Research Report

Imputation of Race and Ethnicity in Health Insurance Marketplace Enrollment Data, 2015–2022 Open Enrollment Periods

Melony E. Sorbero, Roald Euller, Aaron Kofner, Marc N. Elliott

For more information on this publication, visit **www.rand.org/t/RRA1853-1**.

Published by the RAND Corporation, Santa Monica, Calif.

About This Report

Information on the race and ethnicity of individuals enrolled through the Health Insurance Marketplaces is critical for assessing past enrollment efforts and determining whether outreach campaigns should be modified or tailored moving forward. However, approximately one-third of insurance applicants do not complete the race and Hispanic ethnicity questions on the Marketplace application. This report presents the results of imputing race and ethnicity for enrollees from 2015 through 2022 using the modified Bayesian Improved First Name Surname and Geocoding method, developed by the RAND Corporation, which uses surnames, first names, and residential addresses to indirectly estimate race and ethnicity. These findings should be useful to staff at the Office of the Assistant Secretary for Planning and Evaluation and the Centers for Medicare and Medicaid Services.

This research was funded by the Office of the Assistant Secretary for Planning and Evaluation and carried out within the Payment, Cost, and Coverage Program in RAND Health Care.

RAND Health Care, a division of the RAND Corporation, promotes healthier societies by improving health care systems in the United States and other countries. We do this by providing health care decisionmakers, practitioners, and consumers with actionable, rigorous, objective evidence to support their most complex decisions. For more information, see www.rand.org/health-care, or contact:

RAND Health Care Communications
1776 Main Street
P.O. Box 2138
Santa Monica, CA 90407-2138
(310) 393-0411, ext. 7775
RAND_Health-Care@rand.org

Acknowledgments

We extend thanks to Keith Branham, Kenneth Finegold, Lucy Chen, and Cinthya Alberto at the Office of the Assistant Secretary for Planning and Evaluation and Christine Eibner and Paul Koegel from the RAND Corporation for reviewing a draft of this report and providing helpful comments. We also thank Kathleen Call and Giovann Alarcon from SHADAC at the University of Minnesota and Justin Timbie from RAND for their reviews of this report.

Summary

Information on the race and ethnicity of individuals enrolling through the Health Insurance Marketplaces is critical for assessing past enrollment efforts, determining whether outreach campaigns should be modified or tailored moving forward, and identifying where to target outreach activities. However, approximately one-third of insurance applicants do not complete the race and Hispanic ethnicity questions on the Marketplace application.

The RAND Corporation's modified Bayesian Improved First Name Surname and Geocoding (BIFSG) method uses surnames, first names, and residential addresses to indirectly estimate race and ethnicity. We used 2015–2022 data from the Centers for Medicare and Medicaid Services Multidimensional Insurance Data Analytics System (MIDAS), which contains person-year level data for Marketplace enrollees. The surname and first name for each individual were used to estimate initial probabilities for each of the six mutually exclusive racial and ethnic groups: American Indian/Alaskan Native (AI/AN); Asian American, Native Hawaiian, and Pacific Islander (AANHPI); Black; Hispanic; Multiracial; and White. Geocoded address information was used to refine these estimations and generate final probabilities.

Self-reported race and ethnicity were missing for 32.5 percent of the 71,610,609 records across the eight years of MIDAS enrollment data (2015 through 2022). Using enrollees' records from other years to replace missing race and ethnicity reduced the level of missingness to 23.5 percent. Enrollees who self-reported race and ethnicity were more likely to be AANHPI than nonreporting enrollees for whom race and ethnicity were imputed (9.4 percent versus 6.7 percent) or White (59.5 percent versus 49.3 percent) and less likely to be Black (10.9 percent versus 15.7 percent) or Hispanic (17.9 percent versus 26.1 percent). When combining self-reported race and ethnicity data with the imputed race and ethnicity probabilities for enrollees who did not report their race and ethnicity, we estimated that 8.7 percent of Marketplace enrollees were AANHPI; 0.6 percent were AI/AN; 12.0 percent were Black; 19.8 percent were Hispanic; 1.8 percent were Multiracial; and 57.1 percent were White.

Based on conventional standards for C-statistics, the ability of the modified BIFSG to differentiate AANHPI, Black, Hispanic, and White enrollees from other groups was "excellent." It did not reach an "acceptable" level for AI/AN or Multiracial enrollees.[1] Currently, we recommend that modified BIFSG-imputed race and ethnicity not be used to make inferences about AI/AN or Multiracial enrollees.

[1] A C-statistic of 0.7 is considered "acceptable"; 0.8 is considered "strong" (Hosmer and Lemeshow, 2000); and 0.9 or higher is considered "excellent" (per authors).

This report describes the modified BIFSG and the steps involved in its application; presents the results of the imputation; and provides an assessment of the algorithm's performance, both overall and for subgroups of Marketplace enrollees.

Contents

Figures and Tables

Figures

Tables

Chapter 1. Introduction

Information on the race and ethnicity of individuals enrolling through the Health Insurance Marketplace is critical for assessing past enrollment efforts, determining whether outreach campaigns should be modified or tailored moving forward, identifying where to target outreach activities, and multiple other potential policy uses. However, approximately one-third of insurance applicants do not complete two optional questions on race and Hispanic ethnicity on the application.

When self-reported race and ethnicity information is missing, other information about an individual can be used to infer race and ethnicity with some range of uncertainty, such as surnames, first names, and addresses (as we will explain later in this chapter), with each characteristic contributing meaningfully to the identification of each group.

For the first time, in 2007, the U.S. Census Bureau released a tabulation of more than 6 million unique surnames with their associated percentages in six race and ethnicity groups. The tabulation was based on self-reported data from the 2000 Census and subsequently updated for the 2010 Census (U.S. Census Bureau, undated). Surnames are particularly useful for distinguishing people who identify as Hispanic and Asian American, Native Hawaiian, and Pacific Islander (AANHPI) from other racial and ethnic groups, which is highlighted by the fact that more than 90 percent of people with the very common last names of Garcia, Martinez, and Rodriguez (Table 1.1) self-identified as Hispanic. Similarly, 94 percent of people with the surname Kim self-identified as AANHPI (not included in Table 1.1).

Table 1.1. Race and Ethnicity Percentages for the Ten Most-Common Surnames in the United States, 2010 Census

Name	Rank	American Indian/Alaskan Native (AI/AN)	AANHPI	Black	Hispanic	Multiracial	White
Smith	1	0.89	0.50	23.11	2.40	2.19	70.90
Johnson	2	0.94	0.54	34.63	2.36	2.56	58.97
Williams	3	0.82	0.46	47.68	2.49	2.81	45.75
Brown	4	0.87	0.51	35.60	2.52	2.55	57.95
Jones	5	1.00	0.44	38.48	2.29	2.61	55.19
Garcia	6	0.47	1.41	0.45	92.03	0.26	5.38
Miller	7	0.66	0.54	10.76	2.17	1.77	84.11
Davis	8	0.82	0.49	31.60	2.44	2.45	62.20
Rodriguez	9	0.18	0.57	0.54	93.77	0.18	4.75
Martinez	10	0.51	0.60	0.49	92.91	0.22	5.28

NOTE: The data source is the 2010 Census (U.S. Census Bureau, undated). Percentages are calculated from responses to questions on race and Hispanic ethnicity and the names that are provided when people complete the Census form. Percentages may not sum to 100 because of rounding.

Information on residential addresses can be geocoded to Census block groups and linked to racial and ethnic distributions from the most recently available decennial Census. These data are particularly useful in distinguishing Black and White individuals who frequently reside in racially segregated neighborhoods. However, the predictive power of an address does vary by location. For example, Detroit, Michigan, is a highly segregated city, while Las Vegas, Nevada, is much less segregated (Frey, 2018).

The Bayesian Improved Surname and Geocoding (BISG) method, developed by the RAND Corporation, uses both surnames and residential addresses to indirectly estimate race and ethnicity of health plan enrollees (Elliott et al., 2008; Elliott et al., 2009; Fremont et al., 2016). In brief, a surname is used to estimate the prior probabilities for each of six mutually exclusive racial and ethnic groups for each individual: AANHPI, AI/AN, Black, Hispanic, Multiracial, and White.[2] Geocoded address information is used to refine these estimations and generate posterior probabilities.

Validation studies found that BISG has an average accuracy of 93 percent by the common area under the curve (AUC) measure in commercial populations (Elliott et al., 2013; Grundmeier et al., 2015); performance may be higher in Medicare populations. RAND researchers have adapted the BISG methodology to Medicare data by incorporating additional administrative data, including Centers for Medicare and Medicaid Services (CMS) administrative data on race and

[2] Individuals reporting Hispanic ethnicity are categorized as Hispanic. All race categories are restricted to individuals not reporting Hispanic ethnicity.

ethnicity, in the imputation methodology, which results in an AUC accuracy of 99 percent for Black, 98 percent for AANHPI, 96 percent for White, and 95 percent for Hispanic Medicare beneficiaries (Dembosky et al., 2019; Haas et al., 2019; CMS, Office of Minority Health, 2021).

Including first names in the BISG further improves the method's accuracy (Voicu, 2018; Haas et al., 2019; CMS, Office of Minority Health, 2021), reducing false negative rates by up to 3.3 percent and false positive rates by up to 4 percent, depending on the racial and ethnic group (Voicu, 2018). The addition of first names to the BISG uses data that contain the estimated probability of each of the six racial and ethnic groups for 4,250 unique first names that were drawn from almost 2.5 million mortgage applications in 2007 and 2010 along with self-reported race and ethnicity from the applications (Tzioumis, 2018). For example, 96 percent of people with the first name Andreas self-identified as White, while 83 percent of people with the first name Andres self-identified as Hispanic. When a first name is included in the race and ethnicity imputation method, it is referred to as the *Bayesian Improved First Name Surname and Geocoding* (BIFSG) method. Additional RAND refinements to the BISG methodology improved the use of compound and rare surnames, as described in Chapter 2, in addition to the inclusion of first names, and also improved its accuracy (Haas et al., 2019). The version used for this report, which includes both use of first names and improved use of surnames, is referred to as the *modified BIFSG*.

This project built on 2017 work from Justin Timbie performed for the Office of the Assistant Secretary for Planning and Evaluation (ASPE) that applied the BISG Marketplace enrollment data from to 2014 through 2016. Our report describes the results of applying the modified BIFSG to individuals who enrolled in health insurance plans through the Health Insurance Marketplaces during the open enrollment periods (OEPs) from 2015 through 2022. Imputed probabilities of race and ethnicity from the modified BIFSG are used to fill in missing race and ethnicity information for enrollees who do not self-report race and ethnicity in any year they enrolled.

In Chapter 2, we describe the data used to impute race and ethnicity and the steps involved in implementing the modified BIFSG methodology. In Chapter 3, we present the results of performing the imputation and assess the accuracy of the imputation compared with self-reported race and ethnicity. In Chapter 4, we highlight potential uses for imputed race and ethnicity, discuss limitations of the modified BIFSG methodology, and describe refinements that could be implemented in the future.

Chapter 2. Methods

In this chapter, we describe the data used to impute race and ethnicity estimates. We then describe the steps involved in implementing the modified BIFSG.

MIDAS Data

Our project used an extract of data from the Multidimensional Insurance Data Analytic System (MIDAS), which contains person-year–level records for enrollees in health plans offered by Federally Facilitated Marketplaces and purchased through HealthCare.gov or State-Based Marketplaces (also referred to as *State-Based Exchanges*) that use the federal platform. Both Federally Facilitated Marketplaces and State-Based Exchanges on the federal platform are included in this project. For simplicity, we refer to them jointly as *Marketplaces* throughout the remainder of this report. MIDAS captures purchases and plan selections made during OEPs, during special enrollment periods, and in response to qualifying events, such as marriages, divorces, and new births. The states participating in the Marketplaces varied by year (see Appendix A). The analyses for this project included plan selections made during the OEPs for 2015 through 2022. Information for the 2021 COVID-19 Special Enrollment Period (SEP) is included in Appendix B.

MIDAS contains a variety of information about enrollees and the plans they select, such as self-reported race and ethnicity information collected using separate questions on race and Hispanic ethnicity. Enrollees are not required to complete these questions when they purchase insurance through Marketplaces, resulting in these fields having more missing data than other variables in MIDAS.

Imputing Race and Ethnicity

We used two approaches to impute race and ethnicity when these questions were not completed by Marketplace enrollees: replacement with self-reported data from another year and imputations using the modified BIFSG.

Replacing with Self-Reported Race and Ethnicity from Another Year

When self-reported race and ethnicity information was missing in a given year, we assessed whether self-reported race and ethnicity information was available for the enrollee in a different year because we could have up to eight years of data for each enrollee. If we had self-reported race and ethnicity from a different year, this value replaced the missing race and ethnicity. We refer to this as *missing replacement* in this report. If we had more than one race and ethnicity value for an enrollee, which occurred for approximate 0.5 percent of enrollees, we examined

prior years of enrollment data and used the self-reported race and ethnicity that were most proximate to the year for which we were missing self-reported race and ethnicity. If all prior years were missing race and ethnicity information, we used the most proximate subsequent year with available self-reported race and ethnicity. See Appendix C for additional information on the changes in self-reported information.

Indirect Race and Ethnicity Imputation with the Modified BIFSG

The main steps required to generate indirect race and ethnicity estimates are as follows:

1. **Prepare the race and ethnicity variables**. Existing self-reported race and ethnicity data from separate questions on race and Hispanic ethnicity were grouped into six mutually exclusive categories (AI/AN, AANHPI, Black, Hispanic, Multiracial, and White).[3] Indirect estimates of race and ethnicity can be generated for these six categories only. The Multiracial category consists of people who reported two or more of the following: AI/AN, AANHPI, Black, or White. People who report Hispanic ethnicity and one race category are not classified as multiple races. Thus, the first step is to map the existing self-reported race and ethnicity data in MIDAS into the six mutually exclusive groups using the following rules:
    a. Enrollees who reported Hispanic ethnicity were categorized as Hispanic regardless of races reported.
    b. Non-Hispanic respondents who reported two or more races were categorized as Multiracial, with the exception of those who selected two or more from among the following options (but no other race): Chinese, Filipino, Guamanian or Chamorro, Hawaiian, Indian, Japanese, Korean, Multiple Asian, Multiple Pacific Islander, Other Asian, Other Pacific Islander, Samoan, or Vietnamese. These enrollees were categorized as AANHPI.
    c. Non-Hispanic respondents who reported exactly one race were categorized as AANHPI, AI/AN, Black, or White, according to their responses.
2. **Geocode address-related variables to derive each enrollee's Census block group**. The BISG methodology was developed using race and ethnicity data from the 2010 Census reported by Census block group, which is the most granular level of geography available to the public. Data from the 2020 Census were not publicly available when our work was performed. We geocoded address information to derive the most granular geographic information supported by the available data for each record;
    a. When full addresses were available, we geocoded these addresses to longitude and latitude using ArcGIS. Longitude and latitude were then mapped to the 12-digit Federal Information Processing System (FIPS) code corresponding to each enrollee's Census block group. In a small number of cases, we had to map to an intersection rather than an exact address. For individuals with incomplete address information or whose address could not be mapped to a longitude and latitude, we used less precise available ZIP code information.

[3] All race and ethnicity categories other than Hispanic are non-Hispanic.

b. If a full address was either not available or could not be mapped to longitude and latitude (e.g., addresses in new developments), we cross-walked nine-digit ZIP codes to the 12-digit FIPS codes. The cross-walk file contains ZIP codes and 12-digit FIPS codes corresponding to the geographic center of each ZIP code. When nine-digit ZIP code were unavailable, we cross-walked five-digit ZIP codes to the centroid of the Census tract and used U.S. Census Bureau data for the Census tract. Using ZIP codes produces less accurate indirect race and ethnicity estimates because the ZIP code centroid may not correctly identify an enrollee's true Census block group, and race and ethnicity patterns may differ considerably across Census block groups within a ZIP code. Based on our experience, nine-digit ZIP codes offer a significant improvement over five-digit ZIP codes and provide a similar level of accuracy as full addresses because both full address and nine-digit ZIP codes can be mapped to Census block groups, while five-digit ZIP codes can be mapped to Census tracts.

3. **Create "clean" versions of enrollee surnames and first names.** This step facilitates merging the data to a Census data set containing surname-specific race and ethnicity percentages for thousands of surnames for the six race and ethnicity groups with the data to the list of first names. For hyphenated or compound surnames, we first removed hyphenations, concatenated the components, and attempted to match the concatenated surname to the Census list. If this was unsuccessful, we then attempted to match each component to the Census list and kept the set of six race and ethnicity probabilities associated with each component name matched. We used the highest Hispanic probability among the matched components of the surname and then rescaled the means of the surname components for the other race and ethnicity probabilities so the sum of the set of six probabilities was 100 (Haas et al., 2019). The surname file includes a row with race and ethnicity probabilities for all other surnames not appearing in the file, which is used when a surname does not match to names in the data. In the event that a first name was not available or did not map to one of the 4,250 names included in the first-name data set, race and ethnicity probability values for "all other first names" were used in the imputation.
4. **Generate uncalibrated race and ethnicity probabilities.** The modified BIFSG methodology applies Bayes' Theorem to update surname-based prior probabilities of each racial and ethnic group using first-name–based probabilities and address-based probabilities to produce posterior probabilities that combine the surname, first name, and address information (Voicu, 2018). Specifically, the modified BIFSG algorithm calculates

$$p(r|s,f,g) = \frac{p(r|s) * p(f|r) * p(g|r)}{\sum_{r=1}^{6} p(r|s) * p(f|r) * p(g|r)},$$

where $p(r|s, f, g)$ is the posterior probability of being in a specific racial and ethnic group r conditional on a specific surname (s), first name (f), and geographic location (g); $p(r|s)$ is the probability that the person is a specific race and ethnicity r conditional on the person's surname s; $p(f|r)$ is the probability of a specific first name conditional on identifying as a specific race and ethnicity r; $p(g|r)$ is the probability that a person resides in a specific geographic area g conditional on identifying as a specific race r; and the

denominator is the summation of the described factors over the six racial and ethnic categories.

5. **Calibrate race and ethnicity probabilities.** The underlying probabilities that relate surnames, first names, and addresses to racial and ethnic groups are based on national averages (surname and first name) and block group averages (address). The racial and ethnic distribution among Marketplace enrollees for a specific surname and block group may differ from the U.S. average. We calibrated the imputations to better match the racial and ethnic distribution of Marketplace enrollees based on the information from those enrollees who did self-report race and ethnicity. Calibrating imputations to the distribution of self-reported race and ethnicity among Marketplace enrollees helps make the imputed results better reflect the observed data, but it also assumes that the unobserved racial and ethnic distribution of nonreporters (conditional on surname and block group) is more similar to Marketplace enrollees residing in the block group who self-report race and ethnicity than to all residents in the block group, which is the reference population for the BISG. However, calibration does not assume that the distribution of surnames and block groups is the same among nonreporters as Marketplace enrollees who self-report race and ethnicity. Although the distribution of imputed race and ethnicity among those reporting race and ethnicity is expected to mirror their self-reported race and ethnicity, the distribution of imputed race and ethnicity among nonreporters may be different because of differences in their surnames and where they live.

The BISG was designed as a set of six race and ethnicity probabilities rather than as single classifications. Classification-based assignments are less accurate than using the race and ethnicity probabilities directly at the population level (or as the basis of a formal multiple imputation) and may overestimate the probabilities of race and ethnicities with higher prevalence (e.g., White) and underestimate the probabilities of race and ethnicities with lower prevalence (e.g., AANHPI), resulting in biased estimates (McCaffrey and Elliott, 2008). Thus, we do not recommend using single classification-based imputations based on the largest probability across the six mutually exclusive racial and ethnic groups.

Assessing the Accuracy of Modified BIFSG Performance

To assess the accuracy of the modified BIFSG algorithm, we applied the algorithm not only to records missing self-reported race and ethnicity but also to records with self-reported race and ethnicity, including those records in which missing race and ethnicity were replaced for the enrollees using self-reported information from another year. Comparing Marketplace self-reported race and ethnicity data with imputed race and ethnicity enables the assessment of the performance of the modified BIFSG methodology in the Marketplace population. We examined two properties of the imputed results: calibration and discrimination (the ability to differentiate between groups).

Calibration refers to the agreement between a model's predicted outcome and observed outcomes. With a well-calibrated prediction algorithm, the means for the six sets of race and

ethnicity probabilities closely match the means of the self-reported race and ethnicity with missing replacement for individuals who complete these questions when enrolling in a Marketplace plan.

An algorithm that differentiates well will also produce a higher probability for an individual's true racial and ethnic group than for all other racial and ethnicity groups. We assessed discrimination using the C-statistic, which is the result of performing an AUC analysis. To conduct the AUC analysis, we fit six separate logistic regression models—one for each racial and ethnic group—in which each dependent variable is a binary indication for the specific racial and ethnic group versus all other groups (1 if in the specific group; 0 otherwise) with independent variables that were each enrollee's set of six race and ethnicity probabilities. We calculated the C-statistic for each racial and ethnic category and an overall C-statistic. The C-statistic ranges from 0.5 (no better than chance) to 1.0 (predicts perfectly). In general contexts, a C-statistic of 0.7 is considered "acceptable," 0.8 is considered "strong" (Hosmer and Lemeshow, 2000), and 0.9 or higher is considered "excellent" (per authors).

We examined calibration and discrimination both overall and within strata defined by categories of enrollee characteristics, including age (0–17 years, 18–34 years, 35–64 years, and 65 and older), Census division (East North Central, East South Central, Middle Atlantic, Mountain, New England, Pacific, South Atlantic, West North Central, and West South Central), and percentage of Marketplace plans' enrollees self-reporting race and ethnicity (fewer than 60 percent of the plans' enrollees self-report race and ethnicity, and at least 60 percent of the plans' enrollees self-report race and ethnicity) to assess the method's accuracy within enrollee subgroups.

Chapter 3. Results

The analyses included 71,610,609 records across the eight years of MIDAS enrollment data, from 2015 through 2022. The annual number of records ranged from 8,250,833 in 2021 to 10,255,632 in 2022. Individuals were enrolled through Marketplaces for a mean of 4.1 years (see Appendix D for additional results).

Self-Reported Race and Ethnicity in MIDAS

Self-reported information on race and ethnicity was missing for 32.5 percent of records, ranging from 26.3 percent in 2019 to 38.7 percent in 2022 (see Figure 3.1; see also Table E.1 in Appendix E). Among the records with self-reported race and ethnicity information, White enrollees constituted the large group (40.8 percent of all records), followed by Hispanic enrollees (11.8 percent), Black enrollees (7.0 percent) and AANHPI enrollees (6.4 percent). Multiracial and AI/AN enrollees each constituted fewer than 2 percent of enrollees (Table 3.1).

Using enrollees' records from other years to replace missing race and ethnicity reduced missingness to 23.5 percent, approximately a 28-percent reduction in missingness (see Figure 3.1; see also Table E.2 in Appendix E). Although replacing missing self-reported race and ethnicity with other years of self-reported data increased the percentage of enrollees identified as AANHPI, Black, Hispanic and White, it had little effect on the percentage identified as AI/AN or Multiracial.

Figure 3.1. Missing Self-Reported Race and Ethnicity Among Marketplace Enrollees

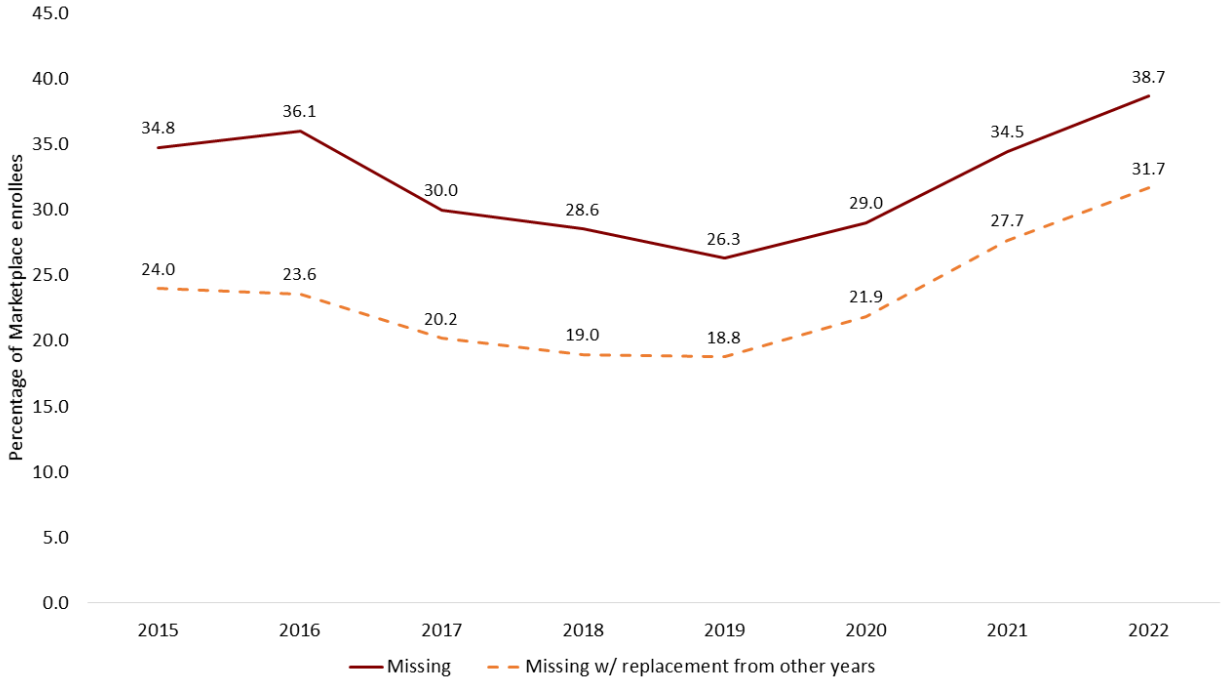

Table 3.1. Self-Reported Race and Ethnicity Prior to Implementing Modified BIFSG Imputation

	Self-Reported Race and Ethnicity Without Missing Replacement		Status of Records Missing Self-Reported Race and Ethnicity After Missing Replacement		Self-Reported Race and Ethnicity with Missing Replacement		Share Represented by Records with Successfully Replaced Missing Race and Ethnicity
Race and Ethnicity	***N***	**Percentage**	***N***	**Percentage**	***N***	**Percentage**	**Percentage**
Missing	23,266,952	32.49	16,841,149	72.38	16,841,149	23.52	0.00
AANHPI	4,559,680	6.37	577,350	2.48	5,137,030	7.17	11.24
AI/AN	282,386	0.39	21,746	0.09	304,132	0.42	7.15
Black	5,006,122	6.99	938,756	4.03	5,944,878	8.30	15.79
Hispanic	8,443,451	11.79	1,351,985	5.81	9,795,436	13.68	13.80
Multiracial	860,396	1.20	130,955	0.56	911,351	1.38	14.37
White	29,191,622	40.76	3,405,011	14.63	32,596,633	45.52	10.45
Total	71,610,609	100.00	23,266,952	100.00	71,610,609	100.00	8.97

Name and Address Data Used for Modified BIFSG Imputation

The MIDAS data contained first name and surname information for all records. We were able to match first names to the first name file and obtain first name–specific race and ethnicity probabilities for 80.7 percent of records. The remaining 19.3 percent of records received race and ethnicity probabilities associated with "all other first names." We were able to match the surnames for 87.5 percent of records to the surname file. The remaining 12.5 percent of records received race and ethnicity probabilities for "all other surnames."

The MIDAS data contained address information for all records. More than 99.9 percent of the records in MIDAS were successfully geocoded. Full address, the most complete and precise geographic information available, was used for 95.8 percent of records (Table 3.2) and mapped to a latitude and longitude and, ultimately, to a Census block group. The remaining records were geocoded using five-digit ZIP codes (3.2 percent) or nine-digit ZIP codes (1.0 percent), which were mapped to the centroid of Census tracts and Census block groups, respectively. A small number of records were geocoded using street intersections. Fewer than 4,000 records across the seven years of data were unable to be geocoded. Although the number of records that could not be geocoded was small in each year, it steadily increased over time from 239 in 2018 to 1,053 in 2022. This suggests that the more-recent data may include a small number of addresses in newly constructed communities that were not recognized by the geocoding software.

Table 3.2. Results of Geocoding Enrollee Residential Addresses

Level of Geocoding	*N*	Percentage
Full address	68,581,861	95.77
9-digit ZIP code	708,497	0.99
5-digit ZIP code	2,316,278	3.23
Street intersection	91	0.00
Not geocoded	3,882	0.00
Total	71,610,609	100.00

Although we were able to geocode more than 99.9 percent of records, a small number of these records could not be matched to the Census data (n = 3,421; 0.005 percent of successfully geocoded records). In these cases, either the Census block group was created after the 2010 Census or the person resided outside the 50 states and the District of Columbia. An additional 2,253 records had addresses that were geocoded and mapped to a Census block group or Census tract that had zero residents in the Census data, which happens in nonresidential areas, such as industrial parks or ZIP codes used only for post office boxes. We were unable to impute race and ethnicity for these two groups of records in addition to the 3,882 records we were unable to

geocode (the records could have self-reported race and ethnicity information).[4] Using the modified BIFSG, we imputed race and ethnicity for 71,601,053 records (over 99.9 percent of records in the data set). Table 3.3 summarizes our ability to impute race and ethnicity by whether such self-reported information was provided, missing data were replaced by self-reported data from another year or were missing for all years of enrollment. In subsequent tables, information on records for which we were unable to impute race and ethnicity will appear in the modified BIFSG columns in the row labeled "Missing."

Table 3.3. Ability to Impute Race and Ethnicity by Self-Reported Race and Ethnicity Status

	Successfully Imputed Race and Ethnicity	Unable to Impute Race and Ethnicity	Total
Self-reported race and ethnicity	48,336,686	6,971	48,343,657
Missing race and ethnicity replaced with self-reported data from another year	6,425,162	641	6,425,803
Missing race and ethnicity after missing replacement	16,839,205	1,944	16,841,149
Total	71,601,053	9,556	71,610,609

Modified BIFSG Imputation Results

Using the modified BIFSG, we imputed race and ethnicity for 71,601,053, as noted above. First, we compared self-reported and imputed race and ethnicity among the 54,761,848 records with either self-reported or replaced race and ethnicity (hereto referred to as self-reported race and ethnicity with missing replacement). Table 3.4 shows that the means of the probability based modified BIFSG imputations for each of the six race and ethnicity categories matches the distribution of the self-reported race and ethnicity means exactly. This is because we calibrated the imputations to the overall distribution of self-reported race and ethnicity in the sample. Therefore, the matching of the distributions is expected and indicates that calibration was successful. Appendix F presents the uncalibrated results, which are very similar to the calibrated results.

[4] The total number of records for which we were unable to impute race and ethnicity was 9,556.

Table 3.4. Comparison of Overall Self-Reported and Modified BIFSG-Imputed Racial and Ethnic Distributions Among Enrollees with Self-Reported Race and Ethnicity

Race and Ethnicity	Self-Reported Race and Ethnicity with Missing Replacement		Modified BIFSG-Imputed Results		
	N	Percentage	Estimated *N*	Percentage	C-statistic
AANHPI	5,137,030	9.38	5,134,761	9.38	0.96
AI/AN	304,132	0.56	303,137	0.55	0.62
Black	5,944,878	10.85	5,944,703	10.85	0.95
Hispanic	9,795,436	17.88	9,795,232	17.88	0.96
Multiracial	991,351	1.81	991,315	1.81	0.68
White	32,596,633	59.52	32,592,699	59.52	0.94
Missing	0	0	7,612	0.01	
Total	54,769,460	100.00	54,769,460	100.00	0.94

The AUC analysis produced C-statistics of 0.96 for AANHPI, 0.62 for AI/AN, 0.95 for Black, 0.96 for Hispanic, 0.68 for Multiracial, and 0.94 for White groups, resulting in a weighted average of 0.94 across all groups. Thus, the ability of the modified BIFSG to differentiate AANHPI, Black, Hispanic, and White enrollees from other groups is excellent. It did not reach an "acceptable" level for AI/AN or Multiracial enrollees. Currently, we recommend that modified BIFSG-imputed race and ethnicity not be used to make inferences about AI/AN enrollees or Multiracial enrollees.

As shown in Table 3.5, the racial and ethnic distribution for those who self-report race and ethnicity (columns on the right) differed from the imputed race and ethnicity for those who did not report race and ethnicity (column on the left) in multiple ways. Enrollees who self-reported race and ethnicity were more likely to be White than nonreporting enrollees for whom race and ethnicity were imputed (59.5 percent versus 49.3 percent) or AANHPI (9.4 percent versus 6.7 percent), and less likely to be Black (10.9 percent versus 15.7 percent) or Hispanic (17.9 percent versus 26.1 percent).

Table 3.5. Comparison of Race and Ethnicity Distribution for Enrollees Who Do and Do Not Self-Report Race and Ethnicity

Race and Ethnicity	Self-Reported Race and Ethnicity with Missing Replacement		Modified BIFSG-Imputed Results for Non-Self-Reporters	
	N	Percentage	Estimated *N*	Percentage
AANHPI	5,137,030	9.38	1,122,373	6.66
AI/AN	304,132	0.56	87,205	0.52
Black	5,944,878	10.85	2,643,536	15.70
Hispanic	9,795,436	17.88	4,397,238	26.11
Multiracial	991,351	1.81	294,715	1.75
White	32,596,633	59.52	8,294,137	49.25
Missing	0	0	1,944	0.01
Total	54,769,460	100.00	16,841,149	100.00

NOTE: Sum of estimated *N* for race and ethnicity categories may not match the total because of rounding.

Table 3.6 presents the racial and ethnic distribution for the entire sample of 71.6 million Marketplace enrollees. These results combine self-reported data for those who reported their race and ethnicity and the imputed race and ethnicity for those who did not self-report. The combined results estimated that 0.6 percent of Marketplace enrollees were AI/AN; 8.7 percent were AANHPI; 12.0 percent were Black; 19.8 percent were Hispanic; 1.8 percent were Multiracial; and 57.1 percent were White.

Table 3.6. Racial and Ethnic Distribution of the Overall Marketplace Enrollee Sample

Race and Ethnicity	Self-Reported Race and Ethnicity with Missing Replacement, Excluding Records with Missing Race and Ethnicity		Self-Reported Race and Ethnicity with Missing Replacement and Including Records with Missing Race and Ethnicity		Combined Self-Reported and Modified BIFSG-Imputed Results	
	N	Percentage	*N*	Percentage	Estimated *N*	Percentage
AANHPI	5,137,030	9.38	5,137,030	7.17	6,259,403	8.74
AI/AN	304,132	0.56	304,132	0.42	391,337	0.55
Black	5,944,878	10.85	5,944,878	8.30	8,588,414	11.99
Hispanic	9,795,436	17.88	9,795,436	13.68	14,192,674	19.82
Multiracial	991,351	1.81	991,351	1.38	1,286,066	1.80
White	32,596,633	59.52	32,596,633	45.52	40,890,770	57.10
Missing	0	0	16,841,149	23.52	1,944	0.00
Total	54,769,460	100.00	71,610,609	100.00	71,610,609	100.00

NOTE: Sum of estimated *N* for race and ethnicity categories may not match the total because of rounding.

Imputed Results for Enrollee Subgroups

Next, we report comparisons of self-reported with imputed race and ethnicity for the nearly 55 million Marketplace enrollees who self-reported race and ethnicity for subgroups of enrollees; this includes the records for which we were able to replace missing race and ethnicity with self-reported data from another year for the enrollee. These analyses enable us to identify variation in race and ethnicity across subgroups, whether the imputed racial and ethnic distributions are well-calibrated to the self-reported distributions for different subgroups and whether the modified BIFSG differentiates well between racial and ethnic groups within each subgroup.

We report when the mean percentage in a racial and ethnic group was slightly over- or underestimated by the modified BIFSG (i.e., the mean imputed percentage for a racial and ethnic group based on the modified BIFSG was different from the self-reported mean percentage by 1.0 to 2.0 percentage points) or was over- or underestimated (mean difference of 2.0 to 3.0 percentage points). Overestimation occurs when the modified BIFSG-imputed percentage is larger than the self-reported percentage; underestimation occurs when the modified BIFSG-imputed percentage is smaller than the self-reported percentage.

Tables 3.7 through 3.10 compare the distributions of self-reported and modified BIFSG-imputed race and ethnicity by age categories. Table 8 presents results for Marketplace enrollees aged 0–17 years. Table 3.8 present results for enrollees aged 18–34 years. Table 3.9 presents results for enrollees aged 35–64 years, and Table 3.10 presents results for enrollees aged 65 years and older. Age was missing for 201 enrollees who self-reported race and ethnicity, and they were excluded from these tables.

The percentage of Marketplace enrollees self-identifying as Hispanic or AANHPI was substantially larger among enrollees 65 and older than among younger enrollees, and the percentage of enrollees self-identifying as White was substantially lower among enrollees 65 and older. The percentage of enrollees self-identifying as Multiracial was higher among younger enrollees than older enrollees.

The modified BIFSG overestimates the probability of being a Black enrollee and slightly underestimates the probability of being a Multiracial enrollee in the 0–17 age group. It slightly overestimates the probability of being a White enrollee in the 18–34 age group. The modified BIFSG overestimates the probability of being a White enrollee, underestimates the probability of being an AANHPI enrollee, and slightly underestimates the probability of being a Black enrollee in the 65-and-older age group. The C-statistics indicate that modified BIFSG imputations differentiate well between the racial and ethnic groups for each age group. However, the C-statistics do vary, ranging from an average C-statistic for those ages 0 to 17 of 0.90 to a high of 0.97 for those 65 and older.

Table 3.7. Comparison of Self-Reported and Modified BIFSG-Imputed Racial and Ethnic Distributions Among Marketplace Enrollees, Ages 0–17 Years

Race and Ethnicity	Self-Reported Race and Ethnicity with Missing Replacement		Modified BIFSG-Imputed Results Among Self-Reporters		
	N	Percentage	Estimated *N*	Percentage	C-statistic
AANHPI	483,408	9.90	497,388	10.19	0.96
AI/AN	46,766	0.96	35,078	0.72	0.80
Black	282,170	5.78	400,880	8.21	0.91
Hispanic	853,602	17.49	787,696	16.14	0.92
Multiracial	188,497	3.86	99,932	2.05	0.67
White	3,027,232	62.01	3,059,702	62.68	0.89
Missing	0	0.00	999	0.01	
Total	4,881,675	100.00	4,881,675	100.00	0.90

NOTE: Sum of estimated *N* for race and ethnicity categories may not match the total because of rounding.

Table 3.8. Comparison of Self-Reported and Modified BIFSG-Imputed Racial and Ethnic Distributions Among Marketplace Enrollees, Ages 18–34 Years

Race and Ethnicity	Self-Reported Race and Ethnicity with Missing Replacement		Modified BIFSG-Imputed Results Among Self-Reporters		
	N	Percentage	Estimated *N*	Percentage	C-statistic
AANHPI	1,332,,677	9.42	1,336,205	9.45	0.96
AI/AN	80,890	0.57	78,156	0.55	0.62
Black	1,722,724	12.18	1,687,228	11.93	0.95
Hispanic	2,598,115	18.37	2,504,262	17.70	0.95
Multiracial	361,519	2.56	264,823	1.87	0.67
White	8,050,673	56.91	8,274,043	58.49	0.93
Missing	0	0.00	1,881	0.01	
Total	14,146,598	100.00	14,146,598	100.00	0.93

NOTE: Sum of estimated *N* for race and ethnicity categories may not match the total because of rounding.

Table 3.9. Comparison of Self-Reported and Modified BIFSG-Imputed Racial and Ethnic Distributions Among Marketplace Enrollees, Ages 35–64 Years

Race and Ethnicity	Self-Reported Race and Ethnicity with Missing Replacement		Modified BIFSG-Imputed Results Among Self-Reporters		C-Statistic
	N	Percentage	Estimated *N*	Percentage	
AANHPI	3,147,239	8.95	3,139,121	8.93	0.97
AI/AN	174,940	0.50	187,981	0.53	0.60
Black	3,876,159	11.03	3,799,761	10.81	0.96
Hispanic	6,153,020	17.50	6,313,385	17.96	0.97
Multiracial	434,038	1.23	616,631	1.75	0.66
White	21,371,125	60.79	21,095,039	60.00	0.95
Missing	0	0.00	4,603	0.01	
Total	35,156,521	100.00	35,156,521	100.00	0.95

NOTE: Sum of estimated *N* for race and ethnicity categories may not match the total because of rounding.

Table 3.10. Comparison of Self-Reported and Modified BIFSG-Imputed Racial and Ethnic Distributions Among Marketplace Enrollees, 65 Years And Older

Race and Ethnicity	Self-Reported Race and Ethnicity with Missing Replacement		Modified BIFSG-Imputed Results Among Self-Reporters		C-statistic
	N	Percentage	Estimated *N*	Percentage	
AANHPI	173,701	29.72	162,043	27.72	0.98
AI/AN	1,536	0.26	1,920	0.33	0.89
Black	63,795	10.92	56,802	9.72	0.96
Hispanic	190,675	32.62	189,864	32.49	0.98
Multiracial	7,296	1.25	9,926	1.70	0.74
White	147,462	25.23	163,782	28.02	0.95
Missing	0	0.00	129	0.02	
Total	584,465	100.00	584,465	100.00	0.97

NOTE: Sum of estimated *N* for race and ethnicity categories may not match the total because of rounding.

Tables 3.11 through 3.19 compare the distributions of self-reported and modified BIFSG-imputed race and ethnicity within the nine U.S. Census Bureau divisions: Table 3.11 presents results for Marketplace enrollees residing in the East North Central Division; Table 3.12 presents results for enrollees residing in the East South Central Division; Table 3.13 presents results for enrollees residing in the Middle Atlantic Division; Table 3.14 presents results for enrollees residing in the Mountain Division; Table 3.15 presents results for the New England Division; Table 3.16 presents results for the Pacific Division; Table 3.17 presents results for the South Atlantic Division; Table 3.18 presents results for the West North Central Division; and Table 3.19 presents results for the West South Central Division. The states included in each Census division are presented in Appendix G.

The racial and ethnic distributions varied substantially across Census divisions. The percentage of Marketplace enrollees self-identifying as AANHPI was highest among enrollees residing in the Pacific, West South Central, and Middle Atlantic divisions. The percentage of enrollees self-identifying as Black was highest among enrollees in the South Atlantic and East South Central divisions. The percentage of enrollees self-identifying as Hispanic was highest among enrollees in the South Atlantic, West South Central, and Mountain divisions. The percentage of Marketplace enrollees self-identifying as White was largest among enrollees in the New England, West North Central, and East North Central divisions.

In general, calibration was quite good for each Census division. None of the Census divisions had a difference between the percentage based on the modified BIFSG imputation and self-reported race and ethnicity of greater than 3.0 percentage points for any racial and ethnic group. The percentage of White enrollees was slightly underestimated for three Census divisions (East South Central, Middle Atlantic, and West North Central), underestimated for two Census divisions (East North Central and New England), and overestimated for one Census division (West South Central). The percentage of Hispanic enrollees was slightly overestimated for two Census divisions (East North Central and East South Central). The percentage of Black enrollees was slightly overestimated for three Census divisions (Mountain, New England, and Pacific). The percentage of Multiracial enrollees was slightly underestimated in the Pacific Division.

The C-statistics indicate that modified BIFSG imputations differentiate well between the racial and ethnic groups for each Census division. However, the C-statistics varied, ranging from an average C-statistic for those residing in the New England Division of 0.84 to a high of 0.95 for those residing in the Middle Atlantic and West South Central divisions. Although the overall C-statistics for the Census divisions indicated the modified BISFG had “excellent” discrimination, discrimination was “poor” (less than 0.7) for AI/AN enrollees for five of the nine Census divisions and for Multiracial enrollees in all Census divisions except the West South Central Division, where it was “acceptable.”

Table 3.11. Comparison of Self-Reported and Modified BIFSG-Imputed Racial and Ethnic Distributions Among Marketplace Enrollees Residing in the East North Central Census Division

Race and Ethnicity	Self-Reported Race and Ethnicity with Missing Replacement		Modified BIFSG-Imputed Results Among Self-Reporters		C-statistic
	N	Percentage	Estimated *N*	Percentage	
AANHPI	573,141	7.18	587,129	7.35	0.95
AI/AN	20,889	0.26	41,856	0.52	0.54
Black	450,761	5.65	506,824	6.35	0.94
Hispanic	433,001	5.42	523,346	6.56	0.92
Multiracial	127,491	1.60	143,515	1.80	0.64
White	6,378,008	79.89	6,180,499	77.42	0.93
Missing	0	0	122	0.00	
Total	7,983,291	100.00	7,983,291	100.00	0.92

NOTE: Sum of estimated *N* for race and ethnicity categories may not match the total because of rounding.

Table 3.12. Comparison of Self-Reported and Modified BIFSG-Imputed Racial and Ethnic Distributions Among Marketplace Enrollees Residing in the East South Central Census Division

Race and Ethnicity	Self-Reported Race and Ethnicity with Missing Replacement		Modified BIFSG-Imputed Results Among Self-Reporters		C-statistic
	N	Percentage	Estimated *N*	Percentage	
AANHPI	225,023	6.13	219,600	5.98	0.97
AI/AN	11,229	0.31	19,060	0.52	0.65
Black	724,990	19.75	711,121	19.37	0.95
Hispanic	119,713	3.26	161,114	4.39	0.91
Multiracial	58,021	1.58	75,985	2.07	0.61
White	2,532,036	68.97	2,484,021	67.67	0.93
Missing	0	0.00	111	0.00	
Total	3,671,012	100.00	3,671,012	100.00	0.93

NOTE: Sum of estimated *N* for race and ethnicity categories may not match the total because of rounding.

Table 3.13. Comparison of Self-Reported and Modified BIFSG-Imputed Racial and Ethnic Distributions Among Marketplace Enrollees Residing in the Middle Atlantic Census Division

Race and Ethnicity	Self-Reported Race and Ethnicity with Missing Replacement		Modified BIFSG-Imputed Results Among Self-Reporters		C-statistic
	N	Percentage	Estimated *N*	Percentage	
AANHPI	400,738	11.98	405,151	12.11	0.98
AI/AN	2,967	0.09	12,932	0.39	0.53
Black	222,289	6.65	230,868	6.90	0.96
Hispanic	336,546	10.06	352,766	10.55	0.95
Multiracial	50,581	1.51	56,303	1.68	0.67
White	2,331,307	69.71	2,286,331	68.36	0.95
Missing	0	0.00	77	0.00	
Total	3,344,428	100.00	3,344,428	100.00	0.95

NOTE: Sum of estimated *N* for race and ethnicity categories may not match the total because of rounding.

Table 3.14. Comparison of Self-Reported and Modified BIFSG-Imputed Racial and Ethnic Distributions Among Marketplace Enrollees Residing in the Mountain Census Division

Race and Ethnicity	Self-Reported Race and Ethnicity with Missing Replacement		Modified BIFSG-Imputed Results Among Self-Reporters		C-statistic
	N	Percentage	Estimated *N*	Percentage	
AANHPI	238,161	6.78	229,343	6.53	0.95
AI/AN	31,594	0.90	31,135	0.89	0.73
Black	61,239	1.74	100,801	2.87	0.90
Hispanic	608,914	17.34	593,695	16.91	0.93
Multiracial	77,662	2.21	61,516	1.75	0.68
White	2,493,793	71.02	2,494,232	71.03	0.91
Missing	0	0.00	641	0.02	
Total	3,511,363	100.00	3,511,363	100.00	0.91

NOTE: Sum of estimated *N* for race and ethnicity categories may not match the total because of rounding.

Table 3.15. Comparison of Self-Reported and Modified BIFSG-Imputed Racial and Ethnic Distributions Among Marketplace Enrollees Residing in the New England Census Division

Race and Ethnicity	Self-Reported Race and Ethnicity with Missing Replacement		Modified BIFSG-Imputed Results Among Self-Reporters		C-statistic
	N	Percentage	Estimated *N*	Percentage	
AANHPI	23,167	3.03	22,957	3.00	0.95
AI/AN	1,656	0.22	4,129	0.54	0.56
Black	5,687	0.74	16,035	2.10	0.85
Hispanic	16,304	2.13	23,859	3.12	0.85
Multiracial	10,614	1.39	12,304	1.61	0.62
White	707,939	92.50	685,995	89.63	0.85
Missing	0	0.00	89	0.01	
Total	765,367	100.00	765,367	100.00	0.84

NOTE: Sum of estimated *N* for race and ethnicity categories may not match the total because of rounding.

Table 3.16. Comparison of Self-Reported and Modified BIFSG-Imputed Racial and Ethnic Distributions Among Marketplace Enrollees Residing in the Pacific Census Division

Race and Ethnicity	Self-Reported Race and Ethnicity with Missing Replacement		Modified BIFSG-Imputed Results Among Self-Reporters		C-statistic
	N	Percentage	Estimated *N*	Percentage	
AANHPI	145,371	11.95	145,090	11.92	0.96
AI/AN	11,941	0.98	12,081	0.99	0.72
Black	12,402	1.02	33,574	2.76	0.86
Hispanic	80,704	6.63	88,757	7.29	0.88
Multiracial	49,981	4.11	29,413	2.42	0.67
White	916,438	75.31	906,468	74.49	0.89
Missing	0	0.00	1,454	0.12	
Total	1,216,837	100.00	1,216,837	100.00	0.88

NOTE: Sum of estimated *N* for race and ethnicity categories may not match the total because of rounding.

Table 3.17. Comparison of Self-Reported and Modified BIFSG-Imputed Racial and Ethnic Distributions Among Marketplace Enrollees Residing in the South Atlantic Census Division

Race and Ethnicity	Self-Reported Race and Ethnicity with Missing Replacement		Modified BIFSG-Imputed Results Among Self-Reporters		C-statistic
	N	Percentage	Estimated *N*	Percentage	
AANHPI	1,919,750	9.18	1,929,573	9.23	0.96
AI/AN	49,303	0.24	92,552	0.44	0.66
Black	3,370,754	16.12	3,245,053	15.52	0.95
Hispanic	4,868,280	23.28	4,757,090	22.75	0.97
Multiracial	358,146	1.71	379,208	1.81	0.68
White	10,346,248	49.47	10,506,687	50.24	0.94
Missing	0	0.00	2,318	0.01	
Total	20,912,481	100.00	20,912,481	100.00	0.94

NOTE: Sum of estimated *N* for race and ethnicity categories may not match the total because of rounding.

Table 3.18. Comparison of Self-Reported and Modified BIFSG-Imputed Racial and Ethnic Distributions Among Marketplace Enrollees Residing in the West North Central Census Division

Race and Ethnicity	Self-Reported Race and Ethnicity with Missing Replacement		Modified BIFSG-Imputed Results Among Self-Reporters		C-statistic
	N	Percentage	Estimated *N*	Percentage	
AANHPI	175,560	5.11	175,832	5.11	0.96
AI/AN	26,501	0.77	25,954	0.76	0.71
Black	178,687	5.20	198,297	5.77	0.95
Hispanic	168,933	4.91	195,380	5.68	0.91
Multiracial	65,721	1.91	62,931	1.83	0.65
White	2,822,197	82.10	2,777,967	80.81	0.90
Missing	0	0.00	1,238	0.04	
Total	3,437,599	100.00	3,437,599	100.00	0.90

NOTE: Sum of estimated *N* for race and ethnicity categories may not match the total because of rounding

Table 3.19. Comparison of Self-Reported and Modified BIFSG-Imputed Racial and Ethnic Distributions Among Marketplace Enrollees Residing in the West South Central Census Division

Race and Ethnicity	Self-Reported Race and Ethnicity with Missing Replacement		Modified BIFSG-Imputed Results Among Self-Reporters		C-statistic
	N	Percentage	Estimated *N*	Percentage	
AANHPI	1,436,119	14.47	1,420,087	14.31	0.98
AI/AN	148,052	1.49	63,436	0.64	0.74
Black	918,069	9.25	902,129	9.09	0.95
Hispanic	3,163,041	31.86	3,099,225	31.22	0.97
Multiracial	193,134	1.95	170,143	1.71	0.72
White	4,068,667	40.99	4,270,502	43.02	0.94
Missing	0	0.00	1,562	0.02	
Total	9,927,082	100.00	9,927,082	100.00	0.95

NOTE: Sum of estimated *N* for race and ethnicity categories may not match the total because of rounding.

Last, we examined the modified BIFSG's calibration by plans' percentage of enrollees who did not self-report race and ethnicity because of the possibility that plans with lower percentages of self-reporting enrollees might differ in ways that affect the accuracy of the modified BIFSG algorithm. Table 3.20 shows the distributions of self-reported and modified BIFSG-imputed race and ethnicity within plans in which fewer than 60 percent of enrollees self-reported race and ethnicity, while Table 3.21 presents information for plans with at least 60 percent of enrollees self-reporting race and ethnicity. Nineteen percent of Marketplace enrollees who self-reported race and ethnicity were in a plan in which fewer than 60 percent of the enrollees were self-reporters. The percentage of Marketplace enrollees self-identifying as Black or Hispanic was larger than 60 percent of the enrollees self-reporting, and the percentage reporting as White was smaller in plans with fewer than 60 percent of enrollees self-reporting race and ethnicity than in plans with more than 60 percent of enrollees self-reporting. This is consistent with disproportionate nonreporting by Black and Hispanic enrollees.

Calibration was very good for both groups of plans. Among plans with fewer than 60 percent of enrollees self-reporting race and ethnicity, the percentage of White enrollees was slightly overestimated, and the percentage of Black enrollees was slightly underestimated. There were no substantial differences among contracts with at least 60 percent of enrollees self-reporting race and ethnicity. The C-statistic was very high for both groups of health plans (0.96 for plans with fewer than 60 percent self-reporters versus 0.93 for plans with at least 60 percent self-reporters).

Table 3.20. Comparison of Self-Reported and Modified BIFSG-Imputed Racial and Ethnic Distributions Among Marketplace Enrollees in Plans with Fewer Than 60 Percent of Enrollees Self-Reporting Race and Ethnicity

Race and Ethnicity	Self-Reported Race and Ethnicity with Missing Replacement		Modified BIFSG-Imputed Results Among Self-Reporters		C-statistic
	N	Percentage	Estimated *N*	Percentage	
AANHPI	1,072,937	9.23	1,066,414	9.18	0.97
AI/AN	17,334	0.15	44,447	0.38	0.60
Black	1,853,678	15.95	1,702,163	14.65	0.96
Hispanic	3,877,002	33.37	3,770,250	32.45	0.97
Multiracial	151,681	1.31	181,184	1.56	0.72
White	4,646,487	39.99	4,853,983	41.78	0.95
Missing	0	0.00	678	0.00	
Total	11,619,119	100.00	11,619,119	100.00	0.96

NOTE: Sum of estimated *N* for race and ethnicity categories may not match the total because of rounding.

Table 3.21. Comparison of Self-Reported and Modified BIFSG-Imputed Racial and Ethnic Distributions Among Marketplace Enrollees in Plans with at Least 60 Percent of Enrollees Self-Reporting Race and Ethnicity

Race and Ethnicity	Self-Reported Race and Ethnicity with Missing Replacement		Modified BIFSG-Imputed Results Among Self-Reporters		C-statistic
	N	Percentage	Estimated *N*	Percentage	
AANHPI	4,064,093	9.42	4,068,347	9.43	0.96
AI/AN	286,798	0.66	258,690	0.60	0.71
Black	4,091,200	9.48	4,242,540	9.83	0.95
Hispanic	5,918,434	13.72	6,024,982	13.96	0.95
Multiracial	839,670	1.95	810,132	1.88	0.66
White	27,950,146	64.77	27,738,716	64.28	0.94
Missing	0	0.00	6,934	0.02	
Total	43,150,341	100.00	43,150,341	100.00	0.93

NOTE: Sum of estimated *N* for race and ethnicity categories may not match the total because of rounding

Chapter 4. Discussion

Summary

Self-reported race and ethnicity are collected when individuals enroll in health plans. However, approximately one-third of insurance applicants do not complete two optional questions on race and ethnicity on the Marketplace application. The probabilities of six mutually exclusive racial and ethnic categories (AI/AN, AANHPI, Black, Hispanic, Multiracial, and White) were imputed using the modified BIFSG. Race and ethnicity imputations using the modified BIFSG are highly accurate for API, Black, Hispanic, and White enrollees. As shown in our AUC analyses, the predictive accuracy of the modified BIFSG is lower for AI/AN and Multiracial enrollees. As a result, to ensure adequate precision of estimates, based on our experience, we caution against making inferences for AI/AN or Multiracial enrollees. Furthermore, the accuracy was lower for children and young adults than for older enrollees and varied by Census division. Accuracy was high for plans regardless of the plan level of missingness of their race and ethnicity data.

Potential Uses for Imputed Race and Ethnicity

Information on the race and ethnicity of individuals enrolling through the Health Insurance Marketplace is critical for assessing past enrollment efforts. The Department of Health and Human Services could use self-reported race and ethnicity augmented by imputed race and ethnicity when self-reported information is unavailable to identify and address disparities in health insurance enrollment. For example, the department could compare Marketplace enrollment by race and ethnicity with estimates of expected enrollment to identify subgroups for which enrollment is lagging and to determine whether outreach campaigns should be modified or tailored moving forward to better target these populations. This information could also be used to identify enrollees that could benefit from tailored materials. For example, if reenrollment is lower among Hispanic enrollees than other groups, tailored reenrollment materials that are culturally sensitive could be sent to enrollees with a probability of being Hispanic that is above a prespecified cutoff, such as 75 percent. As an alternative, materials could be sent to a prespecified number of enrollees with the highest probability of being Hispanic.

ASPE could also use race and ethnicity probabilities to understand whether the plan selection patterns vary by race and ethnicity, for example, choice of plan metal level, channel used to purchase Marketplace plans, and the extent to which brokerage assistance or navigators supported plan selections. ASPE could also explore whether plan costs vary by race and ethnicity.

Limitations

Our analyses were limited in a few ways. First, the lack of information on enrollee gender in our MIDAS extract meant that we were unable to assess how gender affects the accuracy of the imputation among Marketplace enrollees. However, BISG and Medicare BISG accuracy varies little by gender, especially when first name is included in the imputation, which further reduces variation by gender (Elliott et al., 2009; Haas et al., 2019). Thus, we do not expect much variation in accuracy by gender in Marketplace enrollees.

Second, the probabilities of race and ethnicity for first names were developed from a large sample of mortgage applications. Given the variable rates of home ownership in the United States by race and ethnicity, first names that are more common among non-White racial and ethnic groups may be underrepresented in the data. The states participating in the Federally Facilitated Marketplace (or State-Based Marketplaces using the Federal Platform [HealthCare.gov]) varied by year, with states contributing between one and eight years of data. In addition, Medicaid expansion occurred during the same period in some of the included states, reducing the number of Marketplace enrollees and potentially changing the distribution of enrollees' demographic characteristics. For these reasons, we only reported results by year in the appendixes, with the exception of the percentage of records missing self-reported race and ethnicity.

Last, in our approach, we first examined prior years using the closest year with self-reported data on race and ethnicity; if none of the prior years had such data, we then looked at the subsequent years. A small of percentage (0.5 percent) of Marketplace enrollees self-reported inconsistent race and ethnicity across years. Although enrollees self-identifying as Multiracial accounted for only 1.4 percent of records in our data, the Multiracial category accounted for more than 50 percent of inconsistencies in self-reported race and ethnicity (additional details presented in Appendix C). These inconsistencies likely contributed to the low C-statistic for the Multiracial group.

Potential Opportunities for Modified BIFSG Refinement

Although the modified BIFSG performed well, there is room for improvement, particularly in the identification of enrollees who are AI/AN and Multiracial. There are potential opportunities to further enhance the algorithm. There are additional variables in the MIDAS data that could be incorporated into the BIFSG. By including enrollee age, we could adjust estimates for generational changes in the distribution of race and ethnicity and address the observed differences in calibration by age. MIDAS contains information on membership in federally recognized American Indian tribes, which affects eligibility for cost-sharing reductions. Eligibility varies by tribe, but being 1/16 AI (6.25 percent) is a common requirement, and this information could be used to refine imputations. In addition, updating imputations using the 2020 Census data, which recently became available, may improve the accuracy of the

imputation, particularly for the Multiracial group.[5] The percentage of the U.S. population self-identifying as Multiracial increased from 2.9 percent in the 2010 Census to 10.2 percent in the 2020 Census, with the increase stemming from multiple factors, such as changing demographics in the United States and improvements in question design (Jones et al., 2021).

[5] Such analyses would, however, need to account for the introduction of statistical noise to 2020 Census data via the "differential privacy" procedure, which is a newly implemented mathematical approach to safeguard privacy, particularly for those who reside in a small area and whose race or ethnicity differs from that of their neighbors (U.S. Census Bureau, 2021).

Appendix A. States Participating in the Federally Facilitated Marketplaces and State-Based Marketplaces Using Federal Platform

Federally Facilitated Marketplaces are organized markets developed by the U.S. Department of Health and Human Services for the purchase of qualified health insurance plans in states that have opted not to build their own Marketplace, or website for offering ACA-compliant individual insurance. States that operate their own State-Based Marketplaces may also opt to use the federal platform (HealthCare.gov) while retaining other state administrative functions creating a State-Based Marketplace–federal platform. The states with Federally Facilitated Marketplaces or State-Based Marketplace–federal platforms vary by year with some states participating in all years since 2015 (Table A.1). California, Colorado, Connecticut, Washington, D.C., Idaho, Maryland, Massachusetts, Minnesota, New York, Rhode Island, Vermont, and Washington each operated its own SBM that did not use the federal platform in all years.

Table A.1. States Participating in a Federally Facilitated Marketplace or State-Based Marketplace–Federal Platform

State	2015	2016	2017	2018	2019	2020	2021	2022
Alaska	X	X	X	X	X	X	X	X
Alabama	X	X	X	X	X	X	X	X
Arizona	X	X	X	X	X	X	X	X
Arkansas	X	X	X	X	X	X	X	X
Delaware	X	X	X	X	X	X	X	X
Florida	X	X	X	X	X	X	X	X
Georgia	X	X	X	X	X	X	X	X
Hawaii		X	X	X	X	X	X	X
Iowa	X	X	X	X	X	X	X	X
Illinois	X	X	X	X	X	X	X	X
Indiana	X	X	X	X	X	X	X	X
Kansas	X	X	X	X	X	X	X	X
Kentucky			X	X	X	X	X	
Louisiana	X	X	X	X	X	X	X	X
Maine	X	X	X	X	X	X	X	
Michigan	X	X	X	X	X	X	X	X
Mississippi	X	X	X	X	X	X	X	X
Missouri	X	X	X	X	X	X	X	X

State	2015	2016	2017	2018	2019	2020	2021	2022
Montana	X	X	X	X	X	X	X	X
North Carolina	X	X	X	X	X	X	X	X
North Dakota	X	X	X	X	X	X	X	X
Nebraska	X	X	X	X	X	X	X	X
New Hampshire	X	X	X	X	X	X	X	X
New Jersey	X	X	X	X	X	X		
New Mexico	X	X	X	X	X	X	X	
Nevada	X	X	X	X	X			
Ohio	X	X	X	X	X	X	X	X
Oklahoma	X	X	X	X	X	X	X	X
Oregon	X	X	X	X	X	X	X	X
Pennsylvania	X	X	X	X	X	X		
South Carolina	X	X	X	X	X	X	X	X
South Dakota	X	X	X	X	X	X	X	X
Tennessee	X	X	X	X	X	X	X	X
Texas	X	X	X	X	X	X	X	X
Utah	X	X	X	X	X	X	X	X
Virginia	X	X	X	X	X	X	X	X
West Virginia	X	X	X	X	X	X	X	X
Wisconsin	X	X	X	X	X	X	X	X
Wyoming	X	X	X	X	X	X	X	X

Appendix B. 2021 COVID-19 Special Enrollment Period

In accordance with an Executive Order issued by President Joseph Biden in January 2021, CMS provided a SEP in response to the COVID-19 public health emergency that enabled qualifying consumers to enroll in Marketplace plans from February 15, 2021, through May 15, 2021.[6] Subsequently, the SEP was extended through August 15, 2021.[7]

MIDAS included 2,069,596 records for the 2021 SEP. Self-reported information on race and ethnicity was missing for 48.9 percent of records, which is substantially higher than the missingness observed during OEPs (Table B.1). White enrollees comprised the largest group of self-reported race and ethnicity (28.2 percent of all records), followed by Hispanic enrollees (11.4 percent), Black enrollees (7.0 percent), and AANHPI enrollees (3.1 percent). Multiracial and AI/AN enrollees each comprised fewer than 2 percent of enrollees.

Using enrollees' records from 2021 and other years' OEP data to replace missing race and ethnicity reduced missingness to 45.7 percent, just a 6.5 percent reduction in missingness (Table B.2). Replacing missing self-reported race and ethnicity with other years of self-reported data slightly increased the percentage of enrollees identified as Black, Hispanic, and White.

Table B.1. Self-Reported Race and ethnicity Prior to Implementing Modified BIFSG Imputation

	Self-Reported Race and Ethnicity Without Missing Replacement		Status of Records Missing Self-Reported Race and Ethnicity After Missing Replacement		Self-Reported Race and Ethnicity with Missing Replacement		Share Represented by Records with Successfully Replaced Missing Race and Ethnicity
Race and Ethnicity	***N***	**Percentage**	***N***	**Percentage**	***N***	**Percentage**	**Percentage**
Missing	1,011,400	48.87	946,040	93.54	946,040	45.71	0.00
AANHPI	64,145	3.10	4,135	0.41	68,280	3.30	6.06
AI/AN	8,402	0.41	263	0.03	8,665	0.42	3.04
Black	144,990	7.01	15,855	1.57	160,845	7.77	9.86
Hispanic	234,987	11.35	17,037	1.68	252,024	12.18	6.76
Multiracial	22,160	1.07	1,068	0.11	23,228	1.12	4.60
White	583,512	28.19	27,002	2.67	610,514	29.50	4.42

[6] CMS, "2021 Special Enrollment Period in Response to the COVID-19 Emergency," fact sheet, January 28, 2021.

[7] U.S. Department of Health and Human Services, "2021 Special Enrollment Period Access Extended to August 15 on HealthCare.gov for Marketplace Coverage," press release, March 23, 2021.

Race and Ethnicity	Self-Reported Race and Ethnicity Without Missing Replacement		Status of Records Missing Self-Reported Race and Ethnicity After Missing Replacement		Self-Reported Race and Ethnicity with Missing Replacement		Share Represented by Records with Successfully Replaced Missing Race and Ethnicity
	N	Percentage	*N*	Percentage	*N*	Percentage	Percentage
Total	2,069,596	100.00	1,011,400	100.00	2,069,596	100.00	3.16

Table B.2. Results of Geocoding Enrollee Residential Addresses

Level of Geocoding	*N*	Percentage
Full address	1,921,480	92.84
9-digit ZIP code	21,617	1.04
5-digit ZIP code	126,277	6.10
Street intersection	3	0.00
Not geocoded	219	0.01
Total	2,069,596	100.00

Although we were able to geocode 99.9 percent of records, a small number of these records could not be matched to the Census data (n = 53; 0.003 percent of successfully geocoded records). In these cases, either the Census block group was created after the 2010 Census or the person resided outside the 50 states and the District of Columbia. An additional 97 records had addresses that were geocoded and mapped to a Census block group or Census tract that had zero residents in the Census data, which happens in nonresidential areas, such as industrial parks or ZIP codes used only for post office boxes. We were unable to impute race and ethnicity for these two groups of records in addition to the 219 records we were unable to geocode (the records could have self-reported race and ethnicity information).[8] Using the modified BIFSG, we imputed race and ethnicity for 2,069,227 records (over 99.9 percent of 2021 SEP records). Table B.3 summarizes our ability to impute race and ethnicity by whether such self-reported information was provided; missing data was replaced by self-reported data from another year or missing for all years of enrollment. In subsequent tables, information on records for which we were unable to impute race and ethnicity will appear in the modified BIFSG columns in the row labeled "Missing."

[8] The total number of records for which we were unable to impute race and ethnicity was 369.

Table B.3. Ability to Impute Race And Ethnicity by Self-Reported Race and Ethnicity Status

	Successfully Imputed Race and Ethnicity	Unable to Impute Race and Ethnicity	Total
Self-reported race and ethnicity	1,057,980	216	1,058,196
Missing race and ethnicity replaced with self-reported data from another year	65,356	4	65,360
Missing race and ethnicity after missing replacement	945,891	149	946,040
Total	2,069,227	369	2,069,596

Modified BIFSG Imputation Results

Using the modified BIFSG, we imputed race and ethnicity for 2,069,227 individuals, as noted above. First, we compare self-reported and imputed race and ethnicity among the 1,123,556 records with either self-reported or replaced race and ethnicity (hereto referred to as self-reported race and ethnicity with missing replacement). Table B.4 shows that the means of the probability-based modified BIFSG imputations for each of the six race and ethnicity categories matches the distribution of the self-reported race and ethnicity means exactly. This is because we calibrated the imputations to the overall distribution of self-reported race and ethnicity in the sample. Therefore, the matching of the distributions is expected and indicates that calibration was successful.

Table B.4. Comparison of Overall Self-Reported and Modified BIFSG Imputed Racial and Ethnic Distributions Among Enrollees with Self-Reported Race And Ethnicity

Race and Ethnicity	Self-Reported Race and Ethnicity with Missing Replacement		Modified BIFSG Imputed Results		
	N	Percentage	Estimated *N*	Percentage	C-statistic
AANHPI	68,280	6.08	68,233	6.07	0.96
AI/AN	8,665	0.77	8,636	0.77	0.81
Black	160,845	14.32	160,830	14.31	0.95
Hispanic	252,024	22.43	252,003	22.43	0.95
Multiracial	23,228	2.07	23,228	2.07	0.69
White	610,514	54.34	610,406	54.33	0.93
Missing	0	0	220	0.02	
Total	1,123,556	100.00	1,123,556	100.00	0.93

The AUC analysis produced C-statistics of 0.96 for AANHPI, 0.81 for AI/AN, 0.95 for Black, 0.95 for Hispanic, 0.69 for Multiracial, and 0.93 for White groups, resulting in a weighted average of 0.93 across all groups. Thus, the ability of the modified BIFSG to differentiate AANHPI, Black, Hispanic, and White enrollees from other groups is excellent. It is "acceptable" for AI/AN and "marginally acceptable" for Multiracial enrollees.

The racial and ethnic distribution for those who self-report race and ethnicity (columns on the right) differed from the imputed race and ethnicity for those who did not report race and ethnicity (column on the left) in multiple ways (Table B.5). Enrollees who self-reported race and ethnicity were more likely to be White than nonreporting enrollees for whom race and ethnicity were imputed (54.3 percent versus 41.3 percent) or AANHPI (6.1 percent versus 4.1 percent), and less likely to be Black (14.3 percent versus 18.5 percent) or Hispanic (22.4 percent versus 33.8 percent).

Table B.5. Comparison of Race and Ethnicity Distribution for Enrollees Who Do and Do Not Self-Report Race and Ethnicity

	Self-Reported Race and Ethnicity with Missing Replacement		Modified BIFSG Imputed Results for Non-Self-Reporters	
Race and Ethnicity	*N*	Percentage	Estimated *N*	Percentage
AANHPI	68,280	6.08	38,398	4.06
AI/AN	8,665	0.77	5,685	0.60
Black	160,845	14.32	174,581	18.45
Hispanic	252,024	22.43	319,768	33.80
Multiracial	23,228	2.07	17,208	1.82
White	610,514	54.34	390,251	41.25
Missing	0	0	149	0.02
Total	1,123,556	100.00	946,040	100.00

NOTE: Sum of estimated *N* for race and ethnicity categories may not match the total because of rounding

Table B.6 presents the racial and ethnic distribution for the entire sample of 2021 SEP Marketplace enrollees. These results combine self-reported data for those who reported their race and ethnicity and the imputed race and ethnicity for those who did not self-report. The combined results estimated that 0.7 percent of Marketplace enrollees were AI/AN, 5.2 percent were AANHPI, 16.2 percent were Black, 27.6 percent were Hispanic, 2.0 percent were Multiracial, and 48.4 percent were White.

Table B.6. Racial and Ethnic Distribution of the Overall Marketplace Enrollee Sample

Race and Ethnicity	Self-Reported Race and Ethnicity with Missing Replacement, Excluding Records with Missing Race and Ethnicity		Self-Reported Race and Ethnicity with Missing Replacement and Including Records with Missing Race and Ethnicity		Combined Self-Reported and Modified BIFSG Imputed Results	
	N	Percentage	*N*	Percentage	Estimated *N*	Percentage
AANHPI	68,280	6.08	68,280	3.30	106,678	5.15
AI/AN	8,665	0.77	8,665	0.42	14,350	0.69
Black	160,845	14.32	160,845	7.77	335,426	16.21
Hispanic	252,024	22.43	252,024	12.18	571,792	27.63
Multiracial	23,228	2.07	23,228	1.12	40,436	1.95
White	610,514	54.34	610,514	29.50	1,000,765	48.36
Missing	0	0	946,040	45.71	149	0.01
Total	1,123,556	100.00	2,069,596	100.00	2,069,596	100.00

NOTE: Sum of estimated *N* for race and ethnicity categories may not match the total because of rounding.

Appendix C. Inconsistencies in Self-Reported Race and Ethnicity Across Years

Approximately 0.5 percent (n = 127,810) of Marketplace enrollees self-reported inconsistent race and ethnicity across the years they were included in the data, representing 143,061 changes in self-reported race and ethnicity. We examined all the unique combinations of race and ethnicity for those enrollees who changed their self-reported race and ethnicity across years. We categorized changes as single race and ethnicity to a different single race and ethnicity, single race and ethnicity to Multiracial, and Multiracial to single race and ethnicity (Table C.1). Table C.2 presents the most-common changes in race and ethnicity. We report these for all individuals with inconsistent self-reported race and ethnicity. We also report separately for enrollees who had self-reported race and ethnicity for all enrolled years (107,719 changes) and enrollees who were missing self-reported race and ethnicity in at least one year (35,342 changes).

Table C.1. Summary of Changes in Self-Reported Race and Ethnicity

Changes in Race and Ethnicity	All Changes (*N* = 143,061) (%)	Changes Among Enrollees Who Self-Reported Race and Ethnicity in All Years Enrolled (*N* = 107,719) (%)	Changes Among Enrollees Missing Self-Reported Race and Ethnicity in at Least One Year (*N* = 35,342) (%)
Single race and ethnicity to single race and ethnicity	49.3	48.0	53.4
Single race and ethnicity to Multiracial	31.1	32.4	27.2
Multiracial to single race and ethnicity	19.6	19.6	19.4

Table C.2. Most-Common Changes in Self-Reported Race and Ethnicity

Most-Common Changes in Race and Ethnicity	All Changes (%)	Changes Among Enrollees Who Self-Reported Race and Ethnicity in All Years Enrolled (%)	Changes Among Enrollees Missing Self-Reported Race and Ethnicity in at Least One Year (%)
White to Multiracial	18.9	19.1	18.3
White to Hispanic	18.4	19.3	15.6
Hispanic to White	13.3	12.8	14.9
Multiracial to White	11.1	11.4	10.2
Black to Multiracial	4.6	4.5	4.7
AANHPI to Multiracial	4.3	4.3	4.2
Multiracial to Black	3.5	3.2	4.3
AI/AN to Multiracial	2.6	3.0	1.4
White to Black	2.6	2.2	3.5
AANHPI to White	2.3	2.0	3.0
All other changes	18.4	18.2	19.9

Appendix D. Years of Enrollment by Race and Ethnicity

Table D.1 reports the number of years that enrollees appeared in the Marketplace data across the eight years included in our analyses, 2015 through 2022. This table is limited to states that participated in Federally Facilitated Marketplaces or State-Based Marketplaces that use the federal platform in all eight years. For each racial and ethnic group, the percentage of records in the year columns sum to 100. The results indicate that most enrollees are enrolled in Marketplace plans for multiple years. The slightly lower mean years of enrollment among AI/AN, Black, and Multiracial enrollees suggest that these groups may be the most challenging to replace missing race and ethnicity with self-reported data from another year.

Table D.1. Distribution of Years Enrolled in Marketplaces by Race and Ethnicity (%)

Race and Ethnicity	1	2	3	4	5	6	7	8	Mean Years
AANHPI	10.4	11.5	11.9	11.4	11.6	11.3	12.1	19.7	4.83
AI/AN	17.9	17.4	15.7	13.1	10.1	9.1	8.0	8.7	3.83
Black	22.4	18.3	14.6	11.2	9.2	7.8	7.6	8.9	3.65
Hispanic	18.7	15.8	13.7	11.3	10.0	9.0	9.1	12.2	4.02
Multiracial	18.5	16.6	14.4	11.7	10.3	9.1	8.7	10.7	3.94
White	16.0	15.9	14.2	12.0	10.7	9.5	9.3	12.3	4.12
All	17.0	15.8	14.0	11.7	10.4	9.3	9.3	12.4	4.10

Appendix E. Self-Reported Race and Ethnicity Prior to Implementing Modified BIFSG Imputation, by Year

Table E.1. Self-Reported Race and Ethnicity Without Missing Replacement with Other Years of Data

Race and Ethnicity	Missing		AI/AN		AANHPI		Black		Hispanic		Multiracial		White		All	
	N	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%
All Years	23,266,952	32.49	282,386	0.39	4,559,680	6.37	5,006,122	6.99	8,443,451	11.79	860,396	1.20	29,191,622	40.76	71,610,609	100.00
2015	3,074,330	34.79	25,085	0.28	456,858	5.17	765,919	8.67	956,390	10.82	83,657	0.95	3,475,215	39.32	8,837,454	100.00
2016	3,471,046	36.06	28,737	0.30	533,548	5.54	701,142	7.28	1,006,946	10.46	98,074	1.02	3,785,517	39.33	9,625,010	100.00
2017	2,797,175	30.40	31,663	0.34	600,152	6.52	646,611	7.03	1,056,774	11.49	110,862	1.20	3,957,961	43.02	9,201,198	100.00
2018	2,496,218	28.55	32,619	0.37	633,756	7.25	589,476	6.74	1,033,640	11.82	111,411	1.27	3,846,253	43.99	8,743,373	100.00
2019	2,212,947	26.31	36,880	0.44	638,122	7.59	576,205	6.85	1,037,133	12.33	110,767	1.32	3,798,985	45.17	8,411,039	100.00
2020	2,405,342	29.03	39,477	0.48	597,326	7.21	532,264	6.42	1,020,654	12.32	110,507	1.33	3,580,500	43.21	8,286,070	100.00
2021	2,845,465	34.49	40,935	0.50	521,480	6.32	507,630	6.15	1,030,330	12.49	109,878	1.33	3,195,115	38.72	8,250,833	100.00
2022	3,964,429	38.66	46,990	0.46	578,438	5.64	686,875	6.70	1,301,584	12.69	125,240	1.22	3,552,076	34.64	10,255,632	100.00

Table E.2. Self-Reported Race and Ethnicity with Missing Replacement with Other Years of Data

Race and Ethnicity	Missing		AI/AN		AANHPI		Black		Hispanic		Multiracial		White		All	
	N	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%
All Years	16,841,149	23.52	304,132	0.42	5,137,030	7.17	5,944,878	8.30	9,795,436	13.68	991,351	1.38	32,596,633	45.52	71,610,609	100.00
2015	2,119,693	23.99	28,282	0.32	527,483	5.97	902,110	10.21	1,144,783	12.95	103,036	1.17	4,012,067	45.40	8,837,454	100.00
2016	2,272,719	23.61	32,578	0.34	613,345	6.37	873,134	9.07	1,238,232	12.86	122,215	1.27	4,472,787	46.47	9,625,010	100.00
2017	1,859,249	20.21	34,561	0.38	665,052	7.23	777,475	8.45	1,246,479	13.55	131,108	1.42	4,487,274	48.77	9,201,198	100.00
2018	1,660,915	19.00	34,882	0.40	690,519	7.90	701,301	8.02	1,207,281	13.81	129,918	1.49	4,318,557	49.39	8,743,373	100.00
2019	1,579,762	18.78	39,012	0.46	692,564	8.23	667,683	7.94	1,192,162	14.17	127,597	1.52	4,112,259	48.89	8,411,039	100.00
2020	1,816,080	21.92	41,709	0.50	667,389	8.05	621,916	7.51	1,156,873	13.96	122,865	1.48	3,859,238	46.58	8,286,070	100.00
2021	2,284,970	27.69	43,265	0.52	611,498	7.41	601,775	7.29	1,130,587	13.70	118,440	1.44	3,460,298	41.94	8,250,833	100.00
2022	3,247,761	31.67	49,843	0.49	669,180	6.53	799,484	7.80	1,479,039	14.42	136,172	1.33	3,874,153	37.78	10,255,632	100.00

Table E.3. Combined Self-Reported and Modified BIFSG-Imputed Results for Nonreporters

Race and Ethnicity	Missing		AI/AN		AANHPI		Black		Hispanic		Multiracial		White		All	
	N	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%
All Years	1,944	0.00	391,337	0.55	6,259,403	8.74	8,588,414	11.99	14,192,674	19.82	1,286,066	1.80	40,890,770	57.10	71,610,609	100.00
2015	234	0.00	38,088	0.43	638,513	7.23	1,328,643	15.03	1,555,091	17.60	136,625	1.56	5,140,261	58.16	8,837,454	100.00
2016	253	0.00	42,996	0.45	736,596	7.65	1,280,152	13.30	1,670,595	17.36	161,161	1.67	5,733,257	59.57	9,625,010	100.00
2017	201	0.00	43,354	0.47	783,920	8.52	1,060,322	11.52	1,627,361	17.69	164,926	1.79	5,521,115	60.00	9,201,198	100.00
2018	150	0.00	42,860	0.49	811,746	9.28	930,404	10.64	1,583,919	18.12	160,368	1.83	5,213,926	59.63	8,743,373	100.00
2019	151	0.00	47,304	0.56	820,381	9.75	882,570	10.49	1,609,002	19.13	156,204	1.86	4,895,427	58.20	8,411,039	100.00
2020	176	0.00	51,734	0.62	815,443	9.84	873,405	10.54	1,691,488	20.41	155,294	1.87	4,698,530	56.70	8,286,070	100.00
2021	268	0.00	56,635	0.69	785,380	9.52	929,726	11.27	1,864,605	22.60	159,450	1.93	4,454,769	53.99	8,250,833	100.00
2022	511	0.00	68,367	0.67	867,423	8.46	1,303,192	12.71	2,590,613	25.26	192,040	1.87	5,233,486	51.03	10,255,632	100.00

Appendix F. Calibrated and Uncalibrated Imputation Results

The calibrated modified BIFSG results assume Marketplace enrollees in a Census block group who do not self-report race and ethnicity are more similar to enrollees who self-report race and ethnicity than residents in the overall Census block group.

Table F.1. Comparison of Overall Self-Reported and Modified BIFSG-Imputed Racial and Ethnic Distributions Among Enrollees with Self-Reported Race and Ethnicity, Uncalibrated and Calibrated Results

Race and Ethnicity	Self-Reported Results		Uncalibrated Modified BIFSG-Imputed Results	Calibrated Modified BIFSG-Imputed Results
	N	Percentage	Percentage	Percentage
AI/AN	304,132	0.56	0.59	0.55
AANHPI	5,137,030	9.38	9.29	9.38
Black	5,944,878	10.85	12.22	10.85
Hispanic	9,795,436	17.88	17.26	17.88
Multiracial	991,351	1.81	1.60	1.81
White	32,596,633	59.52	59.02	59.51
Missing	0	0.00	0.01	0.01
Total	54,769,460	100.00	100.00	100.00

Appendix G. U.S. Census Divisions

There are nine Census divisions. The states included in each Census division are reported in Table G.1.

Table G.1. U.S. Census Divisions

Census Division	States
East North Central	Illinois, Indiana, Michigan, Ohio, and Wisconsin
East South Central	Alabama, Kentucky, Mississippi, and Tennessee
Middle Atlantic	New Jersey, New York, and Pennsylvania
Mountain	Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, and Wyoming
New England	Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont
Pacific	Alaska, California, Hawaii, Oregon, and Washington
South Atlantic	Delaware; Florida; Georgia; Maryland; North Carolina; South Carolina; Virginia; Washington, D.C.; and West Virginia
West North Central	Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, and South Dakota
West South Central	Arkansas, Louisiana, Oklahoma, and Texas

Abbreviations

AANHPI	Asian American, Native Hawaiian, and Pacific Islander
AI/AN	American Indian/Alaskan Native
ASPE	Office of the Assistant Secretary for Planning and Evaluation
AUC	area under the curve
BIFSG	Bayesian Improved First Name Surname and Geocoding
BISG	Bayesian Improved Surname and Geocoding
CMS	Centers for Medicare and Medicaid Services
FIPS	Federal Information Processing System
MIDAS	Multidimensional Insurance Data Analytic System
OEP	open enrollment period

References

Centers for Medicare and Medicaid Services, "2021 Special Enrollment Period in Response to the COVID-19 Emergency," fact sheet, January 28, 2021. As of March 31, 2022: https://www.cms.gov/newsroom/fact-sheets/2021-special-enrollment-period-response-covid-19-emergency

Centers for Medicare and Medicaid Services, Office of Minority Health, "Stratified Reporting," webpage, last updated December 1, 2021. As of January 5, 2022: https://www.cms.gov/About-CMS/Agency-Information/OMH/research-and-data/statistics-and-data/stratified-reporting

CMS—*See* Centers for Medicare and Medicaid Services.

CMS, Office of Minority Health—*See* Centers for Medicare and Medicaid Services, Office of Minority Health.

Dembosky, J. W., A. M. Haviland, A. Haas, K. Hamartoma, R. Weech-Maldonado, S. M. Wilson-Frederick, S. Gaillot, and M. N. Elliott, "Indirect Estimation of Race/Ethnicity for Survey Respondents Who Do Not Report Race and Ethnicity," *Medical Care*, Vol. 57, No. 5, May 2019, pp. e28–e33.

Elliott, M. N., K. Becker, M. K. Beckett, K. Hambarsoomian, P. Pantoja, and B. Karney, "Using Indirect Estimates Based on Name and Census Tract to Improve the Efficiency of Sampling Matched Ethnic Couples from Marriage License Data," *Public Opinion Quarterly*, Vol. 77, No. 1, 2013, pp. 375–384.

Elliott, M. N., A. Fremont, P. A. Morrison, P. Pantoja, and N. Lurie, "A New Method for Estimating Race/Ethnicity and Associated Disparities Where Administrative Records Lack Self-Reported Race/Ethnicity," *Health Services Research*, Vol. 43, No. 5, Pt. 1, October 2008, pp. 1722–1736.

Elliott, M. N., P. A. Morrison, A. Fremont, D. M. McCaffrey, P. Pantoja, and N. Lurie, "Using the Census Bureau's Surname List to Improve Estimates of Race/Ethnicity and Associated Disparities," *Health Services and Outcomes Research Methodology*, Vol. 9, No. 2, 2009, pp. 69–83.

Fremont, A., J. S. Weissman, E. Hoch, and M. N. Elliott, *When Race/Ethnicity Data Are Lacking: Using Advanced Indirect Estimation Methods to Measure Disparities*, Santa Monica, Calif.: RAND Corporation, RR-1162-COMMASS, 2016. As of December 23, 2021: https://www.rand.org/pubs/research_reports/RR1162.html

Frey, W. H., "Black-White Segregation Edges Downward Since 2000, Census Shows," Brookings Institution, blog post, December 17, 2018. As of December 8, 2021: www.brookings.edu/blog/the-avenue/2018/12/17/black-white-segregation-edges-downward-since-2000-census-shows/

Grundmeier R. W., L. Song, M. J. Ramos, A. G. Fiks, W. D. Pace, A. Fremont, M. N. Elliott, R. C. Wasserman, and A. R. Localio, "Imputing Missing Race/Ethnicity in Pediatric Electronic Health Records: Reducing Bias with Use of U.S. Census Location and Surname Data." *Health Services Research*, Vol. 50, No. 4, August 2015.

Haas, A., M. N. Elliott, J. W. Dembosky, J. L. Adams, S. M. Wilson-Frederick, J. S. Mallett, S. Gaillot, S. C. Haffer, and A. M. Haviland, "Imputation of Race/Ethnicity to Enable Measurement of HEDIS Performance by Race/Ethnicity," *Health Services Research*, Vol. 54, No. 1, February 2019, pp. 13–23.

Hosmer, D. W., and S. Lemeshow, *Applied Logistic Regression*, 2nd ed., New York: Wiley-Interscience Publication, 2000.

Jones, N., R. Marks, R. Ramirez, and M. Rios-Vargas, "Improved Race and Ethnicity Measures Reveal U.S. Population Is Much More Multiracial," U.S. Census Bureau, August 12, 2021.

McCaffrey, D. F., and M. N. Elliott, "Power of Tests for a Dichotomous Independent Variable Measured with Error," *Health Services Research*, Vol. 43, No. 3, 2008, pp. 1085–1101.

Tzioumis, K., "Demographic Aspects of First Names," *Scientific Data*, Vol. 5, No. 180025, March 2018.

U. S. Census Bureau, "Frequently Occurring Surnames from the 2010 Census," webpage, undated. As of December 16, 2021: www.census.gov/topics/population/genealogy/data/2010_surnames.html

U. S. Census Bureau, "Differential Privacy and the 2020 Census," webpage, July 22, 2021. As of January 5, 2022: https://www.census.gov/library/fact-sheets/2021/differential-privacy-and-the-2020-census.html

U.S. Department of Health and Human Services, "2021 Special Enrollment Period Access Extended to August 15 on HealthCare.gov for Marketplace Coverage," press release, March 23, 2021. As of March 31, 2022: https://www.hhs.gov/about/news/2021/03/23/2021-special-enrollment-period-access-extended-to-august-15-on-healthcare-gov-for-marketplace-coverage.html

Voicu, I., "Using First Name to Improve Race and Ethnicity Classification," *Statistics in Public Policy*, Vol. 5, No. 1, March 2018 pp. 1–13.