Supplementing Missing Self-Reported Race Data with a Probability Distribution in Logistic Regression Models

Stanley Xu[1,2,*], Komal Narwaney[1], Sophia Newcomer[1] and Jason Glanz[1,2]

[1]*The Institute for Health Research, Kaiser Permanente Colorado, Denver, CO, USA*

[2]*School of Public Health, University of Colorado, Aurora, CO, USA*

Abstract: Race is often included as an independent variable in health services research, especially in recent studies of racial and ethnic disparities in health care. Although self-reported information on race exists in large electronic health records (EHR) data, these data are sometimes missing. Recently Bayesian Improved Surname Geocoding method (BISG) is used to estimate the probability distribution of race categories for those with missing information on race. The BISG estimated probability distribution has been used in reporting health care measures but not in statistical modellings with dichotomous events as outcomes. We propose two approaches to accommodate available distribution probability of an independent categorical variable (e.g., race) in logistic regression models: 1) a direct substitution approach and 2) a partial information maximum likelihood estimator (PIMLE). In examining the association between race and up-to-date immunization status of children by three years old from an integrated health care organization, 11.3% of 14,903 children have missing self-reported race information but have BISG estimated probability distribution for the six race/ethnicity categories. We employed the direct substitution approach and PIMLE approach to analyze the under vaccination data. Both approaches included all observations and thus yielded smaller standard errors of estimated coefficients compared to the complete data analyses. Our simulation study showed that the direct substitution approach and PIMLE yielded nearly unbiased coefficient estimates and preserved efficiency when the missing rate of the independent categorical variable was up to 30%.

Keywords: Race and ethnicity, Bayesian Improved Surname Geocoding, up-to-date immunization, direct substitution approach, partial information maximum likelihood estimator.

1. INTRODUCTION

Demographic characteristics such as race are important independent variables (risk factors or predictors) in health services research. Self-reported information on race has been considered as superior to other sources such as observed race [1]. Self-reported information on race often exists in large electronic health records (EHR) data; however, these data are sometimes missing. Reasons for missing race data include the patient declining to report this information or the provider failing to obtain or document this information.

Recent focus on racial and ethnic disparities in health care has encouraged health care organizations to increase their effort to fill in missing race and ethnicity data [2,3]. While ideally race would be self-reported by all patients, other approaches such as the recent Bayesian Improved Surname Geocoding method (BISG) have been developed to compensate for missing information on race [4,5]. BISG utilizes a Bayesian approach to combine racial/ethnic data from last names and geographic units to calculate the probability distribution of race categories for a given individual whose self-reported race/ethnicity is missing in EHR. In 2008, Kaiser Permanente, an integrated health care organization, began applying the BISG algorithm to geocoded member addresses to link to Census Bureau data which describes the racial/ethnic composition of census block groups [6]. Based on members' address and surname analysis using the Census Bureau's list of more than 150,000 surnames and their association with race/ethnicity, an individual's probability distribution is estimated for the following six standardized mutually-exclusive racial/ethnic categories: Asian or Pacific Islander, Black or African American, Hispanic or Latino, American Indian or Alaska Native, Multiracial, and White. In this paper, we use A, B, H, N, M and W to denote these six categories, respectively. Adjaye-Gbewonyo *et al*. [6] validated the classification of race/ethnicity based on the BISG and concluded that BISG may be useful for classifying race/ethnicity of health plan members when needed for health care studies. They also showed that sensitivity and specificity of classification varied by race/ethnic group: using a cutoff of 0.5, sensitivity for A, B, H, N, M and W was 64.4, 71.8, 71.0, 0.0, 0.3, and 85.2 percent, respectively; specificity was 99.6, 91.1, 99.0, 100.0,100.0, and 76.8 percent, respectively. It remains of great interest in health care research how to use the newly available probability distribution information without classifying race/ethnicity along with existing self-reported race information in analyzing EHR data.

*Address correspondence to this author at the Kaiser Permanente Colorado, Institute for Health Research, Denver Highlands, 10065 E. Harvard Ave., Suite 300, Denver CO 80231, USA; Tel: (303) 614-1200; Fax: (303)-614-1225; E-mail: stan.xu@kp.org

E-ISSN: 1929-6029/15

Analyzing outcomes with covariates (e.g., race) missing is challenging. Usually multiple imputation can be used to impute missing values of important independent variables such as race in analyzing outcomes [7-9]. In a recent study examining the association between initial antihyperglycemic therapy and patient-level baseline characteristics, to enable inclusion of patients for whom data were missing on race (17.4%), the authors employed multiple imputations, imputing each missing value five times using site-specific distributions of race [9,10]. With the availability of BISG estimated race probability distribution, other approaches may be more appropriate and more efficient. Recently McCaffrey and Elliott [11] suggested a direct substitution approach and a partial information maximum likelihood estimator approach for a linear regression model with missing binary independent variable but its probability distribution is available.

This paper focuses on dichotomous dependent variable (e.g., up-to-date immunization status of children) and independent categorical variable (i.e., race). We propose logistic regression models that use 1) the direct substitution approach and 2) a partial information maximum likelihood estimator when missing independent variable is categorical (e.g., race). We then demonstrate these two methods in analyzing up-to-date immunization status of children by three years old in an integrated health care organization. We also conduct simulation study to evaluate these two approaches regarding their bias and efficiency.

2. STATISTICAL METHODS

2.1. Statistical Models with a Categorical Variable as an Independent Variable without Missing

Suppose a categorical independent variable z has K levels. Let y_i be the dependent variable for the ith subject where i=1 to n. For example, y_i=1 if the immunization status of a child is up-to-date and y_i=0 if the immunization status of a child is not up-to-date. Also let $\mu_{1(i)} = \alpha_r + I_{iz}\alpha_k + x_i\beta$, where α_r is the coefficient for the reference category of the categorical independent variable, a_k is the coefficient for category $k(k \neq r), k = 1$ to $K-1, I_{iz}$ is the indicator variable for the categorical independent variable equal to 1 if $z = k$ and equal to zero if $z \neq k, x_i$ is a row vector of covariates (does not include z) and β is a column of corresponding coefficients. Then the probability of $y_i = 1$ can be written as $p_{1(i)}(y_i = 1) = \frac{exp(\mu_{1(i)})}{1 + exp(\mu_{1(i)})}$ in a logistic regression model. The following log likelihood for the ith subject can be obtained

$$ll_{1(i)} = y_i \log(p_{1(i)}) + (1 - y_i)\log(1 - p_{1(i)}) \quad (1)$$

In analyzing binary outcome with independent categorical variables, a categorical independent variable appears in the analytic dataset as a single variable. Statistical software (e.g., SAS) creates a design matrix in initiating analytic procedures so that the parameters in equation (1) can be estimated accordingly [12]. For example, if a classification variable z has K levels, then its main effect has degrees of freedom $(K-1)$, and the design matrix has $(K-1)$ columns that correspond to the $(K-1)$ levels of z. Overall log likelihood across subjects can be obtained and can then be maximized to obtain the maximum likelihood estimates (MLE).

2.2. A Direct Substitution Approach when some Values of a Categorical Independent Variable are Missing but Supplemented with Probability Distribution

We propose statistical models for binary outcomes with some of the independent categorical variable missing but supplemented with probability distribution. Let d_k denote the probability for category level k, where $\sum d_k$=1 for each individual. In analyzing data with probability distribution for missing values of a categorical variable, a similar design matrix as the one without missing information must be created. For demonstration purposes, let the *r*th category be the reference group, r∈(*k*). For the *i*th individual with missing value for a categorical variable, we let $\pi_i = \sum d_{ik}\alpha_k$ where α_k is the coefficient for the *k*th category and $k \neq r$ in order to avoid a full rank matrix problem. Let $\mu_{2(i)} = \alpha_r + \pi_i + x_i\beta$, where definitions of α_r, x_i and β remain the same as in (1), then $p_{2(i)}(y_i = 1) = \frac{exp(\mu_{2(i)})}{1 + exp(\mu_{2(i)})}$. Again, for the Kaiser Permanente example, let A, B, H, N, M and W denote Asian and Pacific, Black, Hispanic, Native American, Multiracial, and White races. For example, if a member has missing self-reported race value but has the following probability distribution for his/her race, d_{iA} = 0.2, d_{iB} = 0.1, d_{iH} = 0.5, d_{iM} = 0.15, d_{iN} = 0.0, d_{iW} = 0.05, using White as the reference group, then $\pi_i = 0.2\alpha_A + 0.1\alpha_B + 0.5\alpha_M + 0\alpha_N$. The following log likelihood can be obtained for an individual with missing self-reported race,

$$ll_{2(i)} = y_i \log(p_{2(i)}) + (1 - y_i)\log(1 - p_{2(i)}) \quad (2)$$

For those with self-reported race information available, log likelihood values can be obtained as in (1). Then the overall log likelihood values across individuals can be calculated and can be maximized to obtain MLEs of αs. This approach is similar to the direct substitution method proposed by McCaffrey and Elliott [11] for a linear model using predicted probabilities for a dichotomous independent variable rather than the actual variable.

2.3. A partial Information Maximum Likelihood Estimator when some Values of a Categorical Independent Variable are Missing but Supplemented with Probability Distribution

McCaffrey and Elliott [11] also proposed a partial information maximum likelihood estimator (PIMLE) for the linear model with dichotomous independent variable. In this paper, we derive the PIMLE for logistic regressions with categorical independent variables. Let p_{ik} represent the probability of outcome $(y = 1)$ if an individual belongs to *k*th category, $p_{ik}(y_i = 1) = \frac{exp(\mu_{ik})}{1 + exp(\mu_{ik})}$, where $\mu_{ik} = \alpha_r + \alpha_k + x_i\beta$, the log partial information likelihood for subject *i* with missing categorical variable is,

$$ll_{3(i)} = \sum_{k=1}^{K} d_{ik}[y_i \log(p_{ik}) + (1 - y_i)\log(1 - p_{ik})] \quad (3)$$

where $k \neq r$. For those with race information available, log likelihood values can be obtained as in (1). Then the overall log likelihood values across individuals can be calculated and can be maximized to obtain MLEs of αs. MLEs from models (2) and (3) can be obtained in SAS PROC NLMIXED using general (*ll*).

3. AN EXAMPLE

Under vaccination of young children is a public health challenge in the US and worldwide [13]. Recent outbreaks of vaccine-preventable diseases such as measles are an apparent result of under vaccination in some communities in the US [14,15]. Examining the association between race and childhood immunization is of interest to vaccination researchers and policy makers [16]. In examining the association between race and up-to-date vaccination status by three years of age, 14,903 children born in Kaiser Permanente Colorado (KPCO) between January 1st 2004 and December 31 2009 and with three years of continuous enrollment were followed for their up-to-date immunization status by three years. The up-to-date immunization status was assessed using CDC's National Immunization Survey combined series completion criteria of receiving 4 doses of diphtheria, tetanus and pertussis (DTaP), 3 doses of polio, 1 dose of measles, mumps and rubella (MMR), 3 or 4 doses of *Haemophilus influenza* type b(Hib) (based on whether they received the conjugate vaccine), 3 doses of Hepatitis B, 1 dose of varicella, and 4 doses of Pneumococcal conjugate (PCV) by 3 years of age [17]. Among 14,903 children, 11.3% had missing self-reported race information but had probability distribution for the six race/ethnicity categories as computed by the BISG algorithm; 13,602 children (91.27%) were up-to-date at 3 years of age. The probability distributions for the six race/ethnicity categories were similar between those with and without self-reported race information (Table **1**).

Table 1: Mean (standard Deviation) of BISG Estimated Probabilities for the Six Race Categories by the Availability of Self-Reported Race

	Available race N=13217	Missing self-reported race N=1686
White	0.76 (0.34)	0.75 (0.34)
Black	0.04 (0.12)	0.04 (0.11)
Asian Pacific	0.04 (0.16)	0.04 (0.16)
Hispanic	0.15 (0.31)	0.15 (0.31)
Native	0.01 (0.02)	0.01 (0.03)
Multiple race	0.01 (0.02)	0.01 (0.02)

We conducted three analyses to the under vaccination data: a) complete data analysis which included only those individuals who have self-reported race. A conventional logistic regression model was employed to obtain odds ratios (OR) with White as the reference category; b) entire population analysis with the direct substitution approach which included those with self-reported race and those without self-reported race but with probability distribution of the six categories; c) entire population analysis with PIMLE approach which included those with self-reported race and those without self-reported race but with probability distribution of the six categories. For analyses b) and c), models (2) and (3) in Section 2 were fit. SAS programs for fitting these two models were provided in Appendix A.

In the complete data analyses, the point estimate of OR for Native race was very large and the range of 95% confidence intervals is extremely wide, indicating estimation instability due to low prevalence of the

Native race in the population (Table **2**). With inclusion of those with missing self-reported race, both the direct substitution and the PIMLE yielded stable estimation of ORs and confidence intervals for the Native race category. Comparing to the results from complete data analyses, the direct substitution and PIMLE yielded comparable point estimates of ORs for other race categories (Black, Asian Pacific, Hispanic and Multiracial races) with White as the reference group although ORs from both direct substitution and PIMLE were slightly underestimated for Black, Asian Pacific and Hispanic. In general, the 95% confidence intervals of ORs from the direct substitution and PIMLE are narrower than those from the complete data analyses. This is consistent with the fact that the direct substitution and PIMLE approach included all subjects in the analyses.

4. SIMULATION STUDY AND RESULTS

We also conducted a simulation study to evaluate the performance of the direct substitution method and PIMLE using the complete data set (N= 13,217). We used the following simulation strategy as in Xu *et al.* [18]. Briefly, while keeping the covariates (gender and race variables) in the complete data set, we used the coefficients from the complete data analyses of the Kaiser under vaccination data to simulate the outcome (up-to-date vaccination) based on the following probabilistic model

$$prob(y_i = 1 \mid \hat{\alpha}_r, \hat{\alpha}_k, \hat{\beta}) = \frac{\exp(\hat{\alpha}_r + I_{iz}\hat{\alpha}_k + gender\hat{\beta})}{1 + \exp(\hat{\alpha}_r + I_{iz}\hat{\alpha}_k + gender\hat{\beta})}$$

where $y_i = 1$ if a child's immunization status is up-to-date and $y_i = 0$ if not, I_{iz} is an indicator variable for race categories, I_{iz}=1 if $z = k$, otherwise I_{iz}=0 with White being the reference; the estimated intercept, $\hat{\alpha}_r$=2.318; $\hat{\alpha}_k$ were estimated coefficients for race categories with $\hat{\alpha}_A = 0.677$, $\hat{\alpha}_B = 0.0.270$, $\hat{\alpha}_H = 0.182$, $\hat{\alpha}_N = 11.212$, and $\hat{\alpha}_M = 0.332$, indicating that White is more likely under vaccinated in this population. The coefficient for gender (gender =1 if male) is $\hat{\beta} = 0.028$. Note that exponentiation of these estimated coefficients results in ORs from the complete data analyses in Table **2**.

For each Monte Carlo sample, we randomly assigned missing self-reported race value in the complete dataset in which both self-reported race and BISG estimated probability distribution of race categories were available. For each rate of missing self-reported race, 5000 random samples were generated from the complete dataset by randomly assigning missing self-reported race. We then conducted four analyses: 1) an analysis which excluded those individuals with missing self-reported race; 2) the direct substitution approach; 3) the PIMLE approach and 4) multiple imputation. For the multiple imputation approach, each missing race was imputed five times using the distribution of race in the complete dataset as in Raebel *et al.* [9]; five separate models were fit; then coefficient estimates and their standard errors were pooled from these five models [9,10]. Six rates of missing self-reported race were evaluated: 0% (without missing race information), 10%, 20%, 30%, 50% and 70%. The analytic results without missing race information (footnotes in Table **3**) served as gold standards for comparing the results from the four analyses. For convenient comparison among these three analytic approaches, we reported mean coefficients and mean of standard errors of coefficients instead of ORs and their confidence intervals to evaluate bias and efficiency of these two methods.

Table **3** showed the mean coefficients (mean standard errors) in the simulation study. As expected,

Table 2: Odds Ratios (95% Confidence Intervals) of Gender and Race Categories*

	Complete data analyses (n=13217)	The direct substitution (n=14903)	PIMLE (n=14903)
Male gender	1.03 (0.91 1.16)	1.04 (0.93 1.16)	1.04 (0.92 1.16)
Black	1.31 (0.93 1.84)	1.29 (0.93 1.79)	1.27 (0.93 1.73)
Asian Pacific	1.97 (1.34 2.89)	1.95 (1.36 2.79)	1.90 (1.34 2.70)
Hispanic	1.20 (1.02 1.41)	1.16 (0.99 1.35)	1.15 (0.99 1.34)
Native	7.40×10^4 (0.00 5.87×10^{163})	3.53 (0.32 39.0)	2.13 (0.42 10.77)
Multiple race	1.39 (0.97 2.01)	1.44 (1.00 2.07)	1.42 (1.00 2.02)

PIMLE: partial information maximum likelihood estimator.
*Reference for gender is Female and reference group for the racial/ethnic categories is White.

Table 3: Simulation Study: Mean Coefficients (Mean Standard Errors) from 5000 Replicates by Different Missing Rates

Methods	Parameters	10% missing	20% missing	30% missing	50% missing	70% missing
Exclude those with missing race	Intercept	2.32 (0.05)	2.32 (0.06)	2.32 (0.06)	2.32 (0.07)	2.32 (0.09)
	Male gender	0.03 (0.07)	0.03 (0.07)	0.03 (0.08)	0.03 (0.09)	0.03 (0.12)
	Black	0.29 (0.19)	0.29 (0.20)	0.29 (0.21)	0.30 (0.25)	0.31 (0.33)
	Asian Pacific	0.70 (0.21)	0.70 (0.22)	0.70 (0.24)	0.71 (0.29)	0.75 (0.38)
	Hispanic	0.18 (0.09)	0.18 (0.09)	0.19 (0.10)	0.19 (0.12)	0.19 (0.15)
	Native	12.01 (332.39)	11.92 (341.91)	11.95 (375.08)	12.15 (491.49)	11.79 (603.43)
	Multiple race	0.35 (0.20)	0.35 (0.21)	0.35 (0.23)	0.36 (0.27)	0.38 (0.35)
Direct substitution	Intercept	2.32 (0.05)	2.32 (0.05)	2.32 (0.05)	2.32 (0.05)	2.32 (0.05)
	Male gender	0.03 (0.06)	0.03 (0.06)	0.03 (0.06)	0.03 (0.06)	0.03 (0.06)
	Black	0.28 (0.18)	0.28 (0.19)	0.28 (0.19)	0.29 (0.21)	0.29 (0.23)
	Asian Pacific	0.69 (0.20)	0.69 (0.21)	0.69 (0.21)	0.69 (0.22)	0.69 (0.24)
	Hispanic	0.18 (0.09)	0.18 (0.09)	0.19 (0.09)	0.17 (0.09)	0.17 (0.10)
	Native	7.82 (7.25)	5.63 (4.93)	4.60 (3.95)	3.51 (3.07)	2.83 (2.68)
	Multiple race	0.35 (0.20)	0.35 (0.21)	0.35 (0.22)	0.36 (0.27)	0.37 (0.34)
PIMLE	Intercept	2.32 (0.05)	2.32 (0.05)	2.33 (0.05)	2.34 (0.05)	2.35 (0.05)
	Male gender	0.03 (0.06)	0.03 (0.06)	0.03 (0.06)	0.03 (0.06)	0.03 (0.06)
	Black	0.27 (0.17)	0.25 (0.17)	0.23 (0.17)	0.20 (0.17)	0.16 (0.17)
	Asian Pacific	0.67 (0.20)	0.65 (0.20)	0.63 (0.20)	0.58 (0.20)	0.54 (0.20)
	Hispanic	0.18 (0.08)	0.17 (0.09)	0.17 (0.09)	0.15 (0.09)	0.14 (0.09)
	Native	1.49 (1.34)	0.94 (0.95)	0.67 (0.77)	0.37 (0.60)	0.22 (0.51)
	Multiple race	0.33 (0.19)	0.32 (0.20)	0.31 (0.21)	0.28 (0.23)	0.23 (0.25)
Multiple imputation	Intercept	2.32 (0.05)	2.33 (0.05)	2.34 (0.05)	2.36 (0.05)	2.37 (0.05)
	Male gender	0.03 (0.06)	0.03 (0.06)	0.03 (0.06)	0.03 (0.06)	0.03 (0.06)
	Black	0.26 (0.19)	0.23 (0.19)	0.20 (0.20)	0.15 (0.22)	0.09 (0.23)
	Asian Pacific	0.61 (0.21)	0.53 (0.22)	0.45 (0.22)	0.31 (0.22)	0.18 (0.23)
	Hispanic	0.16 (0.09)	0.15 (0.09)	0.13 (0.10)	0.10 (0.11)	0.06 (0.11)
	Native	10.10 (306.10)	8.55 (280.02)	7.12 (244.09)	4.97 (197.48)	3.56 (154.85)
	Multiple race	0.30 (0.20)	0.27 (0.21)	0.23 (0.22)	0.17 (0.23)	0.10 (0.24)

PIMLE: partial information maximum likelihood estimator; the mean coefficients (standard errors) from the complete data analysis are 2.32 (0.05), 0.03 (0.06), 0.28 (0.18), 0.70 (0.20), 0.18 (0.08), 12.02 (313.59), and 0.34 (0.19) for Intercept, Male gender, Black, Asian Pacific, Hispanic, Native, and Multiple race, respectively.

when observations with missing race were excluded, the mean standard errors of estimated coefficients increased for gender and all categories of race while the mean coefficients remained nearly unbiased. The Direct Substitution approach produced the same coefficients and standard errors for intercept and gender as those without race information missing due to no loss of observations. It yielded nearly unbiased coefficients and similar standard errors to gold standard for all race categories for the rates of missing self-reported race up to 30%. When the rates of missing self-reported race increased to 50% and 70%, the coefficients remained unbiased but their standard errors were overestimated slightly except for the Native category.

Similar to the direct substitution approach, the PIMLE approach produced the same coefficients and standard errors for intercept and gender due to no loss of observations. With increasing rates of missing self-reported race, the PIMLE approach yielded underestimated coefficients for all categories while the estimates of standard errors remained consistent with those without race information missing. When 50% of self-reported race were missing, the estimated coefficient for Black decreased 40% (from 0.28 to

0.20); for Asian Pacific, the estimated coefficient decreased 17.1% (from 0.7 to 0.58); for the Hispanic, the estimated coefficient decreased 16.7% (from 0.18 to 0.15); for multiracial, the estimated coefficient decreased 7.8% (from 0.34 to 0.28).

The results using the multiple imputation approach were also reported in Table **3**. While the standard errors of coefficients changed slightly with rates of missing race increasing, the estimated coefficients decreased significantly with missing rates of race increasing. Compared to the PIMLE approach, the multiple imputation approach underestimated coefficients more.

5. DISCUSSION

We proposed two approaches to accommodate available distribution probability of an independent categorical variable (e.g., race) in logistic regression models when some of the independent categorical variable missing: 1) the direct substitution approach and 2) a partial information maximum likelihood estimator. These two methods included all observations and thus yielded smaller standard errors of estimated coefficients in analyzing up-to-date immunization status of children by three years old. Our simulation study showed that, when the missing rate of the independent categorical variable was up to 30%, the direct substitution approach and PIMLE yielded coefficient estimates and their standard errors similar to those without race missing. For a given missing rate of race, the multiple imputation approach yielded the most biased coefficient estimates due to the fact that it just used the raw probabilities of race categories in the complete dataset.

When the missing rate was 50% or higher, the direct substitution produced greater standard errors of the categorical variable's coefficients and thus the efficiency decreased. However the standard errors were still less than those from the analyses that excluded observations with missing values. The PIMLE approach underestimated coefficients of the categorical variable's coefficients when the missing rate was 50% or higher. Thus the direct substitution approach is preferred when the missing rate was 50% or higher.

There are some limitations in this study. First, the sample size in both example application and simulation study is large. For a large sample size data, the impact of different missing rates of the independent categorical variable may not be dramatic, especially on the standard errors of coefficients. The performance of these two approaches for small and medium size of datasets is unknown. A missing rate of 30% may result in significant bias of coefficient estimation and less efficiency (larger standard error) for a small or medium dataset. Second, we used the estimated coefficients of race categories from the KPCO example to simulate the outcome; thus the performance may differ when the effects of race categories on a dichotomous outcome differ.

The process of these two newly proposed methods is relatively simpler and easier to implement with SAS codes provided in Appendix A than multiple imputation. Our simulation showed that both the direct substitution and PIMLE improved efficiency by accommodating the available BISG estimated probability distribution and yielded nearly unbiased coefficient estimates when the rate of missing categorical variable is not higher (e.g., less than 30%) while multiple imputation using raw distribution of race categories yielded biased coefficient estimates.

ACKNOWLEDGEMENTS

This study was supported by the Centers for Disease Control and Prevention *via* contract 200-2002-00732 (the Vaccine Safety Datalink Project) with America's Health Insurance Plans. Xu was also supported by NIH/NCRR Colorado CTSI Grant Number UL1 RR025780. We are grateful to Matthew Daley for reviewing the manuscript and providing useful comments.

APPENDIX A

SAS codes for analyzing under immunization data with missing self-reported race

```
*PR_. are indicator variables (0 or 1) for available race categories;

*BISG_PR_. are estimated probabilities for missing race information.

*****************************************;

*Direct substitution method using all data;

*****************************************;

proc nlmixed data=underimmunization;

parms alpha=0.2 beta_A=0.6 beta_B=0.1 beta_H=0.1 beta_N=0.2 beta_M=0.1 beta_sex=0.2;
```

```
if missing_race=0 then xbeta=alpha + (PR_BLACK)*beta_B +(PR_HISP)*beta_H + (PR_API)*beta_A + (PR_AIAN)*beta_N + (PR_MULTI)*beta_M +sex*beta_sex;

else xbeta=alpha + (BISG_PR_BLACK)*beta_B +(BISG_PR_HISP)*beta_H + (BISG_PR_API)*beta_A + (BISG_PR_AIAN)*beta_N + (BISG_PR_MULTI)*beta_M +sex*beta_sex;

p=exp(xbeta)/(1+exp(xbeta));

loglike=y*log(p)+(1-y)*log(1-p);

model y~general(loglike);

ods output ParameterEstimates=direct (keep=Parameter Estimate StandardError);

run;

*****************************************;

**PIMLE method using all data;

*****************************************;

proc nlmixed data= underimmunization;

parms alpha=0.2 beta_A=0.6 beta_B=0.1 beta_H=0.1 beta_N=0.2 beta_M=0.1 beta_sex=0.2;

xbeta_A=alpha+ beta_A+sex*beta_sex;

xbeta_B=alpha+ beta_B+sex*beta_sex;

xbeta_H=alpha+ beta_H+sex*beta_sex;

xbeta_N=alpha+ beta_N+sex*beta_sex;

xbeta_M=alpha+ beta_M+sex*beta_sex;

xbeta_W=alpha+sex*beta_sex;

py_A=exp(xbeta_A)/(1+exp(xbeta_A));

py_B=exp(xbeta_B)/(1+exp(xbeta_B));

py_H=exp(xbeta_H)/(1+exp(xbeta_H));

py_N=exp(xbeta_N)/(1+exp(xbeta_N));

py_M=exp(xbeta_M)/(1+exp(xbeta_M));

py_W=exp(xbeta_W)/(1+exp(xbeta_W));

if missing_race=0 then loglike=PR_API*(y*log (py_A)+(1-y)*log(1-py_A))+
PR_BLACK*(y*log(py_B)+(1-y)*log(1-py_B))+PR_HISP*(y*log(py_H)+(1-y)*log(1-py_H))

+PR_AIAN*(y*log(py_N)+(1-y)*log(1-py_N))+PR_MULTI*(y*log(py_M)+(1-y)*log(1-py_M))+PR_White*(y*log(py_W)+(1-y)*log(1-py_W));

else loglike=BISG_PR_API*(y*log(py_A)+(1-y)*log(1-py_A))+ BISG_PR_BLACK*(y*log(py_B)+(1-y)*log(1-py_B))+BISG_PR_HISP*(y*log(py_H)+(1-y)*log(1-py_H))

+BISG_PR_AIAN*(y*log(py_N)+(1-y)*log(1-py_N))+BISG_PR_MULTI*(y*log(py_M)+(1-y)*log(1-py_M))+BISG_PR_White*(y*log(py_W)+(1-y)*log(1-py_W));

model y~general(loglike);

ods output ParameterEstimates=PIMLE(keep= Parameter Estimate StandardError);;

run;
```

REFERENCES

[1] Boehmer U, Kressin NR, Berlowitz DR, Christiansen CL, Kazis LE, Jones JA. Self-reported vs administrative race/ethnicity data and study results. Am J Public Health 2002; 92: 1471-2. http://dx.doi.org/10.2105/AJPH.92.9.1471

[2] Bilheimer LT, Sisk JE. Continue collecting adequate data on racial and ethnic disparities in health: the challenges. Health Affairs 2008; 27: 383-91. http://dx.doi.org/10.1377/hlthaff.27.2.383

[3] Institute of Medicine 2009. Race, ethnicity, and language data: standardization for health care quality improvement.Washington, DC: The National Academies Press.

[4] Elliott MN, Fremont A, Morrison PA, Pantoja P, Lurie N. A new method for estimating race/ethnicity and associated disparities where administrative records lack self-reported race/ethnicity. Health Serv Res 2008; 43: 1722-36. http://dx.doi.org/10.1111/j.1475-6773.2008.00854.x

[5] Elliott MN, Morrison P, Fremont A, McCaffrey D, Pantoja P, Lurie N. Using the census bureau's surname list to improve estimates of race/ethnicity and associated disparities. Health Services and Outcomes Research Methodology 2009; 9: 69-83. http://dx.doi.org/10.1007/s10742-009-0047-1

[6] Adjaye-Gbewonyo D, Bednarczyk RA, Davis RL, Omer SB. Using the Bayesian Improved Surname Geocoding Method (BISG) to create a working classification of race and ethnicity in a diverse managed care population: a validation study. Health Serv Res 2014; 49: 268-83. http://dx.doi.org/10.1111/1475-6773.12089

[7] van der Heijden GJ, Donders AR, Stijnen T, Moons KG. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. J Clin Epidemiol 2006; 59: 1102-9. http://dx.doi.org/10.1016/j.jclinepi.2006.01.015

[8] Janssen KJ, Donders AR, Harrell FE Jr, Vergouwe Y, Chen Q, Grobbee DE, Moons KG. Missing covariate data in medical research: to impute is better than to ignore. J Clin Epidemiol 2010; 63: 721-7. http://dx.doi.org/10.1016/j.jclinepi.2009.12.008

[9] Raebel MA, Xu S, Goodrich GK, Schroeder EB, Schmittdiel JA, Segal JB, O'Connor PJ, Nichols GA, Lawrence JM, Kirchner HL, Elston Lafata J, Butler M, Newton KM, Steiner JF. Initial antihyperglycemic drug therapy among 241 327 adults with newly identified diabetes from 2005 through 2010: a surveillance, prevention, and management of diabetes mellitus (SUPREME-DM) study. Ann Pharmacother 2013; 47: 1280-91. http://dx.doi.org/10.1177/1060028013503624

[10] Horton NJ, Kleinman KP. Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. The American Statistician 2007; 61: 79-90. http://dx.doi.org/10.1198/000313007X172556

[11] McCaffrey DF, Elliott MN. Power of tests for a dichotomous independent variable measured with error. Health Serv Res 2008; 43: 1085-101. http://dx.doi.org/10.1111/j.1475-6773.2007.00810.x

[12] SAS Institute Inc 2011. Base SAS® 9.3 Procedures Guide. Cary, NC: SAS Institute Inc.

[13] Glanz JM, Newcomer SR, Narwaney KJ, Hambidge SJ, Daley MF, Wagner NM, McClure DL, Xu S, Rowhani-Rahbar A, Lee GM, Nelson JC, Donahue JG, Naleway AL, Nordin JD, Lugg MM, Weintraub ES. A population-based cohort study of under vaccination in eight managed care organizations across the United States. Archives of Pediatrics & Adolescent Medicine. JAMA Pediatrics 2013; 167: 274-281. http://dx.doi.org/10.1001/jamapediatrics.2013.502

[14] Sugerman DE, Barskey AE, Delea MG, Ortega-Sanchez IR, Bi D, Ralston KJ, Rota PA, Waters-Montijo K, Lebaron CW. Measles outbreak in a highly vaccinated population, San Diego, 2008: role of the intentionally undervaccinated. Pediatrics 2010; 125: 747-55. http://dx.doi.org/10.1542/peds.2009-1653

[15] Omer SB, Enger KS, Moulton LH, Halsey NA, Stokley S, Salmon DA. Geographic clustering of nonmedical exemptions to school immunization requirements and associations with geographic clustering of pertussis. Am. J. Epidemiol 2008; 168: 1389-96. http://dx.doi.org/10.1093/aje/kwn263

[16] Luman ET, Ching PL, Jumaan AO, Seward JF. Uptake of varicella vaccination among young children in the United States: a success story in eliminating racial and ethnic disparities. Pediatrics 2006; 117: 999-1008. http://dx.doi.org/10.1542/peds.2005-1201

[17] Centers for Disease Control and Prevention. National, state, and local area vaccination coverage among children aged 19-35 months — United States, 2011. *Morbidity and Mortality Weekly Report (MMWR)* 2012; 61: 689-696. Available from: http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6135a1.htm

[18] Xu S, Schroeder EB, Shetterly S, Goodrich GK, O'Connor PJ, Steiner JF, Schmittdiel JA, Desai J, Pathak RD, Neugebauer R, Butler MG, Kirchner L, Raebel MA. Accuracy of hemoglobin A1c imputation using fasting plasma glucose in diabetes research using electronic health records data. Statistics, Optimization & Information Computing 2014; 2: 93-104.

Received on 08-05-2015 Accepted on 15-07-2015 Published on 19-08-2015

http://dx.doi.org/10.6000/1929-6029.2015.04.03.2