Published in final edited form as:
J Immigr Minor Health. 2011 December ; 13(6): 1099–1109. doi:10.1007/s10903-010-9415-8.

The Advantage of Imputation of Missing Income Data to Evaluate the Association Between Income and Self-Reported Health Status (SRH) in a Mexican American Cohort Study

Anthony B. Ryder, **Anna V. Wilkinson**, **Michelle K. McHugh**, **Katherine Saunders**, **Sumesh Kachroo**, **Anthony D'Amelio Jr.**, **Melissa Bondy**, and **Carol J. Etzel**
Department of Epidemiology, UT MD Anderson Cancer Center, Houston, TX 77030, USA

Carol J. Etzel: cetzel@mdanderson.org

Abstract

Missing data often occur in cross-sectional surveys and longitudinal and experimental studies. The purpose of this study was to compare the prediction of self-rated health (SRH), a robust predictor of morbidity and mortality among diverse populations, before and after imputation of the missing variable "yearly household income." We reviewed data from 4,162 participants of Mexican origin recruited from July 1, 2002, through December 31, 2005, and who were enrolled in a population-based cohort study. Missing yearly income data were imputed using three different single imputation methods and one multiple imputation under a Bayesian approach. Of 4,162 participants, 3,121 were randomly assigned to a training set (to derive the yearly income imputation methods and develop the health-outcome prediction models) and 1,041 to a testing set (to compare the areas under the curve (AUC) of the receiver-operating characteristic of the resulting health-outcome prediction models). The discriminatory powers of the SRH prediction models were good (range, 69–72%) and compared to the prediction model obtained after no imputation of missing yearly income, all other imputation methods improved the prediction of SRH (P<0.05 for all comparisons) with the AUC for the model after multiple imputation being the highest (AUC = 0.731). Furthermore, given that yearly income was imputed using multiple imputation, the odds of SRH as good or better increased by 11% for each $5,000 increment in yearly income. This study showed that although imputation of missing data for a key predictor variable can improve a risk health-outcome prediction model, further work is needed to illuminate the risk factors associated with SRH.

Keywords

Self-rated health; Missing income data; Data imputation techniques; Mean substitution; Multiple imputation; Minority health

Introduction

Missing data often occur during cross-sectional surveys, longitudinal or experimental studies and have the potential of reducing the sample size available for analysis thereby introducing statistical biases that possibly lead to invalid inferences, inflated type 1 error (i.e., incorrectly rejecting the null hypothesis when the hypothesis is true) or lower power [1]. Further, the impact of missing income data on health outcomes in racially diverse

Correspondence to: Carol J. Etzel, cetzel@mdanderson.org.

populations has been understudied thus undermining health disparities research. To address potential problems associated with missing data and its impact on health outcomes in minority populations, exploring the relationship between economic indicators (i.e., income and socioeconomic status) and a key predictor of health status (i.e., self-rated health (SRH)) is a necessary step when conducting health disparities research.

To investigate the relationship between missing data and health status in a minority population we conducted a study in two stages. First we analyzed data from an on-going Mexican American Cohort Study (MACS) to determine which variables could be associated with SRH, a key predictor of morbidity and mortality. Our analyses revealed education, gender, income and age could be potentially associated to self-reported health (SRH) status. Based on this finding, we hypothesized that the imputation of missing data and yearly income would enhance the predictive power of SRH among study participants.

Therefore to illustrate our hypothesis, we first give a brief overview of the types of missing data and some statistical methods frequently used to address such data. In view of that, the purpose of this study was to identify, using four statistical techniques, how imputing missing income data affects the association between income and self-reported health (SRH) status, a robust predictor of morbidity and mortality among a minority population.

Types of Missing Data

Missing data vary depending upon the pattern of missingness and the nature of the underlying assumptions. Although there are numerous causes of missing data, the mechanisms that lead to all of them can be grouped into three main categories: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). Data are defined as MCAR when the probability of missingness for a variable of interest is independent of all other observed and unobserved variables [2, 3]. For example, individuals may refuse to report their annual household income, but their refusal may not be dependent neither on how much they earn, nor on their age nor their gender. From this perspective, the process of acquiring complete data from participants can be regarded as a simple random sample drawn from the entire set of observations. Data are MAR when the probability of missingness is independent of the true value of that variable when controlling for one or more variables [2, 3]. In this case, the probability of missing “yearly income” may be related to gender (e.g., women may be more likely not to report income compared to men), but within each gender class, the probability of missing “yearly income” is unrelated to how much someone earns. MCAR and MAR data are considered *ignorable* because missing values can be imputed without inducing any systematic bias [2, 3].

Non-ignorable data include those data that are MNAR in which the probability of missing depends on some unobserved data values or the magnitude of the missing value [4]. For example, people who make a low “yearly income” or high “yearly income” maybe less likely to report income compared to middle-income earners because they do not want to admit how much they make; therefore, a participant’s refusal to respond to a question about the amount of his or her wage could be linked to the amount of his/her income. *Non-ignorable* data cannot be readily imputed because they generate a systematic bias that affects point estimates, standard deviations and correlation coefficients [2, 3].

Addressing Missing Data

Several methods of handling *ignorable* missing data have been described in the literature [5] including listwise deletion (LD) in which any unit (or participant) with a missing value for a variable is dropped from the analysis. Several statistical packages run this technique by default. Listwise deletion is easy to interpret if the data is considered MCAR. The caveats of

this approach however, are that the sample size decreases proportionately to the number of missing data, hence yielding a lack of representativeness of the sample for the population of interest. Additionally, standard errors increase and therefore, the power to detect a statistical difference decreases [2, 3]. Nonetheless, there are several available methods to impute missing data points to overcome the problems generated by the listwise deletion approach: single (mean and regression methods) and multiple imputation [3, 5].

In single imputation, each missing value in the data matrix is filled with one replacement value. The sample size is thus brought back to its theoretical maximum resulting in no loss of power. Some common forms of single imputation include *mean substitution*, *regression-based single imputation* and *regression-based single imputation with error term*. In the *mean substitution* approach, the mean value of a variable is calculated using all available values and any missing value is replaced by the mean. Although easy to implement, this method results in a lesser variability after imputation than what existed prior to imputation. The *regression-based single imputation* method involves regressing the variable with missing values onto other available variables using all non-missing data. A missing value is replaced by its predicted value from the resulting regression. Like the mean-substitution method, the regression-based single imputation is easy to implement and may also result in reduced variability. Further, a regression-based single imputation may introduce systematic bias in the analysis [3, 5].

To restore uncertainty around each imputed value and hence restore the same degree of variability within the imputed variable, a value representing random error can be added to the regression-based single imputation. The random error value can be generated as a zero-mean normally distributed variate with variance based on the mean square error of the regression. In theory, this approach of *regression-based single imputation with error term* performs better than all the previous ones because it reduces the artificial inflation seen in the relationship between the variable with missing values and the other variables in the model. Unfortunately, it yields a biased standard error [2, 6].

Multiple imputation (MI) differs from single imputations in that it yields a set of several plausible replacement values for each missing data point. This method offers the advantage of producing unbiased estimates and valid statistical inferences that appropriately account for the uncertainty associated with each newly imputed value [7]. Furthermore, MI produces standard errors that reflect the inherent uncertainty present in imputing missing values. Like the other techniques, MI also requires that the data be either MCAR or MAR [3]. Although much of the comparisons between MI and single imputation methods are associated with the distribution of the imputed variable, it is important to consider how these methods affect further inference of the imputed variable and other factors related with the imputed variable. Numerous statistical packages exist to perform MI, including but not limited to S-PLUS, R, SOLAS, STATA, SAS, LogXact, and Winbugs [5, 8]. More recently, Mplus released a new version (6.0) which generates imputation for both independent and identically distributed and for clustered continuous and ordered data [9].

Importance of Socioeconomic Status and Self-Rated Health

Previous studies have highlighted the importance of socioeconomic status (SES), of which income is a central component, as a significant determinant of various health outcomes, including self-rated health (SRH) [10–16]. For example, income has been extensively examined to better understand existing social and racial disparities in health outcomes [11, 12, 16]. However, investigating the role of SES and the accurate reporting of income or obtaining correct and complete information remains a challenge. Factors previously noted (i.e., such as women may be more likely not to report income compared to men; or people who make a low “yearly income” or high “yearly income” maybe less likely to report

income compared to middle-income earners because they do not want to admit how much they make) contribute to a high proportion of missing data for the income variable that leads to a reduction in the analyzable study population and eventually affects the analysis and final inferences. In order to maintain sample size, the variable could also be excluded from all analyses which also limit final inferences as well.

Self-rated health (SRH) is one of the most consistent predictors of morbidity and mortality [15, 17–19]. Regardless of ethnicity, individuals who perceive their health as fair or poor tend to report more physician visits per year, spend more days per year in bed, and are more likely to die prematurely than are individuals who typically rate their health as good or better [15, 18, 19].

Past studies have examined variations in the proportions of missing data in different racial groups, but most have focused on clinical variables, not SES or income data. In a population-based study of postpartum women in California, Kim et al. [11] reported that both US born Latinas as well as Latina immigrants have a higher proportion of missing income data compared to their white counterparts. More importantly, because studies looking at SRH and missing SES data have focused on the population at large [20] the generalizability of results to minority populations may be questionable since the amount of missing data is higher among minority populations [11]. Most existing studies involving minorities have employed only one type of data imputation technique and have not compared their results to other possible imputation methods. Yet the higher proportion of missing data for household (HH) income among minority and immigrant respondents compared to whites [11, 21] leads to a reduction in the analyzable study population, and thereby may lead to biased results and invalid inferences.

Materials and Methods

Data Source: Mexican American Cohort Study

Data for this analysis are derived from participants in a population-based study of Mexican-American households conducted by the University of Texas MD Anderson Cancer Center's Department of Epidemiology. The Mexican American Cohort Study (MACS) was created to serve as a long-term resource to study the health of Mexican Americans and to examine bio-behavioral determinants (e.g., smoking and obesity) of disease (including cancer). Participants in the MACS included self-identified adults of Mexican origin who were enrolled as the primary contact from the household into the MACS. Participants were recruited between July 1, 2002 and December 31, 2005 from 16 neighborhoods in south central and southeast Houston using a variety of recruitment methods that included random-digit dialing, intercept, door-to-door block walking and through network contacts. The neighborhoods were selected if more than 80% of the residents were of Mexican origin and all were low income according to the 2000 US census. For more detailed information see Wilkinson et al. [22]. The University of Texas MD Anderson Cancer Center Institutional Review Board approved all aspects of this research.

Study Population

The variable "country of birth" was self-reported as either "United States (US)" or "Mexico." The variable "years lived in the US" represents the number of years the Mexican-born study participants had lived in the US. "Education level" comprised the following categories: less than high school, high school graduate, and some college. Smoking status was self-reported and categorized as follows: *never smokers*, those who had smoked <100 cigarettes in their lifetime; *former smokers*, those who had quit smoking >1 year before interview; or *current smokers*, those who had quit smoking within the past 12 months or

who currently smoked at time of interview. Pack-years for current and former smokers were calculated as follows: [(years smoked) × (the average number of cigarettes smoked per day)/20]. For example, a person who smoked 10 cigarettes a day for 30 years has a (10 × 30)/20 = 15 pack year smoking history.

Body mass index (BMI) was calculated as the ratio of weight (*kg*) to the square of the height (*m*) at the time of enrollment. On the basis of the Centers for Disease Control and Prevention's (CDC) BMI classification scheme, individuals were grouped according to their BMI as *under-weight* (<18.5), *normal* (18.5 to 24.9), *overweight* (25.0 to 29.9), or *obese* (≥30.0) [23]. Furthermore, the MACS respondents self-reported up to four physician-diagnosed diseases or conditions. Respondents reported a wide range of conditions, however the four most prevalent chronic conditions (i.e., hypertension, diabetes, cancer, and heart disease) accounted for overwhelming majority of the conditions reported. Therefore, in our study, participants were asked to self report a prior diagnosis of any cancer, hypertension, heart disease, or diabetes mellitus. A composite index was then computed per participant with a code that ranged from 0 (no self-reported prior diagnosis) to 4 (the four conditions were self reported by the same participant). The composite index for physical health status was defined as an aggregate of all self-reported and prior co-morbidities ranging from 0 (no reported prior diagnoses) to 4 (prior diagnosis of any cancer, hypertension, heart disease, or diabetes mellitus).

Annual household (HH) income was defined as annual HH income before taxes over the past year and was reported as <$5,000; $5,000–$9,999; $10,000–$14,999; $15,000–$24,999; $25,000–$34,999; $35,000–$44,999; $45,000–$54,999; $55,000–$64,999; $65,000–$74,999; and >$75,000. Annual HH income was considered missing if the participant responded "Don't know" or "Prefer not to answer."

The outcome variable of interest, SRH was assessed using the question, "In general would you rate your health as, excellent, very good, good, fair or poor?" Variations of this one item measure have been used for over half a century to assess health status in population-based research and the measure is widely accepted as reliable and valid. Self-rated health has been examined as a continuous variable, an ordinal categorical variable as well as a dichotomous variable; regardless of how it is analyzed, results are consistent [17]. Therefore to be consistent with previous research based on the MACS [22], SRH was dichotomized into good or better and fair or poor perceptions.

Statistical Analyses

Although the MACS began in July 2001, the SRH item was introduced in July 2002. Therefore, of the 4,916 participants enrolled by December 31 2005, 4,306 were asked the item probing SRH. To better clarify the effect of missing yearly income on the health-outcome prediction model, all observations with missing values for any variable other than income were excluded, yielding a final data set of 4,162 participants (overall set). Seventy-eight participants were excluded due to missing data on height or weight variables, 33 due to missing data on smoking status, and the remainder due to missing data on education level. The baseline characteristics and the odds ratio were drawn from the analyses of the overall set and then divided into a training set (to derive the yearly income imputation methods and develop the health-outcome prediction models) and a testing set (for area under the curve (AUC) comparison of the resulting health-outcome prediction model after imputation). Data for 75% of the participants (N = 3,121) were randomly assigned to the training set; data for the remaining 25% (N = 1,041) were randomly assigned to the testing set. Descriptive statistical analyses were used to characterize the study population. Pearson's χ^2 test was used to analyze the distribution differences between the health-perception groups. Student's

t test was used to evaluate the differences between the two health-perception groups for all continuous variables. All *P* values were two-sided.

Imputation of Missing Data

For imputation purposes, the yearly income was re-coded by using the median of the income class intervals (i.e., the median of $20,000 was used for the interval $15,000–$24,999). We used four different methods to impute missing income data. Three of these methods were based on single imputation as previously discussed, in which each missing value in the data matrix was filled with one replacement value: mean substitution (MS), regression-based single imputation (RB), and regression-based single imputation with error term (RBE). For the MS method, missing yearly income was replaced with mean yearly income calculated from those participants with complete yearly income data. For the RB method, we regressed yearly income on age, gender, country of birth, years lived in the US, and education level (variables hypothesized to be associated with yearly income) and replaced each missing yearly income value with its predicted value. The imputation of yearly income for the RBE method was the same as that of the RB method except that we added a random error term to each predicted yearly income value such that the random error term followed a normal distribution with a mean of zero and variance equal to the mean square error of the RB model [2, 6]. Listwise deletion (LD), MS, RB, and RBE were performed using STATA version 9 (Stata Corporation, College Station, TX).

Because the missing data on income may be due to a myriad of causes that could be due to mechanisms that lead to both ignorable and non-ignorable data, we used multiple imputation (MI) as the fourth imputation method under the assumption that the data was ignorable (i.e., MCAR or MAR) and would yield better results. We performed MI with Winbugs 1.41 software, using a Bayesian approach including Markov chain Monte Carlo estimation and a Gibbs sampler technique. To determine the number of imputations needed, we used the formula of relative efficiency [24]. $(1+\frac{\gamma}{m})^{-1}$, where γ is the percentage of missing information for the quantity being estimated, and *m* is the number of imputations.

In this study, we considered a relative efficiency for five imputations with a noninformative prior for income. Imputed values were drawn from a posterior predictive distribution of the missing values conditional on the observed values. The Winbugs software produced the initial values for five parallel Markov chain Monte Carlo sampling chains and created five datasets after an initial “burn-in” cycle of 2,000 iterations each, followed by a cycle of 20,000 iterations. Each chain was thinned by 10 to reduce the autocorrelation among the parameter estimates. From all five newly created datasets, Winbugs automatically calculated per observation the average of the five imputed values for yearly income. S-PLUS was used to reassemble the final complete dataset for analysis using the averaged imputed yearly income across all five imputation sets.

Analyses were conducted on the overall dataset (base-lines), the training dataset (model building) and testing dataset (model prediction) using Stata, version 9 (Stata Corporation, College Station, TX); S-PLUS, version 7 (Insightful Corporation, Seattle, WA); Winbugs, version 1.41 (Winbugs, Imperial College School of Medicine, London, and MRC Biostatistics Unit, Cambridge, United Kingdom); and Number Cruncher Statistical System NCSS 2007 (Kaysville, Utah).

Imputation Methods and the Development of the Health-Outcome Prediction Models

We used logistic regression analysis to build the five health-outcome prediction models, one model for each imputation method and a model for which no imputation of yearly income was used, commonly referred to as listwise deletion (LD) in statistical software packages. In

LD individuals with missing income data are excluded. For our models, the response variable, SRH, was re-coded as 0 for fair or poor and 1 for good or better. In the training data set, we calculated odds ratio (ORs) and 95% confidence intervals (CIs) for the evaluation of SRH by yearly income, and adjusted for gender, age, years lived in the US, country of birth, education, BMI, and smoking history. In these models, yearly income was included in its linear form and after log transformation.

We used the testing data set to calculate the specificity and sensitivity of the resulting health-outcome prediction models, constructed receiver operating characteristic (ROC) curves, and calculated the area under the curve (AUC) statistic to estimate each model's ability to discriminate between poor and good health perception. We then compared the AUCs for each health-outcome prediction model with the one obtained through LD (reference model) by using the method described by Hanley and McNeil [25] and Hintz [26].

Results

Demographic characteristics of the 4,162 Mexican American participants are shown in Table 1. Most participants were women (87.3%), born in Mexico (69.5%), had not completed their high school education (61.5%), never smoked (75.2%) and had no previous history of cancer, hypertension, heart disease, or diabetes mellitus (73.9%). A total of 2,376 participants (57.1%) rated their health as good or better.

Participants who perceived themselves to be in good or better health had a statistically significant younger mean age (40.18 years) than their peers with fair or poor SRH (46.39 years; $P<0.0001$). Among Mexico-born participants, the years lived in the US were significantly higher for those with fair or poor SRH (47.74 years) compared to those who rated their health as good or better (41.78 years; $P<0.0001$).

The mean BMI was significantly lower among participants with good or better SRH (29.46) than among those with fair or poor SRH (31.03; $P<0.0001$). Consistent with this result, a higher percentage (83.3%) of overweight or obese participants rated their health as fair or poor (based on CDC categories for BMI) compared to those with a good or better SRH (76.2%; $P<0.0001$).

A higher percentage (72.6%) of participants who rated their health as fair or poor had less than a high school education compared to their peers who rated their health as good or better (53.3%). The percentage of participants with more than a high school education was higher in the group with a good or better SRH (26.8%) than it was in the group with fair or poor SRH (13.2%; $P<0.0001$). Of those participants who perceived their health as good or better, 85.1% had no prior co-morbidities compared with 59.1% among those who perceived their health as fair or poor ($P<0.0001$).

Among the responses on yearly income obtained from the 2,376 participants with good or better SRH, we noted 1,506 observations and 870 missing values. Similarly, among the responses on yearly income from the 1,786 participants with a poor SRH, we noted 925 observations and 861 missing values. Although the median yearly income was $20,000 for both groups, the overall percentage of missing data was 41.6% (36.6% among those with a good or better SRH and 48.2% among those with fair or poor SRH).

Table 2 compares yearly income of those with good versus those with poor SRH within the training and testing sets, without and after imputation. Irrespective of the imputation method, the mean yearly income was always significantly greater in those with a good or better SRH than it was in those with fair or poor SRH (in the training set, $P<0.0001$ for all

methods; in the testing set, $P<0.001$ for the MS imputation and $P<0.0001$ for LD, RB, RBE and MI), demonstrating that degree of difference in yearly income among the two SRH groups does not change after imputation.

The main effects of the SRH prediction models without and after income imputation are summarized in Table 3. These results are derived using the combined data (training plus testing set), as the ORs between these models and those derived from the training set were similar, but CIs using the combined data were more precise. Although SRH was not significantly associated with gender, country of birth, or smoking status in any of the models, all other results were in the expected directions.

Age was indirectly associated with good or better SRH rating (OR = 0.97) across all models. Individuals with more than a high school education were more likely (OR range 1.80–2.08 across the models) to rate their health as good or better compared to those with less than a high school education. Individuals who were obese were more likely to rate their health as fair or poor (OR range 0.67–0.71 across the models) compared to those who were normal or underweight. Likewise, those participants who had at least one reported co-morbidity were more like to rate their health as fair or poor compared to those with no co-morbidities.

Regardless of the imputation method, there was a significant association between SRH and yearly income. In the situation in which missing yearly income was not imputed (LD method), the odds of SRH as good or better increased by 5% for each $5,000 increment in yearly income; similar to the results obtained after fitting the prediction model with the MS and RB imputation methods. However, when yearly income was imputed using MI, the odds of SRH as good or better increased by 11% for each $5,000 increment in yearly income.

We also compared the prediction of SRH for the logistic models that included the variable “yearly income” in its linear form and after a logarithmic transformation. In the analyses of the five logistic models, either form of yearly income significantly improved the fit of the health-outcome prediction model when compared respectively with a reduced model excluding yearly income ($P<0.0001$; Table 3). However for SRH prediction, the logistic model with yearly income in its linear form provided a better fit than did the logistic model using the logarithmic transformation (data not shown).

Table 4 compares the discriminatory power for each SRH prediction model without and after imputation. The AUCs (all 0.724) of the MS, RB, RBE models were significantly higher compared to the AUC (0.696) of the LD model (P values 0.03, respectively), suggesting that the single imputation of yearly income improved the prediction of SRH. The prediction model obtained after MI of yearly income had the highest AUC (0.731, 95% CI = 0.699–0.760) which was significantly higher compared to the LD method ($P = 0.007$) and also the other imputation methods ($P<0.05$; data not shown).

Discussion

In our health-outcome prediction models, we observed that yearly income was significantly associated with SRH status and in the prediction model using MI to impute missing yearly income, the odds of SRH as good or better increased by 11% for each $5,000 increment in yearly income. Our findings support previously published studies that found an association between poor SRH and those at the bottom of the income and SES hierarchy [10, 11, 14–16]. In contrast, a study that examined the relationship between SRH and different indicators of SES such as income, education level, public assistance, material deprivation and subjective social standing among pregnant Latinas found no association between income and SRH [27]. Such contradictory findings suggest that the relation between income and

minority population health warrants continued research and the development of more parsimonious risk prediction models.

The discriminatory powers of the SRH prediction models were good [28] (range, 69–73%) and compared to the model obtained from the LD method, all other imputation methods improved the prediction of SRH with multiple imputation of yearly income resulting in the model with the highest discriminatory power. The decrease in the available sample size during LD may have decreased the power to detect a statistical difference, thus it caused the LD method to yield lower, albeit still good, discriminatory power. Since MI models are credited for producing unbiased estimates for the imputed variable of interest, as well as, account for the presence of the uncertainty needed around the newly imputed value, we also wanted to see how MI of a missing variable (yearly income) would result in an optimal prediction model of an associated factor (health status).

In our study, we used a *noninformative* prior for MI imputation of yearly income and observed that the SRH prediction model built after MI imputation of missing yearly income resulted in the best prediction model (in terms of AUC) for SRH compared to other income imputations. However, further studies are warranted to examine the effects of both non-informative and informative priors for imputation of missing yearly income and how MI of yearly income with an informative prior may further improve the discriminatory power of an SRH prediction model. Although we utilized MI imputation to handle missing income data, there exist alternative model-based approaches to handling missing data under the MAR assumption that we could have used. These include, but are not limited to full information maximum-likelihood (FIML), the EM algorithm or fully Bayesian estimation. The literature is rich with articles comparing the pros and cons of model-based and model-free imputation methods. However, we choose the MI approach due to its ease of use and its robustness.

Based on our literature search, we found very few studies that have looked at missing income data in minority populations. Kim et al. [11] used data from a population-based postpartum survey from California, and reported that both US born Latinas as well as Latina immigrants have higher proportion of missing data on income compared to their white counterparts. A possible indication from this could be that higher proportion of missing income data in minority populations could be due to the majority of the missing responses come from females, who may not be aware of their total earnings. It is also possible that these females work as daily wageworkers and do random jobs (with different pay rates) when these jobs become available, which further increases the chances of not remembering their actual income. Thus, it is important to conduct further research to investigate the factors that could be affecting the responses obtained from females from minority groups.

A study comparing nutrient intake among Mexican American and Non-Hispanic white women has also reported that the proportion of study participants with missing data was much higher in Mexican American women as compared to Non-Hispanic white women [21]. Hence, one of the questions that researchers need to focus on is how effective are the currently available data collection instruments for minority populations, especially when the complexity of questions increases. The responses may be affected differently when factors like poverty, low education, acculturation stress, “discriminatory experiences”, “immigration experiences” and psychological stress come into play [29, 30].

Very few minority studies have looked at different methods of data imputation and SRH [20]; therefore our study is unique because we examined different imputation techniques for addressing missing income data in a minority population and applied the results to develop a risk health-outcome prediction model. Given the fact that the proportion of minority and immigrant populations in the US continues to increase and that these groups are typically of

low SES, with poor access to healthcare, it is timely and warranted to conduct research in these groups. Doing so will help illustrate factors that affect minority and immigrant population health outcomes as well as to improve the currently available methods that address missing data.

Limitations of the study included the self-reporting of income that could result in an increased probability of recall bias. Therefore, it is possible, as in any study working with non-simulated data, that income, when missing, did not meet the MCAR or MAR assumptions. Although problematic for proper imputation of missing income, the causes of missing data may be to due to a myriad of challenges that lead to both ignorable and non-ignorable missing data. Therefore approaches that rely on the assumption of ignorability (i.e., MI) yield better results when compared with other methods such as listwise deletion [31]. Researchers working with complex sampling designs (e.g., data which include strata, probability weights, clustering or replicate weights) should note that using multiple imputation methods are much more complex than single imputation methods. Finally, since the majority of participants in our study were female, our ability to generalize the findings of this study to a sub-population of Mexican American men is limited.

Although “yearly HH income” was a good predictor, in future studies we plan to analyze the interplay of other aspects of SES as well as factors closely associated with SES that impact SRH and may increase the likelihood of meeting the MAR assumption and predicting SRH across generations of Mexican Americans. Specifically we intend to examine determinants of education level, the role of wealth (home ownership), as well as the availability of medical insurance and the overall access to healthcare, which is closely tied to SRH and SES. Finally while the one-item measure of SRH is widely accepted as a valid and reliable measure, it is possible that the item does not measure all aspects of the underlying construct as adequately as a well-constructed multi-item scale.

Conclusion

The results of this study have demonstrated that the MS, RB, and RBE methods, along with the MI method used with Bayesian modeling, have performed better than the LD method in a situation where the ultimate outcome is the development of a risk health-outcome prediction model. Furthermore and regardless of any of the imputation methods applied, the explanatory variable “yearly income” was a good predictor of SRH outcome. However, further research is needed to illuminate other possible aspects of SES that may also predict SRH among minorities populations. Also, studies comparing individual income versus HH income from different sources may also provide interesting insights.

Although the collection of complete data is the ultimate goal of every researcher, missing data, like in this study, are at times inevitable. In such cases, the researcher should evaluate the usefulness of salvaging the incomplete data. Imputation techniques will still play an important role in fixing research data plagued by missing values.

Abbreviations

AUC	Area under the curve
CI	Confidence interval
DF	Degree of freedom
HH	Household
LD	Listwise deletion

LR	*P* value from LR testing
MACS	Mexican American Cohort Study
MAR	Missing at random
MCAR	Missing completely at random
MI	Multiple imputation
MNAR	Missing not at random
MS	Mean substitution
OR	Odds ratio
RB	Regression-based single imputation
RBE	Regression-based single imputation with error term
ROC	Receiver operating characteristic
SD	Standard deviation
SE	Standard error
SES	Socioeconomic status
SRH	Self-rated health

References

1. Raghunathan TE. What do we do with missing data? Some options for analysis of incomplete data. Annu Rev Public Health. 2004; 25:99–117. [PubMed: 15015914]
2. Allison, PL. Missing data. Newbury Park: Sage Publications; 2002.
3. Little, RJ.; Rubin, DB. Statistical analysis with missing data. 2. New York: Wiley; 2002.
4. Streiner DL. The case of the missing data: methods of dealing with dropoouts and other research vagaries. Can J Psychiatry. 2002; 47:68–75. [PubMed: 11873711]
5. Graham JW. Missing data analysis: making it work in the real world. Annu Rev Psychol. 2009; 60:549–76. [PubMed: 18652544]
6. Schafer JL. Multiple imputation: a primer. Stat Methods Med Res. 1999; 8(1):3–15. [PubMed: 10347857]
7. Patrician PA. Multiple imputation for missing data. Res Nurs Health. 2002; 25(1):76–84. [PubMed: 11807922]
8. Horton NJ, Kleinman KP. Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. Am Stat. 2007; 207:61–90.
9. Christensen, LM.; Haug, PJ.; Fiszman, M. Mplus: a probabilistic medical language understanding system. Proceedings from the ACL-02 workshop on Natural language processing in the bio-medical domain; Philadelphia: University of Pennsylvania; 2002.
10. Mackenbach JP, et al. The shape of the relationship between income and self-assessed health: an international study. Int J Epidemiol. 2005; 34(2):286–93. [PubMed: 15561750]
11. Kim S, et al. Potential implications of missing income data in population-based surveys: an example from a postpartum survey in California. Public Health Rep. 2007; 122(6):753–63. [PubMed: 18051668]
12. Lannin DR, et al. Influence of socioeconomic and cultural factors on racial differences in late-stage presentation of breast cancer. JAMA. 1998; 279(22):1801–7. [PubMed: 9628711]
13. Davern M, et al. The effect of income question design in health surveys on family income, poverty and eligibility estimates. Health Serv Res. 2005; 40(5 Pt 1):1534–52. [PubMed: 16174146]

14. Ross NA, et al. Relation between income inequality and mortality in Canada and in the United States: cross sectional assessment using census data and vital statistics. BMJ. 2000; 320(7239): 898–902. [PubMed: 10741994]
15. Subramanian SV, Kawachi I. Income inequality and health: what have we learned so far? Epidemiol Rev. 2004; 26:78–91. [PubMed: 15234949]
16. Banks J, et al. Disease and disadvantage in the United States and in England. JAMA. 2006; 295(17):2037–45. [PubMed: 16670412]
17. Idler EL, Benyamini Y. Self-rated health and mortality: a review of twenty-seven community studies. J Health Soc Behav. 1997; 38(1):21–37. [PubMed: 9097506]
18. Marmot M. The influence of income on health: views from an epidemiologist. Does money really matter? Or is it a marker for something else? Health Aff (Millwood). 2002; 21:31–46. [PubMed: 11900185]
19. McGee DL, et al. Self-reported health status and mortality in a multiethnic US cohort. Am J Epidemiol. 1999; 149(1):41–6. [PubMed: 9883792]
20. Engels JM, Diehr P. Imputation of missing longitudinal data: a comparison of methods. J Clin Epidemiol. 2003; 56(10):968–76. [PubMed: 14568628]
21. Abrams B, Guendelman S. Nutrient intake of Mexican-American and non-Hispanic white women by reproductive status: results of two national studies. J Am Diet Assoc. 1995; 95(8):916–8. [PubMed: 7636086]
22. Wilkinson AV, et al. Effects of nativity, age at migration, and acculturation on smoking among adult Houston residents of Mexican descent. Am J Public Health. 2005; 95(6):1043–9. [PubMed: 15914831]
23. US Department of Health, Human Services, C.f.D.C.a. Prevention. Percentage of adults aged ≥ 20 years reporting selected adverse health characteristics by Body Mass Index (BMI) category. MMWR Morb Mortal Wkly Rep. 2006; 55(23):656.
24. Rubin, DB. Multiple imputation for nonresponse in surveys. New York: Wiley; 1987.
25. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology. 1983; 148(3):839–43. [PubMed: 6878708]
26. Hintz, J. NCSS, PASS, and GESS. (cited 2008 April 29). 2006. Available from: www.ncss.com
27. Stewart AL, et al. Race/ethnicity, socioeconomic status and the health of pregnant women. J Health Psychol. 2007; 12(2):285–300. [PubMed: 17284493]
28. Bewick V, Cheek L, Ball J. Statistics review 13: receiver operating characteristic curves. Crit Care. 2004; 8(6):508–12. [PubMed: 15566624]
29. Farley T, et al. Stress, coping, and health: a comparison of Mexican immigrants, Mexican-Americans, and non-Hispanic whites. J Immigr Health. 2005; 7(3):213–20. [PubMed: 15900422]
30. Finch BK, Vega WA. Acculturation stress, social support, and self-rated health among Latinos in California. J Immigr Health. 2003; 5(3):109–17. [PubMed: 14512765]
31. Muthén B, Kaplan D, Hollis M. On structural equation modeling with data that are not missing completely at random. Psychometrika. 1987; 52:431–62.

Table 1

Baseline characteristics of participants in the University of Texas MD Anderson Cancer Center Mexican American Cohort Study

Characteristic	Total (*N* = 4,162)	SRH		*P* value[a]
		Good or better (*N* = 2,376)	Fair or poor (*N* = 1,786)	
Age: mean ± SD (years)	42.85 ± 14.87	40.18 ± 14.45	46.39 ± 14.67	<0.0001
Gender: *N*(%)				
Women	3,634 (87.3)	2,051 (86.3)	1,583 (88.6)	0.026
Men	528 (12.7)	325 (13.7)	203 (11.4)	
Country of birth: *N*(%)				
Mexico	2,893 (69.5)	1,579 (66.5)	1,314 (73.6)	<0.0001
USA	1,269 (30.5)	797 (33.5)	472 (26.4)	
Years lived in the US[b]: mean ± SD	16.74 ± 12.08	15.41 ± 11.58	18.34 ± 12.48	<0.0001
BMI: mean ± SD	30.13 ± 6.23	29.46 ± 5.99	31.03 ± 6.43	<0.0001
BMI by CDC category: *N*(%)				
Underweight	27 (0.7)	18 (0.8)	9 (0.5)	<0.0001
Normal	835 (20.0)	546 (23.0)	289 (16.2)	
Overweight	1,376 (33.0)	828 (34.8)	548 (30.7)	
Obese	1,924 (46.2)	984 (41.4)	940 (52.6)	
Smoking status: *N*(%)				
Active smoker	493 (11.8)	296 (12.5)	197 (11.0)	<0.0001
Former smoker	538 (12.9)	265 (11.2)	273 (15.3)	
Never smoked	3,131 (75.2)	1,815 (76.3)	1,316 (73.7)	
Education level: *N*(%)				
<High school graduate	2,562 (61.5)	1,266 (53.3)	1,296 (72.6)	<0.0001
High school graduate	726 (17.4)	473 (19.9)	253 (14.2)	
>High school graduate	874 (20.9)	637 (26.8)	237 (13.2)	
Health index: *N*(%)[c]				
0	3,078 (73.9)	2,022 (85.1)	1,056 (59.1)	<0.0001
1	757 (18.1)	277 (11.7)	480 (26.9)	
2	282 (6.7)	67 (2.8)	215 (12.0)	
3+	45 (1.0)	10 (0.4)	35 (2.0)	
Yearly income group: *N*(%)				
<$5,000	151 (3.6)	74 (3.1)	77 (4.3)	
$5,000–$9,999	260 (6.2)	130 (5.5)	130 (7.3)	
$10,000–$14,999	407 (9.7)	237 (10.0)	170 (9.5)	
$15,000–$24,999	646 (15.5)	404 (17.0)	242 (13.5)	
$25,000–$34,999	479 (11.5)	305 (12.8)	174 (9.7)	
$35,000–$44,999	228 (5.4)	158 (6.6)	70 (3.9)	
$45,000–$54,999	121 (2.9)	92 (3.9)	29 (1.6)	
$55,000–$64,999	53 (1.2)	37 (1.6)	16 (0.9)	
$65,000–$74,999	43 (1.0)	31 (1.3)	12 (0.7)	

Characteristic	Total (*N* = 4,162)	SRH		*P* value[a]
		Good or better (*N* = 2,376)	Fair or poor (*N* = 1,786)	
$75,000 ($80,000)	43 (1.0)	38 (1.6)	5 (0.3)	
Total responders	2,431 (58.4)	1506 (63.4)	925 (51.8)	
Missing values	1,731 (41.6)	870 (36.6)	861 (48.2)	

SRH self-rated health, *BMI* body mass index, *CDC* Centers for Disease Control and Prevention

[a] Values from two-sided Student's *t* test for continuous variables and χ^2 test for categorical variables

[b] Mexico-born participants

[c] Composite index for physical health status = aggregate of all self-reported and prior comorbidities ranging from 0 (no reported prior diagnoses) to 4 (prior diagnosis of any cancer, hypertension, heart disease, or diabetes mellitus)

Table 2

Comparison of yearly income between SRH groups without and after imputation

Imputation method	SRH				*P* value
	Good or better		Fair or poor		
	N	Yearly income[a] Mean ± SD	*N*	Yearly income Mean ± SD	
Training set (*N* = 3,121)					
LD	1,115	26.39 ± 17.29	706	21.27 ± 14.77	<0.0001
MS	1,770	25.65 ± 13.76	1,351	22.77 ± 10.79	<0.0001
RB	1,770	25.56 ± 14.20	1,351	21.17 ± 11.30	<0.0001
RBE	1,770	25.62 ± 16.75	1,351	21.61 ± 15.52	<0.0001
MI	1,770	25.99 ± 13.74	1,351	22.15 ± 10.72	<0.0001
Testing set (*N* = 1,041)					
LD	391	26.27 ± 16.94	219	21.83 ± 13.98	<0.0001
MS	606	25.70 ± 13.62	435	23.24 ± 10.01	0.001
RB	606	25.40 ± 14.09	435	21.37 ± 10.63	<0.0001
RBE	606	25.16 ± 16.33	435	20.90 ± 14.98	<0.0001
MI	606	25.88 ± 13.61	435	22.87 ± 9.97	0.0001

LD listwise deletion (no imputation), *MI* multiple imputation, *MS* mean substitution, *RB* regression-based single imputation, *RBE* regression-based single imputation with error term

[a]Yearly income per $1000

Table 3

Multivariable risk model for SRH, odds ratios and associated confidence intervals without and after imputation of yearly income

Risk factor	Imputation method				
	LD	MS	RB	RBE	MI
	OR (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)
Age	0.97 (0.96–0.98)	0.97 (0.97–0.98)	0.98 (0.97–0.98)	0.97 (0.97–0.98)	0.97 (0.97–0.98)
Gender	0.84 (0.63–1.11)	0.88 (0.64–1.00)	0.81 (0.65–1.02)	0.80 (0.65–1.00)	0.81 (0.65–1.01)
Country of birth	0.93 (0.67–1.28)	1.03 (0.81–1.31)	1.05 (0.83–1.34)	1.04 (0.82–1.33)	1.04 (0.82–1.33)
Years lived in the US	1.02 (1.00–1.03)	1.01 (1.00–1.02)	1.01 (1.00–1.02)	1.01 (1.00–1.02)	1.01 (1.00–1.02)
Education					
High school	1.23 (0.96–1.57)	1.40 (1.16–1.69)	1.37 (1.13–1.66)	1.40 (1.16–1.70)	1.36 (1.12–1.64)
>High school	1.80 (1.42–2.28)	2.08 (1.73–2.50)	2.00 (1.66–2.42)	2.08 (1.73–2.51)	1.99 (1.65–2.39)
Smoking					
Former	0.89 (0.62–1.27)	0.82 (0.62–1.07)	0.82 (0.62–1.07)	0.82 (0.62–1.08)	0.80 (0.61–1.27)
Never	1.01 (0.76–1.34)	1.05 (0.84–1.30)	1.05 (0.84–1.30)	1.06 (0.85–1.32)	1.02 (0.82–1.27)
Body mass index					
Overweight	0.96 (0.74–1.24)	0.91 (0.75–1.10)	0.91 (0.75–1.10)	0.91 (0.76–1.11)	0.91 (0.75–1.10)
Obese	0.71 (0.56–0.90)	0.68 (0.57–0.81)	0.68 (0.57–0.82)	0.68 (0.57–0.82)	0.67 (0.56–0.81)
Health index					
1 Comorbidity	0.33 (0.25–0.42)	0.35 (0.29–0.42)	0.35 (0.29–0.42)	0.35 (0.29–0.42)	0.44 (0.28–0.41)
2 Comorbidities	0.17 (0.11–0.25)	0.19 (0.14–0.26)	0.19 (0.14–0.26)	0.19 (0.14–0.26)	0.19 (0.14–0.26)
≥3 Comorbidities	0.15 (0.05–0.40)	0.20 (0.09–0.42)	0.20 (0.09–0.42)	0.19 (0.09–0.41)	0.20 (0.09–0.44)
Income[a]	1.05 (1.02–1.08)	1.05 (1.02–1.08)	1.05 (1.03–1.08)	1.03 (1.02–1.05)	1.11 (1.08–1.14)
P value from LR test	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001

OR odds ratio, *CI* confidence interval, *LD* listwise deletion (no imputation), *MI* multiple imputation, *MS* mean substitution, *RB* regression-based single imputation, *RBE* regression-based single imputation with error term, *HS* High School

[a]Income in increments of $5,000

Table 4

Comparison of discriminatory power as measured by Area under the Receiver-Operator Curve (AUC) of the SRH prediction model without and after imputation of yearly income

Imputation method	AUC	95% CI	*P* value[a]
LD	0.696	(0.664–0.726)	
MS	0.724	(0.692–0.753)	0.03
RB	0.724	(0.692–0.753)	0.03
RBE	0.724	(0.691–0.753)	0.03
MI	0.731	(0.699–0.760)	0.007

LD listwise deletion (no imputation), *MI* multiple imputation, *MS* mean substitution, *RB* regression-based single imputation, *RBE* regression-based single imputation with error term

[a]Compared to LD method