Statistical Methods for Analyzing Multiple Race Response Data

Tommi Gaines

Abstract

Collection of racial data is ubiquitous throughout research as an important measure of the demographic characteristics of the study population. However, the validity of racial data has been a concern, prompting several agencies to modify their measurements by allowing individuals to identify with multiple racial categories. This research aims to add to the current methodology for analyzing multiple race responses as well as single race categories for data generated from the California Health Interview Survey (CHIS). This paper explores three distinct methods for analyzing outcomes that indicate whether individual health behaviors are consistent with goals of the Healthy People 2010 program. One approach uses supplementary data from the Census Bureau and the California Department of Finance to rake multiple-race respondents into single-race categories consistent with the 1977 OMB standards. The second method, following Schenker and Parker (2003), imputes a single race category for multiple race respondents to produce population health estimates. The third method, which we call multiple covariate adjustment, simultaneously controls for indicators of all self-identified race categories (using one group as the referent) in a regression analysis. The three methods are compared with attention focused on inference for the proportion of individuals who meet Healthy People 2010 goals, which has a common interpretation across methods, as well as inference about racial disparities in achieving those goals.

Key Words: Multiracial data; Raking; Imputation; Regression

1. Introduction

Improving the health status of minority racial groups has been the focus of national health goals and planning in the United States over the past few decades (US Department of Health and Human Services, 1979; 1985; 1991; 2000). This is because race is an important indicator of disparity in health care delivery and health outcomes such as excess mortality, morbidity, and disability. However the validity of racial data has been a concern since they were thought to incorrectly reflect the racial diversity of a wide variety of people. As a result several agencies have modified their standards of racial data collection to adjust for a changing racial and ethnic profile. Prior standards for the tabulation and presentation of racial data have typically followed federal guidelines discussed in the Office of Management and Budget (OMB) 1977 statistical policy directive (OMB, 1997). This policy defined racial categories as: White, Black, Native American or Alaska Native (AIAN) and Asian or Pacific Islander (API). The OMB revised these federal standards in 1997 allowing an individual to identify with more than one racial group, thus eliminating the idea of mutually exclusive single-race categories (OMB, 1997). Therefore under the revised standards a total of 31 possible racial categories exist.

Although these revisions offer a wider choice for racial identification than previously available, the resulting data pose analytic dilemmas for the researcher. This is because the revised system inhibits compatibility between different data collection systems, presents difficulty with studying trends overtime, and can lead to insufficient sample sizes to generate statistically reliable estimates. The first two issues are directly related to other datasets that report statistics by single race category only, whereas the last is due to rare multiracial groups in a population.

This paper will to add to the current methodology for analyzing data collected from multiracial individuals through comparing different statistical methods for analyzing multiple race responses as well as single race categories for data generated from the California Health Interview Survey (CHIS). This is a stratified sample that generates population based estimates of health outcomes across racial groups. The advantages and disadvantages of these methods will be explored by investigating how racial identification affects our understanding of health disparities among racial groups for all of California in context of health goals specified in Healthy People 2010 (HP2010). This program provides a framework for health promotion and disease prevention.

The next section will provide a description of the CHIS data set that is used for investigating the proposed methods. Section 3 discusses three methods to analyze multiple race response data, for which two approximate the size of the single race groups under the 1977 OMB standards to produce population health estimates and one simultaneously controls for the effects of all self-identified race categories on the estimation of health outcomes in a regression analysis. All three methods are used to estimate the proportion of Californians that experience a health outcome in context of HP2010 while adjusting for sociodemographic variables. The three proposed analyses differ in that two allocate multiracial individuals into a single race category creating a dataset with mutually exclusive racial categories whereas the third method preserves the multiracial status of the individual. Section 4 illustrates the application of these methods with the CHIS dataset by comparing the estimates generated under the three methods and conclude with a brief discussion in Section 5.

2. Data Source

The California Health Interview Survey is a biennial telephone interview survey with the first wave starting in 2001. The objective is to provide estimates of the health status of Californians through the collection of data on health, demographic, and economic characteristics. The sample design is a two-stage stratified random-digit-dial telephone survey. The first stage consists of randomly sampling telephone numbers generated for 44 predefined geographic areas that correspond to 41 individual California counties and 3 areas that are groupings of smaller California counties. The second stage involves the random sampling of one adult among all adults living in the household. A total of 56, 270 adults aged 18 years and older were sampled.

3. Methods

3.1 Raking Adjustment

Raking is a statistical method that is primarily used is to adjust the survey estimates for undercoverage and response biases by attaching weights to the survey data using known population totals (Deming & Stephan, 1940; Deville & Sarndal, 1992; Brick 2003). In general, this weighting procedure uses auxiliary data from a supplementary source, such as a larger survey or census. The advantages of this method are to reduce the bias and variance of the estimates, force totals to match external totals, and adjust for sources of error.

Raking is performed by adjusting survey weights so that the marginal totals of the adjusted data agree with the population total from the marginal distribution of one dimension (or variable). The next step is to adjust the resulting weights to agree with the

population totals for the second marginal distribution. This process continues by alternating between all dimensions in a cross-classification table. The algorithm iterates until convergence that is until the sum of the adjusted data simultaneously agree with the population totals for all the marginal distributions within a specified tolerance level. A formal mathematical description of computing the weights at each iteration t, in a two variable situation, is as follows:

$$\widetilde{w}_{ij} = \hat{N}_{ij} \qquad \text{for t = 0,}$$

$$\widetilde{w}_{ij} = \frac{\widetilde{w}_{ij}^{(t-1)} N_{i.}}{\widetilde{w}_{i.}^{(t-1)}} \qquad \text{for t = odd,}$$

$$\widetilde{w}_{ij} = \frac{\widetilde{w}_{ij}^{(t-1)} N_{.j}}{\widetilde{w}_{.j}^{(t-1)}} \qquad \text{for t = even,}$$

where $\hat{N}_{ij} = \sum_{k \in (i,j)} d_{(i,j)k}$ is the unadjusted estimate of the population total in cell (i,j) and

$d_{(i,j)k} = \frac{1}{\pi_{(i,j)k}}$ where $\pi_{(i,j)k}$ = is the probability of selecting unit k in cell (i,j) or the sum of the sampling weights for persons in the sample falling in the classification corresponding to cell (i,j). This process iterates until convergence.

This procedure is implemented through raking multiple race respondents into single race categories consistent with the 1977 OMB standards that tabulate race as: White, Black, API, and AIAN. The CHIS sampling weights are adjust by introducing revised weights produced by the raking algorithm that uses demographic and county-level data. These weights are constructed to sum to known California population totals that are obtained from the Census Bureau and the California Department of Finance.

Specifically, the marginal counts for California's resident population by race are obtained from the American Community Survey (ACS) that is an annual nationwide survey conducted by the Census Bureau to replace the decennial long from census. The categories for race are all inclusive in that the population totals are tabulated as 'race alone or in combination with one or more other races'. Furthermore, it is assumed that the marginal totals for variables collected in ACS and used in the raking algorithm have negligible error. The California Department of Finance (DOF) is a secondary auxiliary source for population totals according to age. The data are publicly available through the DOF website for which the format of age is from 0 to 100 with 1 year increments.

A total of two demographic variables are used in the raking process that includes race and age. The race variable has 5 levels and is aggregated as race alone or in combination involving the following groups: White, Black, AIAN, API, and other, whereas the age variable has 4 levels: 18-29, 30-44, 45-64, 65+. These variables form a cross-classification table of the CHIS sample for which the CHIS sample weights are adjusted by a factor so that the sum of the adjusted weights simultaneously agrees with the population totals of the demographic variables. To illustrate this method the following quantities are defined:

- d_{ij} = the CHIS sample weighted proportion of Californians in racial category i (i = 1,..,5) and age-class j (j = 1,..,4). The weighted cell estimates are given in Table 3.1.1.
- p_{ij}^{CHIS} = the CHIS sample weighted proportion of Californian's that visited the dentist in racial category i and age-class j. The weighted cell proportions are shown in Table 3.1.2.

The marginal distributions of the cross-classification that forms d_{ij} are used to apply the raking algorithm within each combination of the race and age categories to obtain:

- d_{ij}^{rake} = weighted cell estimates, d_{ij}, raked to the marginal totals obtained from ACS and DOF for race and age. The corresponding cell estimates are given in Table 3.1.3.

The result is a modified weighted proportion of Californians in racial category i with outcome y = 1 (e.g., visited the dentist in the past 12 months) and is termed the CHIS Raked Adjusted proportion given by:

$$\hat{p}_i^{CRA} = \frac{\sum_{j=1}^{4} d_{ij}^{rake} \cdot \hat{p}_{ij}^{CHIS}}{\sum_{j=1}^{4} d_{ij}^{rake}} \qquad (3.1.2)$$

3.2 Multiple Imputation

The dilemma of attempting to combine or compare racial data when classification systems have been revised has been addressed in part by Schenker and Parker (2003) through missing-data methods. Their approach utilizes multiple imputation to generate a distribution of missing values for the single race category. Imputation is a common method for handing missing data by filling-in a value for the missing datum such that complete-data methods of analysis can be applied. With multiple imputation, two or more values are imputed rather than a single value in order to reflect the uncertainty about which value to impute.

The general idea of multiple imputation, as discussed by Rubin (1987), is to replace each missing value with a vector composed of possible values that are independently drawn from a distribution. This distribution reflects assumptions about the

data and the mechanisms creating the missing data. The result is the formation of two or more (usually five to ten) completed data sets for which each data set is analyzed with a complete-data method. The analyses are then combined in a simple way that reflects the extra uncertainty due to having imputed rather than having used actual data (Rubin & Schenker, 1991, 1997; Schafer, 1997).

In terms of the imputation method proposed by Schenker and Parker, an estimated probability of each single-race response is imputed for each multiple-race respondent. This probability is then used to allocate a single race category to the multiracial respondent. Their method is summarized as a two-step procedure that creates a set of 10 imputations for the missing single race category (Bernard et al, 1998). This procedure reflects the variability of primary race given the parameters of the imputation model and the variability due to estimating the parameters.

The two-step procedure is applied to the largest multiple race groups in the CHIS sample as an illustration of the method due to small sample sizes of the other multiracial groups. This includes 3 groups that make up 83% of all multiracial respondents in the 2001 sample involving the following: Black/White, API/White, and AIAN/White. However this approach can be applied to every multiracial combination of adequate sample size. In the first step a logistic regression model is fitted among respondents of a specific multiple race combination for which the outcome, y =1, is a single race category. This model is defined as:

$$log\left(\frac{\pi(x)}{1+\pi(x)}\right) = \beta_0 + \beta_1 + \cdots + \beta_p x_p + \varepsilon \qquad (3.2.1)$$

The predictors, x_j for j = 1,…,p, included in the model are Hispanic ethnicity, gender, born in the US versus foreign born, household income measured by federal poverty

level (FPL) below 200% versus at least 200%, educational attainment of high school or less versus more than high school, median household income for county of residence, and racial composition of county of residence. The independent variable for racial composition measures the percentage of Black residents in the Black/White model, the percentage of API residents in the API/White model, and the percentage of AIAN residents in the AIAN/White model.

Once that model has been generated logistic regression coefficients are drawn from their approximate posterior distribution. This distribution is a multivariate normal given by:

$$\beta \sim MVN\left(\hat{\beta}, \hat{\Sigma}\right) \tag{3.2.2}$$

where the mean, $\hat{\beta}$, is the estimate of β, a vector of the logistic regression coefficients and whose covariance matrix, $\hat{\Sigma}$, is the estimated variance-covariance matrix of $\hat{\beta}$. Both $\hat{\beta}$ and $\hat{\Sigma}$ are estimated from the logistic regression model fitted in (3.2.1) to each particular multiracial combination, for example the model of all Black/White biracial respondents.

The second step is to compute the probability of primary race category for every individual with a missing primary race by using the logistic regression coefficients drawn from the distribution specified in (3.2.2). The person specific probability of the i^{th} individual is given by:

$$\pi_i = \frac{exp(x_i'\beta)}{1+exp(x_i'\beta)} \tag{3.2.3}$$

After the 2-step procedure has been carried out a multiracial person is reallocated to a single race category and complete data methods of analysis are applied. The

assignment of a multiracial individual to a particular single race group can be thought of as a Bernoulli coin flip where the chance of the coin randomly selecting a particular race is determined from π_i calculated in (3.2.3).

In the CHIS survey all multiracial persons were asked to select a single race category that best identified oneself. Responses included one of the single race categories that made up the multiracial combination, other, both/all/multiracial, refused, none of these, or don't know. In the logistic regression model to predict primary race the binary outcome consisted of a specific single race category or 'other'. The remaining responses were treated as missing and the 2-step procedure was applied to impute a single race category among these multiracial individuals.

Details of the 2-step procedure can be illustrated, for example, by considering the Black/White respondents in the CHIS sample. A logistic regression model is fitted to this specific biracial combination to predict the probability of being Black, White, or Other. Initially a logistic regression model is fitted where the outcome is Other versus (Black or White). The probability of being Other is calculated for each person through the 2-step procedure. A subsequent logistic regression model is fitted where the outcome is Black versus White among all Black/White respondents that did not identify their primary race as Other. The probability of Black is then computed via the 2-step procedure. As a result each Black/White respondent with a missing primary race has a probability of being Black, White, or Other assigned to them. These probabilities are used to randomly assign a single race to that person.

Table 3.2.1 displays the results of fitting separate logistic regression models to the 3 multiracial groups. Many of the covariates in Table 3.2.1 are not significantly predictive

of primary race as was found by Schenker and Parker. The covariates that are predictive differ across the racial groups with the exception of Hispanic ethnicity that has a positive association for all biracial models versus Other.

Complete data methods of analysis are applied following the implementation of the 2-step procedure. This involves the utilization of CHIS sample weights to account for differential probabilities of selection in order to produce unbiased population estimates. For a survey of n subjects, the weighted proportion of individuals that have seen the dentist is calculated for each single race group. These groups consist of: White, Black, API, AIAN, and Other. For a particular racial group, the weighted proportion is defined as:

$$\hat{p}_k^{MI} = \frac{\sum_{i=1}^{n} w_{ik} y_{ik}}{\sum_{i=1}^{n} w_{ik}} \qquad (3.2.4)$$

where w_{ik} is the inverse probability of selecting individual i of racial group k and y_{ik} = 1 if the i^{th} individual of the k^{th} racial group saw the dentist within the past 12 months and zero otherwise. The corresponding standard errors are calculated through the jackknife methods.

3.3 Multiple Covariate Adjustment

The methods discussed in the prior sections involve the allocation of multiracial persons into a single race group. This is followed by a separate analysis using the multiracial responses assigned to a single-race group as well as the single race responses to calculate the prevalence of a particular health outcome by racial subgroup. As a result, a weighted estimate of the crude or unadjusted proportion of Californians from group k

with outcome y = 1 (e.g., the having seen the dentist at least once during the past year) is calculated.

An alternative to the estimating the crude proportion is to use the conventional regression approach of including dummy variables in a multiple logistic regression analysis. In this situation the model would simultaneously control for the effects of all the self-identified race categories in the CHIS sample on the outcome y = 1. Table 3.3.1 displays the estimated proportion of individuals that visited the dentist across different independent variables among 5 racial categories. Overall those who visited the dentist are younger, of higher income, more educated, carry insurance, and live in a metropolitan area. However differences between racial groups are evident with Whites having a greater proportion of individuals seeing the dentist in general across all categories of the independent variables. The variations of these socio-deomgraphic variables across racial groups can be adjusted for in a logistic regression model. The independent variables considered in the analysis include: (1) race: White, Black, API, AIAN, Other, (2) age (in years), (3) annual household income classified according to the federal poverty level (FPL): <100%, 100-199%, 200% - 299%, ≥300% of FPL , (4) education: less than high school, high school graduate, and any college education, (5) insurance status: currently uninsured, uninsured for any of the past 12 months, and insured for all of the past 12 months, (6) OMB classification of a Metropolitan Statistical area (MSA): metropolitan and non-metropolitan.

In addition to controlling for socio-demographic factors, the logistic model can also address the multiracial status of survey respondents. This is accomplished through incorporating dummy variables to reflect each self-identified racial group in the CHIS

sample. For example, suppose that the j[th] independent variable, x_j, has k_j levels where the $k_j - 1$ design variables are denoted as D_l and the respective coefficients are denoted as β_l, $l = 1,2...,k_j - 1$. The logit fitted to the data with p independent variables where the j[th] independent variable is nominal is:

$$\log\left(\frac{\mu}{1-\mu}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \sum_{l=1}^{k_j-1} \beta_l D_l + \beta_p x_p + \varepsilon \qquad (3.3.1)$$

Therefore, to obtain the proportion of individuals by racial category with a particular health outcome involves the estimation of μ in model (3.3.1). For example, the proportion of Black individuals that visited the dentist is estimated from the logistic probability, $\hat{\mu}$ = p(Y=1| race, age, education, income, insurance status, MSA). To do so by racial groups involves the incorporation of dummy variables where the j[th] independent variable is racial category with $k_j = 5$ levels. Ignoring ethnicity, there are a total of 4 single race categories in the 1977 OMB standards for racial data collection. However, an additional category considered in the model is "other" which can be thought of as a residual category for those that did not identify with one of the four. Therefore a total of 4 design variables, $k_j - 1$, are included in the model with 'other' treated as the reference group. These design variables differ from the usual manner in that they are not mutually exclusive. Rather each racial category is defined as race alone or in combination and therefore will overlap one another. For example, the total sample size is n = 56,037 but from Table 3.3.1 we see that the sum of sample sizes across the 5 racial categories exceeds this total.

Another aspect of the fitted model is to incorporate the complex sample design of CHIS involving stratification and sample weights. From the fitted logit we can predict

the logistic proportion of individuals with outcome y = 1 among observations i = 1,…, n as:

$$\hat{\mu}_i = exp(x_i'\hat{\beta})/1 + exp(x_i'\hat{\beta}) \quad (3.3.2)$$

where $\hat{\beta}$ is the value that maximizes the weighted likelihood estimator defined as:

$$l(\beta) = \prod_{i=1}^{n} \left[\frac{exp(x_i'\hat{\beta})}{1+exp(x_i'\hat{\beta})}\right]^{w_i y_i} \left[1 - \frac{exp(x_i'\hat{\beta})}{1+exp(x_i'\hat{\beta})}\right]^{w_i(1-y_i)} \quad (3.3.3)$$

and w_i is the sample weight associated to each observation (Korn & Graubard, 1999).

The predicted logistic proportions are then averaged across the k racial groups to produce the model adjusted proportion of individuals with outcome y = 1:

$$\hat{p}_k^{MCA} = \frac{\sum_{i=1}^{n} w_{ik}\hat{\mu}_{ik}}{\sum_{i=1}^{n} w_{ik}} \quad (3.3.4)$$

The regression coefficients for the logistic regression model are given in Table 3.3.2. After adjusting for the socio-demographic variables Black and AIAN are significant predictors in the model. An additional model was generated to investigate whether an additive effect of race appropriately represents the data. Under this situation, it is assumed that the impact of race on the proportion of Californians with outcome y = 1 is independent of the multiracial status. However including interactive effects between each of the 4 design variables representing the four racial groups permits the effect of race on the outcome to vary across single and multiracial persons. This model included the same independent variables as shown in Table 3.3.2 however the interactive effects among the racial groups were included (model not shown). A likelihood ratio test between the two models did not support including these interactive effects ($\chi^2_{(6)} = 10.37$ and p-value = 0.11). Furthermore, the area under the ROC curve was compared between the two models and did not exhibit substantial improvement with either model but was

approximately 0.67 for both. This area was used as another statistical comparison since it represents the probability that a randomly chosen subject with outcome y = 1 is correctly identified by the logistic model with greater likelihood of having outcome y = 1 than a randomly chosen subject without outcome y = 1 (Hanley & McNeil, 1982). Of additional note is the non-linear transformation of age. The fractional polynomial method was used to investigate whether age is linear in the logit and it was determined that the squared inverse of age was significantly better than a linear term for age.

3.4 Variance Estimation

The methods presented in the previous section estimates the proportion of individuals with outcome y = 1 by accounting for the weights and stratification of the CHIS sample. The multiple imputation and covariate adjustment involve a non-linear function of the data through predictions from a logistic regression model to eventually generate a weighted proportion. The multiple imputation method does so through a logistic model to predict primary race that is subsequently used in the calculation of the weighted proportion; whereas the covariate adjustment utilizes a weighted logistic regression to derive a weighted predicted proportion from the model. The raking adjustment is less complex but involves a modification to the initial CHIS sampling weights. Therefore, replication methods are applied in estimating the variance for the proportion of individuals with outcome y = 1.

The corresponding variance for the estimates produced by the multiple imputation (3.2.4) and multiple covariate adjustment (3.3.4) are calculated using the jackknife method. The jackknife variance estimator is calculated by excluding observations in the

i[th] PSU of stratum h for the k[th] racial group. This increases the sample weights of the remaining observations in stratum h by a factor of $m_h/(m_h-1)$, and is proceeded by calculating $\hat{p}_{(hi)}$ of the new data set. The variance estimator is given by:

$$\sum_{h=1}^{H} \frac{m_h-1}{m_h} \sum_{i=1}^{m_h} \left(\hat{p}_{(hi)} - \hat{p}\right)^2 \qquad (3.4.1)$$

The bootstrap is used for variance estimation in the raking adjustment. The bootstrap is calculated by generating artificial data sets that are of the same size and structure of the original where sampling is done repeatedly with replacement (Efron, 1979; 1982). Korn and Graubard, (1999) illustrate the bootstrap estimate of the variance in context of survey data as follows:

$$\widehat{var}_{BS}\left(\widehat{\Theta}\right) = \frac{1}{B-1} \sum_{b=1}^{B} \left(\widehat{\Theta}_b^* - \widehat{\Theta}^*\right)^2 \qquad (3.4.2)$$

where

$$\widehat{\Theta}^* = \frac{1}{B} \sum_{b=1}^{B} \widehat{\Theta}_b^*$$

and B is the number of replicated data sets.

4. Application

The data for the applying the three methods is taken from the 2001 California Health Interview Survey (CHIS). The advantages of using CHIS are that survey participants are allowed to identify with more than one racial group and it can be used to measure the states progress towards achieving the national goals specified in HP2010. During the first wave approximately 4.6% identified as multiracial in the 2001 survey. Among the multiracial respondents, an additional question is asked requesting them to

designate a race that best identifies them. The utility of CHIS is to whether California is meeting the goal of eliminating health disparities as outlined in Healthy People 2010. Therefore, the application of the three methods is illustrated in context of the oral health goal of Healthy People 2010. The current goal is that 56% of the population uses the oral health care system each year and their statistics document AIAN and Blacks see the dentist less frequently than other racial groups (US Department of Health, 2000).

The proportions stratified by racial group were estimated from equations (3.1.2), (3.2.4), and (3.3.4). An additional estimate was generated to compare with the other three. This estimate is similar in that it is a weighted proportion of visiting the dentist where the weights are derived from the CHIS sample. However, it differs in that there is an additional racial group termed multiracial and this proportion does not account for the reallocation or adjusting of the multiracial status. Table 4.1 displays the estimates generated under each method. The standard errors for the CHIS weighted proportion, $\hat{p}^{CHIS}$, multiple covariate adjustment, $\hat{p}^{MI}$, and multiple covariate adjustment, $\hat{p}^{MCA}$, were obtained through the jackknife method. A bootstrap standard error was calculated for the raking adjustment, $\hat{p}^{CRA}$. We can see that the estimates generated by the different methods closely agree with each other. However noticeable differences are observed in the standard errors with the multiple covariate adjustment having smaller standard errors across all racial groups relative to the other methods. Overall, there is a disparity among the AIAN and Other racial group in the utilization of the oral health care system. These individuals see the dentist less frequently than the other groups. However, California has exceeded the HP2010 national goal across every racial group as estimated under each method.

5. Discussion

This paper described three distinct methodologies for analyzing multiple race as well as single race responses. The methods illustrated in this paper are best applied to analyses where the primary objective is to describe the role of race on the outcome of interest. For example, the raking adjustment and multiple imputation allocate a multiracial individual into a single race category in context of investigating health disparities across racial groups. Although these two methods do not preserve the multiracial status of the individual both attempt to avoid problems that can arise when aggregating all multiracial persons into a single racial category sometimes referred to as the "2+ Races" category.

In Public Health, we perform analyses by racial groups with the goal of identifying potential problems and to gain an understanding of areas for improving health in the population. Therefore, we need data that not only identify groups suffering from health inequities but also identify the reasons behind these inequities. Yet understanding the mechanisms that lead to these disparities is highly complex in a category that aggregates all multiracial individuals. This is because of the heterogeneity of other demographic variables associated with health outcomes. It has previously been shown that the proportion of multiracial individuals with a particular educational attainment or labor force participation is dependent on the specific multiple race combination (Snipp, 2006, personal communication). According to the 2004 American Community Survey (ACS), approximately 16% of non-Hispanic White/AIAN, 10% of non-Hispanic White/Black , 6% of non-Hispanic White/Asian, and 21% of non-Hispanic White/Other individuals have less than a high-school education. The 2004 ACS also estimates that

approximately 64% of non-Hispanic White/AIAN, 75% of non-Hispanic White/Black , 74% non-Hispanic White/Asian, and 70% of non-Hispanic White/Other individuals are participating in the civilian labor force. The diversity of demographic variables makes it difficult to produce meaningful inference on the multiracial population. The dilemma here is of obtaining accurate data needed for planning and policy making and to design interventions that target particular racial groups that improve their health.

Comparison of the three methods was limited to the inference made in context of the proportion of individuals meeting the oral health goal of HP 2010. Future work entails a simulation to further evaluate the performance of each method. The sensitivity and robustness of the three methods are checked by fitting an empirical model to a simulated population developed from the CHIS data set. Particularly the simulation will address how each method is affected by varying characteristics of the population including age, income, education, and ratio of single race to multiple race individuals. This will allow for the determination of which method performs better over the situations considered in the simulation.

Tables

Table 3.1.1: Proportion of Californians 18+ by race and age before applying raking algorithm, d_{ij}

Race	18-29	30-44	45-64	65+	Total
White/combo	0.120	0.186	0.194	0.108	0.607
Black/combo	0.016	0.021	0.019	0.007	0.063
API/combo	0.035	0.045	0.034	0.016	0.130
AIAN/combo	0.009	0.011	0.010	0.003	0.033
Other/combo	0.058	0.066	0.034	0.009	0.167
Total	0.238	0.328	0.290	0.144	1

Table 3.1.2: Weighted proportion of Californians that visited the Dentist, p_{ij}^{CHIS}

Total	18-29	30-44	45-64	65+
White/combo	0.70	0.71	0.77	0.69
Black/combo	0.72	0.69	0.67	0.51
API/combo	0.69	0.73	0.71	0.65
AIAN/combo	0.67	0.65	0.64	0.52
Other/combo	0.55	0.58	0.59	0.52

Table 3.1.3: Proportion of Californians 18+ by race and age after applying raking algorithm, d_{ij}^{rake}

Total	18-29	30-44	45-64	65+	Total
White/combo	0.132	0.202	0.217	0.117	0.668
Black/combo	0.017	0.022	0.021	0.008	0.068
API/combo	0.034	0.042	0.033	0.015	0.124
AIAN/combo	0.004	0.005	0.004	0.001	0.015
Other/combo	0.043	0.049	0.026	0.007	0.125
Total	0.230	0.320	0.302	0.148	1

Table 3.2.1: Logistic Regression Predicting Primary Race. Estimated coefficients that are significantly different from zero at the 10 percent level are indicated by a + (positive coefficient) or – (negative coefficient)

Variable	Black/White		API/White		AIAN/White	
	(Black or White) vs. Other	Black vs. White	(API or White) vs. Other	API vs. White	(AIAN or White) vs. Other	AIAN vs. White
Hispanic (Y = 1, N = 0)	+		+	+	+	+
Gender (F = 1, M = 0)						
Born in US (Y = 1, N = 0)		+		-		
FPL (0-199% = 1, ≥ 200 = 0)	+					
Education (≤ HS = 1, >HS = 0)				+		+
Age (continuous)				-		
Median County Household Income						
County percent Black, API, or AIAN		+				+

Table 3.3.1: Univariate Descriptive Statistics of the proportion of Californian's 18+ that have visited the dentist at least once in the past 12 months by racial group (n = 56,037)

Variable	White (se) n = 40,851	Black (se) n = 3,002	API (se) n = 5,465	AIAN (se) n = 2,882	Other (se) n = 6,472
Age					
18-29	0.70 (.01)	0.72 (.03)	0.69 (.02)	0.67 (.05)	0.55 (.02)
30-44	0.71 (.01)	0.70 (.03)	0.73 (.01)	0.63 (.04)	0.58 (.01)
45-64	0.77 (.01)	0.67 (.02)	0.72 (.02)	0.63 (.03)	0.59 (.02)
65+	0.70 (.01)	0.52 (.03)	0.64 (.03)	0.56 (.07)	0.51 (.04)
FPL					
< 100%	0.61 (.01)	0.71 (.03)	0.62 (.02)	0.59 (.04)	0.55 (.01)
100-199%	0.63 (.01)	0.64 (.02)	0.63 (.02)	0.65 (.04)	0.59 (.01)
200-299%	0.70 (.01)	0.71 (.03)	0.73 (.02)	0.70 (.04)	0.66 (.02)
≥ 300%	0.81 (.00)	0.75 (.01)	0.78 (.01)	0.77 (.03)	0.76 (.01)
Education					
< HS	0.57 (.01)	0.59 (.04)	0.54 (.02)	0.62 (.04)	0.54 (.01)
HS	0.70 (.01)	0.70 (.02)	0.71 (.02)	0.69 (.03)	0.68 (.01)
> HS	0.80 (.00)	0.74 (.01)	0.76 (.01)	0.72 (.03)	0.71 (.01)
Insurance					
Current unins	0.53 (.01)	0.50 (.04)	0.55 (.02)	0.53 (.04)	0.49 (.01)
Unins 12 mo	0.62 (.01)	0.67 (.04)	0.62 (.05)	0.62 (.07)	0.55 (.02)
Ins 12 mo	0.78 (.00)	0.74 (.01)	0.76 (.01)	0.72 (.02)	0.68 (.01)
MSA					
Metro	0.74 (.00)	0.71 (.01)	0.72 (.01)	0.67 (.02)	0.61 (.01)
Non-metro	0.70 (.01)	0.73 (.04)	0.68 (.04)	0.68 (.04)	0.60 (.02)

Table 3.3.2: Logistic Regression Analysis for Visiting the Dentist of Californians (n = 56,037)

Variable	Coefficient (SE)	p-value
Race		
White	0.04 (0.04)	0.27
Black	-0.11 (0.05)	0.03
API	-0.02 (0.05)	0.63
AIAN	-0.09 (0.07)	0.22
Other (ref)	1.00	-
Age		
age^{-1}	-26.18 (7.16)	< 0.001
age^{-2}	524.1 (116.63)	< 0.001
FPL		
< 100%	-0.56 (0.05)	< 0.001
100-199%	-0.56 (0.04)	< 0.001
200-299%	-0.40 (0.04)	< 0.001
≥ 300% (ref)	1.00	-
Education		
< HS	-0.65 (0.05)	< 0.001
HS	-0.32 (0.03)	< 0.001
> HS	1.00	-
Insurance		
Current unins	0.70 (0.04)	< 0.001
Unins 12 mo	0.56 (0.06)	< 0.001
Ins 12 mo	1.00	-
MSA		
Metro	0.13 (0.03)	< 0.001
Non-metro	1.00	-

Table 4.1: Proportion of individuals that visited the dentist in the past 12 months by racial group for the three proposed methods with standard errors in parenthesis, n = 56,270

Race	n	$\hat{p}^{CHIS}$	$\hat{p}^{CRA}$	$\hat{p}^{MI}$	$\hat{p}^{MCA}$
White	38760	0.726 (0.003)	0.724 (0.007)	0.725 (0.001)	0.725 (0.003)
Black	2615	0.674 (0.011)	0.672 (0.013)	0.672 (0.003)	0.675 (0.011)
API	5053	0.705 (0.008)	0.703 (0.013)	0.702 (0.002)	0.704 (0.008)
AIAN	935	0.639 (0.023)	0.639 (0.016)	0.616 (0.005)	0.642 (0.021)
Other	6445	0.570 (0.008)	0.571 (0.008)	0.575 (0.002)	0.571 (0.008)
2+ Races	2462	0.648 (0.016)	-	-	-

References:

Barnard, J., Rubin, D.B., Schenker, N., (1998). Multiple Imputation Methods. In Encyclopedia of Biostatistics. Armitage P. & Colton T. (eds). Wiley: Chichester, v.4: 2772-2780.

Brick J.M, Montaquila, J., & Roth, S. Identifying problems with raking estimators. *2003 ASA Proceedings*, Alexandra, VA: American Statistical Association, 710-717.

Deming, W.E. & Stephan, F.F., (1940). On a least squares adjustment of a sample frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11(4): 427-444.

Deville, J.C., & Sarndal, C.E., (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418): 376-382.

Efron, B., (1982). Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7: 1-26

Efron, B., (1982). *The Jackknife, Bootstrap and Other Resampling Plans*. Philadelphia: Society for Industrial and Applied Mathematics

Hanley, J.A., & Mcneil, B.J., (1982). The meaning and use of the area under the receiver operating characteristic (ROC) curve. *Diagnostic Radiology*, v. 143(1): 29-36

Korn, E.I., & Graubard, B.I., (1999). Bootstrap. In Analysis of Health Surveys. Groves, R.M, Kalton, G., et al. (eds). Wiley: New York: 32-33

OMB, (1977). Race and Ethnic Standards for Federal Statistics and Administrative Reporting. Statistical Policy Directive 15.

OMB, (1997). Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity. Federal Register 62FR58781-58790.

OMB, (March 9, 2000). Guidance on the Aggregation and Allocation of Data on Race for Use in Civil Rights Monitoring and Enforcement. Retrieved on August 15, 2006 from http://www.whitehouse.gov/omb/bulletins/b00-02.html.

Rubin, D.B, (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley.

Rubin, D.B., & Schenker, N., (1991). Multiple imputation in health-care databases: An overview and some applications. *Statistics in Medicine*, 10: 585-598.

Schenker, N., & Parker, J.D., (2003). From single-race reporting to multiple-race reporting: using imputation methods to bridge the transition. *Statistics in Medicine*, 22: 1571-1587.

U.S. Department of Health and Human Services (1985). Report of the Secretary's Task Force on Black and Minority Health. U.S. Government Printing Office.

U.S. Department of Health and Human Services (1991). Health People 2000: National health promotion and disease prevention objectives. Washington, DC: U.S. Government Printing Office

U.S. Department of Health and Human Services (2000). Health People 2010. Washington, DC: U.S. Government Printing Office