



# Statistical Methods for Imputing Race and Ethnicity

April | 2024







# Statistical Methods for Imputing Race and Ethnicity

**AUTHORS** Larry Baeder  
Erica Baird, PhD, FSA, MAAA  
Peggy Brinkmann, FCAS, MAAA  
Joe Long, ASA, MAAA  
Caleb Stracke, ASA, MAAA  
Kweweli Togba-Doya  
Gabriele Usan  
Natalie Weaver  
Meseret Woldeyes, MS  
Milliman

**SPONSORS** Diversity, Equity, and Inclusion Strategic  
Research Program  
  
Product Development Section  
  
Reinsurance Section

 **Give us your feedback!**  
Take a short survey on this report. [Click Here](#) 

**Caveat and Disclaimer**

The opinions expressed and conclusions reached by the authors are their own and do not represent any official position or opinion of the Society of Actuaries Research Institute, the Society of Actuaries, or its members. The Society of Actuaries Research Institute makes no representation or warranty to the accuracy of the information.

Copyright © 2024 by the Society of Actuaries Research Institute. All rights reserved.

## CONTENTS

<b>Section 1: Introduction .....</b>	<b>5</b>
<b>Section 2: Background .....</b>	<b>6</b>
2.1 Definitions .....	6
2.2 Performance Metrics .....	6
<b>Section 3: Discussion of Imputation Methods .....</b>	<b>8</b>
3.1 Pre-Bayesian Methods .....	8
3.1.1 Geocoding Only .....	8
3.1.2 Surname Analysis .....	9
3.1.3 Categorical Surname and Geocoding .....	9
3.2 Bayesian Methods .....	10
3.2.1 Bayesian Surname Geocoding .....	10
3.2.2 Bayesian Improved Surname Geocoding .....	11
3.2.3 Medicare Bayesian Improved Surname Geocoding .....	13
3.2.4 Bayesian Improved Surname Geocoding Extensions .....	14
3.2.5 Bayesian Improved First Name Surname Geocoding .....	14
3.2.6 Modified Bayesian Improved First Name Surname Geocoding .....	15
3.2.7 Fully Bayesian Improved Surname Geocoding .....	16
3.2.8 Bayesian Instrumental Regression for Disparity Estimation .....	17
3.3 Predictive Modeling Methods .....	19
3.3.1 Regression .....	19
3.3.2 Natural Language Processing .....	20
3.3.3 Other Machine Learning Algorithms .....	20
<b>Section 4: Discussion of Direct Data .....</b>	<b>22</b>
<b>Section 5: Case Study: Imputation Methods for Race and Ethnicity .....</b>	<b>24</b>
5.1 Dataset and Algorithms .....	24
5.2 Data Cleaning and Preprocessing .....	25
5.3 Performance Metrics .....	29
5.3.1 Actual-to-Expected Distribution .....	29
5.3.2 Probability of Self-Reported Race and Ethnicity Predicted .....	33
5.3.3 Probability of White Race Predicted .....	34
5.3.4 Ratio of True Positives to False Positives .....	35
5.3.5 Area Under the Receiver Operating Characteristic curve (AUC) .....	36
5.3.6 Precision .....	37
5.3.7 Specificity .....	38
5.4: Case Study Summary .....	39
<b>Section 6: Tutorial .....</b>	<b>40</b>
6.1 Input Data .....	40
6.2 Imputation Packages .....	41
6.3 Output .....	42
<b>Section 7: Conclusion .....</b>	<b>43</b>
<b>Acknowledgments .....</b>	<b>44</b>
<b>Appendix A: Summary of Imputation Methods .....</b>	<b>45</b>
<b>Appendix B: Detailed Performance Metrics from Case Study .....</b>	<b>47</b>
B.1 Comparison of Accuracy Metrics Prior to Data Cleaning .....	47
B.2 Ratio of True Positives to False Positives by Cohort .....	48

- Appendix C: Summary of Tutorial Imputation Packages ..... 49**
- Appendix D: Sample Tutorial Output ..... 50**
  - D.1 Calibration Curves .....50
  - D.2 Actual-to-Expected Distribution .....51
  - D.3 Probability of Correct Self-Reported Race and Ethnicity Predicted .....51
  - D.4 Probability of White Race Predicted .....52
  - D.5 Ratio of True Positives to False Positives .....53
  - D.6 Area Under the Receiver Operating Characteristic Curve (AUC) .....53
  - D.7 Precision .....53
  - D.8 Specificity .....54
  - D.9 BIRDIE Outcome Estimates .....54
- References ..... 56**
- About The Society of Actuaries Research Institute ..... 58**

## Section 1: Introduction

While concerns about disparities and disproportionate impacts across race and ethnicity groups in insurance are not new, events in recent years have led to a fresh wave of discussions about racial justice and equality in the United States. Increased focus among the insurance industry and regulatory community on bias and equity in insurance processes and outcomes has spurred new research into disproportionate impact, fairness, and disparate outcome analyses, and the passage of laws like Colorado SB21-169,<sup>1</sup> the stated intent of which is to protect Colorado consumers from unfair discrimination in insurance practices.

However, a lack of consistent data collection or reporting is often an obstacle in the study of disproportionate impacts and equity across race and ethnicity cohorts in an insurance context. Datasets that are useful for research and analysis related to insurance products and services in the United States often do not include variables for race or ethnicity and, when such variables are available, the data is often incomplete or may have restrictions on usage. In property and casualty (P&C) insurance, race and ethnicity data has not been systematically collected (American Academy of Actuaries, 2022) while, in health insurance, race and ethnicity data are often incomplete and inconsistent (Haley, et al., 2022).

In the insurance industry, some of the first uses of statistical methods for imputing or modeling race and ethnicity were in life and health insurance. Fiscella and Fremont (2006) cite multiple studies using surname analyses to assess differences in mortality (Rosenwaike, Hempstead, & Rogers, 1991), cancer incidence (Swallen, West, Stewart, Glaser, & Horn-Ross, 1997), (Swallen, et al., 1998), (Coronado, et al., 2002), and rates of cancer screening (Jacobs & Lauderdale, 2001). In P&C insurance, imputation has been used to analyze various rating variables for correlation and bias, in particular, territory and credit-based insurance scores. For example, the Texas Department of Insurance conducted studies in 2004 and 2005 on the relationship between credit-based insurance scores and ethnicity using a Hispanic surname database (NAIC, 2008). While these first attempts used simple methods to address missing race and ethnicity data, imputation methods have evolved significantly and not all are widely known or understood in the actuarial profession.

This paper describes a range of techniques for developing probabilistic estimates or predictions of individual race and/or ethnicity. The authors apply some methods to data from an insurance application and compare them with a focus on relative accuracy and potential bias. Lastly, the authors apply some of these methods to a simulated dataset to illustrate how to use them in practice.

While this paper does not represent an exhaustive list of possible imputation procedures, the authors believe those presented are a useful starting point for the actuarial profession. The authors do not make any recommendations as to any method's appropriateness for any particular use case because that is situationally dependent. For example, aligning the imputation method's data sources with the population of the dataset to be imputed is critical for accurate results and requires a thorough examination of all relevant data sources. Lastly, these methods fit into a broader context of studying disproportionate impacts and quantifying discriminatory effects. This paper focuses only on the technical aspects of the imputation methods, which should not be used without considering the downstream impacts and consequences.

---

<sup>1</sup> See <https://doi.colorado.gov/for-consumers/sb21-169-protecting-consumers-from-unfair-discrimination-in-insurance-practices>.

## Section 2: Background

### 2.1 DEFINITIONS

Imputation is a term that refers to processes that replace missing data with inferred values or a vector of probabilities. Some sources refer to imputation as “indirect estimation.”

Direct data refers to race and ethnicity information obtained directly from individuals. Direct data is obtained by asking individuals to self-report. For example, the U.S. Census Bureau collects direct data on race and ethnicity (American Academy of Actuaries, 2022).

### 2.2 PERFORMANCE METRICS

Most imputation methods output probabilities of belonging to various cohorts, for example “Black,” “Asian/Pacific Islander,” “Hispanic” and “white/other.” The cohort definitions vary across different datasets and studies. When the output of a method is a vector of probabilities, it can be translated into a classification by selecting the race or ethnicity cohort with the highest probability.

When self-reported data is available, researchers can compare the predictions from an imputation method to the actual data. Several different metrics have been used in the literature to assess and compare imputation method performance, including:

- **Accuracy:** For classification outputs (where each individual is associated with a single imputed race or ethnicity cohort), accuracy is the proportion of individuals where the imputed cohort matches the self-reported race or ethnicity. For continuous outputs such as probabilities, it is measured by the correlation between the estimated probabilities and actual self-reported race or ethnicity. It is important to note that this metric can be misleading in imbalanced datasets where the population distribution by race or ethnicity is not uniform because the accuracy for each cohort may vary significantly.<sup>2</sup> Due to this, it is important to examine the accuracy for each cohort in the study.
- **Error rate:** The overall proportion of individuals where the imputed cohort doesn’t match the self-reported race. It is important to note that, similar to accuracy, this metric can also be misleading in imbalanced datasets, so it is important to examine the error for each cohort in the study.
- **False positives:** Individuals who are classified into one race/ethnicity cohort whose self-reported race is a different cohort. For example, a non-Hispanic individual with an imputed ethnicity of Hispanic would be considered a false positive.
- **False negatives:** Individuals who are classified as not in a race/ethnicity cohort whose self-reported race is that cohort. For example, a Hispanic individual with an imputed ethnicity of something other than Hispanic would be considered a false negative. Note that false positives and false negatives are relevant in the context of one cohort; a false positive in one cohort is also a false negative in another cohort.
- **Precision:** The proportion of individuals with the same imputed race/ethnicity cohort for whom the imputation matches the self-reported race/ethnicity. For example, the proportion of individuals who are predicted to be Hispanic who self-report as Hispanic.

---

<sup>2</sup> For example, if accuracy is 94% for a group that is 90% of the total, and the accuracy is 50% for a group that is 10% of the total, the overall accuracy will be 90% ( $0.90 \times .94 + 0.10 \times 0.50 = 90\%$ ), i.e., the overall accuracy is not sensitive to poor accuracy in smaller subgroups.

- Sensitivity/recall: The true positive rate, i.e., the ability to match the imputed cohort with the self-reported one. For example, the proportion of individuals with self-reported Hispanic ethnicity where the imputed ethnicity is also Hispanic.
- Specificity/selectivity: The true negative rate. For example, the proportion of individuals with a self-reported race/ethnicity other than Hispanic whose imputed race/ethnicity was non-Hispanic.
- Receiver operating characteristic (ROC) curves: A plot of the true positive rate on the vertical axis against the false positive rate on the horizontal axis for various thresholds used for classification.
- Area under ROC curve (AUC): The area under the receiver operating characteristic (ROC) curve. The value can be between 0 and 1, where a higher area measure represents a better ability to distinguish between two cohorts. According to Sorbero, values of 0.7 are considered “acceptable;” 0.8 are considered “strong;” and 0.9 are considered “excellent” (Sorbero, 2022). Also known as the concordance (C) statistic.
- Calibration curve (also known as a reliability diagram): Plot of the frequency of a true positive label on the y-axis by the predicted probability on the x-axis.

Sections 5 and 6 of this paper show examples of these statistics calculated on various datasets and metrics to further illustrate their use, as well as their relative strengths and weaknesses.

## Section 3: Discussion of Imputation Methods

Imputing missing data on race and ethnicity is not a new practice—for example, the U.S. Census Bureau has been using imputation for several population characteristics, including race and ethnicity, since the 1960 Census.<sup>3</sup> Modern methods for imputing race and ethnicity for individuals began in 2008 with the introduction of Bayesian Surname Geocoding (Elliott, Fremont, Morrison, Pantoja, & Lurie, 2008). Since then, researchers have developed variations and refinements, with recent papers exploring the use of machine-learning algorithms and improved data quality.

Most imputation methods are fundamentally similar in that they use data that has race/ethnicity information available to develop conditional distributions using attributes of the individuals, such as name, geolocation, or other characteristics available on the dataset to be analyzed. Methods that use only name and geolocation have an advantage in that conditional distributional data for names and geolocation are readily available. To incorporate other domain-specific information (age, gender, medical conditions, etc.) requires more direct data to develop the conditional distributions but can generate improvements in imputation accuracy.

The remainder of this section provides an overview of various methods in use today and in the recent past. Each section describes the method and particular implementation used by the author, for example, the data used to develop the conditional distributions and the data to which the conditional distributions were applied. However, the methods can be implemented with datasets other than the particular ones used by the authors.

Appendix A contains tables summarizing the inputs, outputs, references, and packages (where available) for each method described in this section.

### 3.1 PRE-BAYESIAN METHODS

#### 3.1.1 GEOCODING ONLY

Geocoding Only (GO) uses individuals' addresses to link to census data about the geographic areas where they live and the associated race/ethnicity distribution of the area (Fiscella & Fremont, 2006). For example, knowing that a person lives in a census block group (a small neighborhood of approximately 1,000 residents<sup>4</sup>), where 90% of the residents are Black, provides useful information for predicting that person's self-reported race/ethnicity (Elliott, Fremont, Morrison, Pantoja, & Lurie, 2008).

Before the introduction of Bayesian imputation methods, GO was a popular method for imputation when names were not available. GO methods at the ZIP code level were used in multiple studies of disproportionate impacts of credit-based insurance scores in P&C insurance in the 1990s and 2000s (NAIC, 2008). In 2006, Fiscella and Fremont noted that using geolocation to estimate the effects of sociodemographic characteristics on health was “relatively new,” but “routinely” used by researchers when direct data are lacking.

GO can be performed at different geographic levels, but the accuracy of GO predictions is expected to increase when smaller, more homogenous units of analysis are used (Krieger, et al., 2002). As Fiscella and

<sup>3</sup> See <https://www.census.gov/newsroom/blogs/random-samplings/2021/08/census-when-demographic-and-housing-characteristics-are-missing.html>.

<sup>4</sup> See [https://www.census.gov/programs-surveys/geography/about/glossary.html#par\\_textimage\\_4](https://www.census.gov/programs-surveys/geography/about/glossary.html#par_textimage_4).



Fremont (2006) found, the accuracy of imputations using GO could vary considerably based on the level of racial and ethnic diversity of a geographic area and the geographic level used (e.g., census tract, census block group, ZIP code, etc.). Several studies surveyed by Fiscella and Fremont suggested GO is unreliable for identifying Hispanic and Asian/Pacific Islander cohorts, although accuracy depends on the distribution of race and ethnicity of the population in the areas assessed. In other words, estimates of race and ethnicity based on geolocation are less reliable in more heterogeneous regions.

It is also important to emphasize that GO is usually not sufficient for drawing conclusions about race and ethnicity at the individual level, but could be sufficiently accurate at the aggregate level, depending on the intended purpose of the imputation. In addition, the definitions of census block groups change over time, so it is important to geocode the data to be imputed with census block definitions consistent with the data used to develop the race/ethnicity distributions.

### 3.1.2 SURNAME ANALYSIS

Surname analysis (SA) infers race/ethnicity cohorts from surnames. Insofar as individuals with a particular surname belong almost exclusively to a particular race or ethnicity cohort, it is possible to identify membership in a cohort by using well-formulated surname dictionaries (Elliott, Fremont, Morrison, Pantoja, & Lurie, 2008).

There are several ways to impute cohorts based on names, such as the use of letter combinations, dictionaries of surnames, and combinations of first, middle, and last names. The U.S. Census Bureau produced Spanish surname lists for each decennial census from 1950 to 1990 (Word & Perkins Jr., 1996). Another surname list was created using the California Department of Public Health birth data (Pérez-Stable, Hiatt, Sabogal, & Otero-Sabogal, 1995). More recent versions of SA rely on a list of popular surnames provided by the U.S. Census Bureau.

Fiscella and Fremont (2006) found several studies suggesting SA produced reasonable predictions for Hispanic and Asian/Pacific Islander cohorts due to more distinctive surnames among these groups, but that it is less accurate for women and individuals with higher socioeconomic status, and no published data was available at the time for using SA to identify Black or white cohorts. Another study found that use of the 1990 Census Spanish surname list to predict whether an individual is Hispanic showed an overall sensitivity (true positive rate) of 79% and specificity (true negative rate) of 90% compared to self-reported ethnicity in a national sample (Perkins, 1993). However, sensitivity and specificity predictions using this method were, respectively, 82% and 92% for men versus 77% and 88% for women. The prevalence of individuals within a cohort in the data being analyzed also has a powerful effect on accuracy; sensitivity and specificity for imputations using the Spanish surname list ranged, respectively, from 88% and 96% in Texas to 34% and 37% in Vermont. For Asian surnames, studies (Lauderdale & Kestenbaum, 2000) showed sensitivities ranging from 74% for Vietnamese individuals to 29% for Filipino individuals when using a list of surnames derived from Social Security records.

Like GO, SA is not sufficient for drawing conclusions about race and ethnicity at the individual level but can be useful for analysis at an aggregate level.

### 3.1.3 CATEGORICAL SURNAME AND GEOCODING

Prior to the development of the Bayesian methods discussed in the next section, a non-Bayesian hybrid method, Categorical Surname and Geocoding (CSG), combined surname and geography information sequentially, first using surname lists to impute race and ethnicity for Asian and Hispanic cohorts, then using geographic distributions to impute either Black or white/other for individuals with surnames not

appearing on the Asian or Hispanic surname lists (Fiscella & Fremont, 2006). An important characteristic of CSG is that it does not produce probabilities as output, like later methods do, instead assigning a categorical imputed race or ethnicity.

## 3.2 BAYESIAN METHODS

### 3.2.1 BAYESIAN SURNAME GEOCODING

To overcome the limitations of the GO, SA, and CSG approaches, Bayesian Surname Geocoding (BSG) (Elliott, Fremont, Morrison, Pantoja, & Lurie, 2008) integrated cohort distributions by surname and geolocation from different datasets using Bayes's theorem. Bayes's theorem provides a framework for calculating conditional probabilities when joint probabilities are not available. Prior to the introduction of BSG, the preponderance of studies on quality of healthcare and patient outcomes were limited to Medicare or Medicaid patients with self-reported race/ethnicity data (Elliott, Fremont, Morrison, Pantoja, & Lurie, 2008).

Bayes's theorem is stated mathematically as the following equation:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

where A and B are events, and

- $P(A|B)$  is the conditional probability of event A given that B occurs
- $P(B|A)$  is the conditional probability of event B given that A occurs
- $P(A)$  is the prior probability of observing A (also known as the marginal probability)
- $P(B)$  is the prior probability of observing B (also known as the marginal probability)

For BSG, Bayes's theorem is applied as follows:

$$P(R|S) = \frac{P(S|R)P(R)}{P(S|R)P(R) + P(S|not R)P(not R)}$$

- Event R is belonging to a specific race/ethnicity cohort
- Event S is having a specific surname
- $P(R|S)$  is the conditional probability of belonging to a specific race/ethnicity cohort given a specific surname
- $P(R)$  is the probability of belonging to a specific race/ethnicity cohort based solely on geolocation
- $P(S|R)$  is the conditional probability of having a specific surname, given belonging to a specific race/ethnicity cohort
- $P(not R)$  is the probability of belonging to a race/ethnicity cohort other than R based solely on geolocation
- $P(S|not R)$  is the conditional probability of having a specific surname, given belonging to a race/ethnicity cohort other than R

BSG uses race/ethnicity composition by census block group as an estimate of  $P(R)$  and Asian and Hispanic surname lists to derive  $P(S|R)$  and  $P(S|not R)$ , which are used to update  $P(R)$  given the individual's surname. For example, using a four-cohort distribution (Asian, Hispanic, Black, and non-Hispanic white/other), and

given the following probabilities (all probabilities are for illustrative purposes only and not intended to represent an actual census block group or surname distribution):

- Census block group distribution:
  - $P(\text{Individual is Asian}) = 0.10$
  - $P(\text{Individual is Hispanic}) = 0.50$
  - $P(\text{Individual is Black}) = 0.20$
  - $P(\text{Individual is white/other}) = 0.20$
- Surname distribution:
  - $P(\text{Having an Asian surname} \mid \text{Individual is Asian}) = 0.515$
  - $P(\text{Having an Asian Surname} \mid \text{Individual is not Asian}) = 0.004$

Then, for an individual in this census block group with a surname on the Asian surname list,  $P(\text{Individual is Asian} \mid \text{Individual has an Asian surname}) = (0.10 * .515) / [(0.10 * 0.515) + (1-0.10) * 0.004] = 93.5\%$ . The process would be repeated to estimate  $P(\text{Individual is Hispanic} \mid \text{Individual has an Asian surname})$ , etc., and the final output of the method is the probability of an individual belonging to each of the four cohorts.

In the Elliott, Fremont et al. (2008) paper, the BSG algorithm used census block group distributions from the 2000 Census, the U.S. Census Bureau's Spanish Surname List, and the Lauderdale-Kestenbaum Asian Surname list (Lauderdale & Kestenbaum, 2000). Because conditional probabilities of surnames vary by gender, the BSG algorithm used gender-specific distributions by surname.

Elliott, Fremont et al. (2008) compared BSG to other imputation methods using commercial health plan data provided by the Aetna Health Insurance Company, with self-reported race and ethnicity for 1,973,362 enrollees. The methods used for comparison were GO and CSG. For the comparison, statistical correlations between imputed individual race/ethnicity probabilities and self-reported cohort were examined. BSG and CSG both performed better than GO, and BSG outperformed CSG for the Black cohort.

**Table 1**

**CORRELATION OF INDIVIDUAL PREDICTED RACE OR ETHNICITY WITH SELF-REPORTED RACE OR ETHNICITY**

	Hispanic	Asian	Black	White/other	Weighted Average
GO	0.49	0.34	0.57	0.55	0.53
CSG	0.77	0.65	0.48	0.63	0.63
BSG	0.79	0.67	0.61	0.70	0.70

Source: Elliott, Fremont et al., 2008

The accuracy of BSG is closely linked to the accuracy and completeness of the surname list. While BSG, at the time, improved the accuracy of imputing an individual's race and ethnicity compared to previous methods, it has since been eclipsed by improvements in the use of name data.

### 3.2.2 BAYESIAN IMPROVED SURNAME GEOCODING

Bayesian Improved Surname Geocoding (BISG) is like BSG but uses a different source for surname data and conditions the prior probability of race/ethnicity on surname instead of geolocation (Elliott, et al., 2009). As noted in the limitations for BSG, the accuracy is closely linked to the accuracy and completeness of the surname list. In 2007, the U.S. Census Bureau published a detailed list of surnames classified by self-reported race or ethnicity that appeared at least 100 times in the 2000 U.S. Census Bureau data, representing 89.8% of all individuals enumerated on Census 2000 (Word, Coleman, Nunziata, & Kominski, 2007).

With an improved surname list and more granular data, BISG expanded the number of imputed cohorts from four to six: Hispanic, Black, Asian/Pacific Islander, American Indian/Alaska Native (AI/AN), multiracial, and white. Proportions for the same six cohorts by census block group were again obtained from the 2000 Census. Again, the final output is the probability of an individual belonging to each of the six cohorts.

The BISG imputations were compared to BSG and other methods using the same commercial health plan data used in the BSG paper. Because the output of BSG in Elliott & Morrison (2009) included only four cohorts, a comparison for AI/AN and multiracial was not possible. The biggest improvements in accuracy, as measured by correlation, were observed for cohorts where BSG was the weakest (Asian/PI and Black).

**Table 2**

**CORRELATION OF INDIVIDUAL PREDICTED RACE OR ETHNICITY WITH SELF-REPORTED RACE OR ETHNICITY**

	Hispanic	Asian/PI	Black	AI/AN	Multiracial	White/other	Weighted Average
<b>BSG</b>	0.80	0.69	0.62	NA	NA	0.72	0.70
<b>BISG</b>	0.82	0.77	0.70	0.11	0.02	0.76	0.76

Source: Elliott, Morrison et al., 2009

In addition, the Consumer Financial Protection Bureau (CFPB) assessed BISG in 2014 using self-reported race/ethnicity data for a sample of mortgage applications in 2011 and 2012 (CFPB, 2014). The CFPB study used the same U.S. Census Bureau surname list as the Elliott, Morrison et al. (2009) study and geographic distributions from the 2010 Census with very similar results.

**Table 3**

**CORRELATION BETWEEN PROXY PROBABILITY AND REPORTED RACE AND ETHNICITY**

	Hispanic	Asian	Black	AI/AN	Multiracial	White/other
<b>BISG</b>	0.81	0.83	0.70	0.06	0.05	0.77

Source: CFPB, 2014

While BISG is an improvement in accuracy over BSG using these metrics, it still has limitations, including:

- The exact counts for infrequent surnames in some cohorts are not available or are suppressed altogether due to privacy concerns; these omitted or suppressed surnames may impact less represented race/ethnicity cohorts more than others.
- The methodology performs poorly for identifying self-reported American Indian/Alaska Native and multiracial individuals.
- Despite its superiority relative to other alternatives, BISG is still subject to significant bias and estimation error and may result in overstated disparities in studies of mortgage lending outcomes (Baines & Courchane, 2014). For instance, Baines and Courchane note that, as an individual's FICO score and income increases, BISG is less able to accurately identify Black and Hispanic individuals.
- While noting that BISG probabilities may be relatively less inaccurate than GO and SA, Baines and Courchane (2014) raised concerns about the "objectively high" error rates of the method and how error rates can vary by population. In their results, BISG correctly identified 24.2% of Black consumers, while the CFPB report correctly identified 39% of Black consumers.

While BISG was first developed on a surname list from the 2000 Census, in 2016, the U.S. Census Bureau released an updated list based on the 2010 Census (Comenetz, 2016), which can also be used with BISG.

The updated surname list covers 90.1% of people with surnames recorded in the 2010 Census and introduced edits to address compound/hyphenated names, suffixes and prefixes, and other items.<sup>5</sup>

### 3.2.3 MEDICARE BAYESIAN IMPROVED SURNAME GEOCODING

Medicare Bayesian Improved Surname Geocoding (MBISG) was introduced in 2013 as an improvement to BISG by incorporating data from the Centers for Medicare and Medicaid Services (CMS) (Martino, et al., 2013). CMS's race/ethnicity information for Medicare beneficiaries is primarily derived from self-reported race information collected by the Social Security Administration (SSA); however, this data is limited. For example, for persons assigned a Social Security number (SSN) before 1980, there were only three race/ethnicity response options: Black, white, or other (Haas, et al., 2019). MBISG was developed to supplement the CMS information.

In MBISG, each cohort on the SSA data is associated with a distribution of self-reported race or ethnicity from a large, nationally representative survey of Medicare beneficiaries. These probabilities are combined with BISG probabilities from Elliott, Fremont et al. (2008) using a Bayesian method like the method used within BISG to produce MBISG probabilities (Haas, et al., 2019).

In 2019, an update to MBISG (MBISG 2.0) was proposed with multiple improvements to the data and an improvement to the calibration of the output (Haas, et al., 2019). The first set of improvements incorporated into MBISG 2.0 were targeted at making better use of the data elements used in BISG: surnames, addresses, and race and ethnicity data. The use of surnames was improved by increasing the ability to match individuals with compound names, or surnames formed by a combination of component surnames. Improvement in the use of race/ethnicity data came from stratifying by age because the CMS distributions by race and ethnicity vary by age group. The last proposed improvement to the input data was to include whether an individual is a resident of Puerto Rico in developing the estimate. In MBISG 1.0, only surname and SSA race or ethnicity were used, which underestimated the prevalence of Hispanic beneficiaries in Puerto Rico. The final improvement proposed for MBISG 2.0 was to calibrate the predicted probabilities to match the probabilities in the validation sample, which addressed the undercounting of Hispanic and multiracial beneficiaries. Multinomial logistic regression using additional types of predictors of race/ethnicity<sup>6</sup> was used to improve predictions and calibrate all the probabilities to match the sample.<sup>7</sup> MBISG 2.0 saw improvement in correlation across all race/ethnicity cohorts in comparison to MBISG 1.0, with the largest improvements for the white, Hispanic, and Asian/Pacific Islander cohorts.

Because MBISG is calibrated to a Medicare population, it may not be appropriate for other settings. It requires an initial value for the SSA cohort, which may not be available for many datasets. While MBISG improves the imputations for American Indian/Alaska Native and multiracial beneficiaries, the quality of the predictions for these cohorts remains low, which can add to the inherent uncertainty in an analysis that incorporates these probabilities.

---

<sup>5</sup> See <https://www2.census.gov/topics/genealogy/2010surnames/surnames.pdf>.

<sup>6</sup> Additional predictors included variables for first names, Spanish preference, demographics, and coverage types.

<sup>7</sup> In MBISG 1.0, the imputed probabilities underestimated the proportion of the sample who were Hispanic and multiracial and overestimated the proportion who were white. In MBISG 2.0, an additive approach was used to calibrate the proportion multiracial, and multinomial logistic regression was used to adjust the mean probabilities of the predictions and match the distribution of self-reported race/ethnicity in the sample.

### 3.2.4 BAYESIAN IMPROVED SURNAME GEOCODING EXTENSIONS

Another technique to improve BISG imputation accuracy is to include other information, or covariates, into the race/ethnicity predictions (Imai & Khanna, 2016). In these BISG extensions, the imputation algorithms make use of additional individual-level variables specific to the dataset used for calibration.

For example, to impute race and ethnicity for a Florida voter file, Imai and Khanna (2016) considered covariates for age, gender, and political party registration, which were available in the dataset. Imai and Khanna extended the Bayesian estimation to include the conditional distributions of party registration by race and ethnicity obtained from Gallup polling data. Like BISG, the output is a probability of belonging to a cohort, but with five race and ethnicity cohorts: white, Black, Hispanic, Asian/Pacific Islander, and other, which includes the American Indian/Alaska Native cohort.

Another example of incorporating additional covariates in the Bayesian imputation is a method developed by the U.S. Department of the Treasury to impute race and ethnicity for tax filers for tax policy analysis (Fisher, 2023). The method improves upon BISG estimates, using conditional distributions of race/ethnicity by other variables available on tax forms, such as filing status, age, number of dependents, and gender.

The general idea of using additional information to predict race and ethnicity is a natural extension of the Bayesian methods. While there is potential for improvement in accuracy by incorporating additional information, the main downsides are that 1) additional input variables (and sources for distributions of those variables by race and ethnicity) are required to generalize the method to new data sources, and 2) the additional risk of bias in the imputation if the distributions in the prior data used to calibrate the imputations are significantly different than the population to which the imputation is being applied.

### 3.2.5 BAYESIAN IMPROVED FIRST NAME SURNAME GEOCODING

Bayesian Improved First Name Surname Geocoding (BIFSG) builds upon the BISG by adding first name as a feature that is used in the algorithm in conjunction with surname and geolocation (Voicu, 2018). Like BISG, the output is a probability that can be converted to a classification with six race/ethnicity cohorts (Hispanic, Asian/Pacific Islander, Black, American Indian/Alaska Native, multiracial, and white/other).

The formal equation for calculating this probability is:

$$P(R|G, S, F) = \frac{P(R|S)P(G|R)P(F|R)}{\sum P(R|S)P(G|R)P(F|R)}$$

where:  $P(R|G, S, F)$  is the updated (posterior) probability of having self-reported race/ethnicity R based on geolocation G, surname S, and first name F (Voicu, 2018).

To develop BIFSG, Voicu (2018) used the U.S. Census Bureau surname list and race/ethnicity distributions by census block group based on the 2010 Census.<sup>8</sup> In addition, Voicu (2018) used a list of first names classified by self-reported race and ethnicity drawn from mortgage applications (Tzioumis, 2017).<sup>9</sup>

Voicu (2018) tested BIFSG using proprietary databases of mortgage transactions from multiple lenders between 2012 and 2014, a combined dataset of 279,404 applications. While BIFSG demonstrated improved accuracy on all, the largest improvement was observed for the Black cohort. For the Hispanic and

<sup>8</sup> This list is publicly available at [https://www.census.gov/topics/population/genealogy/data/2010\\_surnames.html](https://www.census.gov/topics/population/genealogy/data/2010_surnames.html).

<sup>9</sup> This list is publicly available at <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/TYJKEZ>.

Asian/Pacific Islander cohorts, surname is already highly predictive, which does not leave much room for improvement with the addition of first name.

**Table 4**

**CORRELATION BETWEEN PROXY PROBABILITY AND SELF-REPORTED RACE AND ETHNICITY (VOICU, 2018)**

	Hispanic	Asian/PI	Black	White/other	Weighted Average
<b>BISG</b>	0.87	0.86	0.71	0.82	0.82
<b>BIFSG</b>	0.88	0.87	0.75	0.84	0.84

Voicu further noted that the improvements for the Black cohort were greatest in geographic areas with greater racial and ethnic diversity. Accuracy differences by gender observed using surname only were reduced when first name was included in the imputation.

It is important to note that, given the variation in rates of home ownership in the United States by race and ethnicity, first names more common to non-white race/ethnicity groups may be underrepresented in the first name data, which was derived from home mortgage applications (Sorbero, Euller, Kofner, & Elliott, 2022).

### 3.2.6 MODIFIED BAYESIAN IMPROVED FIRST NAME SURNAME GEOCODING

In 2022, Sorbero et al. constructed a modified version of BIFSG using the U.S. Census Bureau surname list, demographic distributions by census block group based on the 2010 Census, and first name data from the sample of mortgage applicants. In addition, Sorbero et al. (2022) incorporated additional refinements for compound and rare surnames and recalibrated the output to match the distributions in the self-reported race/ethnicity data on the dataset being analyzed. Race and ethnicity were imputed using a six-level cohort system of probabilities, like BISG.

Sorbero et al. (2022) examined the performance of modified BIFSG on an extract of data from the Multidimensional Insurance Data Analytic System (MIDAS), which contains records at a person-year level for enrollees in health plans offered by Federally Facilitated Marketplaces and purchased through HealthCare.gov or state-based marketplaces.<sup>10</sup> Like BIFSG, modified BIFSG was able to differentiate between Asian/Pacific Islander, Black, Hispanic, and white cohorts reasonably well, but accurately identifying American Indian/Alaska Native and multiracial individuals remained challenging. Performance was also examined by age and region. No comparisons were made to other methods, but Sorbero et al. (2022) examined results by age group and U.S. Census Bureau divisions, which partition the country into nine regions. Accuracy was lower for children and young adults than for older enrollees and varied by census division. The MIDAS extract did not contain information on gender to assess accuracy by gender.

Modified BIFSG has the same limitations as BIFSG regarding representation in the first name data. While the recalibration step helps make the imputed results better reflect the observed data, it assumes that the distribution of the non-reporters is similar to the reporters (rather than all residents in the census block group). Also, accuracy levels for American Indian/Alaska Native and multiracial cohorts remain low.

---

<sup>10</sup> For more details, see [https://www.hhs.gov/sites/default/files/CMS-MIDAS\\_remediated.pdf](https://www.hhs.gov/sites/default/files/CMS-MIDAS_remediated.pdf).

### 3.2.7 FULLY BAYESIAN IMPROVED SURNAME GEOCODING

Fully Bayesian Improved Surname Geocoding (fBISG) seeks to address a couple of recurrent limitations described in the previous methods (Imai, Olivella, & Rosenman, 2022). First, in the publicly available census data, the U.S. Census Bureau will state a value of zero instead of small counts due to privacy concerns. This does not mean there are zero individuals in a race/ethnicity cohort in a geolocation, just that there are very few; fBISG uses a measurement error model so that zero values mean low probability instead of nonexistence.<sup>11</sup> Second, fBISG addresses the problem of missing surnames in the Census Bureau surname list by supplementing the surname list with additional data from voter files from six Southern states. While the Census Bureau indicates its surname lists cover 90% of the population, the missing 10% may also disproportionately impact non-white race/ethnicity cohorts. Despite the name, fBISG also includes extensions of BISG to incorporate first and middle names like BIFSG. The method is available in an open-source package called wru (“Who Are You”), which is available for public use.

Like BISG, fBISG uses the U.S. Census Bureau surname list and demographic distributions by census block group based on the 2010 Census. However instead of mortgage data, Imai et al. built first and middle name lists using data from voter files from six Southern states. This same data is used to supplement the Census Bureau surname files. (Users of the wru package can import the latest available Census Bureau datasets.) The output of the method is a probability of belonging to one of five race/ethnicity cohorts: white, Black, Hispanic, Asian, and other.

Imai, Olivella, and Rosenman (2022) examined the performance of fBISG on the six-state voter data, which contained 37.8 million voter records. The study used AUC as an overall performance metric.

**Table 5**  
**AUC BY METHODOLOGY**

	Area under ROC				
	Hispanic	Asian	Black	White	Other
<b>BISG</b>	0.92	0.82	0.92	0.90	0.59
<b>fBISG with zero-count correction</b>	0.96	0.91	0.94	0.91	0.57
<b>fBISG with additional surname data</b>	0.96	0.91	0.96	0.91	0.58
<b>fBISG with first name</b>	0.97	0.93	0.97	0.94	0.61
<b>fBISG with first and middle name</b>	0.98	0.94	0.98	0.95	0.62

Source: (Imai, Olivella, & Rosenman, 2022).

The performance gains from the zero-count correction were greatest for the Hispanic and Asian cohorts. Imai, Olivella and Rosenman (2022) note that, in the voter dataset, a fifth of Asian voters live in census blocks where the 2010 Census indicated there were no Asian residents, and the proportion varies significantly by state. The expanded name data (surname, first name, middle name) also improved performance.

The fBISG methodology incorporates first and middle name into the prediction to get the full improvement in the results. However, because first and middle names may be more difficult to obtain, the full benefits of the algorithm may not be realized in practice. Another limitation is the possibility of regional bias, as some names may be more prevalent in some regions for certain race/ethnicity cohorts.

<sup>11</sup> The Census Bureau counts by race/ethnicity are treated as a draw from an unknown multinomial distribution.



### 3.2.8 BAYESIAN INSTRUMENTAL REGRESSION FOR DISPARITY ESTIMATION

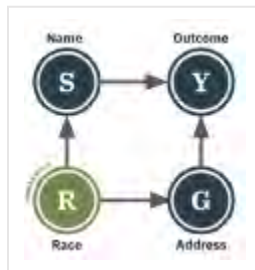
Imputation methods are often used to study racial disparities in contexts where direct data are not available, such as healthcare and financial services. While the methods described above tend to produce well-calibrated predicted probabilities of race and ethnicity, the errors of these methods are often correlated with the outcomes of interest, potentially leading to biased estimates of the race and ethnicity disparities being studied. McCartan et al. (2023) address this problem by introducing a class of models called Bayesian Instrumental Regression for Disparity Estimation (BIRDIE), which use the outcome variable and the predicted probabilities from a Bayesian imputation method (e.g., BISG) to produce less biased estimates of racial disparities in the outcome of interest, along with an updated set of imputed probabilities. These updated probabilities, which incorporate the outcome variable, are also more accurate than the initial probabilities, which were based only on surnames and location.

As McCartan et al. demonstrate, BISG and its extensions tend to produce accurate and well-calibrated estimates of racial probabilities even when assumptions of the BISG methodology do not hold exactly. Thus, if the purpose of the analysis is only to impute probabilities using surname, geolocation, and other covariates, BISG (and its extensions) can be a good option. However, imputed probabilities are often used to weight the outcome data to compare average outcomes. For example, if  $Y$  is the outcome and  $R$  is a specific race cohort:

$$E[Y | R] = \frac{\sum_i Y_i * Prob(R)_i}{\sum_i Prob(R)_i}$$

For example, the average outcome  $Y$  for race/ethnicity cohort Hispanic would be calculated by taking the weighted average of the outcome for all individuals, where the weight for each individual is the estimated probability associated with Hispanic.

If one uses BISG probabilities for the weights, the estimates are biased unless the effect of race is fully mediated by the name and geolocation. That is:



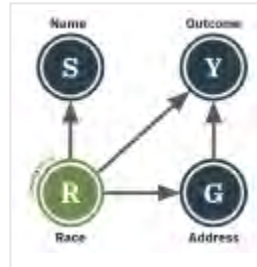
Source: Cory McCartan<sup>12</sup>

This assumption requires that race be conditionally independent of the outcome, given surname and geographic location. The reasonability of this assumption depends on the setting but may often be violated given that race and ethnicity affect many aspects of society besides geolocation and name. For example, say the outcome of interest is voter political party preference. The conditional independence assumption

<sup>12</sup>BIRDIE. BIRDIE: Estimating disparities when race is not observed. Retrieved January 15, 2024, from <https://corymccartan.com/birdie/#:~:text=Bayesian%20Instrumental%20Regression%20for%20Disparity%20Estimation%20%28BIRDIE%29%20i,s,described%20in%20McCartan%2C%20Goldin%2C%20Ho%20and%20Imai%20%282022%29.>

would require that two individuals in the same geographic area, with the same surname, are equally likely to register with the Democratic party, regardless of race.

The BIRDIE methodology addresses this problem by relying on a different assumption: that names are independent of outcomes, conditional on location and race.



Source: Cory McCartan<sup>13</sup>

With this assumption, surname is predictive of the outcome only through its ability to predict race. In other words, after accounting for race, surname has no predictive power of the outcome.

BIRDIE models the outcome variable with fixed effects or mixed-effects regression models.<sup>14</sup> BIRDIE uses BISG or fBISG probabilities and other covariates as inputs and, in the process, produces updated probability estimates that can be more accurate than the input probabilities. BIRDIE models are constructed as follows:

- Generate a set of BISG probability estimates, which are conditional on surname, geolocation, and other observed characteristics.
- Assume a priori that the distribution of the outcome is the same for each race.
- Specify a model of the outcome as a function of race, geolocation, other observed characteristics, and the outcome parameter.
- Compute the posterior distribution of the outcome using the prior distribution, the complete-data outcome model, and the BISG probabilities.

To demonstrate, McCartan et al. applied BIRDIE and other methods to North Carolina voter data to measure the disparity in Democratic party registration between Black and white voters. This dataset includes self-reported individual race, allowing the authors to validate and compare the results of the different methodologies. In this data, the true difference in Democratic party registration was 54.6 percentage points, with Black voters registering Democratic at a much higher rate than white voters. Using BISG probability estimates produced an estimated difference of 30.5 percentage points. The BIRDIE models produced an estimated difference of 49.2 percentage points, much closer to the true value.

<sup>13</sup>BIRDIE. BIRDIE: Estimating disparities when race is not observed. Retrieved January 15, 2024, from <https://corymccartan.com/birdie/#:~:text=Bayesian%20Instrumental%20Regression%20for%20Disparity%20Estimation%20%28BIRDIE%29%20is%20described%20in%20McCartan%2C%20Goldin%2C%20Ho%20and%20Imai%20%282022%29.>

<sup>14</sup> The authors present three alternative model specifications with varying levels of modeling flexibility:

- Complete-pooling model: Estimates a single relationship between outcome and race that does not vary with geolocation or other observed characteristics.
- Saturated (no-pooling) model: Estimates a different relationship for every level of geolocation and other observed characteristics.
- General mixed-effects model: A compromise between the complete-pooling and no-pooling models that maintains the flexibility of the saturated model while allowing information to be shared across levels. This is the method recommended by the authors for general use.

McCartan et al. also examined the updated racial probability estimates produced by the BIRDIE models and used AUC to compare them to the initial BSG estimates. They found that the BIRDIE estimates were more accurate for white and Black voters, about the same for Hispanic, Asian, and Native American voters, and slightly less accurate for “other” voters, compared to the BSG estimates.

The BIRDIE methodology can be implemented using the “birdie” R package. It includes functionality to compute the initial BSG probability estimates and specify the outcome model, as well as extract the updated probabilities.

### 3.3 PREDICTIVE MODELING METHODS

To date, Bayesian methods for imputing race and ethnicity have been the most widely used, but there are other types of algorithms that can be applied to imputation. While these methods show promise, the main limitation is that at least a subset of the data requires race/ethnicity classification to train the model. Additionally, many require additional inputs that limit their applicability or have not shown significant improvement in predictions.

#### 3.3.1 REGRESSION

The use of regression models for missing data imputation is not specific to race/ethnicity imputation, as regression models are a common technique for imputation of missing values where there is a correlation between the variable with the missing values and other variables in the dataset. In the case of imputing race and ethnicity with a regression model, one would use multinomial logistic regression. To implement regression imputation requires race/ethnicity information and other variables on the same dataset. Two examples of using regression to impute race and ethnicity can be found in Xue, Harel, and Aseltine (2019) and Zavez, Harel, and Aseltine (2022).

Xue, Harel, and Aseltine (2019) developed regression models to impute race and ethnicity using birth records from the Connecticut Department of Public Health from 2009 to 2013. Each birth record contained a self-reported race or ethnicity of the mother, as well as an insurance type,<sup>15</sup> an indicator for father missing, and the mother’s age at delivery. Race and ethnicity distributions by census tract and surname were added from the 2010 Census data. Two multinomial logistic regression models were fit: one with variables for census tract and surname race/ethnicity percentages, and a second adding covariates for insurance type, a dummy variable for father missing, and the mother’s age. The models were fit on 5% of the available training data and imputed four race/ethnicity cohorts: white, Black, Hispanic, and other.

The regression models were compared to BSG imputations using 2010 Census data. The regression models had an 81% accuracy of predictions compared to 78% for BSG. The models had much better sensitivity for the Black cohort, 60% to 63%, compared to 39% for BSG. The improvement in sensitivity for the Hispanic cohort was more modest, 71% to 72% compared to 66% for BSG. The sensitivity for the white cohort was slightly worse, 91% versus 93% for BSG.

Zavez, Harel, and Aseltine (2022) also constructed regression models on Connecticut birth records using similar covariates (race/ethnicity percentages for the census tract, insurance type, sex, and age). However, Zavez developed Connecticut-specific first name and surname lists using the birth registry records instead of the census data. Zavez, Harel, and Aseltine also attempted to move the geocoding level from census

---

<sup>15</sup> Insurance types include private insurance, self-/no insurance, Medicaid, and other insurance.

tract to census block, which is a smaller area, but the request was denied by the Census Bureau. The study assumed that 50% of race/ethnicity information was missing at random in the data and used five race/ethnicity cohorts: white, Black, Asian/Pacific Islander, Hispanic, and other.

The models were also validated on Connecticut hospitalization claim data from 2012 to 2017, which had more than 10 million records, with only 1% of self-reported ethnicity and 2% of self-reported race data missing. The regression models with the state-specific name lists performed better than a more generalized version using census data.

Again, to train a regression model for imputation, all the covariates and self-reported race and ethnicity data need to be available on a portion of the dataset, which is not always feasible. Also, it is not clear how much of the accuracy gains observed in these studies were from the use of Connecticut-specific data or the inclusion of other covariates versus the method itself (e.g., regression vs. Bayesian imputation).

### 3.3.2 NATURAL LANGUAGE PROCESSING

The natural language processing (NLP) framework provides another method that can be used to impute race and ethnicity cohorts from names. These methods train a model where the individual's name is the input, and their race/ethnicity cohort probability is the output. Xie (2021) fit an NLP model on a Florida voter registration dataset. Names for Native American and multiracial cohorts were dropped because of lack of data, so a four-level cohort grouping was modeled (Hispanic, Black, Asian/Pacific Islander, and white). An initial large model was trained using Bidirectional Long Short-Term Memory (BiLSTM) architecture, then a "distillation" step was applied to compress the model and obtain a smaller model with fewer parameters and layers for production. Xie also segmented the data by gender because gender is associated with race and ethnicity as well as name. Xie did not do any performance comparisons to other types of models. This model is available in the R package, *rethnicity*.

Chintalapati et al. (2023) also fit long short-term memory (LSTM) models on the Florida voter registration dataset used in Xie (2021). These models are available in the Python package, *ethnicolr*.

### 3.3.3 OTHER MACHINE LEARNING ALGORITHMS

Decter-Frain (2022) investigated whether machine learning (ML) algorithms would be an improved framework over the Bayesian frameworks. They used surnames, first names, middle names, and geocoded locations as input data and tested multiple machine learning methods: multinomial regression, multinomial regression with elastic net penalty, random forests, and gradient boosted decision trees.

The models were trained on data from voter registration records from California, Florida, Georgia, and North Carolina. The data contained over 26 million records, over 5% of the total electorate, but is not a representative sample. Race and ethnicity were self-reported, but only populated for 20% of the California data. The models imputed four race/ethnicity cohorts (white, Black, Hispanic, and Asian) and were trained in a cross-validation manner where the folds were defined as the states, training the model on three states, and testing on the fourth.

The researchers fit two sets of models with different predictor sets. The first set used only 10 probabilistic inputs associated with surname and location and was compared to BISG. When using the minimal set of input variables, the comparison of ML methods to BISG was mixed. The ML methods performed better on AUC for some states and cohorts, but not as well for others.

The second set of models included first and middle name inputs as well. For the models with first and middle name included, the ML models consistently outperformed a Bayesian imputation algorithm for

many groups, based on the AUC metric. The results were similar across the various ML methods and states tested.

**Table 6**  
**AUC BY MODEL FOR CALIFORNIA**

Model	Area Under ROC				
	Hispanic	Asian	Black	White	Other
BISG (+First+Middle)	0.929	0.911	0.900	0.911	0.547
Logistic	0.940	0.949	0.942	0.921	0.556
ElasticNet	0.938	0.951	0.941	0.921	0.554
Random Forest	0.937	0.951	0.943	0.920	0.537
Gradient Boosted Trees	0.938	0.942	0.946	0.917	0.546

Source: Decter-Frain, 2022

In addition, Decter-Frain examined calibration curves by method and found the ML models to be better calibrated than BISG, particularly for Asian and Hispanic voters. The one exception where BISG appeared to be better calibrated was for Asian voters in California. In Florida, BISG was over-calibrated for Hispanic voters and ML methods were under-calibrated. These differences were attributed to differences in the overall population distributions by state.

Chintalapati et al. (2023) fit ML models on the same Florida voter registration dataset used by Xie (2021). However, other types of models were investigated: K-nearest neighbors, random forests, gradient boosted decision trees, and NLP models. The models predict five cohorts: white, Black, Hispanic, Asian/Pacific Islander, and unknown. Multiracial and Native American cohorts were combined with unknown due to small sample size. These models are available in the Python package, *ethnicolr*.

When predicting race and ethnicity with just surnames, the NLP model had the best accuracy, at 81%, and was the top performer for most cohorts. Although performance was poorer for Asian and Black cohorts compared to the others, the addition of first name improved accuracy for these cohorts. Accuracy measures by cohort for K-nearest neighbors were not reported, and no comparisons were made to other imputation methods.

**Table 7**  
**ACCURACY BY MODEL**

Inputs	Model	Accuracy on Hold-out Data					Overall
		Hispanic	Asian	Black	White	Other	
Surname Only	K-Nearest Neighbors	*	*	*	*	*	0.78
	Random Forest	0.66	0.05	0.19	0.70	0.17	0.55
	Gradient Boosted Trees	0.80	0.07	0.12	0.93	0.01	0.75
	NLP	0.84	0.40	0.50	0.91	0.04	0.81
Full Name	K-Nearest Neighbors	*	*	*	*	*	0.73
	Random Forest	0.66	0.22	0.32	0.89	0.03	0.71
	Gradient Boosted Trees	0.22	0.04	0.01	0.98	0.00	0.68
	NLP	0.86	0.63	0.74	0.92	0.07	0.85

\* Accuracy not reported. Source: Chintalapati et al., 2023

Machine learning methods can have a danger of overfitting to a specific dataset and may not be generalizable to broader populations. Decter-Frain (2022) used a conservative out-of-state evaluation strategy, but overfitting is a consistent concern for ML models. The authors suggest that testing multiple methods on a particular dataset may lead to identifying the best method for that particular dataset.

## Section 4: Discussion of Direct Data

Direct data is data collected directly from a primary source. In the context of this paper, direct data refers to self-reported individual race/ethnicity information. One method to obtain direct data is by asking insureds to self-report it, but there are still several challenges and considerations to this approach:

- Legal
  - Historically, it was uncertain whether it was legal to collect on health plan participants (Fiscella & Fremont, 2006).
  - Some state laws would need to be changed to allow insurance companies to collect this data (American Academy of Actuaries, 2022).
- Procedural
  - Even if you start collecting direct race/ethnicity data on new data, you won't have it on previously collected data (Fiscella & Fremont, 2006).
  - Collection can be slow (Elliot et al., 2008).
  - It requires business process changes that would likely result in additional costs that would be passed on to consumers (American Academy of Actuaries, 2022).
- Accuracy and completeness
  - Consumers may decline to participate and/or provide inaccurate information because they may suspect this information could affect their insurance coverage or invade their privacy (American Academy of Actuaries, 2022).
  - Refusal to respond might depend on an individual's race or ethnicity (Fiscella & Fremont, 2006). Voicu (2018) offers evidence that non-reporting of race/ethnicity information is correlated with actual race or ethnicity.
- Social
  - Fear that consumers would assume data was being misused (Fiscella & Fremont, 2006).
  - Racial categories have changed over time and how people self-report can depend on the options presented to them and can change over time.

Another approach is to match self-reported race from external, third-party datasets to the user's dataset using personally identifiable information (PII). There are many vendors of consumer demographic data. However, considerations for this approach include:

- How much data is imputed versus self-reported, and what methods are used for imputation.
- The sources of the self-reported data.
- Is the self-reported data collected in an unbiased setting? (American Academy of Actuaries, 2022)
- Restrictions on use required by the entities collecting the data, for example, there are laws restricting the use of data obtained by/from governmental entities. (American Academy of Actuaries, 2022)
- What categories for race and ethnicity are being captured, and are they appropriate for the intended use? (American Academy of Actuaries, 2022)
- Procedures in place to protect and manage the PII used to do the matching. (American Academy of Actuaries, 2022)
- Ethical and privacy concerns; what if an individual does not want their race or ethnicity disclosed? Is there an option for "prefer not to answer," and will values be imputed for these individuals?<sup>16</sup>

---

<sup>16</sup> <https://www.rti.org/insights/imputing-raceethnicity-part-1#:~:text=Ethically%2C%20we%20should%20be%20concerned%20about%20filling%20in,not%20to%20answer%20is%20a%20valid%20response%20category.>

Self-reported data is collected in several different settings, including healthcare, mortgage lending, and voter registration (in some states). While these datasets can be useful for developing and/or testing imputation methods, they are not generally available to match to other data and focus on particular subsets of the population. Another limitation to completeness is that many individuals may decline to self-report.

The U.S. Census Bureau and the Social Security Administration have the most comprehensive direct data on race and ethnicity. However, the census has a 72-year restriction on individual records (for example, the 1950 Census data was released in 2022), so detailed data is dated, and more recent data is not publicly accessible. Restricted-use government datasets are available to qualified researchers at government agencies, academic institutions, and other entities with approved projects. For example, a 2007 Federal Trade Commission study of the use of credit-based insurance scores in auto insurance obtained race/ethnicity data from the Social Security Administration (Federal Trade Commission, 2007).

## Section 5: Case Study: Imputation Methods for Race and Ethnicity

This section discusses the results of using several imputation methods on actual data. This dataset contained names and addresses, along with self-reported race or ethnicity, so that it could be used to run algorithms using only name and address information and compare the predictions with self-reported race or ethnicity to compare accuracy. This section describes key attributes of the dataset, presents methods for evaluating performance of the imputation methods, and highlights outcomes of the performance comparison.

### 5.1 DATASET AND ALGORITHMS

The authors used a proprietary health insurance dataset, which contained data on approximately 1.4 million lives with insurance coverage in one U.S. state. The geographically concentrated and homogenous coverage nature of the dataset should be kept in mind when reviewing the reported outcomes, because performance may differ materially when considering data from other regions or more diverse populations. The data contained basic demographics, including given name, surname, age, gender, and home address, in addition to self-reported race or ethnicity. The available values of self-reported race or ethnicity were white, Black, Hispanic, Asian/Pacific Islander (API), American Indian/Alaska Native (AI/AN), multiracial, and unknown/other. There are 751,485 distinct lives in the dataset with self-reported race or ethnicity (i.e., excluding values of unknown/other).

Table 8

DATASET DISTRIBUTION BY RACE/ETHNICITY

Self-Reported Race/Ethnicity	Lives	Distribution
Unknown/Other	602,793	44.5%
White	353,359	26.1%
Black	332,533	24.6%
Hispanic	37,526	2.8%
Multiracial	14,587	1.1%
API	8,806	0.7%
AI/AN	4,674	0.3%

Table 9

DATASET DISTRIBUTION OF RACE/ETHNICITY BY AGE

Self-Reported Race/Ethnicity	0-17	18-39	40-64	65+
<b>Count of Individuals</b>				
Unknown/Other	357,782	140,730	77,758	26,523
White	119,338	127,787	77,049	29,185
Black	116,006	124,544	61,182	30,800
Hispanic	21,333	11,696	3,258	1,239
Multiracial	8,348	5,080	1,040	119
API	3,271	2,947	1,645	943
AI/AN	1,797	1,793	898	186
<b>Distribution</b>				
Unknown/Other	57.0%	33.9%	34.9%	29.8%
White	19.0%	30.8%	34.6%	32.8%
Black	18.5%	30.0%	27.5%	34.6%
Hispanic	3.4%	2.8%	1.5%	1.4%
Multiracial	1.3%	1.2%	0.5%	0.1%
API	0.5%	0.7%	0.7%	1.1%
AI/AN	0.3%	0.4%	0.4%	0.2%



The following algorithms were used to impute race or ethnicity for all individuals on the dataset using the Surgeo Python package: first name only (FO), surname analysis (SA), geocoding only (GO), Bayesian Improved Surname Geocoding (BISG), and Bayesian Improved First Name Surname Geocoding (BIFSG). See section 6 below for a technical tutorial demonstrating how to use Surgeo and other imputation packages.<sup>17</sup> Surgeo was used as it provided an off-the-shelf implementation of all five algorithms. Finally, to analyze a method that maximizes the number of individuals the authors were able to assign imputed race/ethnicity probabilities for, the authors added a sixth method that used a hierarchical approach in which the authors assign each individual probability based on the following hierarchy, where individuals only move to the next step in the hierarchy if they are not assigned probabilities at each step:

1. Bayesian Improved First Name Surname Geocoding (BIFSG)
2. Bayesian Improved Surname Geocoding (BISG)
3. Surname analysis (SA)
4. Geocoding only (GO)
5. First name only (FO)

In other words, each individual is first run through the BIFSG algorithm; if they are not assigned an imputed race/ethnicity probability using this algorithm, then they are run through BISG, etc. Therefore, only individuals who did not receive imputed race/ethnicity probabilities from all five algorithms will not receive imputation probabilities. Information on this method is shown in the “hierarchy” column in the tables throughout this section.

## 5.2 DATA CLEANING AND PREPROCESSING

Cleaning data can be an important step in running the imputation algorithms, given the wide degree of variation in individual names. In particular, individuals with compound or hyphenated surnames or given names are less likely to be assigned probabilities by algorithms using such information because these names are less likely to be present in the lists used by the packages. However, the Surgeo package the authors used to impute race and ethnicity for the case study has limited functionality for preprocessing the data for compound or hyphenated names. Therefore, the authors began by running the algorithms on data with no preprocessing other than the basic data cleaning included in the Surgeo package, such as removing spaces or punctuation; the authors did not perform additional data cleaning other than ensuring all input was formatted correctly.

The individuals who are not assigned imputed probabilities include those where information needed to run the algorithm was missing or invalid (e.g., when an address could not be geocoded, or a first name was missing). For the remainder of this section, the authors will use the term “cohort” to refer to the group of individuals with a specific self-reported race or ethnicity who were able to be assigned probabilities by the algorithm under discussion (i.e., who had the necessary data to be run through the algorithm and who did not receive an imputed race of “unknown”).

---

<sup>17</sup> Note that the BIRDIE algorithm was not included in the case study due to lack of an outcome variable to use in training; however, we have included BIRDIE in the tutorial in section 6.

**Table 10**  
**PERCENTAGE OF INDIVIDUALS ASSIGNED PROBABILITIES BEFORE PREPROCESSING**

Self-Reported Race/Ethnicity	FO	GO	SA	BISG	BIFSG	Hierarchy
Unknown/Other	57%	99%	88%	87%	49%	>99%
White	79%	99%	93%	92%	73%	>99%
Black	46%	>99%	96%	95%	44%	>99%
Hispanic	74%	>99%	58%	58%	43%	>99%
Multiracial	53%	98%	91%	90%	48%	>99%
API	53%	99%	86%	85%	46%	>99%
AI/AN	65%	>99%	91%	90%	59%	>99%

As expected, a high proportion of individuals was not assigned probabilities when using the BISG and BIFSG methods without the preprocessing of data, and many of these exclusions were due to individuals with compound or hyphenated names. Therefore, to increase the likelihood of matching these compound and hyphenated names, the authors created multiple records for each individual using the individual pieces of the compound or hyphenated name applied, used the algorithms to impute race/ethnicity probabilities for each record, and then averaged the various probabilities.

As an example, consider an individual John Smith-Doe. To assign imputed race/ethnicity probabilities, the authors would use the name John Smith and the name John Doe and average the probabilities output by each algorithm. The authors note that this approach allows more individuals to have probabilities assigned, but it does not make full use of the information that can be gleaned from knowing an individual has a compound or hyphenated name.

Table 11 shows the percentage of individuals assigned race/ethnicity probabilities using each algorithm after applying this cleaning method.

**Table 11**  
**PERCENTAGE OF INDIVIDUALS ASSIGNED PROBABILITIES AFTER PREPROCESSING**

Self-Reported Race/Ethnicity	FO	GO	SA	BISG	BIFSG	Hierarchy
Unknown/Other	58%	99%	96%	95%	55%	>99%
White	80%	99%	95%	94%	75%	>99%
Black	48%	>99%	98%	98%	46%	>99%
Hispanic	75%	>99%	98%	98%	73%	>99%
Multiracial	54%	98%	95%	93%	51%	>99%
API	56%	99%	89%	88%	51%	>99%
AI/AN	66%	>99%	96%	95%	63%	>99%

As illustrated by the increase in individuals assigned probabilities, particularly for the Hispanic cohort using surname-based algorithms, this step can have a material impact on results. For the remainder of this section, the authors will limit discussion to results after applying these preprocessing methods. Appendix B contains a comparison to results without preprocessing for selected metrics.

As seen in Table 11, the FO and BIFSG algorithms have the lowest percentages of individuals assigned probabilities after preprocessing. This is likely due to the fact that the Surgeo package uses a first names list with these algorithms that was developed using data aggregated from mortgage applications (Tzioumis, 2018), which may not be representative of the individuals in the case study data. Note, that the geolocation and surname lists used by the Surgeo package were developed from the 2010 Census, which are representative of the entire U.S. population. Table 12 shows the percentages of individuals assigned race/ethnicity probabilities by age bands using the FO algorithm after applying preprocessing to names.

**Table 12**  
**PERCENTAGE OF INDIVIDUALS ASSIGNED PROBABILITIES AT EACH AGE RANGE**

Self-Reported Race/Ethnicity	First Name Only Algorithm			
	0-17	18-39	40-64	65+
White	62.8%	84.1%	94.7%	93.4%
Black	31.6%	40.3%	73.0%	86.0%
Hispanic	72.1%	78.2%	81.5%	82.3%
Multiracial	46.2%	61.2%	83.0%	88.2%
API	56.8%	57.8%	60.2%	44.2%
AI/AN	50.6%	69.8%	86.6%	87.1%

As Table 12 shows, the FO algorithm generally demonstrates a higher ability to assign probabilities for older individuals (except for the age 65+ API cohort). The variation in the proportions of the population who are assigned FO probabilities across age bands is especially pronounced for the Black cohort. As mentioned above, the BIFSG algorithm also uses the same first names list as the FO algorithm, therefore, a similar variation in assigned probabilities underlies those results. This information should be kept in mind while reviewing results in this section, as well as when these algorithms are applied in practice, because it demonstrates that the applicability of these algorithms can vary materially depending on the age distribution of the population they are being applied to. Table 13 provides the distribution of individuals (excluding those with unknown or other self-reported race or ethnicity) by age cohort to provide additional context for the metrics in subsequent sections.

**Table 13**  
**DATASET DISTRIBUTION BY AGE RANGE FOR EACH RACE/ETHNICITY (EXCLUDING UNKNOWN/OTHER SELF-REPORTED RACE)**

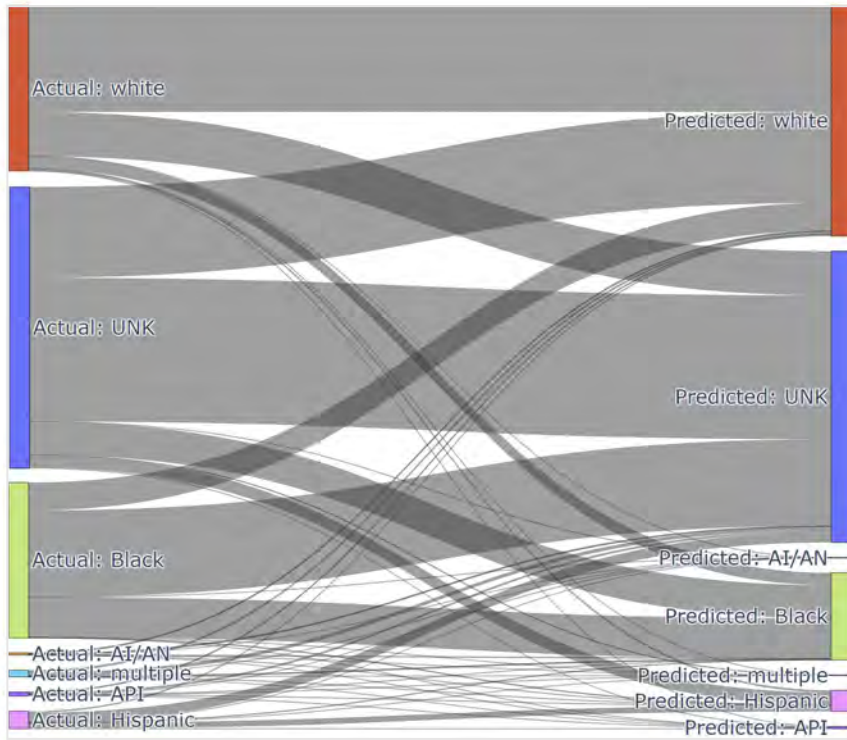
Self-Reported Race/Ethnicity	0-17	18-39	40-64	65+
White	44.2%	46.7%	53.1%	46.7%
Black	43.0%	45.5%	42.2%	49.3%
Hispanic	7.9%	4.3%	2.2%	2.0%
Multiracial	3.1%	1.9%	0.7%	0.2%
API	1.2%	1.1%	1.1%	1.5%
AI/AN	0.7%	0.7%	0.6%	0.3%

As Table 13 shows, the race and ethnicity distribution varies materially by age, with a higher concentration of the Hispanic and multiracial cohorts in the younger age groups. This will impact performance and the ability of the algorithms to assign probabilities and should be considered when reviewing results throughout this section.

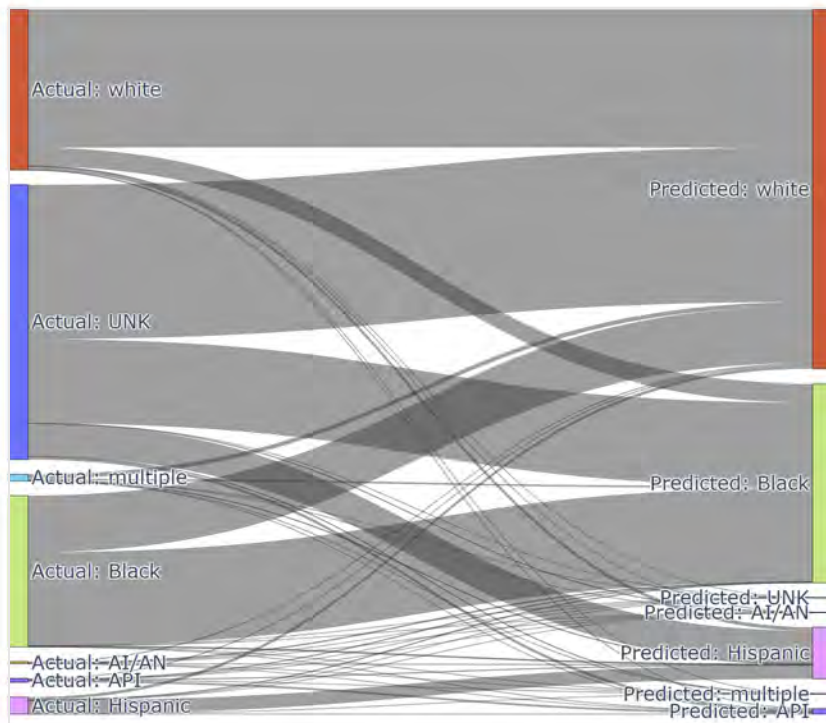
Figures 1 and 2 show the distribution of individuals before and after imputation using BIFSG (the algorithm with the lowest percentage of individuals with imputed race or ethnicity) and hierarchy (the algorithm with the highest percentage of individuals with imputed race or ethnicity), including those who reported other or did not provide race or ethnicity, shown as “UNK” in the figures. These charts are called Sankey diagrams and are intended to provide a visualization of key statistics by self-reported and imputed race and ethnicity that can be found in tabular form throughout the remainder of this section<sup>18</sup>.

<sup>18</sup> See <https://www.data-to-viz.com/graph/sankey.html> for more information about Sankey diagrams and their interpretation.

**Figure 1**  
DISTRIBUTION OF RACE/ETHNICITY BEFORE AND AFTER IMPUTATION USING BIFSG



**Figure 2**  
DISTRIBUTION OF RACE/ETHNICITY BEFORE AND AFTER IMPUTATION USING HIERARCHY



Of particular interest when comparing figures 1 and 2 is the size of the imputed white cohort in figure 2, since it is significantly larger than the cohort self-reporting as white, or the cohort imputed white using the BIFSG method.

Table 14 shows the percentage of individuals who had their race or ethnicity imputed at each step in the hierarchical algorithm.

**Table 14**  
**PERCENTAGE OF INDIVIDUALS ASSIGNED PROBABILITIES AT EACH STEP IN THE HIERARCHY**

Self-Reported Race/Ethnicity	Step 1: BIFSG	Step 2: BISG	Step 3: SA	Step 4: GO	Step 5: FO
Unknown/Other	55%	40%	1%	4%	<1%
White	75%	19%	1%	5%	<1%
Black	46%	51%	<1%	2%	<1%
Hispanic	73%	24%	<1%	2%	<1%
Multiracial	51%	43%	2%	5%	<1%
API	51%	37%	1%	11%	<1%
AI/AN	63%	32%	<1%	4%	<1%
<b>Total</b>	59%	37%	1%	4%	<1%

As shown above, the majority of individuals is assigned probabilities by one of the two Bayesian algorithms, with only the API cohort having less than 90% of individuals imputed by those two steps.

### 5.3 PERFORMANCE METRICS

This section provides an overview of several methods used to evaluate performance of the algorithms that the authors were able to run on the data. Note that the authors show all metrics by race/ethnicity cohort rather than at a population level to avoid misleading metrics based on the distribution of the population by race and ethnicity. The authors also exclude all individuals with unknown or other self-reported race or ethnicity because the authors do not have the information needed to measure performance on that cohort. The authors are including metrics for the multiracial cohort but note that all algorithms tend to perform poorly on this cohort, and the value of grouping all multiracial individuals together for analysis is uncertain.

The results in this case study are specific to the data used for the case study. The authors are not recommending any algorithm, method, or specific use from these results. For any other dataset, deviations from these results are expected. In particular, any differences in the distribution by age, gender, or socioeconomic status may cause material differences in performance. Furthermore, the authors have applied the algorithms using the packages as they are published. The user should consider whether refinements or modifications are necessary based on their data and use case.

#### 5.3.1 ACTUAL-TO-EXPECTED DISTRIBUTION

##### 5.3.1.1 Methods

Here the authors show how the various algorithms perform at predicting the race/ethnicity distribution of the sample. The authors show them as both actual-to-expected ratios (A:E) and calibration curves. The A:E shows the ratio of the actual number of individuals in each cohort to the number that would be predicted to be in that cohort using the imputation algorithms. An A:E above 1 means the algorithm has underpredicted the size of the cohort and an A:E below 1 means the algorithm has overpredicted the size of the cohort.

The authors will look at the A:E, and several other metrics through subsequent subsections, two different ways: the average probability metric, and the true positive rate—classified using highest probability. The average probability metric uses the probabilities output by algorithms directly, calculating the average probability assigned to each individual’s self-reported race or ethnicity by the algorithms. The second method, true positive rate, classified by highest probability, uses a hard classification rule that assigns 100% probability to the race or ethnicity that had the highest associated probability.

These two methods are intended to align with two common ways of using this output. It’s important to note that the two sets of results should not be compared to determine that one method of using the output is more accurate than the other; the two methods are not directly comparable in this way.

Calibration curves visualize the correlation between the predicted probability of a positive and the proportion of actual positives in the response. In a perfectly calibrated model, the calibration curve will be a 45-degree line so, for example, 50% of observations with a predicted probability of 50% would actually have a positive response.

In the context of race/ethnicity imputation algorithms, calibration curves can be generated for each level of race or ethnicity by using an indicator for self-reporting that race or ethnicity as the response. The calibration curves for all races and ethnicities can be plotted on the same set of axes, providing a convenient visual comparison of how the algorithm performs for different races and ethnicities.

The calibration curves below plot the predicted probabilities of each race/ethnicity cohort against the actual distribution of race and ethnicity. For each curve, the population is divided into deciles based on the predicted probabilities associated with the race/ethnicity cohort of interest, shown on the horizontal axis, and this is then compared to the actual proportion of individuals in that group who self-reported that race or ethnicity. It should be noted that some of the points represent relatively small population sizes (in particular, API and AI/AN), since the authors are looking at deciles of predictions on cohorts that are already small. The authors have excluded the multiracial cohort from these calibration curves because the algorithms do not typically assign a wide range of probabilities to the multiracial prediction, meaning that the deciles are not well-defined.

The authors also provide a comparison of the actual-to-expected distribution by age group for the FO algorithm, illustrating the variation in performance as a corollary to Table 12, which illustrated the variation in imputation percentage for this algorithm.

5.3.1.2 Results

Figure 3  
CALIBRATION CURVES

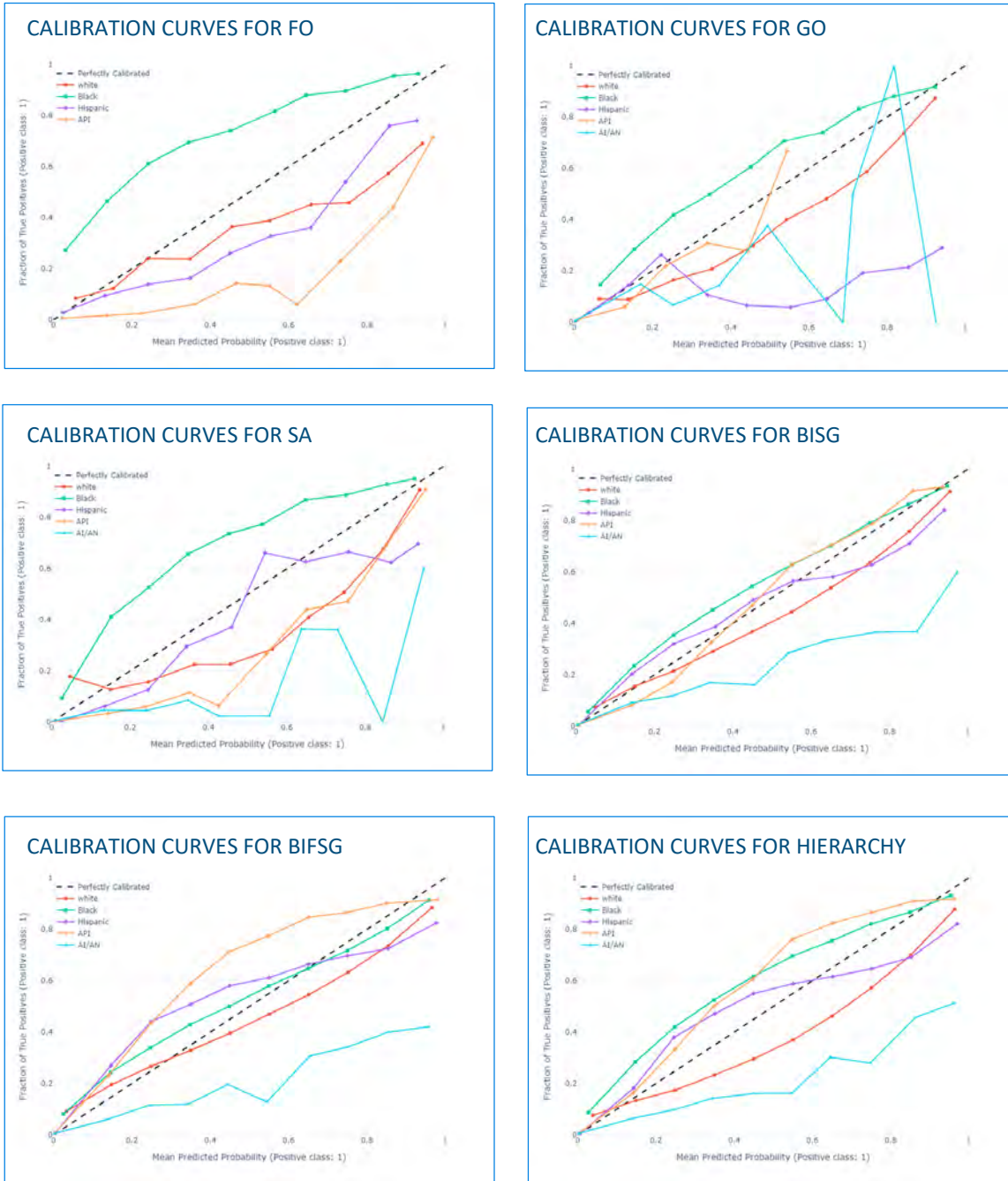


Table 15  
A:E DISTRIBUTION FOR AVERAGE PROBABILITY METHOD

Self-Reported Race/Ethnicity	FO	GO	SA	BISG	BIFSG	Hierarchy
White	0.71	0.77	0.73	0.89	0.91	0.82
Black	4.64	1.44	1.92	1.14	1.16	1.27
Hispanic	0.83	0.97	0.59	0.85	0.98	0.90
Multiracial	8.91	1.38	0.92	1.55	1.48	1.67
API	0.27	0.98	0.60	1.10	1.56	1.42
AI/AN	3.39	1.59	0.76	1.55	2.34	1.97

Table 16  
A:E DISTRIBUTION FOR AVERAGE PROBABILITY METHOD BY AGE BAND

Self-Reported Race/Ethnicity	First Name Only Algorithm			
	0-17	18-39	40-64	65+
White	0.71	0.75	0.71	0.62
Black	4.37	4.59	4.93	4.69
Hispanic	1.25	0.79	0.40	0.28
Multiracial	14.14	10.19	4.08	1.04
API	0.29	0.29	0.25	0.18
AI/AN	3.19	4.08	3.69	1.44

$$\text{Formula: } \frac{\# \text{ of individuals in cohort } R}{\sum \# \text{ of individuals self-reporting } R * A_R}$$

where  $A_R$  is the average probability associated with race/ethnicity cohort  $R$  and where the sum is taken over all race/ethnicities.

Table 17  
A:E DISTRIBUTION FOR CLASSIFICATION METHOD

Self-Reported Race/Ethnicity	FO	GO	SA	BISG	BIFSG	Hierarchy
White	0.63	0.62	0.57	0.87	0.86	0.78
Black	15.02	1.84	4.41	1.15	1.28	1.31
Hispanic	1.30	68.43	0.72	0.76	0.97	0.91
Multiracial	N/A <sup>1</sup>	N/A <sup>1</sup>	1,730.00	261.67	23.39	44.18
API	1.49	395.32	0.88	1.29	1.58	1.60
AI/AN	N/A <sup>1</sup>	121.79	8.32	9.98	13.14	13.02

<sup>1</sup> Values of N/A indicate that no individuals were assigned the corresponding imputed race by that algorithm using the classification method.

$$\text{Formula: } \frac{\# \text{ of individuals in cohort } R}{\# \text{ of individuals with imputed race/ethnicity } R}$$

### 5.3.1.3 Key Observations and Limitations

FO and SA tend to under-predict the size of the Black population, while over-predicting the size of all other cohorts at most levels of predicted probability. In particular, the API cohort is over-predicted by the largest margin at all levels for FO.

BISG produces calibration curves significantly closer to the diagonal for all race/ethnicity cohorts compared to the non-Bayesian algorithms. This indicates that use of BISG to estimate the racial/ethnic composition of an area would result in less biased estimates on the sample than the previous algorithms that used only name or geolocation, but not both.



A notable difference between the calibration curves for BIFSG compared to BISG is on the API cohort, where the BIFSG is showing a clear tendency to understate the size of the population. This aligns with the tables shown in subsection 5.3.2.2 below, which show a deterioration in performance moving from BISG to BIFSG for this cohort.

### 5.3.2 PROBABILITY OF SELF-REPORTED RACE AND ETHNICITY PREDICTED

#### 5.3.2.1 Methods

The authors will next consider a simple and widely published metric: the probability of predicting an individual's self-reported race or ethnicity correctly using the same two methods described in subsection 5.3.1 above.

#### 5.3.2.2 Results

Table 18

#### AVERAGE PROBABILITY ASSOCIATED WITH SELF-REPORTED RACE/ETHNICITY

Self-Reported Race/Ethnicity	FO	GO	SA	BISG	BIFSG	Hierarchy
White	87.2%	68.9%	73.8%	70.3%	80.2%	77.7%
Black	11.9%	39.9%	33.3%	59.6%	56.0%	57.3%
Hispanic	35.4%	9.1%	84.0%	75.0%	73.3%	72.6%
Multiracial	0.2%	1.5%	2.2%	1.6%	1.8%	1.7%
API	27.7%	2.1%	64.3%	57.2%	49.1%	47.5%
AI/AN	0.2%	1.4%	3.0%	4.0%	2.8%	2.9%

$$\text{Formula: } \frac{\# \text{ of individuals self-reporting } R * A_R}{\text{Total individuals in race/ethnicity cohort}}$$

where  $A_X$  is the average probability associated with self-reported race  $X$  for race/ethnicity cohort  $X$ .

Note: For this formula and all formulas below, "individuals" refers to all individuals for whom sufficient data was available to run the given algorithm and excluding individuals with an imputed race or ethnicity of "unknown."

For example, the value of 11.9% for the FO algorithm on the Black cohort can be interpreted to mean that, if the algorithm were run on only the Black cohort, the average probability associated with Black would be 11.9%, or that it would estimate 11.9% of that population to be Black.

Table 19

#### TRUE POSITIVE RATE – CLASSIFIED USING HIGHEST PROBABILITY

Self-Reported Race/Ethnicity	FO	GO	SA	BISG	BIFSG	Hierarchy
White	97.6%	89.2%	92.5%	78.7%	87.7%	86.0%
Black	5.7%	39.7%	18.8%	66.7%	58.4%	61.9%
Hispanic	38.5%	0.1%	93.6%	90.5%	78.3%	79.9%
Multiracial	0.0%	0.0%	0.0%	<0.1%	0.3%	0.1%
API	25.9%	0.1%	69.2%	65.6%	55.1%	54.2%
AI/AN	0.0%	0.2%	2.7%	3.3%	2.4%	2.4%

$$\text{Formula: } \frac{\# \text{ of individual whose imputed race/ethnicity matches self-reported}}{\text{Total individuals in race/ethnicity cohort}}$$

For example, the value of 5.7% for the FO algorithm on the Black cohort indicates that, if a user were to run the algorithm on only the Black cohort, and each individual's imputed race was assumed to be the race with the highest probability assigned by the algorithm, 5.7% of the population would be imputed as Black.

### 5.3.2.3 Key Observations and Limitations

The authors can see that the true positive rate—classified using highest probability—produces more accurate predictions for the white cohort for every algorithm under this metric, but for other race/ethnicity cohorts, the results are more mixed. In particular, performance is more likely to degrade for the non-Bayesian algorithms and for the AI/AN cohort (though the accuracy of all algorithms is extremely low for the AI/AN cohort using either method).

The authors also note that combining geolocation and surname degrades performance for certain cohorts—namely, white, Hispanic, and API. This may be due to the specifics of the data used in the authors’ evaluation but may also reflect the variation in geographic racial/ethnic homogeneity for these cohorts.

As discussed in more detail below, care should be taken when drawing conclusions from this. For example, algorithms that classify a large percentage of the population as white would appear to perform very well on the white cohort but may not be very useful in measuring disproportionate impacts or understanding the composition of a population.

## 5.3.3 PROBABILITY OF WHITE RACE PREDICTED

### 5.3.3.1 Methods

In addition to analyzing the ability of the algorithms to correctly predict self-reported race and ethnicity, the authors consider the likelihood associated with predicting that an individual is white across all races and ethnicities. This metric is important to consider because all algorithms perform highest on the white population, which can be misleading and skew metrics given the high prevalence of white individuals in the population; furthermore, overestimating the prevalence of white individuals in a population could affect any study of disparities or bias significantly.

### 5.3.3.2 Results

Table 20

#### AVERAGE PROBABILITY ASSOCIATED WITH WHITE RACE/ETHNICITY

Self-Reported Race/Ethnicity	FO	GO	SA	BISG	BIFSG	Hierarchy
White	87.2%	68.9%	73.8%	70.3%	80.2%	77.7%
Black	77.5%	52.2%	59.5%	37.7%	41.7%	40.0%
Hispanic	56.0%	62.3%	10.6%	17.7%	22.0%	21.5%
Multiracial	80.8%	64.3%	63.9%	56.6%	65.8%	61.2%
API	60.6%	66.1%	19.4%	22.7%	31.2%	31.6%
AI/AN	82.0%	61.0%	60.6%	51.1%	63.3%	57.6%

$$\text{Formula: } \frac{\# \text{ of individuals self-reporting } R * A_W}{\text{Total individuals in race/ethnicity cohort}}$$

where  $A_W$  is the average probability associated with white for race/ethnicity cohort  $R$ .

**Table 21**  
**RATE OF WHITE IMPUTATIONS – CLASSIFIED USING HIGHEST PROBABILITY**

Self-Reported Race/Ethnicity	FO	GO	SA	BISG	BIFSG	Hierarchy
White	97.6%	89.2%	92.5%	78.7%	87.7%	86.0%
Black	90.4%	60.2%	80.0%	32.6%	40.9%	37.5%
Hispanic	60.7%	85.4%	5.6%	7.6%	20.3%	18.2%
Multiracial	92.8%	82.1%	82.2%	60.6%	71.0%	66.7%
API	68.9%	84.8%	20.7%	19.8%	31.3%	32.3%
AI/AN	92.8%	79.3%	76.0%	52.4%	68.4%	61.2%

$$\text{Formula: } \frac{\text{\# of individual with imputed race white}}{\text{Total individuals in race/ethnicity cohort}}$$

### 5.3.3.3 Key Observations and Limitations

Tables 20 and 21 show that, although the method of classifying individuals using highest probability produces a higher true positive rate, it also produces a higher or similar false positive rate for all algorithms and cohorts except Hispanic. This illustrates the importance of considering multiple performance metrics when evaluating algorithms, depending on the intended use.

## 5.3.4 RATIO OF TRUE POSITIVES TO FALSE POSITIVES

### 5.3.4.1 Methods

This metric represents the odds that the imputation algorithm results in a race or ethnicity consistent with an individual's self-reported race or ethnicity. A ratio of 1 indicates the odds are 50-50 that the imputation is correct, meaning the imputation does not perform better than random. The higher the ratio, the better the imputation algorithm is at matching self-reported race or ethnicity. In particular, a ratio less than 1 means that individuals in that group are more likely to be identified incorrectly than correctly, which indicates a high likelihood of introducing bias in any analysis.

The authors looked at the ratios of true positives to false positives using the same two methods described in the prior sections, focusing on white predictions for the reasons noted above. Table 22 represents the average imputed probability associated with white for the cohort with a self-reported race of white, compared to the average imputed probability associated with white for all other cohorts. Table 23 represents the probability that an individual has a self-reported race of white if the imputed race or ethnicity assigned by using the highest probability prediction is white, compared to the probability that the individual does not have a self-reported race of white.

The authors focus on the white cohort in this section because the distinction between white people and people of color is often of primary importance when performing bias and disparity analysis. Appendix B provides detail on other cohort predictions.

### 5.3.4.2 Results

Table 22

#### ODDS OF MATCHING SELF-REPORTED RACE USING PROBABILITIES ASSOCIATED WITH WHITE

Self-Reported Race/Ethnicity	FO	GO	SA	BISG	BIFSG	Hierarchy
White	1.64	1.13	1.17	1.66	2.72	1.76

$$\text{Formula: } \frac{\# \text{ of individuals self-reporting white} * A_w}{\sum \# \text{ of individuals self-reporting cohort } R * A_R}$$

where  $A_x$  is the average probability associated with white for race/ethnicity cohort  $X$  and the sum is taken over all cohorts except white.

Table 23

#### RATIO OF TRUE POSITIVES TO FALSE POSITIVES FOR INDIVIDUALS PREDICTED TO BE WHITE

Imputed Race/Ethnicity	FO	GO	SA	BISG	BIFSG	Hierarchy
White	1.59	1.23	1.11	2.16	3.02	2.07

$$\text{Formula: } \frac{\# \text{ of individuals with imputed race white and self-reported race white}}{\sum \# \text{ of individuals with imputed race white and self-reported race } R}$$

where the sum is taken over all cohorts except white.

### 5.3.4.3 Key Observations and Limitations

The table shows that the FO algorithm has a low ratio relative to the BIFSG algorithm on the white cohort—contrasted with the fact that it has the highest true positive rate for that cohort (the first row of table 19), it also has a high false positive rate (table 21) because it assigns a high probability of being white to a large swath of the population, regardless of self-reported race or ethnicity. Further, in contrast to the prior metrics shown, combining surname and geolocation in BISG or BIFSG always improves the ratio compared to surname analysis or geocoding only.

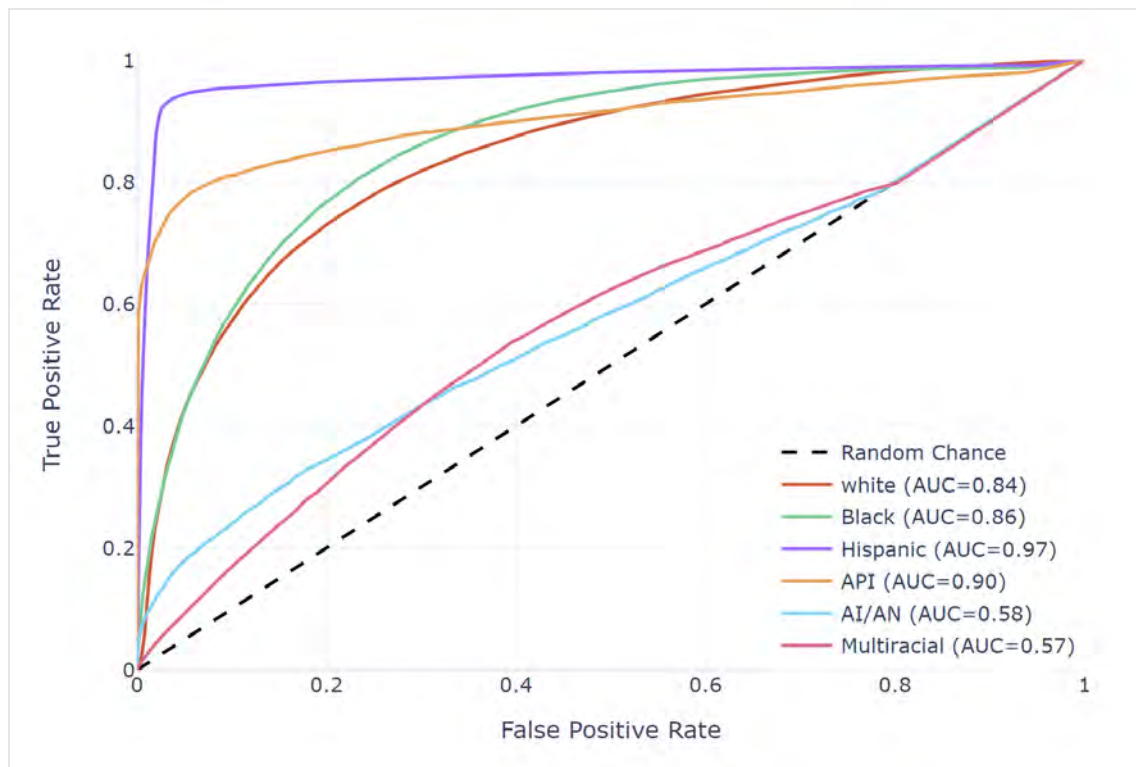
## 5.3.5 AREA UNDER THE RECEIVER OPERATING CHARACTERISTIC CURVE (AUC)

### 5.3.5.1 Methods

The receiver operating characteristic (ROC) curve is a plot of the true positive rate on the vertical axis against the false positive rate on the horizontal axis for various thresholds used for classification. The area under the ROC curve (AUC) is a metric describing the ROC curve. A higher AUC represents a better ability to distinguish between two cohorts. The AUC measures how well a probability model can distinguish positives from negatives in a binary response. It can take values from 0 to 1, where an AUC of 1 means the model can perfectly distinguish between positives and negatives and an AUC of 0.5 means the model is no better than a coin flip. AUC values below 0.5 are theoretically possible but are uncommon as this would mean the model actually does worse at distinguishing outcomes than a random guess.

### 5.3.5.2 Results

**Figure 4**  
ROC CURVES FOR HIERARCHY METHOD



Users should be aware that a model can have a high AUC but perform poorly by other metrics.

**Table 24**

#### AUC

Self-Reported Race/Ethnicity	FO	GO	SA	BISG	BIFSG	Hierarchy
White	0.67	0.73	0.75	0.82	0.84	0.84
Black	0.68	0.74	0.79	0.85	0.85	0.86
Hispanic	0.80	0.74	0.97	0.98	0.97	0.97
Multiracial	0.49	0.58	0.52	0.60	0.51	0.57
API	0.73	0.71	0.89	0.91	0.89	0.90
AI/AN	0.51	0.61	0.53	0.63	0.57	0.58

### 5.3.5.3 Key Observations and Limitations

Similar to the metrics shown in subsection 5.3.4 above, despite mixed results for BISG and BIFSG in the rate of true positives or average probability of correct self-reported race or ethnicity (see subsection 5.3.2), BISG and BIFSG are an improvement on (or similar to) the performance of the SA algorithm using this metric. The authors also note that all algorithms show performance on the AI/AN cohort that is only slightly better than 0.50.

## 5.3.6 PRECISION

### 5.3.6.1 Methods

Precision, also known as positive predicted value, is the proportion of predicted positives that actually had a positive response. It measures the relevance of positive predictions. Precision and recall/sensitivity (true

positive rate) are often reported together and combine to present a fuller picture of the model performance.

In the context of race/ethnicity imputation methods, the precision for each level of race or ethnicity is the proportion of people predicted to be that race or ethnicity who actually self-report as that race or ethnicity.

### 5.3.6.2 Results

**Table 25**  
**PRECISION**

Self-Reported Race/Ethnicity	FO	GO	SA	BISG	BIFSG	Hierarchy
White	61.4%	55.2%	52.6%	68.4%	75.1%	67.4%
Black	86.0%	73.3%	82.9%	76.6%	74.9%	81.3%
Hispanic	50.1%	9.0%	67.2%	69.2%	75.8%	72.5%
Multiracial	0.0%	0.0%	0.0%	11.5%	6.3%	6.4%
API	38.6%	31.8%	61.1%	84.5%	87.3%	86.4%
AI/AN	N/A <sup>1</sup>	28.9%	22.4%	33.2%	31.6%	31.5%

<sup>1</sup>Values of N/A indicate no positive predictions for that cohort and algorithm.

$$\text{Formula: } \frac{\text{\# of individuals with imputed race/ethnicity R who self-report that race/ethnicity}}{\text{Total \# of individuals with imputed race/ethnicity R}}$$

### 5.3.6.3 Key Observations and Limitations

The two cohorts of particular interest for this metric are the Black and API cohorts. In the Black cohort, FO and SA produce the highest precision of all methods though, in subsection 5.3.1 above, this is driven by a tendency to underestimate the size of the Black cohort (thereby decreasing the denominator of the metric). For the API cohort, BISG and BIFSG are a striking improvement over other algorithms when using this metric.

## 5.3.7 SPECIFICITY

### 5.3.7.1 Methods

True negative rate is the proportion of negative observations with negative predictions. It measures the ability of the imputation to detect negative responses among all the data. In the context of race/ethnicity imputation algorithms, the true negative rate can be calculated for each level of race or ethnicity and thought of as the probability of a specific race or ethnicity not being imputed among individuals who do not self-report that race or ethnicity.

### 5.3.7.2 Results

Table 26  
SPECIFICITY

Self-Reported Race/Ethnicity	FO	GO	SA	BISG	BIFSG	Hierarchy
White	14.1%	35.9%	28.2%	68.8%	60.7%	63.1%
Black	99.5%	88.4%	96.8%	83.2%	90.2%	88.7%
Hispanic	97.6%	99.9%	97.6%	97.8%	98.4%	98.4%
Multiracial	>99.9%	>99.9%	>99.9%	>99.9%	99.9%	>99.9%
API	99.6%	>99.9%	99.5%	99.9%	99.9%	99.9%
AI/AN	100.0%	>99.9%	99.9%	>99.9%	>99.9%	>99.9%

**Formula:** 
$$\frac{\# \text{ of individuals with imputed race/ethnicity other than R who did not self-report that race/ethnicity}}{\text{Total \# of individuals with imputed race/ethnicity other than R}}$$

### 5.3.7.3 Key Observations and Limitations

Contrary to many other metrics considered, the specificity is highest for the Hispanic, API, and AI/AN cohorts across all algorithms. This indicates a tendency not to assign high probabilities to these cohorts for individuals who do not self-report as those cohorts.

## 5.4: CASE STUDY SUMMARY

This section summarizes which algorithms achieved the highest performance using each metric throughout the case study, by race/ethnicity. It is important to keep in mind that performance on a user's dataset may vary materially from the performance metrics presented in the case study. In addition, the optimal metric for a given situation will take into account the intended purpose of the imputation, so the choice of algorithm for a specific use case should take into account the characteristics of the data and the consequences of misclassifications. It should also be kept in mind that identifying an algorithm as the highest performing for a given cohort and metric does not indicate that the performance is acceptable, particularly when considering the multiracial and AI/AN cohorts.

Table 27  
SUMMARY OF METRICS

Metric	White	Black	Hispanic	Multi-racial	API	AI/AN
Coverage <sup>1</sup>	GO	GO	GO	GO	GO	GO
<b>Methods Using Probabilities Directly</b>						
A to E distribution (Table 15)	BIFSG	BISG	BIFSG	SA	GO	SA
Avg Probability of self-reported (Table 18)	FO	BISG	SA	SA	SA	BISG
Avg Probability of white (Table 20)	N/A	BISG	SA	BISG	SA	BISG
Odds of matching self-reported (Table 37)	BIFSG	Hierarchy	BIFSG	BIFSG	BIFSG	BIFSG
<b>Methods Classifying Individuals Using Highest Probability</b>						
A to E distribution (Table 17)	BISG	BISG	BIFSG	BIFSG	SA	SA
True positive rate (Table 19)	FO	BISG	SA	BIFSG	SA	BISG
Rate of white imputations (Table 21)	N/A	BISG	SA	BISG	BISG	BISG
Area under the curve (Table 24)	Hierarchy	Hierarchy	BISG	BISG	BISG	BISG
Precision (Table 25)	BIFSG	FO	BIFSG	BISG	BIFSG	BISG
Specificity (Table 26)	BISG	FO	GO	FO	GO	FO
Ratio of true to false positives (Table 38)	BIFSG	FO	BIFSG	BISG	BIFSG	BISG

<sup>1</sup> Coverage measures the percent of the population for whom the algorithm assigned race and ethnicity probabilities. Note that the hierarchy method achieved the highest coverage for all race and ethnicity cohorts, but it is excluded from this metric since it was intentionally designed to maximize coverage.

## Section 6: Tutorial

To help readers become familiar with how to implement the imputation methods described above, this section provides a tutorial that uses Python and R packages to impute race and ethnicity on simulated data. The code and simulated dataset can be downloaded from the Society of Actuaries website. The remainder of this section serves as a guide to the data and code and reviews the output from various imputation methods run on the simulated dataset.

### 6.1 INPUT DATA

The tutorial dataset is simulated purely for demonstration purposes. Although the authors made an effort to match the U.S. national distribution by race and ethnicity, the tutorial data should not be used to draw any conclusions about the true population or accuracy of any particular method. Rather, the purpose of the tutorial dataset is to show an example of the input data and code for imputation and analyzing imputation output.

The authors created the tutorial dataset by first creating fictitious names, combining a dataset with first names and gender with a dataset of frequently occurring surnames.<sup>19</sup> The authors used these names to simulate self-reported race or ethnicity with the `ethnicolr` package.<sup>20</sup> Next, the authors assigned ZIP Code Tabulation Areas (ZCTAs) based on the distribution of race and ethnicity by ZCTA from the U.S. Census Bureau and ZCTAs were mapped to census tracts, block groups, and blocks. Age was randomly assigned based on a U.S. national distribution. The authors also simulated a hypothetical outcome variable (variable name = “target”) from normal distributions with different means by race and ethnicity.

Tables 28 and 29 show a sample of five records from the tutorial dataset and summary statistics.

**Table 28**  
**TUTORIAL SAMPLE INPUT RECORDS**

Variable	1	2	3	4	5
first_name	Ronald	Paul	Roy	Stella	Kenneth
last_name	Kim	Loo	Madrid	Bunting	Ma
middle_name	Thomas	Clay	Leonard	Lorraine	Christopher
gender	M	M	M	F	M
race	asian	asian	asian	asian	asian
ZCTA	92880	94122	23228	27502	1003
age	42	28	59	60	45
GEOID_block	060650406202002	060750327007001	510872004113006	371830534331002	250158204001038
GEOID_blockgroup	000060650406202	000060750327007	000510872004113	000371830534331	000250158204001
GEOID_tract	000006065040620	000006075032700	000051087200411	000037183053433	000025015820400
Target	10.1105164	7.367597	7.60920374	8.360537	10.4004808

For display, this table has been transposed from the usual format, so that each column corresponds to a single record and each row corresponds to a data element.

<sup>19</sup> Gender by Name dataset from the UC Irvine Machine Learning Repository and Frequently Occurring Surnames from the 2010 Census (all surnames occurring 100 or more times in the 2010 U.S. Census).

<sup>20</sup> Full Name Florida voter registration-based model.



**Table 29**  
**TUTORIAL DATASET SUMMARY STATISTICS**

Self-Reported Race/Ethnicity	Records	Distribution	Average Target
White	57,322	57.3%	9.9
Black	11,766	11.8%	8.6
Hispanic	19,624	19.6%	6.1
API	5,854	5.9%	8.5
Other	5,434	5.4%	7.4

As discussed in the case study, data preparation is an important step for imputation packages to work well. Due to the nature of the tutorial dataset, the authors did not include code for data cleaning as the data cleaning needed will vary from dataset to dataset. Some packages include tools for some data cleaning tasks, but the specific data prep process will depend on the format of the raw data and the requirements of the packages being used.

## 6.2 IMPUTATION PACKAGES

The following packages are used to impute race and ethnicity within the tutorial code:

1. Surgeo (Python), which implements first name only (FO), geocoding only (GO), surname analysis (SA), Bayesian Improved Surname Geocoding (BISG), and Bayesian Improved First Name Surname Geocoding (BIFSG).
2. Ethnicolr (Python), which provides pretrained neural network models that can be applied to new data.
3. Wru (R), which implements the BISG and fBISG methods proposed in Imai, K. & Khanna, K. (2016).
4. BIRDIE (R), which implements Bayesian Instrumental Regression for Disparity Estimation as described in McCartan et al. (2022) and BISG and fBISG algorithms described in Imai et al. (2022).

Appendix C contains a table summarizing the methods available for each package tested, as well as distribution sources, types of geolocations used (e.g., census block, ZCTA), and the types of outputs produced.

### 6.3 OUTPUT

The output of the imputations is an array of probabilities attached to the input dataset. Table 34 shows a sample from the BIRDIE BISG function.

**Figure 30**

**TUTORIAL SAMPLE OUTPUT RECORDS**

Variable	1	2	3	4	5
first_name	Ronald	Paul	Roy	Stella	Kenneth
last_name	Kim	Loo	Madrid	Bunting	Ma
middle_name	Thomas	Clay	Leonard	Lorraine	Christopher
gender	M	M	M	F	M
race	asian	asian	asian	asian	asian
	...	...	...	...	...
nh_white	0.00307	0.00829	0.08556	0.88415	0.0182
nh_black	0.00105	0.00015	0.01375	0.06463	0.00147
hispanic	0.00633	0.00961	0.85528	0.01206	0.00346
asian	0.97266	0.90893	0.02511	0.00534	0.95015
american_indians_alaska_native	0.00001	0	0.00181	0.00054	0.00003
other	0.01687	0.07302	0.01849	0.03328	0.02668

For display, this table has been transposed from the usual format, so that each column corresponds to a single record and each row corresponds to a data element.

When self-reported data is available, it is possible to calculate performance metrics. The tutorial code includes logic to calculate the same metrics shown in the case study above. A sample of performance metrics from the BIRDIE tutorial is included in Appendix D. Again, given the nature of the tutorial data, it is important not to make any inferences about imputation algorithm performance based on these results.

## Section 7: Conclusion

Imputation is a powerful tool for studying disproportionate impact, unfair discrimination, and equity. As it grows in use by the insurance industry, it is important that actuaries understand the limitations and appropriate uses of various methods and data used for imputation and give careful thought to the validity of the distributions and methods being used for the population under study.

The authors have reviewed several methods for imputing race and ethnicity data and prepared a case study to illustrate how an off-the-shelf package performs on a particular sample dataset and the importance of examining multiple performance metrics by cohort. This study, and other studies, show how accuracy can vary from dataset to dataset and the strengths and weaknesses of various methods and data sources. While imputation methods have improved greatly, the field is continuing to develop. It is the authors' hope that this paper will help advance the development of best practices for implementation and use of imputation methods.



**Give us your feedback!**

Take a short survey on this report.

[Click Here](#)

 **SOA**  
**Research**  
INSTITUTE

## Acknowledgments

The researchers' deepest gratitude goes to those without whose efforts this project could not have come to fruition: the Project Oversight Group and others for their diligent work overseeing, reviewing, and editing this report for accuracy and relevance.

Project Oversight Group members:

Dorothy Andrews, Ph.D., ASA, MAAA, CSPA

Brian Bayerle, FSA, MAAA

Stephen Cameron, FSA, MAAA

Amine Elmeghni, FSA, MAAA, MSc

Jean-Marc Fix, FSA, MAAA

Hannah Kraus, ASA, MAAA

Tim Luedtke, FSA, MAAA

Ian McCulla, FSA, MAAA

Andrew Melnyk

Min Mercer, FSA

Murali Niverthi, FSA, MAAA

Renee West, FSA, MAAA

At the Society of Actuaries Research Institute:

Lisa Schilling, FSA, EA, FCA, MAAA, Senior Research Actuary

## Appendix A: Summary of Imputation Methods

Table 31  
SUMMARY TABLE OF INPUT AND OUTPUTS

Method	Input Variables					Outputs
	First Name	Middle Name	Surname	Geo-location	Other	
Geocoding Only (GO)				X		Variable
Surname Analysis (SA)			X			Variable
Categorical Surname and Geocoding (CSG)			X	X		Variable
Bayesian Surname Geocoding (BSG)			X	X		White/other, Black, Hispanic, Asian
Bayesian Improved Surname Geocoding (BISG)			X	X		White, Black, Hispanic, API, AI/AN, multiracial
Medicare Bayesian Improved Surname Geocoding (MBISG)	X		X	X	X	White, Black, Hispanic, API, AI/AN, multiracial
Bayesian Improved Surname Geocoding Extensions (BISG Ext)			X	X	X	Variable
Bayesian Improved First Name Surname Geocoding (BIFSG)	X		X	X		White, Black, Hispanic, API, AI/AN, multiracial
Modified BIFSG	X		X	X	X	White, Black, Hispanic, API, AI/AN, multiracial
Fully Bayesian Improved Surname Geocoding (fBISG)	X	X	X	X		White, Black, Hispanic, Asian
Bayesian Instrumental Regression for Disparity Estimation (BIRDIE)	(i)	(i)	X	X	X	White, Black, Hispanic, Asian, Native, Other
Regression	X		X	X	X	Variable
Natural Language Processing (NLP)	X		X			White, Black, Hispanic, API, AI/AN, multiracial
Machine Learning (ML)	X	X	X	X		White, Black, Hispanic, API, Other

(i) User can provide imputed probabilities from other packages that use first and/or middle names. BIRDIE includes functions to create BISG and fBISG estimates.

**Table 32**  
**SUMMARY TABLE OF PACKAGES AND REFERENCES**

Method	References	Packages	Notes
GO	Fiscella & Fremont, 2006 Krieger, et al., 2002	Surgeo (Python)	Older method. Not accurate enough for individual predictions, should only be used in aggregate.
SA	Fiscella & Fremont, 2006 Elliott, Fremont et al., 2008 Lauderdale & Kestenbaum, 2000	Surgeo (Python) WRU (R) BIRDIE (R)	Older method. Not accurate enough for individual predictions, should only be used in aggregate.
CSG	Fiscella & Fremont, 2006		Older method of combining surname and geolocation. Categorical predictions only, not probabilities.
BSG	Elliott, Fremont, et al., 2008		Older method, improvement to CSG but more limited surname lists.
BISG	Elliott, Morrison, et al., 2009 Baines & Courchane, 2014	Surgeo (Python) WRU <sup>®</sup> BIRD <sup>®</sup> (R)	Structurally similar to BSG, with better surname data and more cohorts imputed.
MBISG	Martino et al., 2013 Haas et al., 2019	Non-public	Improved use of data elements of BISG. Calibrated to Medicare population.
BISG Ext	Imai & Khanna, 2016 Fisher, 2023		Other predictors: age, gender, political party registration, number of dependents, tax filing status.
BIFSG	Voicu, 2018 Sorbero, et al., 2022	Surgeo (Python) WRU (R)	Significant improvement for Black cohort.
Modified BIFSG	Sorbero, et al., 2022		Added refinements for compound and rare surnames.
fBISG	Imai, Olivella & Rosenman, 2022	WRU (R) BIRDIE (R)	Added measurement error to geolocations with zero population. Created new first name list.
BIRDIE	McCartan et al., 2023	BIRDIE (R)	Can incorporate distributions specified by the user besides name and geolocation. Utilizes a correlated target variable to update the race/ethnicity prediction.
Regression	Xue, Harel & Aseltine (2019) Zavez, Harel & Aseltine (2022)		Other predictors: insurance type, missing father, mother's age. Connecticut-specific first name and last name lists.
NLP	Xie (2021)	Rethnicity (R) Ethnicolr (Python)	No comparisons to Bayesian methods.
ML	Decter-Frain (2022) Chintalapati et al. (2023)	Ethnicolr (Python)	Concerns about overfitting.

## Appendix B: Detailed Performance Metrics from Case Study

Section 5 contained key performance metrics from the case study analysis; this appendix provides further detail on many of these items.

### B.1 COMPARISON OF ACCURACY METRICS PRIOR TO DATA CLEANING

In section 5, the authors focused on performance metrics using data that had been preprocessed prior to running through the algorithms to enhance the ability to assign probabilities to individuals with compound or hyphenated names. Here the authors show a comparison to metrics prior to this preprocessing, illustrating the importance of this step.

Table 33

#### AVERAGE PROBABILITY ASSOCIATED WITH SELF-REPORTED RACE/ETHNICITY

Self-Reported Race/Ethnicity	FO	SA	GO	BISG	BIFSG	Hierarchy
White	87.2%	74.1%	68.9%	70.5%	80.4%	77.6%
Black	12.0%	33.3%	39.9%	59.7%	56.2%	57.1%
Hispanic	35.3%	83.3%	9.1%	74.1%	72.7%	46.5%
Multiracial	0.2%	2.2%	1.5%	1.6%	1.8%	1.7%
API	25.7%	65.6%	2.1%	58.4%	49.3%	46.9%
AI/AN	0.2%	3.1%	1.4%	4.2%	2.9%	2.9%

Table 34

#### TRUE POSITIVE RATE – CLASSIFIED USING HIGHEST PROBABILITY

Self-Reported Race/Ethnicity	FO	SA	GO	BISG	BIFSG	Hierarchy
White	97.6%	92.9%	89.2%	78.9%	87.8%	86.1%
Black	5.8%	18.9%	39.7%	66.7%	58.7%	61.6%
Hispanic	38.4%	90.6%	0.1%	88.2%	77.4%	46.7%
Multiracial	0.0%	0.0%	0.0%	<0.1%	0.3%	0.1%
API	23.5%	70.5%	0.1%	66.9%	55.8%	53.6%
AI/AN	0.0%	2.8%	0.2%	3.5%	2.5%	2.4%

Comparing the above results to the results in subsection 5.3.2.2 above shows that the impact on accuracy is minimal for most cohorts, but it increases the accuracy for the Hispanic cohort by 26 and 33 percentage points, for the average probability and true positive rate, respectively.

Table 35

#### AVERAGE PROBABILITY ASSOCIATED WITH WHITE RACE

Self-Reported Race/Ethnicity	FO	GO	SA	BISG	BIFSG	Hierarchy
White	87.2%	68.9%	74.1%	70.5%	80.4%	77.6%
Black	77.7%	52.2%	59.5%	37.6%	41.6%	40.2%
Hispanic	56.1%	62.3%	11.5%	18.4%	22.4%	37.9%
Multiracial	80.8%	64.3%	64.3%	56.9%	66.1%	61.5%
API	62.3%	66.1%	19.3%	22.5%	31.8%	32.4%
AI/AN	81.9%	61.0%	62.3%	52.2%	64.4%	58.7%

**Table 36**  
**RATE OF WHITE PREDICTIONS – CLASSIFIED USING HIGHEST PROBABILITY**

Self-Reported Race/Ethnicity	FO	GO	SA	BISG	BIFSG	Hierarchy
White	97.6%	89.2%	92.9%	78.9%	87.8%	86.1%
Black	90.6%	60.2%	80.0%	32.6%	40.8%	37.8%
Hispanic	60.8%	85.4%	8.1%	9.1%	20.6%	45.3%
Multiracial	92.8%	82.1%	82.8%	60.8%	71.3%	63.4%
API	71.1%	84.8%	20.7%	19.6%	31.6%	33.5%
AI/AN	92.8%	79.3%	78.5%	53.8%	69.6%	63.4%

Comparing these results to those shown in subsection 5.3.3.2 highlights the impact of data cleaning on this metric that is once again large for the Hispanic cohort, particularly when measured on using the highest probability to classify individuals.

## B.2 RATIO OF TRUE POSITIVES TO FALSE POSITIVES BY COHORT

Tables 37 and 38 show the ratio of true positives to false positives for each race/ethnicity cohort using the same data (after preprocessing). See subsection 5.3.4.1 for a description of this metric.

As described in section 5, this represents the ability of the models to accurately categorize individuals in each cohort—higher ratios represent a better ability to distinguish between individuals of the specified race or ethnicity and individuals not of that specified race or ethnicity. A ratio of 1 would indicate the model does no better than a random guess, and ratios below 1 indicate worse predictive ability than a random guess.

**Table 37**  
**ODDS OF MATCHING SELF-REPORTED RACE USING PROBABILITIES**

Self-Reported Race/Ethnicity	FO	GO	SA	BISG	BIFSG	Hierarchy
White	1.64	1.13	1.17	1.66	2.72	1.76
Black	1.24	1.34	1.78	2.15	1.83	2.65
Hispanic	0.42	0.10	0.98	1.75	2.54	1.89
Multiracial	0.02	0.02	0.02	0.07	0.07	0.03
API	0.08	0.02	0.64	1.68	3.27	2.07
AI/AN	0.01	0.02	0.02	0.07	0.07	0.06

**Table 38**  
**RATIO OF TRUE POSITIVES TO FALSE POSITIVES**

Imputed Race/Ethnicity	FO	GO	SA	BISG	BIFSG	Hierarchy
White	1.59	1.23	1.11	2.16	3.02	2.07
Black	6.14	2.74	4.84	3.27	2.98	4.34
Hispanic	1.00	0.10	2.05	2.24	3.13	2.64
Multiracial	N/A <sup>1</sup>	N/A <sup>1</sup>	N/A <sup>1</sup>	0.13	0.07	0.07
API	0.63	0.47	1.57	5.44	6.89	6.37
AI/AN	N/A <sup>1</sup>	0.41	0.29	0.50	0.46	0.46

<sup>1</sup>Values of N/A indicate no false positives for that cohort and algorithm.

The BISF and BIFSG algorithms perform very well on the API and Hispanic cohorts using this metric, meaning that individuals with a high imputed probability associated with API or Hispanic have a high probability of being API or Hispanic, which is consistent with the specificity metrics shown in subsection 5.3.7.2.



## Appendix C: Summary of Tutorial Imputation Packages

Package	Language	Methods								
		SA	GO	FO	S + F (i)	BISG	BIFSG	fBISG	fBIFSG	BIRDIE
Surgeo	Python	X	X	X		X	X			
Ethnicolr	Python	X			X					
WRU	R	X				X	X (ii)	X	X	
BIRDIE	R	X				X	(iii)	X	(iii)	X

(i) Surname plus first name only.

(ii) Wru also allows use of middle names.

(iii) Not implemented in BIRDIE, but user can provide estimated probabilities produced from other methods/packages.

	Data Sources	Geography Levels	Outputs
<b>Surgeo</b>	2010 Census  Demographic aspects of first names from mortgage applications.	ZCTA/ZIP Code State census tract (for BISG)	Six cohorts: White, Black, API, Native, Multiple, Hispanic
<b>Ethnicolr</b>	Ethnicolr1: 2000/2010 Census  Ethnicolr2: 2020 Census Florida and North Carolina voter registration	NA – models not based on geographic information	Multiple models trained on four-cohort and five-cohort data. Four cohorts: White, Black, Hispanic, Asian Five cohorts: White, Black, Hispanic, Asian, Other
<b>WRU</b>	2010 or 2020 Census	Census geolocations: county, tract, block group, block, place	Five cohorts: White, Black, Hispanic, Asian, Other
<b>BIRDIE</b>	Flexible: Can use decennial census or 1- and 5-year ACS surveys	State ZCTA/ZIP Code	Six cohorts: White, Black, Hispanic, Asian, American Indian, and Alaska Native, Other

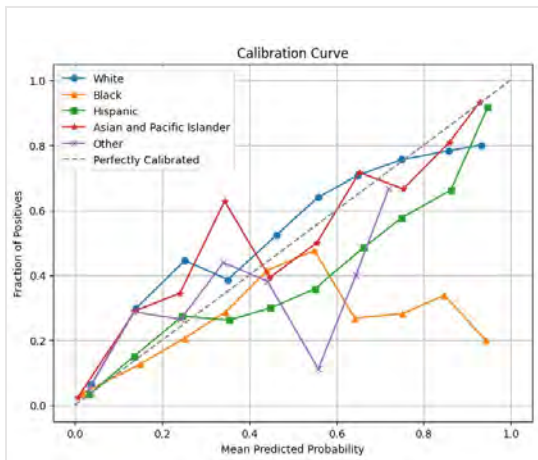
## Appendix D: Sample Tutorial Output

Performance metrics from the BIRDiE tutorial are shown below. Results on the tutorial data should not be used to draw any conclusions about the true population or accuracy of any particular method.

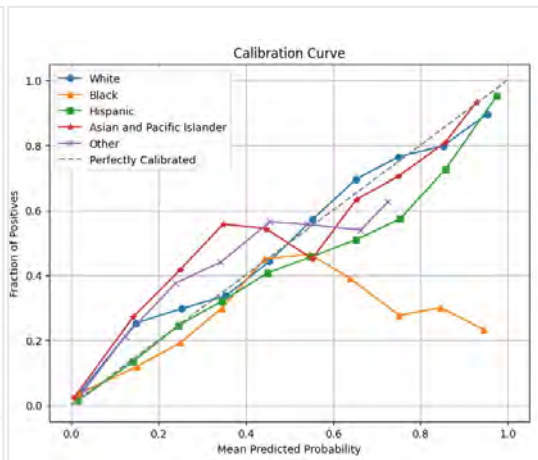
### D.1 CALIBRATION CURVES

A calibration curve below the 45-degree line indicates that the model overpredicts the probability, while a calibration curve above the 45-degree line indicates that the model underpredicts the probability.

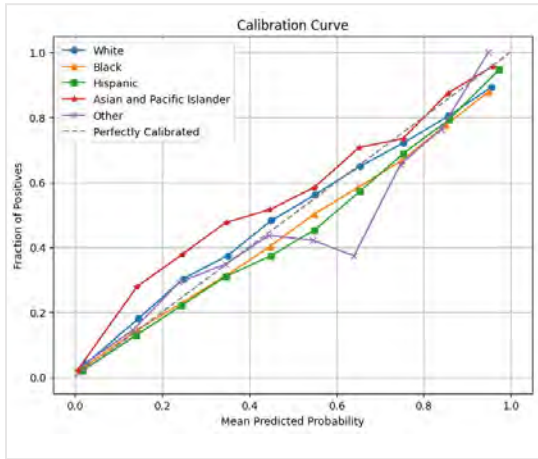
**Figure 5**  
**BIRDIE TUTORIAL — CALIBRATION CURVES**  
 SA



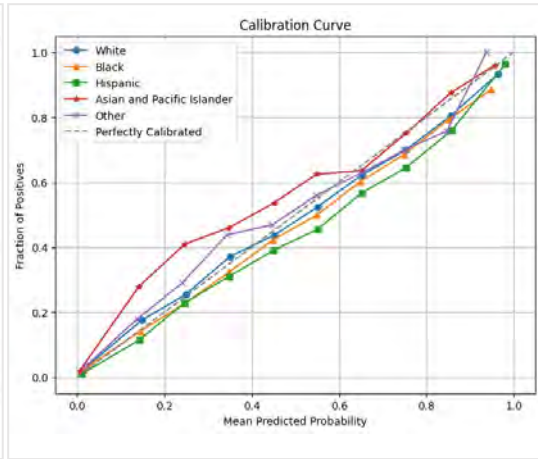
BIRDIE IMPROVED SA



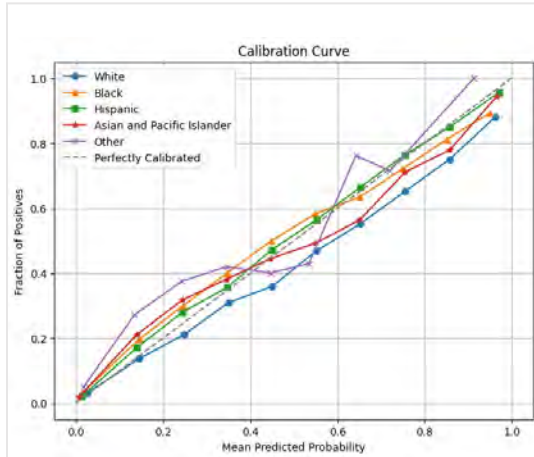
BISG



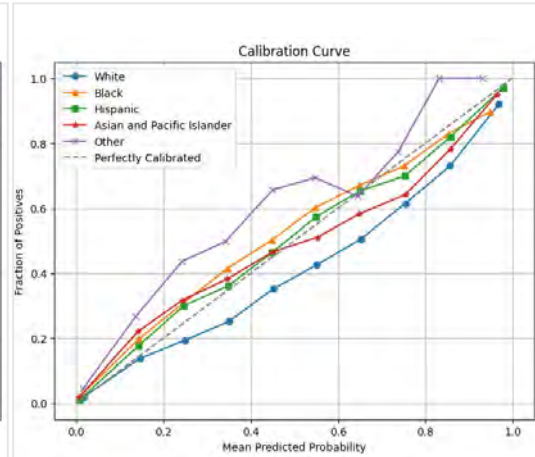
BIRDIE IMPROVED BISG



FBISG



BIRDIE IMPROVED fBISG



D.2 ACTUAL-TO-EXPECTED DISTRIBUTION

Table 39  
BIRDIE TUTORIAL - ACTUAL-TO- EXPECTED FOR AVERAGE PROBABILITY METHOD

Imputed Race/Ethnicity	SA	BISG	fBISG	BIRDIE Improved SA	BIRDIE Improved BISG	BIRDIE Improved fBISG
White	0.98	0.96	0.90	0.99	0.97	0.91
Black	0.91	0.99	1.19	0.91	0.99	1.18
Hispanic	0.95	0.96	1.03	0.95	0.96	1.01
API	1.45	1.42	1.25	1.41	1.37	1.22
Other	1.57	1.51	2.66	1.53	1.46	2.49

Table 40  
ACTUAL-TO-EXPECTED FOR CLASSIFICATION METHOD

Imputed Race/Ethnicity	SA	BISG	fBISG	BIRDIE Improved SA	BIRDIE Improved BISG	BIRDIE Improved fBISG
White	0.80	0.87	0.84	0.82	0.89	0.85
Black	2.40	1.16	1.49	2.16	1.13	1.43
Hispanic	1.02	1.00	1.03	0.98	0.96	1.00
API	1.45	1.50	1.39	1.44	1.46	1.35
Other	66.27	12.49	27.87	35.52	8.06	17.14

D.3 PROBABILITY OF CORRECT SELF-REPORTED RACE AND ETHNICITY PREDICTED

The true positive rate can be calculated for any imputation method that directly classifies observations or produces probabilities of a positive and applies a classification threshold. The analogous metric for the probabilities is the average probability of being associated with the correct self-reported race or ethnicity.

**Table 41**  
**AVERAGE PROBABILITY ASSOCIATED WITH CORRECT RACE AND ETHNICITY**

Self-Reported Race/Ethnicity	SA	BISG	fBISG	BIRDIE Improved SA	BIRDIE Improved BISG	BIRDIE Improved fBISG
White	75%	83%	87%	78%	86%	89%
Black	29%	51%	46%	30%	53%	48%
Hispanic	80%	83%	80%	86%	89%	87%
API	46%	52%	55%	48%	53%	57%
Other	5%	8%	5%	7%	12%	8%

**Table 42**  
**TRUE POSITIVE RATE – CLASSIFIED USING HIGHEST PROBABILITY**

Self-Reported Race/Ethnicity	SA	BISG	fBISG	BIRDIE Improved SA	BIRDIE Improved BISG	BIRDIE Improved fBISG
White	93%	93%	95%	94%	94%	96%
Black	16%	57%	48%	20%	59%	51%
Hispanic	84%	87%	86%	89%	92%	91%
API	55%	56%	59%	55%	58%	60%
Other	<1%	4%	2%	1%	7%	4%

#### D.4 PROBABILITY OF WHITE RACE PREDICTED

For each level of race or ethnicity, one can calculate the rate at which individuals in that cohort are predicted to be white by the imputation algorithm. This metric is important because imputation algorithms that overestimate the prevalence of white individuals can bias subsequent analyses.

**Table 43**  
**AVERAGE IMPUTED PROBABILITY ASSOCIATED WITH WHITE RACE**

Self-Reported Race/Ethnicity	SA	BISG	fBISG	BIRDIE Improved SA	BIRDIE Improved BISG	BIRDIE Improved fBISG
White	75%	83%	87%	78%	86%	89%
Black	61%	40%	47%	59%	37%	44%
Hispanic	16%	12%	15%	9%	6%	8%
API	28%	25%	27%	28%	25%	27%
Other	65%	65%	69%	55%	54%	59%

**Table 44**  
**POSITIVE RATE FOR WHITE PREDICTIONS – CLASSIFIED USING HIGHEST PROBABILITY**

Self-Reported Race/Ethnicity	SA	BISG	fBISG	BIRDIE Improved SA	BIRDIE Improved BISG	BIRDIE Improved fBISG
White	93%	93%	95%	94%	94%	96%
Black	83%	42%	51%	78%	38%	47%
Hispanic	14%	11%	12%	10%	6%	7%
API	29%	27%	28%	30%	27%	28%
Other	77%	74%	78%	70%	64%	70%

### D.5 RATIO OF TRUE POSITIVES TO FALSE POSITIVES

The ratio of true positives to false positives describes the odds that the imputed race/ethnicity group is consistent with the individual's self-reported race or ethnicity.

Table 45

RATIO OF AVERAGE PROBABILITY OF CORRECT SELF-REPORTED RACE AND ETHNICITY TO AVERAGE PROBABILITY OF INCORRECT RACE OR ETHNICITY

Imputed Race/Ethnicity	SA	BISG	fBISG	BIRDIE Improved SA	BIRDIE Improved BISG	BIRDIE Improved fBISG
White	3.00	4.85	6.70	3.55	6.07	8.42
Black	0.41	1.06	0.85	0.43	1.14	0.92
Hispanic	3.91	4.85	4.08	5.99	7.88	6.61
API	0.86	1.07	1.24	0.91	1.13	1.32
Other	0.05	0.08	0.05	0.08	0.14	0.09

Table 46

RATIO OF TRUE POSITIVES TO FALSE POSITIVES

Imputed Race/Ethnicity	SA	BISG	fBISG	BIRDIE Improved SA	BIRDIE Improved BISG	BIRDIE Improved fBISG
White	2.87	4.19	3.80	3.23	5.03	4.41
Black	0.60	1.90	2.45	0.78	2.06	2.72
Hispanic	6.33	6.62	7.71	6.86	7.95	9.98
API	4.01	5.41	4.46	3.76	5.51	4.36
Other	0.37	0.86	1.13	1.13	1.36	1.93

### D.6 AREA UNDER THE RECEIVER OPERATING CHARACTERISTIC CURVE (AUC)

A higher AUC represents a better ability to distinguish between cohorts.

Table 47

AUC

Self-Reported Race/Ethnicity	SA	BISG	fBISG	BIRDIE Improved SA	BIRDIE Improved BISG	BIRDIE Improved fBISG
White	0.80	0.89	0.89	0.86	0.93	0.93
Black	0.80	0.91	0.91	0.81	0.92	0.92
Hispanic	0.93	0.97	0.97	0.98	0.99	0.99
API	0.83	0.90	0.92	0.87	0.92	0.93
Other	0.47	0.62	0.63	0.70	0.74	0.77

### D.7 PRECISION

Precision for each level of race or ethnicity is the proportion of people predicted to be that race or ethnicity who self-report that race or ethnicity.

**Table 48**  
**PRECISION**

Self-Reported Race/Ethnicity	SA	BISG	fBISG	BIRDIE Improved SA	BIRDIE Improved BISG	BIRDIE Improved fBISG
White	74%	81%	79%	76%	83%	82%
Black	37%	65%	71%	44%	67%	73%
Hispanic	86%	87%	89%	87%	89%	91%
API	80%	84%	82%	79%	85%	81%
Other	27%	46%	53%	53%	58%	66%

### D.8 SPECIFICITY

True negative rate is the proportion of negative observations with negative predictions. It measures the ability of the imputation to detect negative responses among all the data.

**Table 49**  
**SPECIFICITY**

Self-Reported Race/Ethnicity	SA	BISG	fBISG	BIRDIE Improved SA	BIRDIE Improved BISG	BIRDIE Improved fBISG
White	56%	70%	67%	61%	75%	71%
Black	97%	96%	97%	97%	96%	97%
Hispanic	97%	97%	97%	97%	97%	98%
API	99%	99%	99%	99%	99%	99%
Other	99.9%	99.8%	99.9%	99.9%	99.7%	99.9%

### D.9 BIRDIE OUTCOME ESTIMATES

Although the focus of this paper is not to look at how to analyze differences in outcomes, the BIRDIE package was developed to model unbiased estimates of outcomes by race and ethnicity. As mentioned above, the data used in this tutorial is not real-world data and the outcome variable predicted doesn't have any meaning. The authors have included the example here to illustrate this additional application of BIRDIE.

**Table 50**  
**PROBABILITY-WEIGHTED AVERAGE OUTCOME BY GROUP**

Self-Reported Race/Ethnicity	SA	BISG	fBISG	BIRDIE with SA probabilities	BIRDIE with BISG probabilities	BIRDIE with fBISG probabilities	Actual Outcome
White	9.90	9.90	9.90	9.49	9.61	9.58	9.90
Black	8.61	8.62	8.62	9.27	8.90	8.86	8.62
Hispanic	6.13	6.13	6.13	6.46	6.38	6.35	6.14
API	8.50	8.49	8.49	8.45	8.44	8.41	8.50
Other	7.40	7.39	7.39	8.78	8.59	8.56	7.37

**Table 51**  
**AVERAGE OUTCOME BY GROUP – CLASSIFIED USING HIGHEST PROBABILITY**

Self-Reported Race/Ethnicity	SA	BISG	fBISG	BIRDIE with SA probabilities	BIRDIE with BISG probabilities	BIRDIE with fBISG probabilities	Actual Outcome
White	9.40	9.49	9.47	9.43	9.55	9.51	9.90
Black	9.17	8.88	8.80	9.18	8.86	8.79	8.62
Hispanic	6.50	6.49	6.44	6.52	6.46	6.41	6.14
API	8.45	8.46	8.44	8.46	8.43	8.40	8.50
Other	8.68	8.23	8.19	8.51	8.28	8.11	7.37

## References

- American Academy of Actuaries. (2022). *Sourcing Protected Class Information in P&C Insurance*.
- Baines, A. P., & Courchane, M. J. (2014). *Fair Lending: Implications for the Indirect Auto Finance Market*.
- CFPB. (2014). *Using publicly available information to proxy for unidentified race and ethnicity*.
- Chintalapati, R., Laohaprapanon, S., & Sood, G. (2023). *Predicting Race and Ethnicity From the Sequence of Characters in a Name*.
- Comenetz, J. (2016). *Frequently Occurring Surnames in the 2010 Census*.
- Coronado, G. D., Koepsell, T. D., Thompson, B., Schwartz, S. M., Wharton, R. S., & Grossman, J. E. (2002). *Assessing Cervical Cancer Risk in Hispanics*.
- Decter-Frain, A. (2022). "How Should We Proxy for Race/Ethnicity? Comparing Bayesian Improved Surname Geocoding to Machine Learning Methods".
- Elliott, M. N., Fremont, A., Morrison, P. A., Pantoja, P., & Lurie, N. (2008). *A New Method for Estimating Race/Ethnicity and Associated Disparities Where Administrative Records Lack Self-Reported Race/Ethnicity*.
- Elliott, M. N., Morrison, P. A., Fremont, A., McCaffrey, D. F., Pantoja, P., & Luri, N. (2009). *Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities*.
- Federal Trade Commission. (2007). *Credit-Based Insurance Scores: Impacts on Consumers of Automobile Insurance*.
- Fiscella, K., & Fremont, A. M. (2006). *Use of Geocoding and Surname Analysis to Estimate Race and Ethnicity*.
- Fisher, R. (2023). *Estimation of Race and Ethnicity by Re-Weighting Tax Data*.
- Haas, A., Elliott, M. N., Dembosky, J. W., Adams, J. L., Wilson-Frederick, S. M., Mallett, J. S., . . . Haviland, A. M. (2019). *Imputation of race/ethnicity to enable measurement of HEDIS performance by race/ethnicity*.
- Haley, J. M., Dubay, L., Garrett, B., Caraveo, C. A., Schuman, I., Johnson, K., . . . DePoy, B. (2022). *Collection of Race and Ethnicity Data for Use by Health Plans to Advance Health Equity*.
- Imai, K., & Khanna, K. (2016). *Improving Ecological Inference by Predicting Individual Ethnicity from Voter Registration Records*.
- Imai, K., Olivella, S., & Rosenman, E. T. (2022). *Addressing census data problems in race imputation via fully Bayesian Improved Surname Geocoding and name supplements*.
- Jacobs, E. A., & Lauderdale, D. S. (2001). *Receipt of Cancer Screening Procedures Among Hispanic and Non-Hispanic Health Maintenance Organization Members*.
- Krieger, N., Chen, J. T., Waterman, P. D., Soobader, M.-J., Subramanian, S., & Carson, R. (2002). *Geocoding and Monitoring of US Socioeconomic Inequalities in Mortality and Cancer Incidence: Does the Choice of Area-Based Measure and Geographic Level Matter?: The Public Health Disparities Geocoding Project*.
- Lauderdale, D. S., & Kestenbaum, B. (2000). *Asian American Ethnic Identification by Surname*.



- Martino, S. C., Weinick, R. M., Kanouse, D. E., Brown, J. A., Haviland, A. M., Goldstein, E., . . . Elliott, M. N. (2013). *Reporting CAHPS and HEDIS Data by Race/Ethnicity for Medicare Beneficiaries*.
- McCartan, C., Goldin, J., Ho, D. E., & Imai, K. (2023). *Estimating Racial Disparities When Race is Not Observed*.
- NAIC. (2008). *Review of the Use of Credit-Based Insurance Scoring By Insurers*.
- Pérez-Stable, E. J., Hiatt, R. A., Sabogal, F., & Otero-Sabogal, R. (1995). *Use of Spanish surnames to identify Latinos: comparison to self-identification*.
- Perkins, R. C. (1993). *Evaluating the Passel-Word Spanish Surname List: 1990 Decennial Census Post Enumeration Survey Results*.
- Rosenwaike, I., Hempstead, K., & Rogers, R. G. (1991). *Using surname data in U.S. Puerto Rican mortality analysis*.
- Sorbero, M. E., Euller, R., Kofner, A., & Elliott, M. N. (2022). *Imputation of Race and Ethnicity in Health Insurance Marketplace Enrollment Data, 2015–2022 Open Enrollment Periods*.
- Swallen, K. C., Glaser, S. L., Stewart, S. L., West, D. W., Jenkins, C. N., & McPhee, S. J. (1998). *Accuracy of racial classification of Vietnamese patients in a population-based cancer registry*.
- Swallen, K. C., West, D. W., Stewart, S. L., Glaser, S. L., & Horn-Ross, P. L. (1997). *Predictors of misclassification of hispanic ethnicity in a population-based cancer registry*.
- Tzioumis, K. (2017). *Demographic aspects of first names*.
- Voicu, I. (2018). *Using First Name Information to Improve Race and Ethnicity Classification*.
- Word, D. L., & Perkins Jr., R. C. (1996). *Building a Spanish Surname List for the 1990's—A New Approach to an Old Problem*.
- Word, D. L., Coleman, C. D., Nunziata, R., & Kominski, R. (2007). *Demographic Aspects of Surnames from Census 2000*.
- Xie, F. (2021). *Predicting Ethnicity from Names with rethnicity: Methodology and Application*.
- Xue, Y., Harel, O., & Aseltine, R. H. (2019). *Imputing race and ethnic information in administrative health data*.
- Zavez, K., Harel, O., & Aseltine, R. H. (2021). *Imputing race and ethnicity in healthcare claims databases*.

## About The Society of Actuaries Research Institute

Serving as the research arm of the Society of Actuaries (SOA), the SOA Research Institute provides objective, data-driven research bringing together tried and true practices and future-focused approaches to address societal challenges and your business needs. The Institute provides trusted knowledge, extensive experience and new technologies to help effectively identify, predict and manage risks.

Representing the thousands of actuaries who help conduct critical research, the SOA Research Institute provides clarity and solutions on risks and societal challenges. The Institute connects actuaries, academics, employers, the insurance industry, regulators, research partners, foundations and research institutions, sponsors and non-governmental organizations, building an effective network which provides support, knowledge and expertise regarding the management of risk to benefit the industry and the public.

Managed by experienced actuaries and research experts from a broad range of industries, the SOA Research Institute creates, funds, develops and distributes research to elevate actuaries as leaders in measuring and managing risk. These efforts include studies, essay collections, webcasts, research papers, survey reports, and original research on topics impacting society.

Harnessing its peer-reviewed research, leading-edge technologies, new data tools and innovative practices, the Institute seeks to understand the underlying causes of risk and the possible outcomes. The Institute develops objective research spanning a variety of topics with its [strategic research programs](#): aging and retirement; actuarial innovation and technology; mortality and longevity; diversity, equity and inclusion; health care cost trends; and catastrophe and climate risk. The Institute has a large volume of [topical research available](#), including an expanding collection of international and market-specific research, experience studies, models and timely research.

Society of Actuaries Research Institute  
8770 W Bryn Mawr, 10<sup>th</sup> Fl  
Chicago, Illinois 60631  
[www.SOA.org](http://www.SOA.org)