we hope to test different changes to our imputation methodology to improve the results, including the concerning undercounts of Black and Asian adults and the overcount of Hispanic adults that we identified in our benchmarking. These changes include testing using only location to impute race and ethnicity, testing other sampling approaches, using approaches other than raking to produce the race-/ethnicity-by-age counts, and incorporating other data sources alongside the ACS and/or other modeling approaches to produce the probabilities of membership in each racial/ethnic group. We would also like to implement the recommendations in our standards guide to effectively bring community voices into the imputation process, create accountability structures for researchers, and share lessons learned from that process with the field.

Appendix A. Detailed Methodology

For this case study, we imputed a combined race and ethnicity variable on a dataset from a major credit bureau representing a random 2 percent sample of all adults with credit records in the United States. The credit bureau does not publish the race and ethnicity of the people in the data, but it does publish their age and zip code. We used these fields to impute race and ethnicity onto the data in the following steps.

Step 1. Gather Data

We used the key datasets that follow in our imputation methodology.

Credit Bureau Data

We used data provided by a major credit bureau representing 2 percent random sample of all people with credit records in the United States in August 2013. We used the zip code (or county, when zip code was missing) and age variables for imputation. We also used variables on the amount of internal debt and external debt in collections to calculate a binary variable indicating whether each individual had any debt in collections,[26] which we used in our fitness-for-purpose analysis. The credit bureau data included individuals in US territories that are not included in the ACS data.[27] We had to drop these records (n = 4,302) from our analysis.

American Community Survey Data

We used 2011–2015 five-year ACS estimates of population counts by race/ethnicity and age by geography.[28] We used two different tables of five-year ACS data at three levels of geography. We used table B01001, subtables A through I (Sex by Age) and table B03002 (Hispanic or Latino Origin by Race). For each table, we collected data at the ZCTA, county, and national level. The estimates in these tables are reported disaggregated by gender, but we collapsed the gender variable in these data and calculated estimates for the total population because gender is not available in the credit bureau file.[29] We determine the race/ethnicity categories that we would impute based on the available race/ethnicity categories in the ACS data: Non-Hispanic White alone, Non-Hispanic Black or African American alone, Non-Hispanic American Indian and Alaska Native alone, Non-Hispanic Asian alone, Non-Hispanic Native Hawaiian and Other Pacific Islander, Non-Hispanic Some Other Race, Non-Hispanic Two or

More Races, and Hispanic. We used these categories because they are the race and ethnicity breakdowns available from the ACS, though we recognize that these categories may not accurately reflect the way that people in the credit bureau data would self-identify and may conceal important differences within groups.

Our credit bureau data reports the zip code where each person resides. But the ACS only makes available demographic data at the ZCTA level. Zip codes and ZCTAs sometimes overlap perfectly, but not always. To account for this, we used a zip-code-to-ZCTA crosswalk to assign ZCTAs to the credit data. The crosswalk is sourced from the American Academy of Family Physicians Uniform Data System for 2015,[30] and we simply join the file by zip code to the credit bureau data to add a ZCTA column. We used the Wayback Machine to find historical crosswalks but could not find the crosswalk for 2013. The 2015 crosswalk represents the closest in time and greatest degree of overlap among the crosswalks we could find. However, the slight time difference could introduce some error in our ZIP-to-ZCTA mapping. A small number (0.04%) of the ZIP codes in the credit bureau data could not be mapped to ZCTAs using our crosswalk. In these cases, we used the county data (or national, when both ZCTA and county were missing) from the ACS to perform the imputation.

Consumer Financial Protection Bureau Credit-Invisibility Data

The CFPB data showed the percentage of people without a credit record (i.e., people who are "credit invisible") in the United States by race and age (Brevoort, Grimm, and Kambara 2015). We used these data to adjust the ACS population estimates to reflect individuals with a credit record as described in step 2 ("randomly sample") below.

Step 2. Randomly Sample

To accurately reflect the uncertainty associated with the ACS estimates in our imputation process, we first generated a copy of the data by sampling from a normal distribution with the count estimate as the mean and the standard deviation equal to the reported 90 percent margin of error of that count divided by 1.645 (i.e., the ACS-reported standard error). We sampled the following estimates from the ACS.

Hispanic and Non-Hispanic White Populations by Age

The ACS publishes the counts of race/ethnicity by age range for the Hispanic and non-Hispanic white populations; it does not directly publish population counts by age range for other racial and ethnic groups (e.g., non-Hispanic Black people ages 35–44). Therefore, we can directly sample the estimates of race/ethnicity by age range for the Hispanic and non-Hispanic white populations, although we need to use the raking procedure described later in this appendix to produce the estimates of race/ethnicity by age range for other groups. Using these different procedures for different racial/ethnic groups could be a source of bias in our analysis. Because the raking procedure adds a layer of variability to our estimates of the counts of non-Hispanic and nonwhite racial/ethnic groups, our imputations could be less accurate for those groups. We decided to use this approach for two reasons. First, we wanted to use the best information available to produce our estimates, which would mean using the available estimates of race/ethnicity by age for the Hispanic and non-Hispanic white population. Second, we believe that any additional variability caused by raking is unlikely to be consistently biased from the true value in a given direction, which means that producing enough implicates should theoretically produce a consistent estimate.

Total Population by Racial/Ethnic Group

We sampled the total population by race/ethnicity for each geography of interest. These estimates were used as the column constraints for our raking procedure in step 3.

Total Non-Hispanic, Nonwhite Population by Age Range

Because the ACS does not directly publish these estimates, we calculated this ourselves by subtracting the total Hispanic population and total non-Hispanic white population for a given age range from the total population for that age range and then calculated the margin of error for this derived estimate using the ACS formulas.[31] We then used the calculated estimate and margin of error to sample the row constraints for our raking procedure in step 3.

We performed this sampling for the ACS estimates at the ZCTA, county, and national levels. Based on guidance from Census Bureau staff, we truncated any negative counts produced using this sampling procedure (such as in cases of small-count estimates with large standard deviations) at 0. This may skew our samples to the right, as we are left with only positive or zero counts. This was necessary to satisfy the technical constraints for the rest of our imputation procedure.

We also considered using a truncated normal distribution to perform the sampling but found that this more severely inflated the counts of racial/ethnic groups with zero-count estimates from the ACS, as the truncated normal distribution has a low probability mass for values near zero. And because using a truncated normal distribution was not recommended by the census, we decided to use the manual truncation at 0. We recognize that this may underrepresent the true uncertainty associated with zero-count estimates.

The raking procedure we subsequently used in step 3 to calculate the counts by race/ethnicity and age for the non-Hispanic and nonwhite groups required that the row constraints (the count of the non-Hispanic, nonwhite population in different age ranges) and column constraints (total population by racial/ethnic group) summed to the same total amount, representing the total non-Hispanic and nonwhite population across all age and racial/ethnic groups. But by independently sampling the row and column constraints, there was no guarantee that this would be the case (and it almost never was). To ensure that the row and column constraints summed to the same total, we took the sum of the column constraints and set it as the "ground truth" total population, which the row constraints must sum to. We then reweighted the row constraints as shown in the toy example with three racial/ethnic and age groups in table A.1.

TABLE A.1

Hypothetical Examples of Originally Sampled Row and Column Constraints

	Non-Hispanic Black count = 225	Non-Hispanic Asian count = 100	Non-Hispanic Native Hawaiian count = 75
Non-Hispanic, nonwhite 18–21 population = 50			
Non-Hispanic, nonwhite 21–25 population = 100			
Non-Hispanic, nonwhite 26–30 population = 50			

Source: Urban research team.

Notes: Grey cells to be filled in in tables A.3 and A.4 with a demonstration of the raking procedure. These numbers are hypotheticals and not derived from actual analysis.

We saw that the sampled column constraints summed to 400 and the sampled row constraints summed to 200. We reweighted the row constraints so the new totals summed to 400 while maintaining the relative population share of the sampled age groups as shown in table A.2.

TABLE A.2

Hypothetical Examples of Reweighted Sampled Row and Column Constraints

	Non-Hispanic Black count = 225	Non-Hispanic Asian count = 100	Non-Hispanic Native Hawaiian count = 75
Non-Hispanic, nonwhite 18-21 population = 400 * (50/200) = 100			
Non-Hispanic, nonwhite 21-25 population = 400 * (100/200) = 200			
Non-Hispanic, nonwhite 26-30 population = 400 * (50/200) = 100			

Source: Urban research team.

Notes: Grey cells to be filled in in tables A.3 and A.4 with a demonstration of the raking procedure. These numbers are hypotheticals and not derived from actual analysis.

Because we ultimately used the race/ethnicity *proportions* within each age group to perform the imputation and not the counts, we felt that this approach would enable us to meet the requirements of the raking procedure while maintaining the key information from sampling. We decided to use the sum of the column constraints as our “true total” that we would reweight the row constraints to meet—as opposed to reweighting the column constraints to equal the row constraints—because we directly sampled the column constraints from published ACS estimates and margins of error. On the other hand, we sampled each row constraint from a manually derived estimate that relied on three different ACS-reported variables and therefore had higher margins of error.

We considered using a chained sampling procedure to sample the row constraints instead of reweighting. This would have involved sampling each age group sequentially, where the first age group sampled would have a truncated lower bound of 0 and a truncated upper bound of the sum of the column constraints (the total non-Hispanic, nonwhite population). The second age group would then have an updated truncated upper bound of the sum of the column constraints minus the sampled estimate for the first age group, and so forth. But our testing found that the order in which the age groups were sampled significantly influenced the outcome, with the first age groups using up much of the available population and forcing the later age groups to 0. This was especially true for smaller geographies (e.g., ZCTAs), which had larger margins of error for the age group estimates. We felt this would likely skew the accuracy of our imputation results, so we decided to adopt the reweighting approach.

One challenge that we encountered with the reweighting approach was how to handle situations where the sampled column constraints (i.e., race constraints) had a nonzero total, but all of the row

constraints (i.e., age constraints) were sampled with zero estimates, and vice versa. In either case, it would be impossible to use reweighting to make the row and column constraints sum to the same total. In the former case, we decided to resample the age constraints until we had some nonzero results. This could introduce upward bias in the relative population shares of the nonwhite and non-Hispanic race groups by age range by requiring nonzero counts. In the latter case, we decided to set the age constraints and all of the raking cell values to zero. This could introduce downward bias in the relative population shares of the nonwhite and non-Hispanic race groups by age range by forcing the counts to zero. We treated these cases differently, as we chose to regard the sum of the column constraints as the "true total" in the reweighting procedure described above.

Step 3. Rake Counts

We used a technique called iterative proportional fitting, or "raking," to estimate cross-tabulations of race/ethnicity by age, essentially using the existing data on the population by age group and population by race/ethnicity to estimate the full cross-tabulations. As mentioned in the first step ("gather data"), because table B01001 (from the ACS data) does report cross-tabulation estimates for non-Hispanic white by age and Hispanic by age, we only performed this procedure for the remaining non-Hispanic racial categories (excluding non-Hispanic white).

To improve the performance of the raking process, we set initial seeds for the age range by race/ethnicity cells, which were used as the starting point for the algorithm. To set the seeds, we multiplied the count by age group and race alone from B01001 (e.g., Black age 45–54) by the proportion of the overall racial/ethnic population that is non-Hispanic from table B03002 (e.g., the percentage of the Black population that is non-Hispanic), as shown in table A.3, continuing our toy example with three racial/ethnic and age groups.

TABLE A.3

Hypothetical Examples of Initial "Seed" Values for Raking Algorithm

	Non-Hispanic Black count = 225	Non-Hispanic Asian count = 100	Non-Hispanic Native Hawaiian count = 75
Non-Hispanic, nonwhite 18–21 population = 100	Initial seed = 80 (Black population 18–21) * 0.75 (proportion of Black population that is non-Hispanic) = 60	Initial seed = 50 (Asian population 18-21) * 0.9 (proportion of Asian population that is non-Hispanic) = 45	Initial seed = 40 (Native Hawaiian population 18–21) * 0.8 (proportion of Native Hawaiian population that is non-Hispanic) = 32
Non-Hispanic, nonwhite 22–25 population = 200	Initial seed = 100 (Black population 22–25) * 0.75 (proportion of Black population that is non-Hispanic) = 75	Initial seed = 40 (Asian population 22-25) * 0.9 (proportion of Asian population that is non-Hispanic) = 36	Initial seed = 35 (Native Hawaiian population 22–25) * 0.8 (proportion of Native Hawaiian population that is non-Hispanic) = 28
Non-Hispanic, nonwhite 26–30 population = 100	Initial seed = 120 (Black population 26–30) * 0.75 (proportion of Black population that is non-Hispanic) = 90	Initial seed = 40 (Asian population 26–30) * 0.75 (proportion of Asian population that is non-Hispanic) = 36	Initial seed = 25 (Native Hawaiian population 26–30) * 0.8 (proportion of Native Hawaiian population that is non-Hispanic) = 20

Source: Urban research team.
Note: These numbers are hypotheticals and not derived from actual analysis.

Though it is likely incorrect to assume that the share of the total population that is non-Hispanic is constant across age groups, doing so provided informed starting seeds for raking that improved the accuracy of the output relative to random seeds in our tests. However, if this assumption is truer for some racial/ethnic groups than others, this could yield results that are differentially accurate by racial/ethnic group.

We see in table A.3 above that the seed values in each row and column do not sum to the row and column constraints, respectively. The raking procedure iteratively updates those seed values to identify the optimal cell values for the counts of race/ethnicity by age group that sum to the row and column constraints (Lomax and Norman 2016). In brief, one step of the raking algorithm first scales the seed values to sum to the row constraints and then scales the updated values to sum to the column constraints. This step is repeated until the algorithm converges when the total absolute change in cell values between two steps is less than a specified tolerance value. For our analysis, we set this value at 0.10.[32] A possible result of the raking algorithm,[33] using our toy example, is shown in table A.4.

TABLE A.4

Hypothetical Examples of Fitted Values Resulting from Raking Algorithm

	Non-Hispanic Black count = 225	Non-Hispanic Asian count = 100	Non-Hispanic Native Hawaiian count = 75
Non-Hispanic, non-white 18-21 population = 100	65	20	15
Non-Hispanic, non-white 22-25 population = 200	100	65	35
Non-Hispanic, non-white 26-30 population = 100	60	15	25

Source: Urban research team.

Note: These numbers are hypotheticals and not derived from actual analysis.

Step 4. Adjust Counts

Before we were ready to impute, we had to adjust our raked race/ethnicity-by-age-group numbers to exclude the "credit-invisible" population. We used nationally reported credit-invisibility statistics from a CFPB report to adjust the population counts downward. To calculate appropriate adjustments for each age by racial/ethnic group, we used table 8 of that report's appendix, calculating weighted averages to aggregate the more granular age groups in the report to the ACS age groups.[34] We then reduced the counts in our ACS cells by the credit invisible percentage so that the remaining counts represented the credit visible population within each race/ethnicity and age range group and therefore were comparable to the credit bureau data. We show this adjustment using our toy example in table A.5 below, but note that, in addition to performing this adjustment for the raked values, we also adjusted the directly sampled values for non-Hispanic white and Hispanic populations by age.

TABLE A.5

Examples of Adjusted Values for Credit Invisibility

	Non-Hispanic Black	Non-Hispanic Asian	Non-Hispanic Native Hawaiian
Age 18–21	65 * 60% credit visible = 39	20 * 70% credit visible = 14	15 * 60% credit visible = 9
Age 22–25	100 * 70% credit visible = 70	65 * 80% credit visible = 52	35 * 80% credit visible = 28
Age 26–30	60 * 75% credit visible = 45	15 * 80% credit visible = 12	25 * 80% credit visible = 20

Source: Urban research team.

Note: These numbers are hypotheticals and not derived from actual analysis.

These data have the following limitations that could introduce bias into our methodology:

- The CFPB estimates are at the national level. We therefore must assume that these rates hold within age and racial/ethnic groups across the entire United States, which is likely untrue and

may be particularly so for some geographic subgroups, such as rural areas or areas with large immigrant populations.

- The CFPB estimates are from 2010, whereas the credit data are from 2013. If credit invisibility has changed differentially by race and ethnicity in that period, this could be a source of bias.
- The CFPB aggregates data for multiple smaller racial/ethnic groups reported in the ACS into a single "other" category,[35] which we apply to all the constituent groups. This aggregation masks potential heterogeneity in credit invisibility among these groups and is likely more accurate for the larger race/ethnicity groups within this "other" category.
- The age ranges in the CFPB report do not exactly match the age ranges present in the ACS data. For all groups older than 35, the CFPB data report smaller age ranges than the ACS data. For example, the ACS includes the range of 35–44, whereas the CFPB data report credit-invisibility rates for both 35–39 and 40–44. To mitigate this mismatch, we took a simple average across the two age ranges, but this means that older people, especially in geographies containing an uneven distribution within age ranges, will likely have less accurate credit-invisibility rates.

Step 5. Calculate Probabilities and Assign Race

We converted our raked and sampled counts by race/ethnicity, age group, and location (i.e., ZCTA, county, and nation) into proportions and created a vector of race/ethnicity probabilities for each combination of age group and location. We show this calculation for our toy example in table A.6.

TABLE A.6

Hypothetical Examples of Race/Ethnicity Probabilities by Age

	Non-Hispanic Black	Non-Hispanic Asian	Non-Hispanic Native Hawaiian
Age 18–21	39 / 62 = 0.63	14 / 62 = 0.23	9 / 62 = 0.14
Age 22–25	70 / 150 = 0.47	52 / 150 = 0.34	28 / 150 = 0.19
Age 26–30	45 / 77 = 0.58	12 / 77 = 0.16	20 / 77 = 0.26

Source: Urban research team.

Note: These numbers are hypotheticals and not derived from actual analysis.

We then merged these probabilities onto the credit bureau data by the age and location of each observation. In the cases where an individual's ZCTA was missing in the credit bureau data, we used the race/ethnicity data of their age group (if present) for the county (or for the nation, if county was also missing). If an individual's age group was missing in the credit bureau data, we used the race/ethnicity data for all adults at the smallest available geography. We also encountered the case where an individual's age and location was present in the credit bureau data, but the sampling and raking steps produced a total population of zero in the given age group and location, which created missing probabilities for that age group and location. In that case, we first tried to use the probabilities for the population older than 18 for that location, and if that was also missing we used the age-range probabilities for that age range in the next coarser geography until we had nonmissing data.

At this point we had credit bureau data appended with a vector of race/ethnicity probabilities for each record. We then randomly assigned race/ethnicity to each observation in the credit bureau data based on the probabilities in this vector. For example, if an individual's vector was 30 percent Hispanic, 70 percent non-Hispanic white, and 0 percent for all other racial categories, there would be a 30 percent probability of assigning Hispanic and a 70 percent probability of assigning non-Hispanic white as the race/ethnicity for that observation.

Step 6. Creating Multiple Implicates

This process produces a single race and ethnicity variable. Although a single implicate may be fairly accurate at the national level, the variation introduced by the input datasets and imputation process can yield inaccurate results for smaller subpopulations. But when we repeat the process multiple times and average across all the implicates, we should be able to obtain accurate estimates. Moreover, using multiple implicates to calculate estimates from the imputed data allows users to accurately estimate the

variance, as we can account for the variation introduced by the imputation process and measurement error in the input data sources.

For our analysis, we repeated each of steps 1 through 5 as described above 50 times to create 50 different implicates. Each implicate uses a different set of randomly sampled ACS data as described in step 2 (“randomly sample”) as the basis for that specific imputation. Because the ACS datasets for each implicate were created by randomly drawing from the counts and standard errors reported, we ensured each implicate captures variation introduced by the margins of error reported in the ACS data. In addition, by randomly assigning the final race/ethnicity variable based on the generated probabilities in step 5 (“calculate probabilities and assign race”), we ensured that each implicate captures the variation introduced by the imputation process.

Calculating Estimates and Variances Using Multiple Implicates

We used the 50 implicates to calculate several estimates, and the margins of error of those estimates, used in checkpoint 3, as described in the body of this report. To accurately calculate estimates and margins of error using multiple implicates, we followed the methodology for calculating inferences from multiply imputed, partially synthetic datasets in Reiter (2003) and Kim, Drechsler, and Thompson (2019). Reiter (2003) outlines the following four quantities needed to calculate estimates across *m* implicates:

Calculating the estimate:

$$\bar{q}_m = \sum_{i=1}^{m} \frac{q_i}{m}$$

Calculating variance between implicates:

$$b_m = \sum_{i=1}^{m} \frac{(q_i - \bar{q}_m)^2}{m-1}$$

Calculating average variance within each implicate:

$$\bar{v}_m = \sum_{i=1}^{m} \frac{v_i}{m}$$

Calculating the variance of the estimate:

$$T_m = \frac{b_m}{m} + \bar{v}_m$$

Where q_i is equal to the estimate in a single implicate, v_i is equal to the variance of that estimate in a single implicate, and *m* is the number of implicates. The 95 percent confidence interval for the estimate $\bar{q}_m$ can be calculated by taking 1.645 * $\sqrt{T_m}$ where T_m is the variance of the estimate, $\sqrt{T_m}$ is the standard error of the estimate, and 1.645 is the 95 percent critical value.

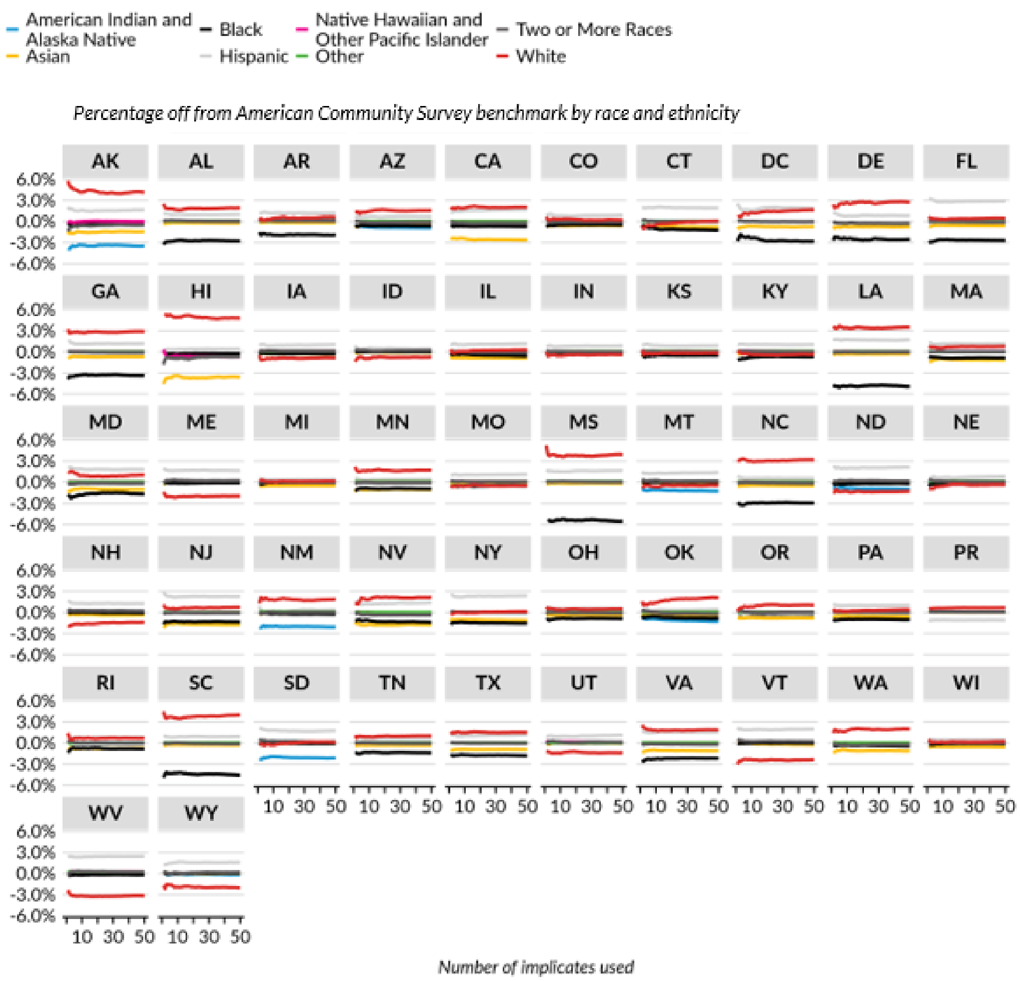
We did encounter a few edge cases when implementing this methodology, which were as follows:

- **Calculating averages of a variable by imputed race/ethnicity and geography.** It is possible that some implicates will assign some observations in a given geography (e.g., ZCTA X) to a given race/ethnicity (e.g., Hispanic), while other implicates assign no observations in that geography to the given race/ethnicity. To calculate, for example, the proportion of Hispanic people in ZCTA X with debt in collections, we decided based on consultations with experts in the field that the best approach is to drop the implicates where no people are assigned Hispanic in ZCTA X when calculating $\bar{q}_m$ and setting *m* in that equation to the number of nonmissing implicates. As a best practice, we recommend analysts report the number of implicates used in the calculation in addition to the estimate and margin of error in any data releases.
- **Calculating population counts by imputed race/ethnicity and geography.** Using the same example above, if we wanted to estimate the population counts by race/ethnicity in ZCTA X, we could use 0 as the implicate estimate q_i of the Hispanic population count for those implicates that do not assign any observations in ZCTA X for the Hispanic group. However, there is no clear guidance on how to calculate the sample variance v_i of the 0-count estimates in those implicates in order to calculate $\bar{v}_m$. For the purposes of this analysis, we decided to impute the "missing" v_i values with one-half of the variance for a count estimate of one, which we felt was a reasonable value in this case. In future analyses, we would test different imputation strategies to verify that our results were robust to this choice.
- **Calculating estimates for very small subpopulations.** When calculating estimates for small subpopulations, such as a single racial/ethnic group at the ZCTA or county level, one might find oneself using a very small number of observations to perform the calculation. As a best practice, we suppressed any estimate where the average number of observations used across implicates was fewer than 50.

Appendix B. Discrepancy from ACS Benchmarks by Number of Implicates

FIGURE B.1

Discrepancies between Implicates and American Community Survey Benchmarks at the State Level



URBAN INSTITUTE

Sources: Authors' analysis of 2013 data from a major credit bureau and 2011–2015 five-year Census Bureau American Community Survey estimates adjusted using Consumer Financial Protection Bureau estimates of credit invisibility and estimates of the adult population calculated from the 2011–2015 American Community Survey microdata from the Integrated Public Use Microdata Series.

Notes: The race and ethnicity values in the credit bureau implicates are estimated via multiple imputation. The major credit bureau does not collect this information. This figure shows only geography and race/ethnicity group combinations where all implicates include at least 50 observations.

FIGURE B.2

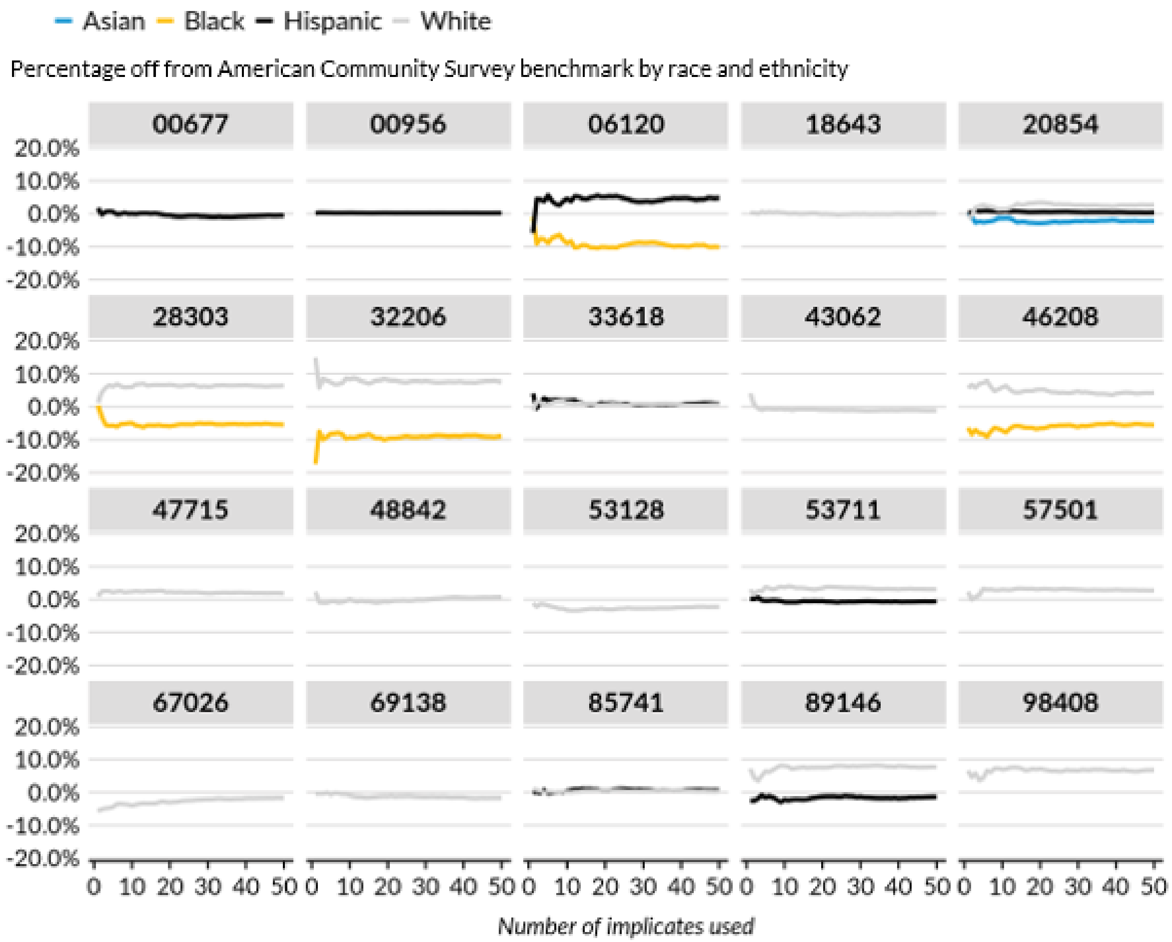
Discrepancies between Implicates and American Community Survey Benchmarks at the County Level for 20 Randomly Selected Counties

URBAN INSTITUTE

Sources: Authors' analysis of 2013 data from a major credit bureau and 2011–2015 five-year Census Bureau American Community Survey estimates adjusted using Consumer Financial Protection Bureau estimates of credit invisibility and estimates of the adult population calculated from the 2011–2015 American Community Survey microdata from the Integrated Public Use Microdata Series.
Notes: The race and ethnicity values in the credit bureau implicates are estimated via multiple imputation. The major credit bureau does not collect this information. This figure shows only geography and race/ethnicity group combinations where all implicates include at least 50 observations.

FIGURE B.3

Discrepancies between Implicates and American Community Survey Benchmarks at the ZCTA Level for 10 Randomly Selected ZCTAs



URBAN INSTITUTE

Sources: Authors' analysis of 2013 data from a major credit bureau and 2011–2015 five-year Census Bureau American Community Survey estimates adjusted using Consumer Financial Protection Bureau estimates of credit invisibility and estimates of the adult population calculated from the 2011–2015 American Community Survey microdata from the Integrated Public Use Microdata Series.

Notes: ZCTA = zip code tabulation area. The race and ethnicity values in the credit bureau implicates are estimated via multiple imputation. The major credit bureau does not collect this information. This figure shows only geography and race/ethnicity group combinations where all implicates include at least 50 observations.

Notes

[1] See “Making the Case for Data Disaggregation to Advance a Culture of Health,” PolicyLink, accessed July 6, 2021, https://www.policylink.org/our-work/community/health-equity/data-disaggregation, and “Why Disaggregating Data by Race Is Important for Racial Equity,” Annie E. Casey Foundation blog, last updated August 18, 2020, https://www.aecf.org/blog/taking-data-apart-why-a-data-driven-approach-matters-to-race-equity/.

[2] Alena Stern, Graham MacDonald, and Khuloud Odeh, “Creating Equitable Technology Programs: A Guide for Cities,” Urban Institute, September 2020, https://apps.urban.org/features/how-to-create-equitable-technology-programs/#intro.

[3] For more on the racial homeownership gap, see Choi and coauthors (2019). For a discussion on the lack of information on race and ethnicity in tax data, see Bearer-Friend (2019). For more on problems with employer credit checks, see Traub and McElwee (2016).

[4] See “American Community Survey (ACS),” US Census Bureau, accessed July 2, 2021, https://www.census.gov/programs-surveys/acs.

[5] Many of the ACS variables we use (such as counts by race and age) are reported disaggregated by gender. We had to combine these to calculate an estimate and margins of error for the total population because the credit bureau data do not include gender.

[6] For more information, see Aaron Glantz and Emmanuel Martinez, “For People of Color, Banks Are Shutting the Door to Homeownership,” Reveal, February 15, 2018, http://revealnews.org/article/for-people-of-color-banks-are-shutting-the-door-to-homeownership/, and Michelle Singletary, “Credit Scores Are Supposed to Be Race-Neutral. That’s Impossible,” *Washington Post*, October 16, 2020, https://www.washingtonpost.com/business/2020/10/16/how-race-affects-your-credit-score/.

[7] American Community Survey race and ethnicity categories include “Non-Hispanic White alone,” “Non-Hispanic Black or African American alone,” “Non-Hispanic American Indian and Alaska Native alone,” Non-Hispanic Asian alone,” “Non-Hispanic Native Hawaiian and Other Pacific Islander,” “Non-Hispanic Some Other Race,” “Non-Hispanic Two or More Races,” and “Hispanic.”

[8] These territories are Palau, the Marshall Islands, the Federated States of Micronesia, Guam, American Samoa, the Northern Mariana Islands, and the US Virgin Islands.

[9] See “2013 Survey of Consumer Finances,” Board of Governors of the Federal Reserve System, May 20, 2021, https://www.federalreserve.gov/econres/scf_2013.htm.

[10] The SCF also represents the population of all households, so using it would have required an adjustment for credit invisibility, which would have been complicated by the fact that our CFPB credit-invisibility data are at the individual level rather than the household level. In addition, the comparability of the SCF statistics and the credit bureau data is uncertain because the questions about holding debt are subject to respondents’ interpretation of what types of debt to include (e.g., whether to include a deferred student loan as debt held by the household).

[11] See, for example, john powell, “Race, Place, and Opportunity,” *American Prospect*, September 21, 2008, https://prospect.org/special-report/race-place-opportunity/, and Katherine Schaeffer, “The Most Common Age among Whites in U.S. Is 58 – More than Double That of Racial and Ethnic Minorities,” Pew Research Center, July 30, 2019, https://www.pewresearch.org/fact-tank/2019/07/30/most-common-age-among-us-racial-ethnic-groups/.

[12] We used a historical version of the Uniform Data System Mapper crosswalk generated in 2015 (which was the closest to 2013 we could find). If the zip code or ZCTA definitions changed between those two years, the ZCTAs

we used may be inaccurate. See "ZIP Code to ZCTA Crosswalk," UDS Mapper, accessed July 2, 2021, https://udsmapper.org/zip-code-to-zcta-crosswalk/.

[13] The zip-code-to-ZCTA crosswalk file could not match the zip codes of 2,066 records in the credit bureau data, effectively forcing us to encode those individuals' ZCTAs as missing. Nearly every record where the zip code field was missing also had county missing. In the cases where county was missing and zip code was present, we used the zip code to impute the county based on the proportions of the zip code population that fell in each overlapping county. This imputed county variable was only used in the analysis of the imputed data, not in the imputation process.

[14] We also considered including variables on whether an individual's ZCTA had no dominant racial/ethnic category constituting more than 50 percent of the ZCTA's total population and whether an individual had any debt in collections, but we dropped these variables because they were highly positively correlated with a majority nonwhite ZCTA population and subprime credit score variables, respectively, and those variables were in turn more highly correlated with the missingness of the age variable.

[15] For the purposes of this analysis, we define a subprime consumer as a person with a credit score of 600 or lower, following the example from Braga and coauthors' "Debt in America: An Interactive Map."

[16] We could test the impact of missing age data by removing the age information for a random sample of rows where that information is present, and testing whether and how the imputation results differ.

[17] We used these reported estimates to calculate the total non-Hispanic, nonwhite population in each age range by subtracting the non-Hispanic white and Hispanic population totals in age range X from the total population in age range X, and using formulas reported by the census to calculate the margin of error for this derived estimate. These calculated totals served as the row totals for the raking step.

[18] We did this because we generated the sampled race/ethnicity totals from a single ACS variable, whereas the sampled age totals were generated from three ACS variables (i.e., [Non-Hispanic, nonwhite population] = [Total Population] – [Non-Hispanic white population] – [Hispanic population]). In other words, we trusted the sampled race/ethnicity totals more than we trusted the age totals.

[19] These groups are Non-Hispanic American Indian and Alaska Native, Non-Hispanic Native Hawaiian and Other Pacific Islander, Non-Hispanic Some Other Race, and Non-Hispanic Two or More Races.

[20] The Integrated Public Use Microdata Series publishes the ACS microdata, which can be used to estimate the share of each racial/ethnic group's population that is 18 and older. The ACS microdata do not include Native Hawaiian and Other Pacific Islander as a separate racial category, so we used the rate for other racial/ethnic groups. For more detail, see Ruggles and coauthors' *IPUMS USA: Version 11.0* [dataset], available at https://doi.org/10.18128/D010.V11.0.

[21] We ran an exploratory test where we collapsed the Native Hawaiian and Pacific Islander, Alaska Native and American Indian, "all other races," and "two or more races" groups into one "other" category before the sampling step to assess whether this improved results, but we found no improvement in accuracy. We did not test the impact of using location alone for imputation.

[22] For more information on how to calculate the uncertainty of estimates derived from multiple implicates, see appendix A.

[23] In our analysis, we consider all of the racial/ethnic groups besides white people people of color, following the example from Braga and coauthors' "Debt in America: An Interactive Map."

[24] We do not publish estimates based on fewer than 50 observations. It is responsible practice to suppress estimates based on small numbers of observations because of concerns about data privacy and accuracy at very small *n* estimates. The precise number of estimates will vary based on your use case.

[25] To calculate this, we identified the number of neighborhoods where any share of the 95 percent confidence interval around the estimate of the proportion of Black adults with debt in collections is greater than or equal to the 90th percentile value for Wayne County of 74 percent. Eighteen ZCTAs in Wayne County met this condition.

[26] We followed the variable definition used in Braga and coauthors' 'Debt in America: An Interactive Map" feature. For more information, see the feature technical appendix.

[27] These are Palau, the Marshall Islands, the Federated States of Micronesia, Guam, American Samoa, the Northern Mariana Islands, and the US Virgin Islands.

[28] See "American Community Survey (ACS)," US Census Bureau, accessed July 2, 2021, https://www.census.gov/programs-surveys/acs.

[29] We use formulas provided by the Census Bureau to calculate an estimate and margins of error for the total population, as the credit bureau data do not include gender. The formulas provided by the Census Bureau to approximate margins of error in such cases are likely to be less accurate for small populations or groups with zero population. This means that, for small populations and smaller racial/ethnic groups, we likely underestimate the uncertainty.

[30] See "ZIP Code to ZCTA Crosswalk," UDS Mapper, accessed July 2, 2021, https://udsmapper.org/zip-code-to-zcta-crosswalk/.

[31] We use formula (1) given in this document as described for aggregating across population subgroups on pages 54–55. Note that in cases where multiple subgroups have zero count estimates, the census recommends using only one of the zero-estimate variance terms in the margin-of-error calculations. See page 21 of this document for more details.

[32] It is also possible that the algorithm may never converge, in which case we specify for the algorithm to stop early after 1,000 iterations. Though this happens fairly rarely, it is more common in cases where there are many zero values in the row or column constraints.

[33] These numbers are produced by the authors as an illustrative example, not by the raking algorithm.

[34] In all cases where the CFPB and ACS age groups did not align, the CFPB age group split the ACS age group in half. For example, the CFPB reports credit invisibility for ages 35–39 and ages 40–44, whereas the ACS reports population for ages 35–44. To calculate credit invisibility for people ages 35–44, we simply took the average of the 35–39 and 40–44 figures. Doing this assumes that the population is uniformly distributed between these two halves of the age range, but we do not have more granular population estimates with which to calculate a weighted average.

[35] These include "Non-Hispanic American Indian and Alaska Native," "Non-Hispanic Native Hawaiian and Other Pacific Islander," "Non-Hispanic Some Other Race," and "Non-Hispanic Two or More Races."

References

Bearer-Friend, Jeremy. 2018. "Should the IRS Know Your Race? The Challenge of Colorblind Tax Data." *Tax Law Review* 73 (1): 1–68.

Brevoort, Kenneth P., Philipp Grimm, and Michelle Kambara. 2015. "Data Point: Credit Invisibles." Washington, DC: Consumer Financial Protection Bureau.

Choi, Jung Hyun, Caitlin Young, Alanna McCargo, Michael Neal, and Laurie Goodman. 2019. *Explaining the Black-White Homeownership Gap*. Washington, DC: Urban Institute.

Kim, Hang J., Jörg Drechsler, and Katherine J. Thompson. 2019. "Synthetic Microdata for Establishment Surveys under Informative Sampling." 2019–07. Washington, DC: US Census Bureau.

Lomax, Nik, and Paul Norman. 2016. "Estimating Population Attribute Values in a Table: 'Get Me Started in' Iterative Proportional Fitting." *The Professional Geographer* 68 (3): 451–61. https://doi.org/10.1080/00330124.2015.1099449.

Reiter, J. P. 2003. "Inference for Partially Synthetic, Public Use Microdata Sets." *Survey Methodology* 29 (2): 181–88.

Traub, Amy, and Sean McElwee. 2016. *Bad Credit Shouldn't Block Employment: How to Make State Bans on Employment Credit Checks More Effective*. New York: Demos.

About the Authors

Alena Stern is a senior data scientist at the Urban Institute studying policy solutions to advance equity and inclusion in cities. Before joining Urban, she worked as a senior program manager with AidData, an Open Cities fellow at the Sunlight Foundation, and a graduate research assistant at the Center for Data Science and Public Policy, where she used machine learning, natural language processing, statistical analysis, and geospatial data to inform the design of government policies and international development programs. Stern holds a BA in economics and international relations from the College of William and Mary and an MS in computational analysis and public policy from the University of Chicago.

Ajjit Narayanan is a data scientist at the Urban Institute, where he works on research to advance equity and inclusion in rural and urban communities. He enjoys building open source data and mapping tools for researchers and advises on tools such as interactive mapping, big data platforms, and data visualization methods to relevant public policy issues. Narayanan also coleads Urban's R Users Group and Mapping Users Group to encourage the use of R and geospatial analysis for statistical analysis, data visualization, and automation. Narayanan holds a BA in economics and a BA in urban studies, along with minors in math and statistics, from the University of Pennsylvania.

Statement of Independence

URBAN INSTITUTE · ELEVATE · THE · DEBATE

500 L'Enfant Plaza SW
Washington, DC 20024

www.urban.org