Health Services and Outcomes Research Methodology (2019) 19:175–195
https://doi.org/10.1007/s10742-019-00200-9

Developing and evaluating methods to impute race/ethnicity in an incomplete dataset

Gabriella C. Silva[1] · Amal N. Trivedi[2] · Roee Gutman[1]

Received: 19 November 2018 / Revised: 10 May 2019 / Accepted: 31 May 2019 / Published online: 8 June 2019

Abstract

The availability of race data is essential for identifying and addressing racial/ethnic disparities in the health care system; however, patient self-reported racial/ethnic information is often missing. Indirect methods for estimating race have been developed, but they usually only consider geocoded and surname data as predictors, may perform poorly among racial minorities, they do not adjust for possible errors for specific datasets, and are unable to provide race estimates for subjects missing some of this information. The objective of this study was to address these limitations by developing novel methods for imputing race/ethnicity when this information is partially missing. By viewing the unobserved race as missing data, we explored different multiple imputation methods for imputing race/ethnicity, and we applied these methods to a subset of Rhode Island Medicaid beneficiaries. Current race imputation methods and newly developed ones were compared using area under the ROC curve statistics and racial composition estimates to identify methods and sets of predictors that yield superior race imputations. Family race was identified as an important predictor and should be included in race estimation models when possible. Bayesian regression models (BRM) provide better race estimates than previously proposed methods. Missing race was multiply imputed using joint modeling and fully conditional specification. Post-imputation analyses showed that fully conditional specification with a BRM is superior to joint modeling for race imputation. The proposed fully conditional specification method is a flexible, effective way of estimating race/ethnicity that allows for propagation of imputation error and ease of interpretation in further analyses.

Keywords Race/ethnicity · Bayesian regression · Multiple imputation · Health care

✉ Gabriella C. Silva
gabriella_silva@brown.edu

[1] Department of Biostatistics, Brown University, 121 South Main Street, Box G-S121-7, Providence, RI 02912, USA

[2] Department of Health Services, Policy and Practice, Brown University, 121 South Main Street, 7th Floor, Providence, RI 02903, USA

1 Introduction

Eliminating health care disparities requires accurate collection of patients' race and ethnicity, but this information is often missing or inaccurate in health care data. Federal policy has therefore prioritized the need to improve the availability of standardized race and ethnicity data. Certain provisions in the Affordable Care Act require all federally funded health programs to collect data on race, ethnicity, primary language, and other demographic characteristics (Ng et al. 2017). The Institute of Medicine (now National Academies of Medicine) has also identified approaches to improve data collection efforts (Ulmer et al. 2009). In addition to federal initiatives, state and local entities have emphasized the need for health care providers to collect race/ethnicity data. For example, starting in January 2007, the state of Massachusetts required all hospitals to report the race/ethnicity of every patient with a visit to the emergency department, an inpatient stay, or an observation unit stay (Ulmer et al. 2009). The Alliance of Chicago Community Health Services established an electronic health record system capable of merging race/ethnicity data with clinical data.

In spite of these ongoing efforts, obtaining self-reported race/ethnicity, the gold standard for racial and ethnic data, has been a challenging and slow endeavor. Patients may infrequently contact the health care system or may experience discomfort with providing race/ethnicity (Adjaye-Gbewonyo et al. 2013). Some providers may mistakenly believe that there are regulatory prohibitions against collecting information on their patients' race/ethnicity. Aetna's Task Force on Racial and Ethnic Disparities in Health Care resulted in the collection of self-reported race/ethnicity for just one-third of their enrollees 4 years after its implementation despite CEO involvement and a considerable allocation of resources towards collecting racial and ethnic information (Hassett 2005).

Reducing racial/ethnic disparities in health care cannot be postponed until self-reported race/ethnicity becomes available for the vast majority of patients. Therefore, some studies have developed indirect measures of race/ethnicity using addresses and surnames to indirectly estimate race/ethnicity. Geocoding involves using individuals' addresses to derive the racial/ethnic composition of their neighborhood. If a considerable portion of an individual's census block group or tract is comprised by members of a particular race/ethnicity, then it increases the probability that this person belongs to that racial/ethnic group as well. Similarly, if census data shows that a high fraction of people sharing a certain surname belong to a particular racial/ethnic group, then the likelihood of someone with this surname also being a member of this racial/ethnic group is high.

Geocoding only (GO) and surname only (SO) estimate the probability of belonging to different racial groups based solely on addresses and surnames, respectively; however, using geocoding and surnames in combination to estimate race/ethnicity is superior (Adjaye-Gbewonyo et al. 2013; Elliott et al. 2008; Fiscella and Fremont 2006). A commonly used method that combines geocoded and surname information is the Bayesian Improved Surname Geocoding (BISG) method (Elliott et al. 2008, 2009). BISG relies on Bayes Theorem and uses geocoded probabilities to update the surname probabilities of belonging to a specific race category.

Although BISG has been shown to outperform existing alternatives, it suffers from a few limitations. The method does not perform well for racial groups that comprise a small portion of the US population, like American Indians (Adjaye-Gbewonyo et al. 2013). For these racial groups, incorporating other possible predictors for race/ethnicity, like the race/ethnicity of other family members or primary household language, could improve accuracy.

Additionally, BISG fails to produce updated race probabilities in cases where geocoded information is unavailable or an individual's surname is uncommon. Prior studies have usually excluded these individuals (Adjaye-Gbewonyo et al. 2013; Consumer Financial Protection Bureau 2014). Another limitation is that the method only produces probabilities of being a member of specific racial/ethnic groups. When using these probabilities in models, their conditional and marginal coefficients cannot be interpreted as the coefficients estimated with categorical race. Lastly, these methods assume the derived probabilities are exact and fail to account for the variability in estimation and across different datasets. Thus, incorporating these probabilities in race-based estimands may underestimate their sampling variability. Ma et al.'s study (2018) explores imputing missing race data using multiple imputation methods; however, the dataset used for their study is de-identified, and so patient-level geocoded and surname information could not be incorporated as predictors of race. ZIP code-level information for racial distribution, income, education, and poverty was used instead. For low-income populations like Medicaid beneficiaries, using ZIP code-level data may not be sufficient to obtain accurate race/ethnicity imputation.

To address these limitations, we view the unobserved race data, along with the missing values for the covariates that can be used to estimate race, as a missing data problem (Rubin 1976; Ma et al. 2018). Using Medicaid enrollment data from a Northeastern state with a substantial fraction of missing race/ethnicity, the objective of this study was to use missing data methods to impute race/ethnicity and compare our approach to imputation methods like BISG that use geocoded and surname probabilities as variables in the imputation model. We also explore incorporating two novel predictors of race/ethnicity into race estimation models: family race/ethnicity and primary household language. Using simulations, we show that multiple imputation methods that include geocoding probabilities, surname probabilities and family race/ethnicity provide the most accurate prediction of race. In addition, these methods are statistically valid and allow for propagation of error in future analysis.

2 Methods

2.1 Data and study population

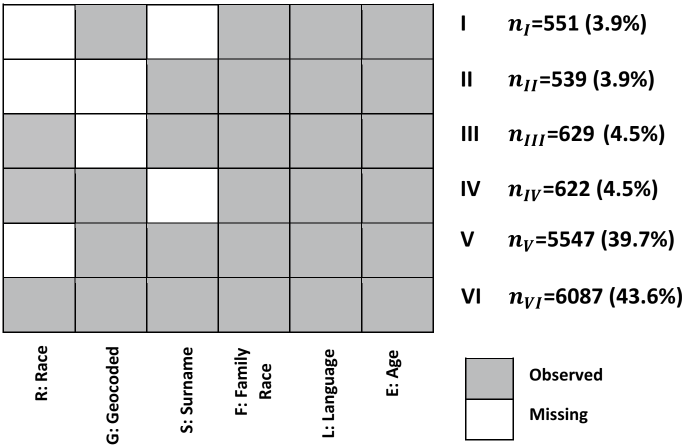
Our dataset is composed of 14,083 Rhode Island Medicaid beneficiaries enrolled in Blue Cross Blue Shield Rhode Island prior to its withdrawal from the Rhode Island Medicaid managed care market. The individuals in our dataset were mostly from low-income backgrounds and included pregnant women, individuals with disabilities, parents and children. The variables in this dataset were derived using Medicaid managed care records and fee-for-service claims between January 1, 2009 and December 31, 2016. In our analysis, we concentrated on the following variables: first and last name, self-reported race/ethnicity, address, age in 2016, primary language, and family ID.

The geocoded probabilities were derived by linking the individual's address to Census block geocoding from the 2011 American Community Survey, and surname probabilities were derived from the Census Bureau's frequently occurring surname list at the 2010 Census. Geocoded probabilities were missing among individuals whose addresses could not be linked to a particular census block group. Missing surname probabilities occurred whenever an individual had a surname reported less than 100 times in the 2010 US Census. Some fields in the 2010 US Census surname list had a reported "(S)", indicating that the percentage of individuals in a particular racial category with that surname was small and it had been excluded for confidentiality reasons. To address this, a small percentage (0.08%) was assigned to these categories and the set of probabilities for this surname was normalized to ensure that it summed to 1. Prior to linking the surnames in the Rhode Island (RI) dataset to the US Census surname list, the last names in the RI dataset were standardized using the recommendations detailed by Word et al. (2008). Individuals missing both surname and geocoded probabilities represented less than 0.8% of the study population; because of the limited information available on them, they were excluded from the analysis. The final dataset contained 13,975 observations.

The race value for each individual, R_i, was recorded as one of five possible racial/ethnic categories: White, Black, American Indian (AI), Hispanic, and Asian Pacific Islanders (API). Race values were missing for 47.5% of individuals. The racial composition for the 13,975 individuals, among those with self-reported race/ethnicity, was 80.6% White, 9.9% Black, 0.7% AI, 6.3% Hispanic, and 2.4% API. These proportions were within 0.1% of the racial composition that was observed for individuals with self-reported race/ethnicity in our original dataset of 14,083 beneficiaries.

For each individual i, we derived six geocoded probabilities ($\mathbf{G}_\mathrm{i}$) and six surname probabilities ($\mathbf{S}_\mathrm{i}$) corresponding to the six racial groups considered in the US Census data: White, Black, AI, Hispanic, API, and Other. In addition, we recorded each individual's age (E_i) and whether or not the primary language spoken in the household was English (L_i). Using subject i's family ID, we identified other subjects who shared the same family ID. We then noted whether or not these subjects had self-reported race/ethnicity and, if so, which of the five racial groups they identified with. Using this information, we defined five family race indicators for subject i corresponding to each of the five racial groups ($\mathbf{F}_\mathrm{i}$). The value of each indicator was set to 1 if any family member self-identified as belonging to one of the five racial groups, and 0 otherwise.

Figure 1 displays the pattern of missing values for the different variables among the 13,975 observations. Most of the observations with missing race (patterns in rows I, II and V) comprise complete geocoded and surname probabilities but unobserved race (row V).



R: Race	G: Geocoded	S: Surname	F: Family Race	L: Language	E: Age		
□	▨	□	▨	▨	▨	I	n_I=551 (3.9%)
□	□	▨	▨	▨	▨	II	n_{II}=539 (3.9%)
▨	□	▨	▨	▨	▨	III	n_{III}=629 (4.5%)
▨	▨	□	▨	▨	▨	IV	n_{IV}=622 (4.5%)
□	▨	▨	▨	▨	▨	V	n_V=5547 (39.7%)
▨	▨	▨	▨	▨	▨	VI	n_{VI}=6087 (43.6%)

▨ Observed
□ Missing

Fig. 1 Visualization of missingness pattern for the 13,975 beneficiaries in the RI dataset. Every row represents a different combination of variables with missing values and includes information on the number and percentage of individuals with this pattern

Although estimating race is the direct interest of the study, rows I and II of Fig. 1 show that missing entries for geocoded and surname probabilities should also be addressed in order to obtain racial estimates for all individuals who did not report race. Cases where race is observed but either geocoded or surname probabilities are missing, specifically the pattern displayed in rows III and IV, are included in the analysis because they may still help inform race/ethnicity for those with missing race/ethnicity values.

2.2 Methods for imputing the missing values

Multiple imputation (MI) is a statistical method developed to handle incomplete datasets when the data is assumed to be missing at random (MAR) (Little and Rubin 2002). MI generates *m* complete datasets in which missing entries in the data are replaced with plausible values (van Buuren 2007). Analysis of each of these complete datasets is performed separately and a final estimate is derived using common combination rules (Rubin 1987). Two general techniques have been proposed to impute multivariate data: joint modeling (JM) and fully conditional specification (FCS). Figure 2 displays the steps to multiply

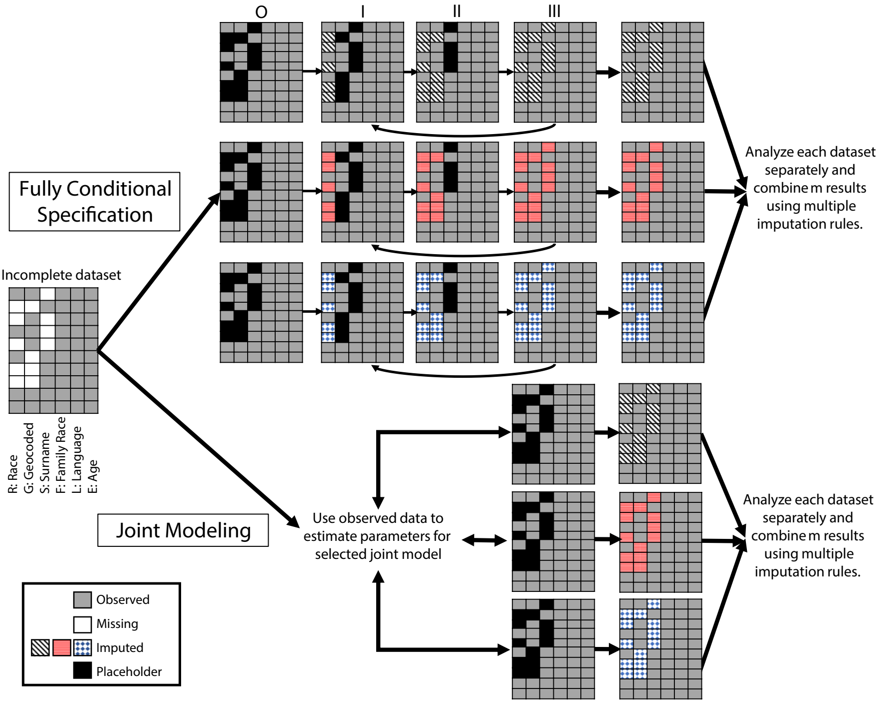

O
I
II
III
Fully Conditional Specification
Incomplete dataset
R: Race
G: Geocoded
S: Surname
F: Family Race
L: Language
E: Age
Analyze each dataset separately and combine m results using multiple imputation rules.
Joint Modeling
Use observed data to estimate parameters for selected joint model
Analyze each dataset separately and combine m results using multiple imputation rules.
Observed
Missing
Imputed
Placeholder

Fig. 2 Flowchart for Multiple Imputation (MI) with m = 3. Fully Conditional Specification (FCS) or Joint Modeling (JM) can be used to generate m imputations. In FCS, we fill in the missing data with starting values (O). Then we estimate the parameters for the multinomial logistic regression model using the complete data and use these to impute race (I). We repeat this for geocoded and surname probabilities using appropriate distributions (II, III). We repeat steps I–III until convergence is reached and consider the final draws as our imputed dataset. JM-GLM (JM-MVN) alternates between estimating the parameters for the general location model (multivariate normal distribution) and using these to impute the dataset. After 100 iterations, we consider these draws as the imputed dataset. For both MI methods, analyses of interest are completed in each dataset and combined using MI rules

impute and analyze an incomplete dataset of ten individuals with $m = 3$ using either FCS or JM. Only two of these individuals have fully observed values for all variables; the other individuals are missing race, geocoded and/or surname information in combinations that fall into one of the patterns described in Fig. 1. We implement JM and FCS on the dataset of Rhode Island Medicaid beneficiaries.

2.2.1 Joint modeling (JM)

JM is a technique for multiply imputing multivariate data that relies on the specification of a parametric multivariate density for the incomplete dataset given a set of model parameters (van Buuren 2007).

The general location model was considered as a possible model to impute an incomplete multivariate dataset like the RI dataset with both continuous and categorical variables (Schafer 2000). Let the data matrix be denoted as $(\mathbf{W}, \mathbf{Z})$, where $\mathbf{W} = (\mathbf{W_1}, \mathbf{W_2}, \ldots, \mathbf{W_P})$ represents the set of P categorical variables in the dataset and $\mathbf{Z} = (\mathbf{Z_1}, \mathbf{Z_2}, \ldots, \mathbf{Z_Q})$ denotes the set of Q continuous ones. In a dataset with n individuals, $\mathbf{W_p} = \{\mathrm{W}_{1p}, W_{2p}, \ldots, W_{np}\}$ while $\mathbf{Z_q} = \{\mathrm{Z}_{1\mathrm{q}}, \mathrm{Z}_{2\mathrm{q}}, \ldots, \mathrm{Z}_{\mathrm{nq}}\}$. The general location model is defined by the marginal distribution of $\mathbf{W}$, which is a multinomial distribution on the cell counts describing the response pattern for the categorical variables, and the conditional distribution of $\mathbf{Z|W}$ which is multivariate normal (Schafer 2000). Here,$\mathbf{W} = (\mathbf{R}, \mathbf{F}, \mathbf{L})$ and $\mathbf{Z} = (\mathbf{G'}, \mathbf{S'}, \mathbf{E})$, where $\mathbf{G'}$ and $\mathbf{S'}$ are the logit-transformed geocoded and surname probabilities. Because the general location model specifies a joint distribution on the whole dataset we can derive the conditional distribution $P(\mathbf{R_i}|\mathbf{F_i}, \mathbf{L_i}, \mathbf{Z_i}, \boldsymbol{\theta}_\mathbf{W}, \boldsymbol{\theta}_\mathbf{Z|W})$, where $\boldsymbol{\theta}_\mathbf{W}$ and $\boldsymbol{\theta}_\mathbf{Z|W}$ are the parameters governing the distribution of $\mathbf{W}$ and $\mathbf{Z|W}$ in the general location model, respectively.

We used Markov chain Monte Carlo techniques (MCMC) to estimate the parameters of the general location model and generate plausible values for missing race. This technique alternates between randomly imputing the missing data under the current parameters' values, and using the completed data to draw a new vector parameter value from the posterior distribution. After a sufficient amount of MCMC iterations, the drawn parameter vector value can be viewed as a random draw from the posterior distribution. Using the drawn vector of parameter values, imputations of the missing data can be generated. In cases where certain combinations of the categorical variables are rare, the restricted general location model was implemented. This version involves specifying restrictions on the parameter sets $\boldsymbol{\theta}_\mathbf{W}$ and $\boldsymbol{\theta}_\mathbf{Z|W}$ (Schafer 2000). JM with the general location model (JM-GLM) was implemented using the R package *mix* (Schafer 2017).

Similar to Ma et al. (2018) in their simulation analysis of the State Inpatient Database, we also considered JM with the multivariate normal distribution, referred to as JM-MVN throughout. Under this method, discrete variables are treated as continuous; thus, in order to impute a categorical variable like race, we create indicators for the different levels of the variables, estimate unnormalized probabilities for each indicator using a multivariate normal distribution, normalize these probabilities, and use the normalized probabilities to sample from a multinomial distribution (Honaker et al. 2018). The R package *Amelia* offers a quick, efficient implementation of JM with the multivariate normal distribution.

2.2.2 Fully conditional specification (FCS)

FCS is a flexible approach for imputing multivariate data that does not require the specification of a joint distribution of all of the variables. Instead, it specifies the multivariate distribution through a set of conditional densities (van Buuren 2007). FCS involves iterating between each variable with missing data and imputing its missing values, given the current values for the other variables.

Formally, let $\mathbf{Y_j}$ denote a variable in the incomplete dataset with missing values. We denote the observed and missing parts of $\mathbf{Y_j}$ as $\mathbf{Y_j^{obs}}$ and $\mathbf{Y_j^{miss}}$, respectively. Here, $\mathbf{Y} = (\mathbf{R}, \mathbf{G'}, \mathbf{S'})$, $\mathbf{Y^{obs}} = (\mathbf{R^{obs}}, \mathbf{G'^{obs}}, \mathbf{S'^{obs}})$ and $\mathbf{Y^{miss}} = (\mathbf{R^{miss}}, \mathbf{G'^{miss}}, \mathbf{S'^{miss}})$. Let $M_{ij} = 1$ if the ith record in $\mathbf{Y_j}$ is observed and 0 otherwise and $\mathbf{M_j} = \{M_{ij}\}$. Then $\mathbf{M} = (\mathbf{M_R}, \mathbf{M_{G'}}, \mathbf{M_{S'}})$ for our dataset. Lastly, let $\mathbf{A}$ denote the set of fully observed variables, which in our analysis comprise a combination of the family indicators, language indicator, and patient's age. FCS specifies the conditional distribution of $P(\mathbf{Y_j}|\mathbf{Y_{-j}}, \mathbf{M}, \mathbf{A}, \boldsymbol{\theta}_\mathbf{j})$ for every $\mathbf{Y_j}$, where $\mathbf{Y_{-j}}$ represents the set $\mathbf{Y}$ without variable $\mathbf{Y_j}$, and $\boldsymbol{\theta}_\mathbf{j}$ are vectors parameter governing this distribution. Imputations are created by iteratively drawing from these conditional distributions (van Buuren 2007).

We defined three conditional models: $P(\mathbf{R}|\mathbf{Y_{-R}}, \mathbf{M}, \mathbf{A}, \boldsymbol{\theta}_\mathbf{R})$, $P(\mathbf{G'}|\mathbf{Y_{-G'}}, \mathbf{M}, \mathbf{A}, \boldsymbol{\theta}_\mathbf{G'})$ and $P(\mathbf{S'}|\mathbf{Y_{-S'}}, \mathbf{M}, \mathbf{A}, \boldsymbol{\theta}_\mathbf{S'})$. The FCS algorithm iterates between drawing the parameters of these models and imputing the missing values (Liu and De 2015). Pseudocode of this algorithm is provided in "Appendix 1".

In FCS, we modeled $P(\mathbf{R}|\mathbf{Y_{-R}}, \mathbf{M}, \mathbf{A}, \boldsymbol{\theta}_\mathbf{R})$ using Bayesian multinomial logistic regression models (BRM) that conditioned on different sets of covariates. The multinomial logistic regression model assumes that the probability of individual i self-identifying with race r is $\frac{exp(\mathbf{x_i^T}\boldsymbol{\beta}_\mathbf{r})}{\sum_{r' \in Y}(exp(\mathbf{x_i^T}\boldsymbol{\beta}_\mathbf{r'}))}$ where $\mathbf{x_i}$ is a vector containing an intercept term and p covariates for individual i, $\boldsymbol{\beta}_\mathbf{r}$ represents the corresponding vector of $p+1$ parameters and $Y = \{White, Black, AI, Hispanic, API\}$ (Kruschke 2011). To complete the Bayesian model we assumed that the prior distributions for all $\boldsymbol{\beta}_\mathbf{r}$'s were independent Normal distributions with zero mean and variance 10^4. Sampling of $\boldsymbol{\beta}_\mathbf{r}$ for each $r \in Y$ was performed using the Polya–Gamma Latent Variable approach, which is a quick and efficient way for sampling from the posterior distribution of a multinomial logistic regression model (Polson et al. 2013a). This approach is implemented in the R package *BayesLogit* (Polson et al. 2013b). In our implementation, we used a burn-in period of 10,000 iterations and then thinned a sample of 2000 $\boldsymbol{\beta}_\mathbf{r}$'s for each $r \in Y$ to obtain 200 final samples from the posterior distribution of each $\boldsymbol{\beta}_\mathbf{r}$. Similar multinomial logistic models are implemented in the R package *mice* (van Buuren and Groothuis-Oudshoorn 2018) and the results were similar to our implementation.

All of the multinomial logistic regression models included geocoded and surname probabilities, and we supplemented these models with different combinations and interactions of the variables: family race indicators, whether the primary household language was English, patient's age, and BISG probabilities. Table 1 includes the full list of the BRMs considered.

To impute the missing geocoded and surname probabilities, the conditional densities $P(\mathbf{G'}|\mathbf{Y_{-G'}}, \mathbf{M}, \mathbf{A}, \boldsymbol{\theta}_\mathbf{G'})$ and $P(\mathbf{S'}|\mathbf{Y_{-S'}}, \mathbf{M}, \mathbf{A}, \boldsymbol{\theta}_\mathbf{S'})$ must also be specified. Although there were six geocoded probabilities and six surname probabilities, only five of each were used because they sum to 1 when they are untransformed. Hence, each of these conditional densities will be modeled using a 5-variate normal regression model. Specifically, $P(\mathbf{G'}|\mathbf{Y_{-G'}}, \mathbf{M}, \mathbf{A}, \boldsymbol{\theta}_\mathbf{G'}) \sim N_5([Y_{-G'}, A]\beta_{G'}, \Sigma_{G'})$ and $P(\mathbf{S'}|\mathbf{Y_{-S'}}, \mathbf{M}, \mathbf{A}, \boldsymbol{\theta}_\mathbf{S'}) \sim N_5([Y_{-S'}, A]\beta_{S'}, \Sigma_{S'})$, where $\boldsymbol{\theta}_\mathbf{G'} = [\beta_{G'}, \Sigma_{G'}]$ and $\boldsymbol{\theta}_\mathbf{S'} = [\beta_{S'}, \Sigma_{S'}]$.

Table 1 Description of BRMs considered

Model name	Description of predictors included
$\boldsymbol{G'}$	G'_i for all $i \in I = \{Black, AI, Hispanic, API, Other\}$
$\boldsymbol{S'}$	S'_i for all $i \in I = \{Black, AI, Hispanic, API, Other\}$
$\boldsymbol{F}$	F_r for all $r \in Y = \{White, Black, AI, Hispanic, API\}$
BISG	$\frac{G_i S_i}{\sum_{r \in I''} G_r S_r}$ for all $i \in I$ where $I'' = \{White, Black, AI, Hispanic, API, Other\}$
$\boldsymbol{G'}+\boldsymbol{S'}$	G'_i and S'_i for all $i \in I = \{Black, AI, Hispanic, API, Other\}$
$(\boldsymbol{G'}+\boldsymbol{S'})^2$	$\boldsymbol{G'}$, $\boldsymbol{S'}$, and all second-order interactions
$\boldsymbol{G'}+\boldsymbol{S'}+L$	$\boldsymbol{G'}$, $\boldsymbol{S'}$, and L
$\boldsymbol{G'}+\boldsymbol{S'}+\mathrm{E}$	$\boldsymbol{G'}$, $\boldsymbol{S'}$, and E
$\boldsymbol{G'}+\boldsymbol{S'}+$**BISG**	$\boldsymbol{G'}$, $\boldsymbol{S'}$, and **BISG**
$\boldsymbol{G'}+\boldsymbol{S'}+\boldsymbol{F}$	$\boldsymbol{G'}$, $\boldsymbol{S'}$, and $\mathbf{F}$
$\boldsymbol{G'}+\boldsymbol{S'}+\boldsymbol{F}+L$	$\boldsymbol{G'}$, $\boldsymbol{S'}$, $\mathbf{F}$, and L
$\boldsymbol{G'}+\boldsymbol{S'}+\boldsymbol{F}+E$	$\boldsymbol{G'}$, $\boldsymbol{S'}$, $\mathbf{F}$, and E
$\boldsymbol{G'}+\boldsymbol{S'}+\boldsymbol{F}+$**BISG**	$\boldsymbol{G'}$, $\boldsymbol{S'}$, $\mathbf{F}$, and **BISG**
$(\boldsymbol{G'}+\boldsymbol{S'}+\boldsymbol{F})^2$	$\boldsymbol{G'}$, $\boldsymbol{S'}$, $\mathbf{F}$, $\sum_{i \in I'} G'_i S'_i$, $\sum_{i \in I'} G'_i F_i$ and $\sum_{i \in I'} S'_i F_i$ where $I' = \{Black, AI, Hispanic, API\}$
$(\boldsymbol{G'}+\boldsymbol{S'}+\boldsymbol{F})^2+\mathrm{L}$	$(\boldsymbol{G'}+\boldsymbol{S'}+\boldsymbol{F})^2$ and L

$\boldsymbol{G'}$ geocoded, $\boldsymbol{S'}$ surname, L language, $\boldsymbol{F}$ family race, E age

2.3 Comparing the methods

2.3.1 Evaluation of prediction models

To determine which of the novel race predictors are useful for predicting race, as well as the proper specification for $P(\mathbf{R}|\mathbf{Y}_{-\mathbf{R}}, \mathbf{M}, \mathbf{A}, \boldsymbol{\theta}_{\mathbf{R}})$ in FCS, the fully observed data that comprised 6087 beneficiaries (row VI of Fig. 1), was split into a training set and a testing set. Because race was missing for 47.5% of Rhode Island Medicaid beneficiaries in the entire dataset, the testing set comprised 47.5% of the 6087 beneficiaries. The training set included n_f individuals and was used to estimate the $\boldsymbol{\beta}_{\mathbf{r}}$'s needed for the BRMs. The testing set included n_e individuals whose race was set to missing and whose predicted probabilities of belonging to each race r were estimated using each of the methods.

In order to examine the sensitivity of the different methods to estimating the racial composition, we selected the training and testing sets under different missing data mechanisms: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) (Little and Rubin 2002). MCAR involved randomly selecting the n_e individuals whose race was set to missing among the $n = 6087$ beneficiaries. MAR and MNAR selected the n_e individuals for the testing set based on the estimated probabilities of having missing race/ethnicity for each of the n subjects. These probabilities were estimated for each subject by constructing a logistic regression model with a missing race/ethnicity indicator as the outcome and the observed covariates as predictors (geocoded and surname probabilities, language, family race indicators, age). Individuals in rows V and VI (Fig. 1) of the data with fully observed covariate information were used to construct this model: those in row V are individuals with missing race/ethnicity while those in rows VI had observed race. In the case of MAR, the probabilities of having missing race/ethnicity depended only on the coefficients of this logistic regression model; for MNAR the probabilities of belonging to the test set depended on both the coefficients of the model and self-reported race. Further details for this splitting procedure, as well as the coefficients used to estimate the probability of having missing race/ethnicity, are provided in "Appendix 2".

To compare the performance of models with different variable configurations, we calculated two statistics among the n_e subjects whose race was set to missing. Let $\mathbf{p_i}=(p_{iWhite},p_{iBlack},p_{iAI},p_{iHispanic},p_{iAPI})$ represent the predicted probabilities of each race for beneficiary i.

The first statistic that we examined is the area under the ROC curve (AUC). This statistic is commonly used in classification problems to compare the predictive performance of models (Fawcett 2006; Hosmer et al. 2013). AUC is the probability that a randomly chosen individual who self-identified as belonging to race $r \in \Upsilon$ would have a higher predicted probability to identify as belonging to race r under this model than a randomly chosen individual who did not identify as belonging to race r. The AUC statistic is derived for each $r \in \Upsilon$ by comparing the vector of predicted probabilities of belonging to race r for the n_e subjects, $\mathbf{p}^r=(p_{1r},p_{2r},\dots,p_{n_e r})$ to an indicator vector describing whether or not each individual i self-reported belonging to race r. Because a Bayesian approach provides a full distribution for the parameters, the distribution of AUCs were computed for the BRM-based models.

The AUC summarized the ability of different models to discriminate between races. However, discrimination alone is insufficient to assess a model's accuracy (Hosmer and Lemeshow 2000). Thus, when assessing the fit of a model it is important to examine the calibration of the model as well. The second set of performance statistics were the racial composition estimates. We estimated the prevalence for each racial group by averaging the expected probabilities across the n_e individuals for each $r \in \Upsilon$. A distribution of these was computed for BRM-based models. We compared these to the reported racial composition of the n_e subjects (Elliott et al. 2008). If the estimated prevalence is close to the reported one regardless of the missing data mechanism, this indicates that the method is robust to different missingness mechanisms.

We report the observed racial compositions for the n_f individuals in the training set in Tables 3 and 4. In the case of MCAR the rates among the n_f training subjects was similar to the observed racial composition in the testing set; when race was MAR and MNAR, these varied from that rates observed among the n_e individuals.

2.3.2 Evaluation of multiple imputation models

The performance of JM and FCS was assessed by imposing the missingness pattern illustrated in Fig. 1 on the subset of fully observed data; that is, approximately 47.5% of races, 8.4% of geocoded probabilities and 8.4% of surname probabilities were set to missing for the 6087 individuals in row VI of Fig. 1. To evaluate the multiple imputation procedure, the specification for $P(\mathbf{R}|\mathbf{Y_{-R}},\mathbf{M},\mathbf{A},\boldsymbol{\theta}_\mathbf{R})$ in the FCS algorithm was selected based on the model with the best AUC's and the closest racial composition statistics. The multivariate normal distributions were used for $P(\mathbf{G'}|\mathbf{G'^{obs}},\mathbf{Y_{-G'}},\mathbf{M},\mathbf{A},\boldsymbol{\theta}_\mathbf{G'})$ and $P(\mathbf{S'}|\mathbf{S'^{obs}},\mathbf{Y_{-S'}},\mathbf{M},\mathbf{A},\boldsymbol{\theta}_\mathbf{S'})$. Upon implementation, all multiple imputation methods generated $m = 25$ completed datasets. Methods were evaluated using the subset of n_e individuals whose race had been set to missing. This subset varied depending on whether race was MCAR, MAR, or MNAR. The selection of these n_e individuals under different missing data mechanisms was identical to that described in Sect. 2.3.1.

Point estimates for the prevalence of each race $r \in \Upsilon$, $\hat{c}_r$, were estimated by averaging the racial composition estimates for the 25 datasets, $\hat{c}_r = \frac{\sum_{k=1}^{25} \hat{c}_{rk}}{25}$, where $\hat{c}_{rk}$ is the prevalence of race $r \in \Upsilon$ in imputed dataset k (Rubin 1987). These were compared to the observed racial composition among the n_e subjects to determine the multiple imputation approach that best captured the observed prevalence. Additionally, variance estimates for racial composition were derived by combining the estimate's average within-imputation variance and the between-imputation variance (Rubin 1987). Formally, the sampling variance of $\hat{c}_r$ for each $r \in \Upsilon$ were estimated by: $Var(\hat{c}_r) = \frac{\sum_{k=1}^{25} \hat{c}_{rk}(1-\hat{c}_{rk})}{25n_e} + (1 + \frac{1}{25})\frac{\sum_{k=1}^{25} (\hat{c}_{rk}-\hat{c}_r)^2}{24}$.

Sensitivity is the percentage of individuals who belong to race r that are correctly identified by the model as belonging to race r. Higher sensitivity across all races indicates superior model performance. The sensitivity of race $r \in \Upsilon$ in imputed dataset k, $\hat{d}_{rk}$, was computed for each dataset and each race by counting the number of times in which estimated race matched the self-reported race and dividing by the total number of individuals among the n_e individuals who reported belonging to that race (n_r). Point estimates for the sensitivity of each race, $\hat{d}_r$, were estimated by averaging the sensitivity of each dataset, $\hat{d}_r = \frac{\sum_{k=1}^{25} \hat{d}_{rk}}{25}$. The sampling variance of this pooled estimate was estimated by: $Var(\hat{d}_r) = \frac{\sum_{k=1}^{25} \hat{d}_{rk}(1-\hat{d}_{rk})}{25n_r} + (1 + \frac{1}{25})\frac{\sum_{k=1}^{25} (\hat{d}_{rk}-\hat{d}_r)^2}{24}$.

Another post-imputation analysis of interest involved computing average ages within each race group. A criterion for determining how well each multiple imputation method performed compared how close within-race average age estimates using imputed race were to within-race average age with the known race.

3 Results

3.1 Evaluation of prediction models

Table 2 presents the AUC values for the five racial categories when the different race estimation methods were applied to the test set where race was assumed to be MCAR. The BRMs that include both geocoded and surname probabilities achieve higher AUC values and are superior to methods that estimate race/ethnicity using these predictors separately (GO, SO, BRM: $\boldsymbol{G'}$, BRM: $\boldsymbol{S'}$). In particular, models including family race/ethnicity, in addition to $\boldsymbol{G'}$ and $\boldsymbol{S'}$, achieve substantially higher AUCs than models excluding family race/ethnicity. Including interaction terms between the geocoded data, surname information, and family race indicators slightly improves the AUC as well. Language, patient's age, and BISG probabilities do not appear to be important predictors of race after accounting for the geocoded and surname probabilities, as their inclusion in the model does not substantially improve model performance.

Table 3 illustrates the reported racial composition among the n_e individuals compared to the prevalence estimates obtained under the various missing data mechanisms.

Table 2 Comparison of the AUC values between previous methods and newly proposed methods

	White	Black	AI	Hispanic	API
Geocoding only (GO)	0.806	0.819	0.570	0.807	0.765
Surname only (SO)	0.825	0.812	0.638	0.867	0.899
BISG	0.866	0.886	0.620	0.896	0.863
BRM					
$\boldsymbol{G'}$	0.802 (0.797, 0.806)[a]	0.801 (0.794, 0.808)	0.514 (0.440, 0.582)	0.803 (0.798, 0.807)	0.783 (0.764, 0.789)
$\boldsymbol{S'}$	0.880 (0.878, 0.882)	0.872 (0.866, 0.876)	0.538 (0.470, 0.598)	0.910 (0.904, 0.913)	0.881 (0.870, 0.894)
$\boldsymbol{F}$	0.898 (0.891, 0.902)	0.941 (0.938, 0.946)	0.911 (0.801, 0.939)	0.873 (0.847, 0.884)	0.886 (0.866, 0.897)
BISG	0.878 (0.875, 0.884)	0.861 (0.847, 0.887)	0.556 (0.446, 0.621)	0.917 (0.911, 0.918)	0.858 (0.846, 0.874)
$\boldsymbol{G'}+\boldsymbol{S'}$	0.881 (0.877, 0.884)	0.877 (0.869, 0.883)	0.538 (0.442, 0.630)	0.909 (0.903, 0.914)	0.906 (0.888, 0.927)
$(\boldsymbol{G'}+\boldsymbol{S'})^2$	0.877 (0.867, 0.885)	0.877 (0.863, 0.888)	0.662 (0.553, 0.744)	0.909 (0.897, 0.919)	0.914 (0.893, 0.933)
$\boldsymbol{G'}+\boldsymbol{S'}+L$	0.885 (0.881, 0.889)	0.886 (0.880, 0.893)	0.579 (0.487, 0.638)	0.901 (0.894, 0.903)	0.929 (0.913, 0.940)
$\boldsymbol{G'}+\boldsymbol{S'}+E$	0.881 (0.877, 0.884)	0.880 (0.872, 0.886)	0.558 (0.439, 0.640)	0.914 (0.909, 0.917)	0.906 (0.889, 0.927)
$\boldsymbol{G'}+\boldsymbol{S'}+$**BISG**	0.883 (0.876, 0.887)	0.882 (0.870, 0.890)	0.543 (0.467, 0.611)	0.902 (0.892, 0.912)	0.923 (0.908, 0.933)
$\boldsymbol{G'}+\boldsymbol{S'}+\boldsymbol{F}$	0.948 (0.946, 0.950)	0.963 (0.959, 0.966)	0.883 (0.813, 0.950)	0.947 (0.944, 0.950)	0.933 (0.917, 0.946)
$\boldsymbol{G'}+\boldsymbol{S'}+\boldsymbol{F}+L$	0.958 (0.954, 0.960)	0.956 (0.953, 0.960)	0.913 (0.856, 0.963)	0.957 (0.953, 0.961)	0.967 (0.951, 0.977)
$\boldsymbol{G'}+\boldsymbol{S'}+\boldsymbol{F}+E$	0.958 (0.954, 0.960)	0.956 (0.952, 0.959)	0.909 (0.846, 0.958)	0.958 (0.955, 0.962)	0.969 (0.962, 0.980)
$\boldsymbol{G'}+\boldsymbol{S'}+\boldsymbol{F}+$BISG	0.957 (0.953, 0.959)	0.948 (0.942, 0.957)	0.883 (0.839, 0.912)	0.958 (0.953, 0.961)	0.959 (0.948, 0.970)
$(\boldsymbol{G'}+\boldsymbol{S'}+\boldsymbol{F})^2$	0.957 (0.953, 0.960)	0.954 (0.948, 0.960)	0.889 (0.843, 0.942)	0.952 (0.947, 0.958)	0.968 (0.958, 0.978)
$(\boldsymbol{G'}+\boldsymbol{S'}+\boldsymbol{F})^2+\mathrm{L}$	0.957 (0.954, 0.960)	0.955 (0.950 0.959)	0.893 (0.844, 0.950)	0.952 (0.946, 0.960)	0.968 (0.947, 0.979)

Race is MCAR

[a] Median (95% CB) reported for BRMs

Regardless of the type of mechanism, BISG overestimates the percentage of Whites and Hispanics in the dataset and underestimates the proportion of Blacks, AI and APIs. The median racial composition estimates in the three BRMs that were examined are very close to the reported ones when race is MCAR. For MCAR, the 95% confidence bounds (CB) for the BRMs that include family race cover the observed statistic across all races. For MAR and MNAR, the 95% CB for the BRM: $(\boldsymbol{G'}+\boldsymbol{S'}+\boldsymbol{F})^2$ that includes family race and interaction terms are all relatively close to the observed value and, in nearly all cases, cover the reported value. Irrespective of the missing data mechanism and set

Table 3 Comparison of racial composition estimates to reported racial composition estimates under various missing data mechanisms

	White	Black	AI	Hispanic	API
MCAR					
Observed (training)	0.806	0.095	0.010	0.068	0.021
Observed[a]	0.819	0.088	0.005	0.062	0.025
Geocoding only (GO)[b]	0.795	0.054	0.005	0.117	0.028
Surname only (SO)	0.725	0.088	0.006	0.153	0.029
BISG	0.859	0.023	0.000	0.106	0.012
BRM: $\boldsymbol{G'}+\boldsymbol{S'}$	0.812 (0.804, 0.822)[c]	0.090 (0.082, 0.099)	0.010 (0.007, 0.014)	0.066 (0.059, 0.072)	0.022 (0.019, 0.025)
BRM: $\boldsymbol{G'}+\boldsymbol{S'}+\boldsymbol{F}$	0.821 (0.813, 0.829)	0.085 (0.079, 0.092)	0.007 (0.005, 0.009)	0.065 (0.059, 0.070)	0.022 (0.020, 0.025)
BRM: $(\boldsymbol{G'}+\boldsymbol{S'}+\boldsymbol{F})^2$	0.821 (0.813, 0.828)	0.086 (0.082, 0.092)	0.007 (0.005, 0.010)	0.061 (0.057, 0.066)	0.024 (0.022, 0.027)
MAR					
Observed (training)	0.859	0.080	0.009	0.034	0.017
Observed	0.761	0.104	0.006	0.099	0.030
Geocoding only (GO)	0.765	0.061	0.006	0.137	0.030
Surname only (SO)	0.669	0.088	0.006	0.198	0.039
BISG	0.807	0.024	0.000	0.151	0.018
BRM: $\boldsymbol{G'}+\boldsymbol{S'}$	0.799 (0.786, 0.809)	0.093 (0.084, 0.103)	0.009 (0.006, 0.012)	0.068 (0.057, 0.078)	0.032 (0.028, 0.037)
BRM: $\boldsymbol{G'}+\boldsymbol{S'}+\boldsymbol{F}$	0.782 (0.766, 0.799)	0.090 (0.081, 0.103)	0.003 (0.002, 0.005)	0.095 (0.080, 0.111)	0.028 (0.021, 0.035)
BRM: $(\boldsymbol{G'}+\boldsymbol{S'}+\boldsymbol{F})^2$	0.774 (0.759, 0.790)	0.093 (0.082, 0.105)	0.004 (0.002, 0.006)	0.099 (0.084, 0.113)	0.030 (0.025, 0.036)
MNAR					
Observed (training)	0.868	0.075	0.008	0.033	0.015
Observed	0.751	0.110	0.007	0.100	0.032
Geocoding only (GO)	0.763	0.062	0.006	0.138	0.030
Surname Only (SO)	0.666	0.090	0.006	0.198	0.040
BISG	0.804	0.026	0.000	0.150	0.019
BRM: $\boldsymbol{G'}+\boldsymbol{S'}$	0.801 (0.785, 0.815)	0.095 (0.085, 0.105)	0.008 (0.005, 0.012)	0.064 (0.055, 0.075)	0.032 (0.027, 0.036)
BRM: $\boldsymbol{G'}+\boldsymbol{S'}+\boldsymbol{F}$	0.781 (0.762, 0.798)	0.094 (0.083, 0.107)	0.003 (0.002, 0.005)	0.095 (0.081, 0.115)	0.026 (0.022, 0.033)
BRM: $(\boldsymbol{G'}+\boldsymbol{S'}+\boldsymbol{F})^2$	0.770 (0.753, 0.788)	0.096 (0.084, 0.108)	0.004 (0.002, 0.006)	0.099 (0.085, 0.115)	0.029 (0.026, 0.034)

[a] Observed racial compositions correspond to those observed in the test set (n_e individuals)

[b] GO, SO, BISG probabilities were normalized to only consider the five races in this study

[c] Median (95% CB) reported for BRMs

of predictors used, BRMs provide racial composition estimates that are closest to the observed racial composition; in particular, with respect to the racial composition statistic, BRM: $(\boldsymbol{G'}+\boldsymbol{S'}+\boldsymbol{F})^2$ that includes family race and interaction terms performed the best.

Table 4 Post-imputation analyses for JM, FCS using m = 25—racial composition estimates

	White	Black	AI	Hispanic	API
MCAR					
Observed (training)	0.806	0.095	0.010	0.068	0.021
Observed[a]	0.819	0.088	0.005	0.062	0.025
JM-GLM[b]	0.751 (0.010)[e]	0.099 (0.007)	0.016 (0.003)	0.109 (0.007)	0.025 (0.004)
JM-MVN[c]	0.650 (0.012)	0.142 (0.009)	0.034 (0.006)	0.122 (0.010)	0.052 (0.007)
FCS[d]	0.819 (0.009)	0.087 (0.006)	0.007 (0.002)	0.064 (0.005)	0.024 (0.003)
MAR					
Observed (training)	0.859	0.080	0.009	0.034	0.017
Observed	0.761	0.104	0.006	0.099	0.030
JM-GLM	0.683 (0.011)	0.100 (0.007)	0.014 (0.003)	0.139 (0.016)	0.064 (0.016)
JM-MVN	0.657 (0.014)	0.142 (0.010)	0.026 (0.005)	0.123 (0.011)	0.051 (0.007)
FCS	0.775 (0.014)	0.092 (0.009)	0.004 (0.002)	0.096 (0.010)	0.032 (0.004)
MNAR					
Observed (training)	0.868	0.075	0.008	0.033	0.015
Observed	0.751	0.110	0.007	0.100	0.032
JM-GLM	0.681 (0.011)	0.099 (0.007)	0.013 (0.003)	0.144 (0.011)	0.062 (0.010)
JM-MVN	0.662 (0.016)	0.141 (0.011)	0.025 (0.004)	0.124 (0.012)	0.047 (0.006)
FCS	0.774 (0.012)	0.094 (0.008)	0.004 (0.002)	0.099 (0.008)	0.029 (0.004)

[a] Observed racial compositions for the n_e individuals in the test set

[b] JM-GLM assumes $\mathbf{W} = (\mathbf{R}, \mathbf{F})$ and $\mathbf{Z} = (\mathbf{G}', \mathbf{S}')$; restricted general location model was implemented

[c] JM-MVN is implemented in *Amelia* package

[d] FCS uses BRM: $(\boldsymbol{G}' + \boldsymbol{S}' + \boldsymbol{F})^2$ to estimate race, multivariate normal distributions for $\mathbf{G}', \mathbf{S}'$

[e] Average (SE) of the $m = 25$ estimates reported

3.2 Evaluation of imputation models

Tables 4, 5, and 6 shows racial composition estimates, average ages, and sensitivity for the multiple imputation models considered, using the variables that were found to contribute substantially to predictive improvements in race estimation: geocoded probabilities, surname probabilities, family race indicators and their interactions. Overall, racial composition estimates for FCS (with BRM: $(\boldsymbol{G}' + \boldsymbol{S}' + \boldsymbol{F})^2$) are closer to the reported values than those for both implementations of JM, and the standard errors (SE) for the FCS model are smaller than those generated by JM-MVN and approximately the same as those in JM-GLM (Table 4). For all three missing data mechanisms, age estimates within each racial group tend to be closer to the observed values using FCS (Table 5). Of the multiple imputation procedures considered, FCS results in higher sensitivity for Whites, Blacks and APIs. For Hispanics, the sensitivity is higher using JM-GLM and for AIs, the sensitivity is slightly higher when using JM-MVN (Table 6).

Table 5 Post-imputation analyses for JM, FCS using m = 25—average ages

	White	Black	AI	Hispanic	API
MCAR					
Observed (training)	26.9	24.5	25.6	29.8	29.3
Observed[a]	26.6	24.3	21.0	28.3	26.6
JM-GLM[b]	26.7 (0.3)[e]	26.7 (1.0)	27.9 (2.8)	24.9 (0.9)	25.8 (1.9)
JM-MVN[c]	26.6 (0.4)	25.9 (1.0)	26.5 (2.2)	26.5 (1.0)	26.4 (1.5)
FCS[d]	26.6 (0.3)	25.1 (1.0)	26.6 (4.9)	26.8 (1.2)	26.9 (2.2)
MAR					
Observed (training)	28.7	26.1	24.4	31.9	27.9
Observed	24.5	22.9	23.7	28.0	27.9
JM-GLM	24.8 (0.3)	24.3 (0.9)	24.9 (2.8)	25.8 (1.0)	23.6 (1.6)
JM-MVN	24.6 (0.4)	24.5 (0.8)	24.2 (2.7)	25.7 (1.0)	25.3 (1.7)
FCS	24.5 (0.3)	23.8 (1.0)	22.3 (4.8)	26.7 (1.0)	27.6 (1.8)
MNAR					
Observed (training)	28.6	26.6	24.7	32.2	28.3
Observed	24.5	22.7	23.4	28.0	27.7
JM-GLM	24.8 (0.3)	24.3 (0.9)	25.5 (2.8)	25.3 (0.8)	23.9 (1.4)
JM-MVN	24.7 (0.4)	24.0 (0.8)	24.2 (2.4)	25.7 (1.0)	25.4 (1.8)
FCS	24.6 (0.3)	23.4 (1.0)	19.9 (4.3)	26.6 (0.9)	27.1 (1.8)

[a] Observed ages for the n_e individuals in the test set

[b] JM-GLM assumes $\mathbf{W} = (\mathbf{R}, \mathbf{F})$ and $\mathbf{Z} = (\mathbf{G}', \mathbf{S}')$; restricted general location model was implemented

[c] JM-MVN is implemented in *Amelia* package

[d] FCS uses BRM: $(\boldsymbol{G}' + \boldsymbol{S}' + \boldsymbol{F})^2$ to estimate race, multivariate normal distributions for $\mathbf{G}', \mathbf{S}'$

[e] Average (SE) of the $m = 25$ estimates reported

Table 6 Post-imputation analyses for JM, FCS using m = 25—Sensitivity

	White	Black	AI	Hispanic	API
MCAR					
JM-GLM[a]	0.865 (0.009)[d]	0.611 (0.039)	0.021 (0.062)	0.718 (0.042)	0.554 (0.066)
JM-MVN[b]	0.735 (0.013)	0.582 (0.046)	0.488 (0.159)	0.469 (0.058)	0.561 (0.070)
FCS[c]	0.939 (0.007)	0.685 (0.036)	0.480 (0.189)	0.600 (0.051)	0.752 (0.068)
MAR					
JM-GLM	0.836 (0.010)	0.594 (0.034)	0.040 (0.053)	0.737 (0.034)	0.754 (0.055)
JM-MVN	0.752 (0.015)	0.517 (0.039)	0.358 (0.149)	0.375 (0.045)	0.488 (0.064)
FCS	0.922 (0.011)	0.603 (0.042)	0.262 (0.124)	0.578 (0.045)	0.806 (0.054)
MNAR					
JM-GLM	0.841 (0.010)	0.596 (0.033)	0.018 (0.041)	0.739 (0.032)	0.721 (0.049)
JM-MVN	0.762 (0.015)	0.531 (0.039)	0.340 (0.127)	0.374 (0.041)	0.468 (0.067)
FCS	0.928 (0.008)	0.608 (0.039)	0.264 (0.128)	0.591 (0.050)	0.756 (0.066)

[a] JM-GLM assumes $\mathbf{W} = (\mathbf{R}, \mathbf{F})$ and $\mathbf{Z} = (\mathbf{G}', \mathbf{S}')$; restricted general location model was implemented

[b] JM-MVN is implemented in *Amelia* package

[c] FCS uses BRM: $(\boldsymbol{G}' + \boldsymbol{S}' + \boldsymbol{F})^2$ to estimate race, multivariate normal distributions for $\mathbf{G}', \mathbf{S}'$

[d] Average (SE) of the $m = 25$ estimates reported

3.3 Application to the entire dataset

Because the FCS approach with BRM: $(\boldsymbol{G}'+\boldsymbol{S}'+\boldsymbol{F})^2$ had the best overall performance in terms of prevalence and sensitivity, we implemented it on the entire dataset of 13,975 observations. Mean (SE) racial prevalence estimates among the 6637 observations that were missing self-reported race/ethnicity using $m=25$ were 0.685 (0.010) White, 0.105 (0.007) Black, 0.006 (0.002) AI, 0.166 (0.008) Hispanic and 0.038 (0.003) API. These are significant differences from the racial composition among the 7338 beneficiaries with self-reported race/ethnicity (Table 7), cautioning us against implementing a complete case analysis and supporting the notion that race values are not missing completely at random. After imputation using FCS, the estimates for the age averages (SE) within each race for the 6637 beneficiaries are: 23.0 (0.2) for Whites, 21.6 (0.6) for Blacks, 23.1 (2.9) for AI, 21.2 (0.5) for Hispanics and 23.9 (1.1) for APIs. The age averages for this group relative to those who self-reported race/ethnicity are significantly lower across all racial/ethnic groups, showing that younger individuals are more likely to be missing race/ethnicity. In addition, those who self-report race/ethnicity are significantly more likely to have family members who report race/ethnicity than those who do not. Imputation of the full dataset using JM-GLM and JM-MVN results in similar conclusions.

4 Discussion

To identify and address disparities in health care, it is important to collect accurate data on patients' race/ethnicity. The gold standard, self-reported race/ethnicity, is often unavailable, and indirect measures of race/ethnicity are often needed. We have viewed this problem from a missing data perspective and proposed two new approaches for multiply imputing race values.

In simulation analyses based on real data, the newly proposed Bayesian multinomial logistic regression models performed better in terms of the discrimination and calibration than previously published methods such as BISG. We recommend incorporating these BRMs into healthcare studies that examine race/ethnicity. Additionally, because models that included family race indicator variables considerably outperformed those that did not, including family race indicators when available could prove valuable. We further demonstrated that including the BISG probabilities in addition to the geocoded and surname probabilities did not significantly improve model performance.

Among the multiple imputation methods considered, the FCS model performed better than the JM models in terms of reporting overall racial composition estimates that were closest to observed ones, as well as in other post-imputation analyses. Although JM-GLM is an elegant and efficient imputation algorithm, it is complex to specify and estimate when there are many variables in the dataset and it may lack the flexibility needed to capture important characteristics of the data (van Buuren 2007). We recommend using FCS for datasets that, in addition to missing racial/ethnic information, are also missing values for other variables. Our findings support the results reported by Seaman and Hughes (2018). We show that with high proportions of missing race/ethnicity values, using multiple imputation with FCS results in better performance compared to misspecified JM-GLM.

The newly proposed methods for estimating race/ethnicity that rely on the multiple imputation framework offer additional advantages over current methods. First, they allow

Table 7 Post-imputation results for FCS on the full dataset compared to results using only those with self-reported race

	White	Black	AI	Hispanic	API
Self-reported race/ethnicity ($n = 7,338$)					
Racial composition	0.806	0.099	0.007	0.063	0.024
Average age	26.9	24.8	23.9	28.9	27.3
Proportion with Family that Reported Race[a]	0.829	0.790	0.774	0.661	0.701
Proportion with Family from Other Race[b]	0.034	0.133	0.491	0.181	0.102
Imputed race/ethnicity ($n = 6,637$)					
Racial composition	0.685 (0.010)	0.105 (0.007)	0.006 (0.002)	0.166 (0.008)	0.038 (0.003)
Average age	23.0 (0.2)	21.6 (0.6)	23.1 (2.9)	21.2 (0.5)	23.9 (1.1)
Proportion with family that reported race	0.373 (0.009)	0.340 (0.028)	0.480 (0.130)	0.275 (0.017)	0.348 (0.036)
Proportion with family from other race	0.130 (0.004)	0.545 (0.017)	0.626 (0.067)	0.549 (0.012)	0.560 (0.025)

[a] Proportion of individuals in each racial/ethnic group that have individuals in the dataset with the same Family ID that *did* report race/ethnicity

[b] Proportion of individuals in each racial/ethnic group that have individuals in the dataset with the same Family ID that reported a different race/ethnicity from what they self-reported or were imputed as

for straightforward estimation and interpretation of race/ethnicity coefficients when race/ethnicity is used as an explanatory variable in regression models. Second, they enable estimation of the error involved in prediction of unknown race/ethnicity values. Third, additional variables that are predictive of subject's race/ethnicity can easily be incorporated.

The new approaches have limitations. First, the methods were only applied and tested on a single large dataset, and results may not be generalizable to other datasets. However, we expect familiar relationships to be recorded in other insurance claims datasets, because children are generally enrolled under their parents' insurance plan. Second, although including primary language was considered in the model selection stage, because of a small number of people who spoke languages other than English, it was not found to have a significant effect. However, should the variable have more predictive ability in other cohorts it can be easily accommodated in imputation-based methods. Third, the set of racial and ethnic categories considered was limited. An increasing portion of individuals in the US identifies as either "Other" or "Multiracial" and excluding these individuals from the analysis may lead to overestimates of model performance. Although BRMs that included family race yielded high AUCs for American Indians, the corresponding sensitivity was low. This may occur if, on average, the model does a good job at estimating a higher probability of belonging to this racial group for someone who is American Indian relative to someone who is not, but regardless assigns them to a different racial/ethnic group because their probability of belonging to another group is greater. Having a greater sample of American Indians, as well as access to other novel predictors, may help improve race imputation in these cases. Nonetheless, our proposed BRM method has performed better than currently available methods within this group of American Indians. Lastly, when race is not missing at random, the proposed methods may result in biased estimates. However, in our example this bias was minimal.

Despite these limitations, the FCS approach detailed here offers a novel and flexible way of estimating race/ethnicity in a dataset in which some persons have missing data. This approach had higher predictive power because it considers other predictors of race/ethnicity aside from geocoded and surname probabilities, it enables propagation of error in race imputation, it allows for estimation within the multiple imputation framework and it can be easily implemented with current software.

5 Online supplementary material

The full code for FCS with a BRM is available on GitHub. A sample dataset (simulated using our observed data) has been included as well.

The link for this is: https://github.com/gsilva2/FCS_raceimputation.

Acknowledgements Research was supported by a Grant (R40 MC 28319) from the Health Resources and Services Administration.

Compliance with ethical standards

Conflict of interest All authors declare they have no conflict of interest.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent The need for informed consent was waived by the institutional review board.

Appendix 1

Pseudocode for multiple imputation with the FCS algorithm is presented below.

```
1. Populate incomplete dataset with initial starting values
2. Draw $\theta_R$ from $P(\theta_R|R^{obs}, Y_{-R}, M, A)$
3. Draw $R^{miss}$ from $P(R|R^{obs}, Y_{-R}, M, A, \theta_R)$ and substitute this into the dataset
4. Draw $\theta_{G'}$ from $P(\theta_{G'}|{G'}^{obs}, Y_{-G'}, M, A)$
5. Draw ${G'}^{miss}$ from $P(G'|{G'}^{obs}, Y_{-G'}, M, A, \theta_{G'})$ and substitute this into the dataset
6. Draw $\theta_{S'}$ from $P(\theta_{S'}|{S'}^{obs}, Y_{-S'}, M, A)$
7. Draw ${S'}^{miss}$ from $P(S'|{S'}^{obs}, Y_{-S'}, M, A, \theta_{S'})$ and substitute this into the dataset
8. Repeat Steps 2–7 until the cycle reaches convergence. The current draws are the set of imputed values.
9. Repeat Steps 1–8 $m$ times to obtain $m$ imputed datasets
```

Appendix 2

To determine the specification for $P(\mathbf{R}|\mathbf{Y}_{-\mathbf{R}}, \mathbf{M}, \mathbf{A}, \boldsymbol{\theta}_{\mathbf{R}})$ and to compare various multiple imputation methods, the observed race for n_e of the $n = 6087$ individuals with fully observed data was set to missing. The remaining set of n_f individuals were used to obtain parameter estimates for the BRMs. Using the final sample of parameter estimates, probabilities of individual i belonging to race $r \in Y$ for each of the n_e individuals, given by $p_{ir} = \frac{exp(\mathbf{x}_\mathbf{i}^\mathbf{T}\boldsymbol{\beta}_\mathbf{r})}{\sum_{r \in Y}(exp(\mathbf{x}_\mathbf{i}^\mathbf{T}\boldsymbol{\beta}_\mathbf{r}))}$, were computed. The probabilities $\mathbf{p_i} = (p_{iWhite}, p_{iBlack}, p_{iAI}, p_{iHispanic}, p_{iAPI})$ for $i \in \{1, 2, \dots, n_e\}$ were compared to the observed races for these individuals using AUC and racial composition.

Individuals in the testing set were those whose race has been set equal to missing. Thus, when determining which set each of these n individuals will be placed in, it is important to specify whether race is missing completely at random (MCAR), missing at random (MAR) or missing not at random (MNAR). We considered all three in our analysis; implementation details for each are available below.

MCAR

The race data is MCAR if the missingness of race is unrelated to any of the study variables; that is, the pattern of race missingness is independent of observable variables (Little and Rubin 2002). In practice, this is a very strong assumption. We implement this assumption by randomly selecting n_e individuals from the group of n. We set $n_e = 2891$ and $n_f = 3196$ such that the percentage of individuals with missing race is 47.5% because this is the percentage of missing race in the full RI dataset.

Table 8 Parameter values used to simulate MAR

Variable	Value (β)
Intercept	2.318
Geocoded	
Black	−2.139
AI	−6.877
Hispanic	−0.529
API	2.325
Other	2.508
Surname	
Black	0.206
AI	−2.486
Hispanic	0.636
API	0.500
Other	1.292
Family race indicators	
White	−2.064
Black	−1.783
AI	−1.604
Hispanic	−1.093
API	−1.796
Language	−0.330
Age	−0.023

MAR

The race data is MAR if the missingness can be explained by variables for which there is complete information. Using the individuals in rows V and VI of Fig. 1, we fit a logistic regression model where the dependent variable is an indicator for whether race is missing and the independent variables are the fully observed geocoded and surname probabilities, family race indicators, language, and age. Using the estimated parameters for this model (intercept, geocoded, surname, family race indicators, language, and age in Table 8), we estimate the probability that race is missing for each of the n individuals with fully observed data. These probabilities are then used to determine whether an individual belongs in the training set or the testing set. In our analysis, $n_f = 3195$ while $n_e = 2892$; hence, race is set to missing for 47.5% of the n beneficiaries considered for this simulation.

MNAR

The race data is MNAR when the missing entries for race depends on the racial groups, even after controlling for other variables with complete information. To simulate MNAR, we fit the same logistic regression model described in the MAR section and also incorporate parameters for the observed race (Table 9). Using this combined set of parameters, we compute the probability that race is missing for each of the n beneficiaries with completely observed data. Similar to before, these are used to determine whether an individual will be

Table 9 Parameter values used to simulate MNAR

Variable	Value (β)
Intercept	2.292
Geocoded	
Black	– 2.139
AI	– 6.878
Hispanic	– 0.529
API	2.325
Other	2.508
Surname	
Black	0.206
AI	– 2.478
Hispanic	0.636
API	0.500
Other	1.292
Family race indicators	
White	– 2.064
Black	– 1.783
AI	– 1.604
Hispanic	– 1.093
API	– 1.796
Language	– 0.330
Age	– 0.023
Observed Race	
Black	0.200
AI	0.300
Hispanic	0.050
API	0.300

placed in the training set or the testing set. In our analysis, $n_f = 3197$ while $n_e = 2890$; the percentage of individuals whose race is set to missing is 47.5%.

Note: The intercepts reported in Tables 8 and 9 were not the intercepts estimated from the logistic regression model. Rather, these were modified so that the percentage of individuals in the test was 47.5% across all missing data mechanisms.

References

Adjaye-Gbewonyo, D., Bednarczyk, R.A., Davis, R.L., Omer, S.B.: 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. **49**(1), 268–283 (2013)

Consumer Financial Protection Bureau: Using publicly available information to proxy for unidentified race and ethnicity : a methodology and assessment. Consumer Financial Protection Bureau, United States (2014)

Elliott, M.N., Fremont, A., Morrison, P.A., 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. **43**(5p1), 1722–1736 (2008)

Elliott, M.N., Morrison, P.A., Fremont, A., McCaffrey, D.F., 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. **9**(2), 69 (2009)

Fawcett, T.: An introduction to ROC analysis. Pattern Recogn. Lett. **27**(8), 861–874 (2006)

Fiscella, K., Fremont, A.M.: Use of geocoding and surname analysis to estimate race and ethnicity. Health Serv. Res. **41**(4 Pt 1), 1482–1500 (2006)

Hassett, P.: Taking on racial and ethnic disparities in health care: the experience at Aetna. Health Aff. **24**(2), 417–420 (2005)

Honaker, J., King, G., Blackwell, M.: Amelia: A program for missing data. R package version 1.7.5 (2018). https://cran.r-project.org/web/packages/Amelia/

Hosmer, D.W., Lemeshow, S.: Applied Logistic Regression. Wiley, Hoboken (2000)

Hosmer, D.W., Lemeshow, S., Sturdivant, R.X.: Applied Logistic Regression. Wiley, Hoboken (2013)

Kruschke, J.K.: Doing Bayesian Data Analysis: A Tutorial with R and BUGS. Academic Press, Burlington, MA (2011)

Little, R.J.A., Rubin, D.B.: Statistical Analysis with Missing Data. Wiley, Hoboken (2002)

Liu, Y., De, A.: Multiple imputation by fully conditional specification for dealing with missing data in a large epidemiologic study. Int. J. Stat. Med. Res. **4**(3), 287–295 (2015)

Ma, Y., Zhang, W., Lyman, S., Huang, Y.: The HCUP SID imputation project: improving statistical inferences for health disparities research by imputing missing race data. Health Serv. Res. **53**(3), 1870–1889 (2018)

Ng, J.H., Ye, F., Ward, L.M., Haffer, S.C.C., Scholle, S.H.: Data on race, ethnicity, and language largely incomplete for managed care plan members. Health Aff. (Project Hope) **36**(3), 548–552 (2017)

Polson, N.G., Scott, J.G., Windle, J.: Bayesian inference for logistic models using Polya-Gamma latent variables (2013a). arXiv:1205.0310

Polson, N.G., Scott, J.G., Windle, J.: BayesLogit (2013b)

Rubin, D.B.: Inference and missing data. Biometrika **63**(3), 581–592 (1976)

Rubin, D.B.: Multiple Imputation for Nonresponse in Surveys. Wiley, Hoboken (1987)

Schafer, J.L.: Analysis of Incomplete Multivariate Data, 1. ed., 1. CRC Press Reprint ed. Monographs on Statistics and Applied Probability, vol. 72. Chapman & Hall/CRC, Boca Raton (2000)

Schafer, J.L.: Mix: Estimation/Multiple imputation for mixed categorical and continuous data. R package version 1.0-10. (2017). https://CRAN.R-project.org/package=mix

Seaman, S.R., Hughes, R.A.: Relative efficiency of joint-model and full-conditional-specification multiple imputation when conditional models are compatible: the general location model. Stat. Methods Med. Res. **27**(6), 1603–1614 (2018)

Ulmer, C., McFadden, B., Nerenz, D.R.: Race, Ethnicity, and Language Data: Standardization for Health Care Quality Improvement. National Academies Academic Press, Washington, D.C. (2009)

van Buuren, S.: Multiple imputation of discrete and continuous data by fully conditional specification. Stat. Methods Med. Res. **16**(3), 219–242 (2007)

van Buuren, S., Groothuis-Oudshoorn, K.: Mice: Multivariate imputation by chained equations in R. J. Stat. Softw. **45**(3), 1–67. (2018). http://www.jstatsoft.org/v45/i03/

Word, D.L., Coleman, C.D., Nunziata, R., Kominski, R.: Demographic Aspects of Surnames from Census 2000. US Census Bureau, Suitland (2008)