# Comparison of Imputation Methods for Race and Ethnic Information in Administrative Health Data

[1]Yishu Xue, [1]Ofer Harel, and [2]Robert Aseltine Jr.
[1]Department of Statistics, University of Connecticut, Storrs, CT 06269, USA
[2]Behavioral Science and Community Health, UConn Health, Farmington, CT 06030, USA

***Abstract*—In the United States of America where there is no national health care, All-Payer Claims Databases provide great resources to investigate and address disparities in access to, utilization, and outcomes of care. Race/ethnicity being missing, however, is a bottleneck on its usage. In most health claim databases Race/ethnicity only observed to 3-5% of the observations, causing a great missing data problem. We try to recover race/ethnicity information for incomplete observations based on studies of the (3%) complete observations. To emulate the data structure, an analysis of birth records from Connecticut is done where the race/ethnicity information is complete, in order to assess competing models performances. While the Connecticut-based full model based on logistic model proposed achieves over 80% prediction accuracy, we are interested in comparing this model performance to more complex machine learning methods and evaluate prediction. An empirical study is presented.**

***Index Terms*—Health insurance claims; Race/ethnicity; Imputation; Missing data; Decision Tree.**

## I. Introduction

In the United States of America (USA) where no national health care is available, All-Payer Claims Databases (APCDs), which are currently established or in formation in 25 states, offer great opportunities to studies of health disparities. However, the majority of patients' self-reported race and ethnicity information is missing in almost all these APCDs. Only approximately 3% of commercially insured beneficiaries have this information available. This missingness greatly limited the usage of APCDs in analysis of racial and ethnic disparities in the utilization and outcomes of care.

To date, the most common missing data procedure is complete case analysis (CCA), in which every observation with incomplete data is being eliminated from the data. It has been shown that CCA unbiased in very limited situations, has a great impact on measures of variance and covariance, and almost always is inefficient. In addition, in situations with large amounts of missing data, most of the data can be eliminated under this procedure. Therefore, this is not a viable solution.

Various indirect methods have been proposed to infer or assign race/ethnicity based on patient information [1]–[4]. Among these approaches, [3], [4] developed Bayesian Surname and Geocoding (BSG) and Bayesian Improved Surname and Geocoding (BISG) methods that combine information from people's surnames and their residential geocoding to produce posterior estimates of probability for one to be in different race/ethnicity groups. These methods, however, have not been implemented in scenarios where a majority of the outcome variable is missing.

To overcome the missing data barrier, and utilize information from other sources than only geocoding and surname, we proposed a multinomial logistic regression approach that calculates the probability vector of individuals to be in different race/ethnic groups. Using the statewide birth registry records of Connecticut (CT), a multinomial logistic regression model is trained on a small subsample of the data, and tested on a testing dataset of 50% size of the entire dataset. The proposed method has been shown to work better than the BSG and BISG methods [5]. Here, we compare the performance of the methods proposed in [5] with popular machine learning algorithms, such as decision trees, and present the results.

## II. Data

In order to evaluate the competing models, we needed to have a complete data. For that we used the statewide birth registry obtained from the Connecticut Department of Public Health. Race is available in this dataset, and therefore we are able to compare the predicted race with the truth. Using the geographical information of Connecticut tracts, and the addresses provided by parents in the birth registry, after subjects with invalid or out-of state addresses are removed, the 162,188 observations, who were born between 2009 and 2013, were geocoded and matched with 827 tracts. Following the common procedure in the field [6], the race/ethnicity for each child were defined to be the self-reported race/ethnicity of the mother. Race/ethnicity was grouped into four big categories: White non-Hispanic (White, 57%), Black non-Hispanic (Black, 13%), Hispanic (22%), and Other (8%).

It is important to note, that more complex specification of the Race and Ethnicity distribution is possible. However, for simplicity and in order to introduced the methodology a simplistic definition of Race and ethnicity is used.

Each mother's surname is associated with a vector of probabilities of being in each of the four groups based on results from the 2010 Census. For example, "Nguyen" gets (1%, 0%, 1%, 98%), and "Smith" gets (71%, 23%, 2%, 4%). If a name is not found in the surname dictionary from 2010 Census, a non-informative (25%, 25%, 25%, 25%) vector is associated with it. This 4-dimensional probability vector included in the predictors of the model. In addition, the distribution of race/ethnicity in Connecticut tracts are obtained from the 2010 Census, and represented in percentages and coded consistent with the four race/ethnicity categories, which makes another

4-dimensional probability vector, and is included in the model as well.

In addition, several other demographic covariates have been identified. The insurance payment method for baby delivery is grouped into four categories: private (60%), Medicaid (35%), None/Self (4%), and Other Insurance (1%). The age of the mother when giving birth is calculated, having a mean of 29.9 and standard deviation of 6.1. A binary variable for fathers' information being missing is introduced. There are 11% father missing cases. These covariates are available in the birth certificate, but they represent the ability to use other covariates in the future when the APCD data are used.

## III. Methods

### A. Proposed Models

The dataset was partitioned into two equally-sized training and testing sets. To mimic the missingness of race/ethnicity, 97% race/ethnicity information in the training data is removed (completely at random), leaving only 3%, i.e., 2,432 informative observations. Next, based on the 3% informative observations, two multinomial logistic regression models were fitted, one using all information available and denoted the CT Based Full (CTBF), and the other using all but the three demographic covariates denoted the CT Based Reduced (CTBR). White category was used as the reference category. Predictions were made using the coefficients obtained from the models on the testing dataset, and compared to the true race/ethnicity. Multiple performance measures were used.

As a comparison, we applied algorithms in machine learning on the same 3% of training dataset. The algorithms considered include neural network, k-nearest neighbors, and decision trees [7].

## IV. Performance Measures

Two performance measures discussed in Elliott et al. [3] are used. Denoting the true prevalence of race/ethnicity in the testing dataset as $(p_w, p_b, p_h, p_o)$, each corresponding to one group, and the predicted prevalence for the testing dataset as $(q_w, q_b, q_h, q_o)$, the weighted measure of error in predicted prevalence is defined as

$$\begin{aligned}\text{Error}_{prevalence} = &|p_w - q_w| \cdot p_w + |p_b - q_b| \cdot p_b \\ &+ |p_h - q_h| \cdot p_h + |p_o - q_o| \cdot p_o.\end{aligned}$$

The other measure used by Elliott et al. [3] is the weighted correlation of each individual's true race/ethnicity and their predicted race/ethnicity, where the weights are the true prevalence, $(p_w, p_b, p_h, p_o)$. For all observations in the testing dataset, a vector of indicators $r_{wt}$ that equals 1 if the true race is White, and 0 otherwise. Another vector of indicators $r_{wp}$ can be obtained for the predictions, equaling 1 if the predicted race is White, and 0 otherwise. The correlation coefficient between $r_{wt}$ and $r_{wp}$ can be obtained and denoted as $\text{corr}_w$. Similarly, we can calculate $\text{corr}_b$, $\text{corr}_h$, and $\text{corr}_o$, as correlation coefficients of black, hispanic and other respectively. The weighted correlation is calculated as:

$$\text{Correlation}_{\text{weighted}} = \text{corr}_w \cdot p_w + \text{corr}_b \cdot p_b + \text{corr}_h \cdot p_h + \text{corr}_o \cdot p_o.$$

TABLE I

Comparison of imputed race/ethnicity to self-reported race/ethnicity for testing subset of CT birth records ($n = 81,094$)

| CT Based Reduced Model (a) | | | | | |
|---|---|---|---|---|---|
| | Self-reported, $n$ | | | | |
| | White | Black | Hispanic | Other | Total |
| Imputed, $n$ | | | | | |
| White | **41,220** | 4,574 | 3,118 | 2,610 | 51,522 |
| Black | 1,708 | **5,032** | 624 | 289 | 7,653 |
| Hispanic | 3,118 | 623 | **14,275** | 337 | 18,353 |
| Other | 324 | 78 | 60 | **3,104** | 3,566 |
| Total | 46,370 | 10,307 | 18,077 | 6,340 | 81,094 |
| Sensitivity, % | 89 | 49 | 79 | 49 | |
| Sepcificity, % | 70 | 96 | 94 | 99 | |
| Cohen's Kappa | 0.62 | | | | |
| Weighted Error | 0.044 | | | | |
| $\text{Correlation}_{\text{weighted}}$ | 0.623 | | | | |
| Correct rate % | 78.47 | | | | |

| CT Based Full Model (b) | | | | | |
|---|---|---|---|---|---|
| | Self-reported, $n$ | | | | |
| | White | Black | Hispanic | Other | Total |
| Imputed, $n$ | | | | | |
| White | **42,330** | 3,078 | 2,971 | 2,388 | 50,767 |
| Black | 1,135 | **6,398** | 703 | 324 | 8,560 |
| Hispanic | 2,479 | 717 | **14,310** | 308 | 17,814 |
| Other | 426 | 114 | 93 | **3,320** | 3,953 |
| Total | 46,370 | 10,307 | 18,077 | 6,340 | 81,094 |
| Sensitivity, % | 91 | 62 | 79 | 52 | |
| Sepcificity, % | 76 | 97 | 94 | 99 | |
| Cohen's Kappa | 0.68 | | | | |
| Weighted Error | 0.037 | | | | |
| $\text{Correlation}_{\text{weighted}}$ | 0.688 | | | | |
| Correct rate % | 81.83 | | | | |

The measures described above are also supplemented with another four common measures of accuracy: sensitivity & specificity [8], Cohen's Kappa [9], and percentage of correct predictions.

## V. Results and Performance Comparison

For the 81,094 observations in the testing dataset, we obtained the predictions given by the CTBF and CTBR, formulated the confusion matrix [10], and computed the performance measures described in Section IV. We report the confusion matrices, together with the performance metrics, in Table I. The results indicates that by combining the information from surnames and census tract race distributions, we are able to predict one's race/ethnicity with around 80% accuracy. Furthermore, upon closer investigation, we found that tract race/ethnicity distributions are highly significant for the prediction of minorities, and the regression coefficients of mother's age and father missing are highly significant for specific race/ethnicity groups, which indicates the helpfulness of additional information other than surname and address.

Using the same set of covariates as the CTBF, a neural network is fitted using the nnet function in the R package caret. The number of hidden units were tuned on values from 1 to 10, and the parameter for weight decay was tuned on

TABLE II
PERFORMANCE OF NEURAL NETWORK MODEL ($n = 81,094$)

| | Self-reported, $n$ | | | | |
|---|---|---|---|---|---|
| | White | Black | Hispanic | Other | Total |
| Imputed, $n$ | | | | | |
| White | **41,912** | 2,686 | 2,941 | 2,300 | 49,839 |
| Black | 1,500 | **6,838** | 767 | 386 | 9,491 |
| Hispanic | 2,519 | 683 | **14,286** | 321 | 17,809 |
| Other | 439 | 100 | 83 | **3,333** | 3,955 |
| Total | 46,370 | 10,307 | 18,077 | 6,340 | 81094 |
| Sensitivity, % | 90 | 66 | 79 | 53 | |
| Sepcificity, % | 77 | 96 | 94 | 99 | |
| Cohen's Kappa | 0.69 | | | | |
| Weighted Error | 0.029 | | | | |
| $\text{Correlation}_{\text{weighted}}$ | 0.690 | | | | |
| Correct rate % | 81.84 | | | | |

TABLE III
PERFORMANCE OF K-NN MODEL ($n = 81,094$)

| | Self-reported, $n$ | | | | |
|---|---|---|---|---|---|
| | White | Black | Hispanic | Other | Total |
| Imputed, $n$ | | | | | |
| White | **42,408** | 3,201 | 3,534 | 2,378 | 51,521 |
| Black | 1,249 | **6,393** | 682 | 329 | 8,653 |
| Hispanic | 2,274 | 611 | **13,777** | 337 | 16,999 |
| Other | 439 | 102 | 84 | **3,296** | 3,921 |
| Total | 46,370 | 10,307 | 18,077 | 6,340 | 81,094 |
| Sensitivity, % | 91 | 62 | 76 | 52 | |
| Sepcificity, % | 74 | 97 | 95 | 99 | |
| Cohen's Kappa | 0.67 | | | | |
| Weighted Error | 0.044 | | | | |
| $\text{Correlation}_{\text{weighted}}$ | 0.676 | | | | |
| Correct rate % | 81.23 | | | | |

TABLE IV
PERFORMANCE OF RANDOM FOREST MODEL ($n = 81,094$)

| | Self-reported, $n$ | | | | |
|---|---|---|---|---|---|
| | White | Black | Hispanic | Other | Total |
| Imputed, $n$ | | | | | |
| White | **41,412** | 2,671 | 3,192 | 2,097 | 49,372 |
| Black | 1,551 | **6,679** | 825 | 412 | 9,467 |
| Hispanic | 2,647 | 740 | **13,850** | 351 | 17,588 |
| Other | 760 | 217 | 210 | **3,480** | 4,667 |
| Total | 46,370 | 10,307 | 18,077 | 6,340 | 81,094 |
| Sensitivity, % | 89 | 65 | 77 | 55 | |
| Sepcificity, % | 77 | 96 | 94 | 98 | |
| Cohen's Kappa | 0.67 | | | | |
| Weighted Error | 0.025 | | | | |
| $\text{Correlation}_{\text{weighted}}$ | 0.672 | | | | |
| Correct rate % | 80.67 | | | | |

TABLE V
PERFORMANCE OF STOCHASTIC GRADIENT BOOSTING MODEL ($n = 81,094$)

| | Self-reported, $n$ | | | | |
|---|---|---|---|---|---|
| | White | Black | Hispanic | Other | Total |
| Imputed, $n$ | | | | | |
| White | **42,381** | 3,164 | 3,079 | 2,374 | 50,998 |
| Black | 1,161 | **6,441** | 900 | 364 | 8,866 |
| Hispanic | 2,364 | 577 | **14,014** | 285 | 17,240 |
| Other | 464 | 125 | 84 | **3,317** | 3,990 |
| Total | 46,370 | 10,307 | 18,077 | 6,340 | 81,094 |
| Sensitivity, % | 91 | 62 | 78 | 52 | |
| Sepcificity, % | 75 | 97 | 95 | 99 | |
| Cohen's Kappa | 0.68 | | | | |
| Weighted Error | 0.039 | | | | |
| $\text{Correlation}_{\text{weighted}}$ | 0.684 | | | | |
| Correct rate % | 81.57 | | | | |

(0, 0.1, 1, 2). Correct prediction rate was used to select the best performing model. The final model was selected to have 3 hidden units, and has 2 as the weight decay parameter. The detailed results are reported in Table II. It can be seen that, compared to the CTBF, the neural network model achieves almost the same prediction accuracy rate. While having slightly improved weighted error and weighted correlation, it still fails to further improve the relatively low sensitivity for the Other category.

Another k-nn model is fitted using the knn3 function, also in the caret package. The number of neighbors, $k$, was tuned on a sequence of values (1, 5, 9, 13, 17, 21, 41, 61, 81). Again, correct prediction rate was used as the criterion for evaluation. The optimal value was selected to have $k = 21$. The results are shown in Table III. Overall the k-nn model provides slightly inferior but still comparable performance than the CTBF. It is worth noticing that the k-nn has lower specificity for White than the CTBF and the neural network model. One possible reason is the fact that White is the dominant class, and prediction of a minority can be significantly biased when there are observations in the White category that resemble it.

In addition to neural network and k-nn, two tree-based models are considered: random forest, and stochastic gradient boosting. A random forest model is fitted using the randomForest function. For the random forest model, as under our hypothetical missing-majority setting, only 3% observations in the training data have observed race and we only have 2,432 such complete observations, in an effort to control over-fitting, 50 trees were used. The number of variables was fixed at 8 for random sample at each split. As seen from Table IV, the random forest model also performs similarly to the proposed CTBF. Compared to the CTBF, it has higher sensitivity for Black, but slightly lower sensitivity for White. The specificities are very close. It is worth noticing that, the random forest model provides better estimation of overall prevalence.

For the stochastic gradient boosting model (GBM), provided by the R package gbm, a 5-fold cross-validation is used. The tree depth was chosen to be 1, i.e., we use decision stumps. The minimum number of observations per node was set to be 10. The learning rate was set to be 0.1. From Table V, the GBM performs better in terms of correct rate than the random forest. Similar to the random forest, it does not achieve better predictions than the CTBF.

## VI. CONCLUSION

We compared performance of CTBF with four popular machine learning procedures: neural network, k-nn, random forest and GBM, when only $3\%$ of race is known and we want to predict the rest. While all four methods provided decent predictions in terms of different measures, none is significantly better than the CTBF.

In addition, training such models usually requires more complex and time consuming computation, while a multinomial logistic regression can be obtained rather quickly on a common laptop. This indicates that the proposed multinomial logistic regression model is able to utilize nearly all information in the dataset in an appropriate way. Compared to more complex machine learning methods, the regression parameters are also informative. They provide insight as to which variables are highly indicative of which race/ethnicity.

## REFERENCES

[1] N. R. Council *et al.*, *Eliminating health disparities: measurement and data needs*. National Academies Press, 2004.

[2] K. Fiscella and A. M. Fremont, "Use of geocoding and surname analysis to estimate race and ethnicity," *Health services research*, vol. 41, no. 4p1, pp. 1482–1500, 2006.

[3] M. N. Elliott, A. Fremont, P. A. Morrison, P. Pantoja, and N. Lurie, "A new method for estimating race/ethnicity and associated disparities where administrative records lack self-reported race/ethnicity," *Health Services Research*, vol. 43, no. 5p1, pp. 1722–1736, 2008.

[4] M. N. Elliott, P. A. Morrison, A. Fremont, D. F. McCaffrey, P. Pantoja, and N. Lurie, "Using the Census Bureaus surname list to improve estimates of race/ethnicity and associated disparities," *Health Services and Outcomes Research Methodology*, vol. 9, no. 2, p. 69, 2009.

[5] Y. Xue, O. Harel, and R. Aseltine Jr, "Imputing race and ethnic information in administrative health data," University of Connecticut, Department of Statistics, Tech. Rep., 2018.

[6] L. R. Mason, Y. Nam, and Y. Kim, "Validity of infant race/ethnicity from birth certificates in the context of US demographic change," *Health Services Research*, vol. 49, no. 1, pp. 249–267, 2014.

[7] M. Kuhn and K. Johnson, *Applied predictive modeling*. Springer, 2013, vol. 26.

[8] M. S. Pepe, *The statistical evaluation of medical tests for classification and prediction*. Medicine, 2003.

[9] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[10] K. M. Ting, *Confusion Matrix*. Boston, MA: Springer US, 2017, pp. 260–260.