DOI: 10.1111/1475-6773.13099

RESEARCH ARTICLE

HSR Health Services Research

Imputation of race/ethnicity to enable measurement of HEDIS performance by race/ethnicity

Ann Haas MS, MPH[1] | Marc N. Elliott PhD[2] | Jacob W. Dembosky MPM[1] | John L. Adams PhD[3] | Shondelle M. Wilson-Frederick PhD[4] | Joshua S. Mallett MS[2] | Sarah Gaillot PhD[5] | Samuel C. Haffer PhD[6] | Amelia M. Haviland PhD[1,7]

[1]RAND Corporation, Pittsburgh, Pennsylvania

[2]RAND Corporation, Santa Monica, California

[3]Kaiser Permanente Center for Effectiveness and Safety Research, Pasadena, California

[4]Office of Minority Health, Centers for Medicare & Medicaid Services, Baltimore, Maryland

[5]Centers for Medicare & Medicaid Services, Baltimore, Maryland

[6]U.S. Equal Employment Opportunity Commission, Washington, District of Columbia

[7]Carnegie Mellon University, Pittsburgh, Pennsylvania

Correspondence
Marc N. Elliott, PhD, RAND Corporation, Santa Monica, CA.
Email: elliott@rand.org

Funding information
This study was funded by the Centers for Medicare & Medicaid Services (CMS) contract GS-10F-0275P/HHSM-500-2014-00399G to the RAND Corporation.

Objective: To improve an existing method, Medicare Bayesian Improved Surname Geocoding (MBISG) 1.0 that augments the Centers for Medicare & Medicaid Services' (CMS) administrative measure of race/ethnicity with surname and geographic data to estimate race/ethnicity.

Data Sources/Study Setting: Data from 284 627 respondents to the 2014 Medicare CAHPS survey.

Study Design: We compared performance (cross-validated Pearson correlation of estimates and self-reported race/ethnicity) for several alternative models predicting self-reported race/ethnicity in cross-sectional observational data to assess accuracy of estimates, resulting in MBISG 2.0. MBISG 2.0 adds to MBISG 1.0 first name, demographic, and coverage predictors of race/ethnicity and uses a more flexible data aggregation framework.

Data Collection/Extraction Methods: We linked survey-reported race/ethnicity to CMS administrative and US census data.

Principal Findings: MBISG 2.0 removed 25-39 percent of the remaining MBISG 1.0 error for Hispanics, Whites, and Asian/Pacific Islanders (API), and 9 percent for Blacks, resulting in correlations of 0.88 to 0.95 with self-reported race/ethnicity for these groups.

Conclusions: MBISG 2.0 represents a substantial improvement over MBISG 1.0 and the use of CMS administrative data on race/ethnicity alone. MBISG 2.0 is used in CMS' public reporting of Medicare Advantage contract HEDIS measures stratified by race/ethnicity for Hispanics, Whites, API, and Blacks.

KEYWORDS
biostatistical methods, HEDIS, Medicare, quality of care/patient safety (measurement), racial/ethnic differences in health and health care

1 | INTRODUCTION

While there is clear evidence of racial/ethnic disparities in health care quality, examining these differences among seniors has been limited by the quality of Medicare administrative race/ethnicity data.[1] Since these administrative data are sometimes the primary data informing federal monitoring of disparities in the care of Medicare beneficiaries, improving their accuracy is necessary to support efforts by the

Centers for Medicare & Medicaid Services (CMS) to measure and reduce disparities.

CMS' administrative racial/ethnic information for Medicare beneficiaries is primarily derived from the Social Security Administration (SSA). For persons assigned a Social Security Number before 1980, there were only three race/ethnicity response options: "Black," "White," or "Other." In 1980, SSA expanded these categories. Prior research suggests that CMS administrative race/ethnicity performs reasonably well in classifying non-Hispanic White and Black beneficiaries, but it misclassifies many Asian/Pacific Islanders (API) and Hispanic beneficiaries as "White" or "Other."[2-4] People who are incorrectly classified by the administrative variable differ systematically from others with the same race/ethnicity, which leads to biased disparity estimates.[5]

One way of addressing imperfect racial/ethnic information is "indirect estimation" methods that supplement or replace imperfect racial/ethnic measures with estimates based on characteristics strongly associated with race/ethnicity. The Institute of Medicine (now the National Academy of Medicine) recommended indirect estimation to monitor health disparities and to target quality-improvement efforts in the absence of direct and accurate race and ethnicity information.[6,7]

Accordingly, researchers modified an existing method for indirectly estimating race/ethnicity from residential address and surname information, known as the Bayesian Improved Surname and Geocoding (BISG),[8] to improve the accuracy of CMS' administrative race/ethnicity variable. BISG uses Bayes' rule to combine US Census information on race/ethnicity by both surname and Census Block Group of residence to produce a set of racial/ethnic probabilities for each person.[8] In this method, each person receives a set of initial probabilities of falling into each of six racial/ethnic groups (White, Black, Hispanic, API, American Indian/Alaska Native [AI/AN], and multiracial) based on the racial/ethnic distributions for their surname, as published by the Census. Addresses are geocoded to Census Block Groups, and the probability of residing in each Census Block Group is calculated for each racial/ethnic group, also using Census data. These two sets of probabilities are combined using Bayes' rule. This method requires an assumption of conditional independence, in this case that the probability of residing in a Block Group given a person's race/ethnicity does not vary by surname. The BISG method has been validated as a tool to impute or to improve imputation of race/ethnicity in a variety of populations.[8-12]

The Medicare-specific adaptation known as Medicare BISG 1.0 (MBISG 1.0)[13,14] combined the BISG racial/ethnic probabilities with CMS race/ethnicity administrative data to produce more accurate indirect estimates of race/ethnicity. The Bayesian method used to aggregate data can only accommodate inputs in the form of a set of six racial/ethnic probabilities. Therefore, in MBISG 1.0, the probability of belonging to each of these six racial/ethnic groups for someone with a given value on the CMS administrative racial/ethnic measure is calculated by linking the administrative field to weighted self-reported race/ethnicity from a large, nationally representative survey of Medicare beneficiaries. This initial set of six administrative probabilities of beneficiary race/ethnicity is based only on CMS' administrative data. Then, each beneficiary's probabilities are calculated independently from surname and address information using BISG. Finally, these two sources of information are combined using a Bayesian method similar to what is used within BISG to produce the set of six MBISG 1.0 probabilities for each person.

MBISG 1.0 underlay the initial reporting of Medicare Advantage (MA) Healthcare Effectiveness Data and Information Set (HEDIS) measures by race/ethnicity within Medicare contracts, by the CMS Office of Minority Health (OMH).[13,15] Compared to the CMS administrative race/ethnicity variable, improvements were very large for Hispanic beneficiaries, moderate for API and White beneficiaries, and small for Black beneficiaries (for whom the CMS administrative variable was already excellent). Correlations between the MBISG 1.0 probabilities and self-reported race/ethnicity are high for these four groups: 0.86 for White, 0.94 for Black, 0.79 for Hispanic, and 0.89 for API. MBISG 1.0 is not recommended for inferences regarding AI/AN or multiracial groups, for which the algorithm's performance is poorer.

While MBISG 1.0 performance was strong, opportunities for improvement remained, especially for Hispanic beneficiaries. The MBISG 2.0 method described here sought to improve the MBISG 1.0 algorithm and underlies current OMH stratified reporting of the HEDIS measures by race/ethnicity.

We organize the description of changes to the race/ethnicity prediction algorithm presented in this article into two sequential phases. The first phase concerns improvements made to the three components of the MBISG 1.0 algorithm (address, surname, and CMS administrative racial/ethnic information). The second phase uses a more flexible framework, multinomial logistic regression, which allows for both additional types of predictors of race/ethnicity and relaxation of the conditional independence assumption used in the application of Bayesian updating used in MBISG 1.0.

2 | METHODS

Validation data came from the 2014 MA and fee-for-service (FFS) Medicare Consumer Assessment of Healthcare Providers and Systems (CAHPS) surveys (41.1 percent response rate), which are annual cross-sectional surveys of Medicare beneficiaries' health care experiences; patterns of survey nonresponse were typical for such surveys.[16] Analytic weights poststratified respondents to match the Medicare population within state (FFS) or contract (MA) on an extensive list of variables, including CMS race/ethnicity, gender, dual eligibility, and zip code-level distributions of income, education, and race/ethnicity using iterative proportional fitting.[17] Race and ethnicity are self-reported using standard items and recoded as non-Hispanic White, Black, Hispanic, API, AI/AN, and multiracial. The 4.1 percent of survey respondents who did not report race/ethnicity were omitted, leaving an analytic sample of 284 627.

We linked survey-reported (CAHPS) race/ethnicity to data from CMS, the US Census, and other sources (as described below) to calculate racial/ethnic probabilities. We evaluated the accuracy of

TABLE 1 Summary of changes from MBISG 1.0 to MBISG 2.0

Issue	Solution
Phase 1	
Improve the three components of the MBISG 1.0 algorithm	
a. Approximately 10% of surnames do not match to Census list and receive a default set of probabilities. Of these unlisted surnames, about 20% are compound surnames.	Split unlisted compound names into component names and combine sets of probabilities from listed component names. If only one component name matches, its set of probabilities is used for surname probabilities. If two or more component names are listed, the highest Hispanic probability is used for the Hispanic surname probability and the means of the race probabilities are rescaled so that the set of surname probabilities sums to 1.
b. Association between CMS administrative race/ethnicity variable and self-reported race/ethnicity may vary by age.	Incorporate age into cross-tabulation.
c. Address probabilities for Puerto Rico are not available; current approach underestimates probability of Hispanic ethnicity.	Develop a set of racial/ethnic probabilities based on self-report for residents of Puerto Rico.
Calibration to population	
d. Means of probabilities underestimate proportion of sample who are Hispanic and multiracial and overestimate proportion White.	Calibration to Medicare population using additive approach for multiracial and multinomial logistic model for other race/ethnicities.
Phase 2	
e. Bayesian framework used to combine data from address, surname, and CMS administrative variable may ignore possible interactive predictive power between elements; conditional independence may not be fully met.	Use multinomial logistic model to allow interactions between existing data elements and between existing and new data elements.
f. Additional data elements may improve racial/ethnic probabilities.	Add such elements, including indicators of first names indicative of API or Hispanic race/ethnicity, gender, low income indicators.

racial/ethnic probabilities using unweighted Pearson correlations with self-reported race/ethnicity. We focus on the four largest racial/ethnic groups (White, Black, Hispanic, and API). Below and in Table 1, we describe the changes made to the MBISG 1.0 algorithm that resulted in MBISG 2.0. In Phase One, we improve MBISG 1.0 within the original Bayesian framework by making better use of the original three data elements—surnames, administrative race/ethnicity data, and residential address. Phase Two uses the improved MBISG 1.0 probabilities from Phase One as inputs, along with additional predictors (first names, Spanish preference, demographics, and health insurance), in a more flexible regression framework.

2.1 | Phase One

Below we describe improvements to the three existing data sources: surname, CMS administrative race/ethnicity, and address information.

2.1.1 | Improving use of surname information

Surnames are linked to 2000 Census distributions of self-reported race/ethnicity by surname for surnames appearing 100 or more times in the Census.[18] Approximately 10 percent of the surnames in our dataset do not appear on this surname list, similar to the proportion in the US population. MBISG 1.0 assigned default probabilities to all unlisted surnames, whereas MBISG 2.0 improves the handling of a subset of unlisted surnames.

Compound surnames are formed as a combination of surnames. While compound names may be listed if 100+ people in the 2000 Census had the exact same compound surname, the majority (95.9 percent) are not listed. In our sample, 18.4 percent of unlisted names were compound names.

We developed a strategy for improving predictions for unlisted compound names. Almost all unlisted compound surnames (99.9 percent) had exactly two component names. We attempted to match each component name to the Census list and kept the set of six probabilities associated with each matching name. For 1.2 percent of compound names, no components matched the Census list; exactly one name matched for 12.0 percent; and exactly two names matched for 86.8 percent.

The listed component names had a high probability of being Hispanic, and beneficiaries with listed component names had a high probability of self-reporting Hispanic ethnicity. The mean Hispanic probabilities were 53.0 percent for those with one listed component names; 60.4 percent of the corresponding beneficiaries self-reported Hispanic ethnicity.

We considered several methods of combining the sets of probabilities into a single set of surname probabilities in cases where more than one component name matched to the Census list. We speculated that even respondents with only one typically Hispanic component of their compound surname would be likely to self-report Hispanic ethnicity. We found a close match to the distribution of self-reported race/ethnicity when we combined the sets of probabilities associated with each matching component name by taking the highest Hispanic probability among listed component names and rescaling the means of the other probabilities so that the final set of surname racial/ethnic probabilities sums to 100 percent. Of cases with two component names that matched the Census, 70.7 percent self-reported Hispanic ethnicity. Taking the means of the two sets of surname probabilities gave an estimate of 60.5 percent Hispanic, whereas using the highest Hispanic probability gave an estimate of 66.9 percent Hispanic.

2.1.2 | Improving use of CMS administrative racial/ethnic data

Our Bayesian updating requires inputs in the form of a set of six racial/ethnic probabilities. CMS' SSA-based administrative racial/ethnic variable has one value for each beneficiary (White, Black, Hispanic, Asian/PI, AI/AN, Other, or Missing). To use this data source, we first convert it to a set of racial/ethnic probabilities corresponding to the distribution of self-reported race/ethnicity in representative, weighted data, here using 2014 MA and FFS CAHPS data.

For MBISG 1.0, these probabilities do not vary by age. MBISG 2.0 sought improvement by allowing these probabilities to vary by beneficiary age, with two motivations. First, a higher proportion of younger beneficiaries are Black or Hispanic due to demographic shifts (see Appendix S1A). Second, age is likely to be related to the inaccuracy of the SSA-based race/ethnicity measure, as it should be correlated with the (unavailable) indicator of a beneficiary's SSA form having been updated or obtained after 1979.

To develop stratified MBISG 2.0 estimates by age, we created six age categories guided by the available sample size: 18-34/35-54/55-64/65-74/75-84/85+ years. For each CMS administrative race/ethnicity category and each age group, we calculated the distribution of self-reported race/ethnicity. To avoid imprecise estimates, we collapsed adjacent age groups to have at least 450 observations.

The resulting proportions are reported separately for each of the CMS administrative racial/ethnic categories in Appendix S1B. In each table, the rows represent the specified age categories, the columns are the self-reported racial/ethnic categories, and the cell values are proportions. For example, someone 18-54 years old and "White" by CMS administrative race/ethnicity would be assigned 87.4 percent White, 0.3 percent Black, 7.2 percent Hispanic, 0.2 percent API, 0.8 percent AI/AN, and 4.1 percent multiracial values. Due to small sample sizes, no age stratification was possible for those listed as AI/AN or unknown in the CMS administrative data. As seen in Appendix S1B, younger beneficiaries listed as "Other" and "White" are more likely to be Hispanic than older beneficiaries in these categories.

2.1.3 | Improving use of residential address

Addresses are geocoded to 12-digit Block Groups and linked to 2010 Census race/ethnicity data. Puerto Rico is not included in this data, so we cannot estimate address-based racial/ethnic probabilities for Puerto Rico using this file. MBISG 1.0 probabilities for Puerto Rico consider only surname and SSA race/ethnicity, underestimating the prevalence of Hispanic ethnicity in Puerto Rico. Therefore, we coded racial/ethnic probabilities for residents of Puerto Rico to probabilities based on self-report of all beneficiaries in Puerto Rico, after Bayesian updating (see Appendix S1C).

2.1.4 | Aggregation of data elements using Bayesian updating and calibration

For both MBISG 1.0 and MBISG 2.0, probabilities based on surname, address, and CMS administrative race/ethnicity are combined in two Bayesian updating steps, as described in the Introduction. MBISG 1.0 probabilities are complete at this point.

The means of MBISG 1.0 predictions slightly undercount Hispanics and multiracial beneficiaries and slightly overcount Whites. For MBISG 2.0, we calibrate probabilities such that mean probabilities are equal to the distribution of self-reported race/ethnicity in the sample.

Because the multiracial group is both underestimated and poorly predicted, we used an additive approach to calibrate this group. We added the difference between percent self-reporting multiracial and mean probability multiracial to each case and rescaled other probabilities to sum to 1. This sets the mean multiracial probability to the mean self-reporting multiracial.

We then calibrate probabilities using multinomial logistic regression, a common means of calibrating three or more probabilities to a national total[8] that applies constant odds ratios rather than a constant arithmetic difference, allowing well-predicted probabilities to increase more than proportionately. We predicted self-reported race/ethnicity using the set of race/ethnicity probabilities; these predictions by definition have a mean equal to the sample distribution of race/ethnicity.

Table 1 summarizes Phase One steps. The rest of this Section 2 describes additional steps taken to calculate MBISG 2.0 probabilities from the improved Phase One MBISG 1.0 probabilities plus additional inputs.

2.2 | Phase Two

The purpose of Phase Two was to evaluate the contributions of additional data elements and the application of a more flexible modeling framework—multinomial logistic regression—in improving the accuracy of MBISG probabilities. The inputs considered for the model are the improved MBISG 1.0 probabilities from Phase One, the individual Phase One subcomponents, and other data elements described below.

We begin with a summary of the Phase One subcomponents and new data elements. Next, we describe the modeling approach that combines these elements into a new set of probabilities

2.3 | Additional data elements considered for Phase Two

2.3.1 | Phase One subcomponents

In addition to the end-of-phase-one probabilities, we considered component racial/ethnic probabilities from each of three data sources used in Phase One (the CMS administrative variable, surname, and address) and the probabilities from the combination of surname and address. We additionally explored poor/missing data indicators for the Phase One subcomponents, since data quality might be correlated with race/ethnicity (see Appendix S1C).

2.3.2 | First names

We explored the use of beneficiary first name. Morrison et al[19] noted that there is generally less information in first names than last names; first name data are less standardized than last name data as one person's first name may appear in different forms. There are some distinctive first names that have a high specificity for one race/ethnicity, but the most popular first names are prevalent for all racial/ethnic groups.

We evaluated first name lists that have been found to predict membership in certain racial/ethnic groups. We matched these group-specific lists to our sample using beneficiary first name. Each first name list corresponded to a single racial/ethnic group; for each such group, we calculated the sensitivity and specificity of the corresponding list. Among beneficiaries whose first name was listed, we calculated the distribution of self-reported race/ethnicity and mean end-of-phase-one probabilities; the discrepancy between the two indicated how much first name information could increase accuracy. Promising first name lists were evaluated in modeling.

2.3.3 | Asian/Pacific Islander first name lists

Lauderdale and Kestenbaum[20] developed first name lists for each of the six largest American API national origin groups (Chinese, Indian, Japanese, Korean, Filipino, and Vietnamese) using SSA administrative data. The number of distinct first names on these lists ranges from 670 for Vietnamese to 3688 for Chinese (see also Wong et al[21]).

A total of 1.3 percent of all respondents and 34.5 percent of API respondents had a first name that matches one list. Specificity for these names is high (99.8 percent). Appendix S1D shows self-reported race/ethnicity and mean of end-of-phase-one probabilities for beneficiaries matching each of the six lists, plus a summary for all beneficiaries matching any list. Having a first name on the Japanese, Korean, or Vietnamese list was associated with a high probability of self-reporting API (about 95 percent), while matching to the Indian, Filipino, or Chinese list was associated with a somewhat lower probability of being API (88 percent, 72 percent, 70 percent, respectively). Matching to the Filipino list was associated with a 23 percent chance first of self-reporting as Hispanic.

The mean end-of-phase-one API probability for those whose first name appears on any of the six API first name lists is lower than the percentage of these people who self-report being API by 13.3 percentage points overall (72.0 percent versus 85.3 percent). The differences vary by national origin group, with the differences especially large for the Japanese (41.3 percentage points) and Korean and Filipino (17.9 percentage points each) lists. This suggests that additional information on race/ethnicity can be gained from these first names. We explored the potential for improving the probabilities by including indicators for these six-first name lists in modeling.

2.3.4 | Hispanic first name list

Morrison et al[19] produced a list of 49 first names strongly identified with Hispanic ethnicity; 1.1 percent of all beneficiaries in our sample and 12.1 percent of Hispanic beneficiaries had a first name on this list. Specificity is high (99.9 percent). Appendix S1D reports the distribution of self-reported race/ethnicity and mean end-of-phase-one probabilities for beneficiaries whose first name matches this list. Ninety-one percent of these beneficiaries reported being Hispanic. The mean end-of-phase-one Hispanic probability is 78.9 percent, 12.2 percentage points lower than the percentage reporting Hispanic ethnicity. This indicates potential gain from adding this information, so we included this indicator in modeling.

2.3.5 | Spanish preference outside of Puerto Rico

Although the CMS administrative files include a Spanish-preference indicator, it is missing for many cases: many beneficiaries have not indicated or been asked their language preference. Because we account for Puerto Rico residence in Phase One, we used a version of Spanish preference that excluded Puerto Rico (1 = Spanish preference and outside of Puerto Rico, 0 = in Puerto Rico, language preference other than Spanish, or missing language preference).

Only 0.4 percent of all beneficiaries and 4.5 percent of Hispanic beneficiaries have an indicator of preferring Spanish outside of Puerto Rico. However, almost everyone (98.6 percent) with this indicator self-reports as Hispanic (see Appendix S1D), and the end-of-phase-one probabilities underestimate this probability by almost 10 percentage points. Therefore, we tested the potential of this variable to improve probabilities in models.

2.3.6 | CMS administrative demographic and coverage variables

There are persistent differences in demographics[22-25] and MA enrollment[22,24] by race/ethnicity. We therefore also evaluated the effect of including several CMS administrative demographic and coverage variables in the model: age (categorized as 18-44, 45-54, 55-64, 65-69, 70-74, 75-79, 80-84, 85-89, ≥90 years), gender, disability (separately for those less than 65 and 65 or older), dually eligible, low income subsidy (LIS) but not dually eligible, and coverage type (FFS only, FFS with Part D coverage [FFS/PDP], MA-only, MA with Part D coverage [MA-PD], and indicators for being in a dual special needs plan [SNP] or chronic condition SNP).

2.4 | Modeling approach

We used multinomial logistic regression in MBISG 2.0 so that we could incorporate both new and prior data elements outside of the Bayesian framework. In this case, we have six outcome categories: the self-reported racial/ethnic groups. The individual-level predictors that we considered (surname, address, etc.) are described above, under "Data elements considered for inclusion in Phase Two."

Multinomial logistic regression automatically calibrates; it guarantees that the mean probabilities of the predictions match the distribution of self-reported race/ethnicity in the sample.

Generalized linear models with categorical outcomes are best summarized via the contributions of variables or groups of variables to improving the log likelihood of the model fit. We therefore evaluated the relative explanatory contributions of blocks of predictors in the final multivariate model by comparing the log likelihoods for models including and excluding those variables in an analysis of deviance.[26]

Our evaluation of models in Phase Two uses 10-fold cross-validation[27] to avoid overfitting, in which incidental associations in the modeling dataset lead to overstating gains in predictive accuracy for new cases. We calculated *percent shrinkage* as (uncross-validated correlations—cross-validated correlations)/(uncross-validated correlations) to estimate how much we would have overstated correlations without cross-validation.

3 | RESULTS

Table 2 describes the validation sample and tests for differences between Whites and each other group. Compared with Whites, higher proportions of Blacks, Hispanics, and AI/AN beneficiaries are younger than 65 (which generally indicates Medicare eligibility due to disability), have lower educational attainment, and report worse general and mental health. A higher proportion of API beneficiaries are in the middle age categories, and a higher proportion report having either less than a high school education or having a bachelor or graduate degree compared with Whites. Dual eligibility for Medicaid (an indicator of low income) is less frequent for Whites than for other groups. Census divisions of residence (nine categories plus a tenth category for US Territories) vary markedly by race/ethnicity. The proportion of Black beneficiaries is highest in the South Atlantic division, the Hispanic proportion is highest in the West South Central and Pacific, the API proportion is highest in the Pacific, and the AI/AN proportion is highest in the West South Central and Mountain regions. Almost all residents of Puerto Rico and the U.S. Virgin Islands are Hispanic.

3.1 | Summary of multinomial logistic model

Based on the investigations of the elements described above, the final MBISG 2.0 multinomial logistic model includes as predictors:

- Main effects for the end-of-phase-one probabilities and two-way interactions between each pair, including quadratic forms
- Subcomponents of the end-of-phase-one probabilities: main effects for five CMS administrative probabilities, main effects for five name-address component probabilities, and five non-White same-group interaction terms between these probabilities
- Additional data elements:
    - Indicators for match to six API first name lists, and interactions between each API group and the API end-of-phase-one probability
    - Indicator for match to Hispanic first name list, and interaction with Hispanic end-of-phase-one probability
    - Indicator for Spanish preferring but not in Puerto Rico, and interaction with the Hispanic end-of-phase-one probability
    - Demographic/coverage variables: gender, disability and younger than 65, dually eligible, and indicators of enrollment in a dual SNP or chronic condition SNP.

Appendix S1E presents coefficients from this final model.

We found that the elements that most contributed to improvement over the end-of-phase-one probabilities were Phase One probability subcomponents (CMS administrative race/ethnicity and name-address probabilities) and their interactions with the Phase One probabilities (Table 3). Each block of predictors significantly improves the model predictions ($P < 0.001$). Adding information on API first names, Hispanic first names, and Spanish preference increased correlations with self-reported race/ethnicity, especially for the API, Hispanic, and White probabilities (results not shown), but marginal gains were small. Further addition of demographic, language preference, and coverage information had only a small marginal impact. Appendix S1F compares the correlation of self-reported race/ethnicity with uncross-validated and cross-validated probabilities from the final multinomial model.

Table 4 shows correlations between self-reported race/ethnicity and different versions of the racial/ethnic probabilities. The first column shows correlations with MBISG 1.0 probabilities. Correlation between self-report and the original racial/ethnic probabilities for the four largest racial/ethnic groups is highest for Blacks (0.94) and lowest for Hispanics (0.79). Phase One changes increased correlation with self-report most for Hispanic (by 0.068) and White (by 0.032) probabilities (comparing columns 1 and 2). Increases for API and Black probabilities were small (0.002 for API and 0.003 Black).

The Phase Two multinomial logistic model, which results in MBISG 2.0, adds 0.014 for White, 0.023 for Hispanic, and 0.029 for API over Phase One (comparing columns 2 and 3; see also column 4). Performance is higher and more uniform in MBISG 2.0 among the four largest groups compared with the MBISG 1.0. Compared with MBISG 1.0 probabilities, MBISG 2.0 adds 0.006 for Black ($r = 0.95$), 0.030 for API ($r = 0.92$), 0.046 for White ($r = 0.90$), and 0.091 for Hispanic ($r = 0.88$). Compared with the MBISG 1.0 probabilities, the updated algorithm through Phase Two also adds 0.072 for AI/AN, though correlations with self-report are still low (0.54). It also adds 0.075 points for multiracial, though predictions are only correlated at 0.12 with self-report.

Table 4 also shows the percentage of error reduced by phase 1 and phase 2. MBISG 2.0 removes about 40 percent of MBISG 1.0 error for Hispanic, 30 percent for White, 25 percent for API, and 9 percent for Black probabilities. For Hispanic, White, and Black probabilities, 55-65 percent of error reduction occurred in Phase One.

TABLE 2 Composition of validation sample[a]

	Overall	White	Black	Hispanic	API	AI/AN	Multiracial
Sample size	284 627	214 519	25 684	27 122	10 257	1334	5711
Age (%)							
18-64	14	11	27***	20***	7***	29***	27***
65-69	24	24	23*	24	25	22	24
70-74	22	23	20***	21***	25**	21	20**
75-79	16	16	13***	14***	19***	15	13***
80-84	12	12	9***	11***	12	8**	8***
85+	13	14	9***	10***	12**	5***	8***
Female	55	55	58***	55	57*	52	54
Education							
8th grade or less	6	3	9***	28***	12***	13***	8***
Some high school	9	7	17***	14***	7	11***	12***
High school graduate	32	33	32	25***	20***	32	28***
Some college/2-y degree	27	28	26***	19***	19***	31	34***
Bachelor degree	11	12	7***	7***	21***	7*	8***
More than 4 y college	15	17	9***	7***	21***	6***	11***
General health status							
Excellent	8	9	5***	8	8	7	8
Very good	27	30	19***	18***	23***	18***	21***
Good	36	36	37	32***	40***	34	31***
Fair	22	20	31***	32***	23***	28***	27***
Poor	6	6	8***	10***	6	14***	13***
Mental health status							
Excellent	24	26	19***	19***	21***	16***	20***
Very good	32	34	27***	23***	31***	28**	27***
Good	28	27	32***	31***	32***	28	29*
Fair	12	10	18***	22***	13***	21***	18***
Poor	3	3	4***	5***	3	7***	5***
Dually eligible (Medicare/Medicaid)	15	10	32***	32***	34***	33***	26***
Low income subsidy deemed	3	2	6***	5***	3	5***	5***
Disabled	16	14	31***	23***	8***	34***	32***
Coverage type							
FFS	71	73	68***	54***	68***	83***	74
MA	29	27	32***	46***	32***	17***	26
Census division[b] and Puerto Rico/Virgin Islands							
New England	5	6	2***	2***	2***	2***	4**

(Continues)

TABLE 2 (Continued)

	Overall	White	Black	Hispanic	API	AI/AN	Multiracial
Middle Atlantic	13	14	14	12***	14	5***	10***
East North Central	15	17	15***	4***	7***	8***	13***
West North Central	7	8	3***	1***	2***	9	6***
South Atlantic	20	19	33***	15***	11***	15*	20
East South Central	7	7	10***	1***	1***	4**	9***
West South Central	10	9	13***	17***	6***	20***	13***
Mountain	6	7	2***	8***	5***	20***	6*
Pacific	14	12	7***	23***	53***	17**	17***
Puerto Rico or Virgins Islands	1	<1	<1***	17***	<1	<1	<1*

Note: Weighted. N = 284 627.

**P* < 0.05, ***P* < 0.01, and ****P* < 0.001 for test of differences between White and each other group.

[a]4.1% of respondents did not report their race/ethnicity and are omitted from this table and all analyses.

[b]Census divisions are created from state of beneficiary residence as follows: New England: CT, MA, ME, NH, RI, VT; Middle Atlantic: NJ, NY, PA; East North Central: IL, IN, MI, OH, WI; West North Central: IA, KS, MN, MO, ND, NE, SD; South Atlantic: DC, DE, FL, GA, MD, NC, SC, VA, WV; East South Central: AL, KY, MS, TN; West South Central: AR, LA, OK, TX; Mountain: AZ, CO, ID, MT, NM, NV, UT, WY; Pacific: AK, CA, HI, OR, WA.

For the API probability, almost all error reduction occurred in Phase Two.

4 | CONCLUSION

Policy makers and researchers have become increasingly interested in understanding and addressing racial/ethnic disparities in health care quality and access. Improvements to Medicare racial/ethnic data facilitate this goal in a large and high-need US population.

The findings described herein support the use of indirect methods to report on race/ethnicity for Medicare beneficiaries. We recognize that self-reported race/ethnicity is the preferred method for collecting race/ethnicity data.[28] However, for administrative data such as Medicare claims data, this information is often unavailable or improperly reported. Indirect methods provide a less burdensome strategy for collecting race/ethnicity data. As described in the Improving Medicare Post-Acute Care Transformation (IMPACT) Act of 2014 Strategic Plan for Accessing Race and Ethnicity Data,[29] CMS could explore the possibility of collecting data from Medicare beneficiaries upon enrollment. While this option would also permit CMS to collect additional demographic data needed to determine quality, resource use, or payment, as well as for measuring and addressing disparities, it would require CMS to modify current enrollment forms or conduct a supplemental mail or electronic survey. There is precedent for this type of supplemental data collection;[4] however, this activity may pose considerable financial constraints on CMS and reporting burden on Medicare beneficiaries. By leveraging existing data systems, we can use indirect methods to report on the race/ethnicity of Medicare beneficiaries.

As the quality of CMS administrative race/ethnicity data changes, it may be necessary to revisit the MBISG 2.0 algorithm. For instance, in 1989 the SSA began "enumeration at birth," which allows parents to apply for Social Security Numbers for their newborns by sending birth certificate information to the SSA. No racial/ethnic information is collected by the SSA in these applications, potentially removing one current source of racial/ethnic data for Medicare applications. Though relationships between self-reported race/ethnicity and predictors have been fairly stable over time, we recommend that CMS periodically re-estimate coefficients using more recent data.

Here, we find that the MBISG 2.0 substantially improves prior methods by using additional data elements and a more flexible modeling framework, removing about a third of the MBISG 1.0's remaining error. Since the largest improvements were in the group with the lowest previous performance (Hispanics), MBISG 2.0 performance is higher and more uniform than its predecessor. The MBISG 2.0 improvements are particularly important for assessing and reducing disparities affecting Hispanic Medicare beneficiaries, a large and growing group. Performance for the Black group remained the highest of any group and improved only modestly.

As with BISG and MBISG 1.0, MBISG 2.0 produces a set of six probabilities for each beneficiary. Although one can convert these probabilities to a categorical variable, for instance by assigning each beneficiary to the group with the highest probability in their set, this approach generally results in less accurate disparity estimates than

TABLE 3 Marginal contribution of predictors in phase 2 multinomial model

Predictors dropped from full model	dF full model—dF model[a]	Deviance (−2 log L model)—(−2 log L full model)	Joint test of significance, *P*-value
Phase 1 probabilities and all interactions, main effects of CMS administrative race/ethnicity probabilities and name-address probabilities, and interactions	220	52 469 413	<0.001
First names, all interactions	70	458 479	<0.001
Demographics	15	85 494	<0.001
Spanish preference outside of PR	10	35 702	<0.001
SNP coverage	10	20 390	<0.001

Notes: Each row represents a model that drops the variables in that row from the full model. Interactions using main effects from different blocks are included in both blocks. These variables are indicated in italics.

[a] Full model has 285 df (excluding five intercepts) and −2 log likelihood = 22 018 912; null model has 0 dF and −2 log likelihood = 85 381 137.

N = 284 627.

Phase 1 probabilities and all interactions, main effects of SSA and name-address and interactions

- Main effects for the five end-of-phase-one probabilities (reference group: White) and fifteen two-way interactions between each pair of probabilities, including quadratic forms
- Main effects for five CMS administrative probabilities, main effects for five name-address component probabilities, and five non-White same-group interaction terms between these probabilities
- Indicator for missing address information
- *Interactions between API Phase One probability and each of six API first name indicators*
- *Interaction between Hispanic Phase One probability and Hispanic first name indicator*
- *Interaction between Hispanic Phase One probability and Spanish-preferring indicator*

First names, all interactions

- Indicators for first name match each of six API first name lists, and *interactions between each API group and the API end-of-phase-one probability*
- Indicator for first name match to Hispanic first name list, and *interaction with Hispanic end-of-phase-one probability*

Demographics

- Gender
- An indicator for disability and age younger than 65
- An indicator for dually eligible

Spanish preference outside of PR

- Indicator for Spanish preferring but not in Puerto Rico, and *interaction with the Hispanic end-of-phase-one probability*

SNP coverage

- Indicators of enrollment in a dual SNP and chronic SNP

the recommended direct use of probabilities as regressors or the use of probabilities within multiple imputation.[30]

Our study has several limitations. Estimates are based on a sample of voluntary respondents; however, the detailed poststratification weighting of survey respondents to represent the Medicare population on many demographic and geographic variables helps the MBISG 2.0 model accurately estimate population-level associations with race/ethnicity. The MBISG 2.0 model is less parsimonious than MBISG 1.0, requiring many more predictors that may not be available to those without access to CMS administrative data. These additional predictors may limit applications to other datasets, so future research could develop intermediate approaches that would retain some of the important predictors unique to version 2.0 while trying to minimize data needs.

We recommend MBISG 2.0 probabilities for general use with CMS data, including estimating racial/ethnic disparities among Black, Hispanic, API, and White groups. Although MBISG 2.0 significantly improves estimates for AI/AN and multiracial beneficiaries, the resulting probabilities for these groups are still not recommended for general use. MBISG 2.0 probabilities have been calculated for all 56 million Medicare beneficiaries as a tool to aid research about Medicare beneficiaries. While survey-based quality measures such as CAHPS surveys and the Health Outcome Survey generally collect self-reported race/ethnicity, the MBISG 2.0 probabilities can be linked to CMS quality measures based on administrative data, such as encounter data, claims, voluntary disenrollment from plans, and mortality data, to estimate and monitor racial/ethnic disparities in health care, identify targets to reduce disparities, and evaluate quality-improvement efforts. The size of many Medicare administrative datasets allows for differences in care by race/ethnicity to be accurately measured in subgroups such as regions and health plans,

TABLE 4 Correlation between self-report and different versions of the racial/ethnic probabilities and percent error reduction of original probabilities during each stage

	MBISG 1.0	End of phase 1	MBISG 2.0	MBISG 2.0-End of phase 1	MBISG 2.0-MBISG 1.0	% of MBISG 1.0 error reduced by phase 1	% of phase 1 error reduced by MBISG 2.0	% of MBISG 1.0 error reduced by MBISG 2.0
White	0.85507	0.88701	0.90145	0.01444	0.04638	20.7%	12.1%	30.3%
Black	0.94082	0.94401	0.94640	0.00239	0.00558	5.2%	4.2%	9.2%
Hispanic	0.78500	0.85276	0.87554	0.02278	0.09054	28.9%	14.4%	39.2%
API	0.88614	0.88780	0.91632	0.02852	0.03018	1.4%	24.3%	25.3%
AI/AN	0.46683	0.46970	0.53851	0.06881	0.07168	0.3%	8.9%	9.2%
Multiracial	0.04817	0.10398	0.12283	0.01885	0.07466	0.9%	0.4%	1.3%

Notes: Percent error reduced is $[(1 - r_{old}^2) - (1 - r_{new}^2)]/(1 - r_{old}^2)$, where r_{old} and r_{new} are the Pearson correlation between self-report and the old and new estimates, respectively.

N = 284 627.

which in turn allows for precise evaluation of interventions and quality-improvement efforts. For example, the MBISG 2.0 probabilities underlie current CMS reporting of HEDIS data by race/ethnicity within MA contracts. The MBISG estimates can be used as predictors in multivariate modeling, which allows for estimation of racial/ethnic disparities after adjusting for other factors, and for estimation of mediation and moderation effects involving race/ethnicity.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: This work has been presented at the Joint Statistical Meeting conference on August 3, 2016 in Chicago, Illinois. This work has also been presented at the Academy Health Annual Research Meeting on June 26, 2017 in New Orleans, Louisiana. The authors would like to thank Biayna Darabidian, BA for help with manuscript preparation.

CONFLICT OF INTEREST

All authors declare no conflict of interest. However, please note that Shondelle M. Wilson-Frederick and Sarah Gaillot are current employees and Samuel C. Haffer is a former employee of the sponsoring agency, the Centers for Medicare & Medicaid Services.

ORCID

Marc N. Elliott http://orcid.org/0000-0002-7147-5535

Amelia M. Haviland http://orcid.org/0000-0003-1068-4031

REFERENCES

1. Filice CE, Joynt KE. Examining race and ethnicity information in medicare administrative data. *Med Care*. 2017;55:e170-e176.
2. Lauderdale DS, Goldberg J. The expanded racial and ethnic codes in the Medicare data files: their completeness of coverage and accuracy. *Am J Public Health*. 1996;86(5):712-716.
3. Waldo DR. Accuracy and bias of race/ethnicity codes in the Medicare enrollment database. *Health Care Financ Rev*. 2004;26(2):61-72.
4. Arday SL, Arday DR, Monroe S, Zhang J. HCFA's racial and ethnic data: current accuracy and recent improvements. *Health Care Financ Rev*. 2000;21(4):107-116.
5. Zaslavsky AM, Ayanian JZ, Zaborski LB. The validity of race and ethnicity in enrollment data for Medicare beneficiaries. *Health Serv Res*. 2012;47(3 Pt 2):1300-1321.
6. Institute of Medicine. *Unequal Treatment: Understanding Racial and Ethnic Disparities in Health Care*. Washington, DC: National Academies Press; 2002.
7. Institute of Medicine. *Race, Ethnicity, and Language Data: Standardization for Health Care Quality Improvement*. Washington, DC: Institute of Medicine; 2009.
8. Elliott MN, Morrison PA, Fremont A, McCaffrey DF, Pantoja P, Lurie N. Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Serv Outcomes Res Methodol*. 2009;9(2):69-83.
9. Adjaye-Gbewonyo D, Bednarczyk RA, Davis RL, Omer SB. Using the Bayesian Improved Surname Geocoding Method (BISG) to create a working classification of race and ethnicity in a diverse

managed care population: a validation study. *Health Serv Res*. 2014;49(1):268-283.

10. Brown DP, Knapp C, Baker K, Kaufmann M. Using Bayesian imputation to assess racial and ethnic disparities in pediatric performance measures. *Health Serv Res*. 2016;51(3):1095-1108.
11. Derose SF, Contreras R, Coleman KJ, Koebnick C, Jacobsen SJ. Race and ethnicity data quality and imputation using U.S. Census Data in an integrated health system. *Med Care Res Rev*. 2012;70(3):330-345.
12. Grundmeier RW, Song L, Ramos MJ, et al. Imputing missing race/ethnicity in pediatric electronic health records: reducing bias with use of U.S. Census location and surname data. *Health Serv Res*. 2015;50(4):946-960.
13. Martino SC, Weinick RM, Kanouse DE, et al. Reporting CAHPS and HEDIS data by race/ethnicity for Medicare beneficiaries. *Health Serv Res*. 2013;48(2 Pt 1):417-434.
14. Elliott MN, Fremont A, Morrison PA, Pantoja P, Lurie N. A new method for estimating race/ethnicity and associated disparities where administrative records lack self-reported race/ethnicity. *Health Serv Res*. 2008;43(5 Pt 1):1722-1736.
15. Centers for Medicare and Medicaid Services. *Racial and Ethnic Disparities in Health Care in Medicare Advantage*. Baltimore, MD: Centers for Medicare and Medicaid Services; 2016.
16. Klein DJ, Elliott MN, Haviland AM, et al. Understanding nonresponse to the 2007 Medicare CAHPS survey. *Gerontologist*. 2011;51(6):843-855.
17. Purcell NJ, Kish L. Postcensal estimates for local areas (or domains). *Int Stat Rev*. 1980;48(1):3-18.
18. Word DL, Coleman CD, Nunziata R, Kominski R. *Demographic Aspects of Surnames from Census 2000*. Washington, DC: U.S. Census Bureau; 2000.
19. Morrison PA, Word DL, Coleman CD. Using First Names to Estimate Racial Proportions in Populations. Population Association of America Annual Meeting; 2001; Washington, DC.
20. Lauderdale DS, Kestenbaum B. Mortality rates of elderly Asian American populations based on Medicare and social security data. *Demography*. 2002;39(3):529-540.
21. Wong EC, Palaniappan LP, Lauderdale DS. Using name lists to infer Asian racial/ethnic subgroups in the Healthcare setting. *Med Care*. 2010;48(6):540-546.
22. Haviland AM, Elliott MN, Weech-Maldonado R, Hambarsoomian K, Orr N, Hays RD. Racial/ethnic disparities in Medicare Part D experiences. *Med Care*. 2012;50(Suppl):S40-S47.
23. Fiscella K, Williams DR. Health disparities based on socioeconomic inequities: implications for urban health care. *Acad Med*. 2004;79(12):1139-1147.
24. Martino SC, Elliott MN, Hambarsoomian K, et al. Racial/ethnic disparities in Medicare beneficiaries' care coordination experiences. *Med Care*. 2016;54(8):765-771.
25. Weech-Maldonado R, Elliott MN, Adams JL, et al. Do racial/ethnic disparities in quality and patient experience within Medicare plans generalize across measures and racial/ethnic groups? *Health Serv Res*. 2015;50(6):1829-1849.
26. Nelder JA, Wedderburn RWM. Generalized linear models. *J R Stat Soc Ser A*. 1972;135(3):370-384.
27. Efron B, Gong G. A leisurely look at the bootstrap, the Jackknife, and cross-validation. *Am Stat*. 1983;37(1):36-48.
28. Office of Management and Budget. *Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity*. Washington, DC: Budget OoMa; 1997.
29. Office of Minority Health. *Report to Congress: Improving Medicare Post-Acute Care Transformation (IMPACT) Act of 2014 Strategic Plan for Accessing Race and Ethnicity Data*. Baltimore, MD: Centers for Medicare & Medicaid Services; 2017.
30. McCaffrey DF, Elliott MN. Power of tests for a dichotomous independent variable measured with error. *Health Serv Res*. 2008;43(3):1085-1101.
31. Bishop YMM, Fienberg SE, Holland PW. *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: M.I.T. Press; 1975.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Haas A, Elliott MN, Dembosky JW, et al. Imputation of race/ethnicity to enable measurement of HEDIS performance by race/ethnicity. *Health Serv Res*. 2019;54:13–23. https://doi.org/10.1111/1475-6773.13099