Imputation for Missing Physiological and Health Measurement Data: Tests and Applications

Matt Jans[1], Steven G. Heeringa[1], Anne-Sophie Charest[2]
[1]Michigan Program in Survey Methodology, Univ. of Michigan, 426 Thompson St. Rm 4050, Ann Arbor, MI 48104
[2]Department of Statistics, Carnegie Mellon University, Baker Hall, Pittsburgh, PA 15213

Abstract
We evaluated alternative approaches to imputation for univariate estimates and multivariate regression analyses of physiological health measures collected in the 2003-2004 National Health and Nutrition Examination Survey (NHANES). From the NHANES public use data files we selected 5041 respondents age 20+ who provided questionnaire or medical exam data. Measures collected at interview (e.g., demographics, self-reported health status) and measures collected at physical examination (e.g., height, weight, blood pressure, cholesterol, hemoglobin, Hematocrit, and iron) were evaluated for rates of item missing data (i.e., item nonresponse). The properties of several imputation methods (including single and multiple imputation) were evaluated with respect to univariate estimates and a regression model using age, sex, race, height, weight, cholesterol, and marital status to predict blood pressure. Only small differences were found between imputation methods, and no major systematic differences between methods were observed. The findings suggest that for the missing data problems considered in our investigation, the specific imputation method makes little difference on univariate and multivariate estimates and standard errors.

Key Words: Health Surveys, Physiological Measurement, Missing Data, Imputation, NHANES

1. Missing Data and Imputation Methods

When survey data are missing at the item level (i.e., individual respondents have not reported values on a subset of items in the survey protocol), an analyst is faced with the decision of whether to impute values for those that are missing. If she chooses to replace missing data, she has two choices; single imputation or multiple imputation (MI). Single imputation methods can range from simple mean imputation, in which the mean of observed values is used as a replacement for missing values, to regression imputation methods in which missing values are predicted from a regression function fitted to the observed data and imputed to missing values in the data matrix. The major difference between single and multiple imputation lies in the number of imputations that are made for any one missing value. In single imputation, one replacement value is created and imputed for each missing observation. In multiple imputation, several missing values are independently imputed, creating multiple (M) data sets and multiple estimates, one for each replication of the imputation process (Rubin, 1976; Little & Rubin, 2002). The multiple estimates are averaged over the M imputations to create a single MI estimate of the statistic of interest. To account for the uncertainty of the imputation process, MI variance estimators that incorporate both within- and between-imputation components are used to account for imputation-related variance in the overall variance of the estimate of interest.

1.1 Types of Missing Data
Data can be missing by design (e.g., only a subset of respondents are asked a subset of survey items), or through some other process related to respondents, interviewers or other survey features. The former is intentional nonresponse, and the latter is unintentional nonresponse. Another feature that defines the nonresponse structure of a study is whether missing data are the result of a random process or a non-random process. Little & Rubin (2002) classify missing data mechanisms as either missing completely at random (MCAR), missing at random (MAR), or not missing at random (NMAR). Under MCAR, data that are missing are not related to any other variable. Under a MAR mechanism, missingness is related to some observed variable(s) but, conditional on the observed data, not the missing values. Under a NMAR mechanism, missingness is related to the missing values (e.g., high values of body weight missing due to overweight individuals not being able to travel to the MEC).

1.2 Imputation Methods

Without employing an imputation method, an analyst is restricted to producing statistics from respondents alone. In multivariate models, this results in listwise (casewise) deletion of cases that may have observed values for some variables, but missing values for others. The analyzable case size is thus restricted to cases that have values for all variables of interest in the analytic model. If data are MCAR (or only MAR, depending on the kind of inference the analyst is employing) estimates will be unbiased. Standard errors will be larger than if there were no missing data due simply to reduced sample size. If missing data are related to the outcome of interest or other predictors of interest, resulting estimates of descriptive statistics or multivariate model coefficients may be biased, depending on the degree and nature of the relationship. Any imputation approach (including no imputation) carries with it an implicit model of the nature of missing data. Below is a brief description of several imputation methods.

Mean Imputation: The simplest way to replace missing data is to replace it with the mean of the observed values. This will, however, lead to an under-estimate of the element variance (s^2) of the imputed variable.

Subclass Mean Imputation: As means can differ significantly across certain subclasses (e.g., demographic categories, such as male and female), imputing the subclass mean for missing values of respondents in a given subclass produces a slightly more accurate imputation than overall mean imputation.

Hot-deck Imputation: Hot-deck imputation imputes values to missing items using observed values for "donor" individuals who are similar to the missing data case on some chosen covariates.

Regression Imputation: Regression imputation is accomplished by building a regression model predicting the observed values of the to-be-imputed variable from a set of observed predictors, and using this model and the observed covariates to predict the variable of interest when it is missing. A random residual can be added to the regression prediction, which introduces a stochastic component in addition to the deterministic values predicted by the regression line.

Multiple Imputation (MI): All methods discussed thus far have assumed that the analyst is imputing one value for each missing item. MI takes into account the notion that any imputed value will probably be wrong, since we can only really estimate a probability distribution for the missing variable. Thus, there is a distribution from which estimates of individual imputation values could be drawn, which produces variance associated with the selection of an individual imputation value. This is accounted for in MI by imputing (M) predicted values (where M defines the number of imputed values per case and thus the number of imputed data sets), and creating an estimate that averages across the M imputations. An imputation variance component is also incorporated in the estimated standard errors and confidence intervals for the sample estimate.

Table 1 outlines some of assumptions and qualities of the different types of imputation just presented.

2. The National Health and Nutrition Examination Survey (NHANES)

2.1 NHANES Structure

The NHANES consists of two phases of data collection. An area probability sample is used to select households and individuals within households. An interview is first conducted, and all interview participants are asked to participate in the medical exam, which takes place in a Mobile Exam Center (MEC). Within the MEC, respondents can refuse specific portions of the exam (e.g., blood draw). In this paper we deal with item nonresponse in interview and MEC measures and unit nonresponse at the MEC stage. We do not specifically evaluate nonresponse at the initial interview recruitment. For more information about the NHANES protocol, see http://www.cdc.gov/nchs/nhanes.htm.

We downloaded and merged the publically-available interview, lab and medical exam data files for the 2003-2004 NHANES sample. The total number cases in the merged file is 10122 from which we selected individuals age 20+ years (n=5041). This is our sample for the current paper. We chose respondents age 20+ for two reasons. First, we wanted to match previous NHANES imputation research as closely as possible (Schafer, Ezatti-Rice, Johnson, Khare, Little, and Rubin, 1996). Also several of the health items we wished to analyze were only asked of individuals age 20 or older.

Of the 5041 adults age 20 years or more, 4742 participated in the MEC (94.1%), and as a result 299 (5.9%) only have questionnaire data. The initial response rate to the NHANES household interview is 72.89% for an overall response rate (interview x MEC) of 68.59%.

NHANES involves a stratified cluster design. NHANES provides weights that are appropriate for analyzing 2-year data for the interview and MEC samples (WTINT2YR and WTMEC2YR respectively). These weights were used in all analyses. In the 2003-2004 data release there are 15 strata and 2 PSU's per strata for a total of 30 SECUS (sampling error calculation units).

Table 1: Comparison of Possible Imputation Techniques

Imputation Method	Missingness Assumption	Benefit	Drawback	Characteristic
No imputation	MCAR*	No extra steps	Casewise deletion	"Deterministic"
Simple overall mean	MCAR	No casewise deletion	Extra step, Attenuates variances and covariances	Deterministic
Subclass mean	MAR** (MCAR within subclass)	No casewise deletion, More "precise" than overall mean	Extra step, Attenuates variances and covariances, Categorical covariates only	Deterministic
Hot-deck	MAR (MCAR conditional on cells)	No casewise deletion	Extra step, Reduced element variance, Categorical covariates only	Stochastic (within range of available donor values) Some implementations are deterministic given a specific sort order
Regression	MAR	No casewise deletion, Allows for use of continuous covariates	Extra Step, Reduced element variance	Deterministic if no error term is included (i.e., predicted values only) Stochastic if residuals are used
Multiple imputation	MAR	No casewise deletion, Imputation variance accounted for	Most complex to implement	Stochastic

*Missing Completely at Random, **Missing at Random

2.2 NHANES Variables of Interest and Missing Data Rates

We selected a small set of NHANES variables that: 1) were intended to be measured for all cases age 20+, 2) represented health behaviors, medical conditions, or important physiological or blood chemistry measurements on NHANES sampled adults. Table 2 lists all the variables used in our imputations and analyses, and their item missing rates. Missing data rates are calculated conditional on interview response for items collected in the interview, and conditional on MEC response for items collected in the MEC. The highest rate was almost 8% for diastolic blood pressure. Other blood measures also had among the highest rates of item nonresponse (around 4-6%). Other self-reported health and demographic missing rates were much lower (2% or less). NHANES makes efforts to collect basic demographics from a respondent's neighbors when they are not obtainable from the respondent, which further reduces item nonresponse for these times.

Table 2: Variables Used in Our Analysis and Missing Data Rates

Interview Variables (N=5041)	Mean/ Prop'n	Miss	Exam Variables (N=4742) (%'s Assume MEC Part'n)	Mean/ Prop'n	Miss
Age (years)	46.31	0%	Standing Height (cm)	167.56	1.96%
Male	47.95%	0%	Body Weight at Exam (kg)	79.92	1.64%
African American	11.23%	0%	Systolic BP at Exam	125.63	7.19%
Mexican American	7.77%				
Married	63.25%	.06%	Diastolic BP at Exam	69.89	7.99%
Education HS or Less	45.26%	.28%	Total Cholesterol (mg/dL)	202.08	5.61%
Poverty Index	2.98	6.29%	Hemoglobin (g/dL)	14.33	4.47%
Self Rep of Heart Attack	5.45%	.22%	Hematocrit (%)	42.44	4.47%
Self Rep of Diabetes (incl. borderline)	8.92%	.06%	Iron, Refrigerated (ug/dL)	84.55	6.16%
Self Rep Height (in)	66.98	1.96%			
Self Rep Weight (lbs)	176.54	1.37%			

3. Methods

The goal of this project was to see how sensitive univariate and multivariate estimates would be to different imputation approaches. We decided to compare 4 imputation techniques reflecting a continuum of "data user sophistication" from no imputation to multiple imputation. There are almost infinite imputation options on this continuum. We felt that in addition to no imputation, the Hot-deck, and two multiple imputation approaches served as reasonable evaluation points because these approaches are commonly used or recommended by advocates of imputation methods. Our rationale was that a naïve user would simply do a complete case analysis. A more advanced use might attempt a hot-deck approach, and an even more sophisticated user would employ a multiple imputation approach.

3.1 Imputation

We employed three imputation methods and compared resulting estimates with each other and with an unimputed (complete case, listwise deletion) analysis. Hot-deck (Stiller & Dalzell, 1998), multiple imputation with IVEware (Raghunathan, Lepkowski, Van Hoewyk, & Solenberger, 2001), and multiple imputation with Mix (Schaefer, 1996; 1997) were used as imputation methods. Code for all the imputations and analyses are available upon request to the lead author.

For each method, we conducted two imputations. The first imputation replaces missing data so that the imputed data set matches the total sample size for the interview (n=5041). Essentially, it imputes for item nonresponse to interview and MEC data as well as unit nonresponse to the MEC. The second imputation replaces missing data so that imputed data set matches the total sample size for the MEC (n=4742). Analyses were run on each of these imputed data sets.

3.1.1 Hot-deck Imputation

For simplicity, clarity, and replicability, we used a previously published algorithm for hot-deck imputation in SAS (Stiller & Dalzell, 1998). We first programmed and implemented this algorithm in SAS and then in R. Results were identical. Our imputation cells were defined by a cross-classification of sex (2 levels), race (3 levels, see Table 2), and age (4 levels, 18-29, 30-44, 45-64, and older than 64).

The hot-deck imputation was performed by sorting the unimputed data file by SECUS (cluster and stratum combinations) within each imputation cell. Within each imputation cell, observed values were added to the hot-deck and used as imputation for missing cases in such a way that the donor and receiver tended to be close on the list (i.e.

came from adjacent SECUS). The algorithm described by Stiller and Dalzell limits the number of times each observation can be a donor to three, but this limit was never attained in our data set due to the small fraction of missing data.

The imputed data set was analyzed in PROC SURVEYMEANS and PROC SURVEYREG to account for the complex (stratified and clustered) design of the NHANES.

3.1.2 IVEware

IVEware (Raghunathan, et al., 2001; http://www.isr.umich.edu/src/smp/ive/) uses a multivariate sequential regression approach to impute missing values whereby rounds of sequential regression predictions are estimated. IVEware can handle as many covariates as the user supplies, and supports continuous, binary, categorical, count, and mixed (a continuous variable where 0 is a meaningful value) data types. In this exercise we used only the 18 variables included in Table 2 as covariates.

IVEware can also accommodate survey weights and cluster codes through its analytic marcos %DESCRIBE and %REGRESS (Raghunathan, et al., 2001). These macros account for survey design features in the same way as PROC SURVEYMEANS and PROC SURVEYREG.

We created and analyzed 5 imputed data sets in IVEware. The IVEware code was extremely simple to program, requiring only intermediate SAS knowledge. With the small number of covariates we used, imputations and analyses took only a few minutes to run on a moderate-performance notebook computer.

3.1.3 Mix

The R Mix routine takes a slightly different approach. It implements the general location model, described in Analysis of Incomplete Missing Data (Schafer, 1997, chapter 9), which accommodates incomplete data sets involving both categorical and continuous variables. It assumes a multinomial distribution for the cells defined by cross-classification of the categorical variables. The conditional distribution of the continuous data given the categorical part is then a multinormal distribution, with different mean vectors for each cell and a covariance matrix assumed constant for all cells.

Due to the limitation of a non-constrained general location model, it was not possible to include all 15 incomplete variables as well as age, gender and race. Attempts at fitting a constrained model failed, sometimes yielding only very cryptic error messages. Instead of dropping age, gender and race from the model, thus loosing very relevant information (and variables that are used in the other imputation methods), we considered all 15 incomplete variables as continuous (as recommended by Schafer, 1996), including the 4 dichotomous ones. Treating categorical variables as continuous for imputation purposes led to imputation of non-sensical values (e.g., 1.5 for sex). A report from the SAS User Group International (SUGI) argues that it is better to use unrounded values to compute means (Ake, 2005). Yet, for ease of interpretation we rounded the values after imputation.

One difficulty with using the Mix Package is that monitoring convergence is not straightforward. An analyst with little experience in this domain will probably not inquire carefully that convergence was achieved. With this in mind, we just assumed that convergence had occurred after 5000 iterations of the EM algorithm. Using this as the starting point for the MCMC, we created the 10 different imputed data sets by subsampling every 100 iterations. Running times for this analysis were similar to that with IVEware.

In Mix we created and analyzed 10 imputed data sets. The data imputed with Mix were analyzed using the %REGRESS and %DESCRIBE macros that are part of the IVEware SAS package.

3.2 Models

We analyzed the impact of imputation method on univariate estimates (means and proportions) and one linear regression model that contained continuous and categorical predictors from the interview and MEC. Specifically we regressed Systolic Blood Pressure on Age, Age^2, Sex, Race, Standing Height, Weight, Cholesterol, and Marital Status. For the regression analyses, the sample size is 5041 in each case. That is, we only report regressions in which we imputed for all individuals who participated in the interview stage of the NHANES. Regressions with only MEC cases show similar results comparatively across imputation methods but are not presented here.

4. Results

4.1 Results across Imputation Methods

Table 3 presents differences in univariate estimates across imputation methods. Columns titled "Interview" have an imputed case base of 5041, and columns titled "MEC" have an imputed case base of 4742.

We looked at univariate estimates of several key health and biomedical measures (Table 3). We also proposed and fit a plausible regression model in which Systolic Blood Pressure is predicted by Age, Sex, Race, Height, Cholesterol and Marital Status (see Table 4). We looked at unimputed and imputed estimates of univariate statistics and regression coefficients with an eye toward differences in substantive interpretation of the findings associated with the imputation method.

The most striking result is that there is very little difference in estimates across imputation methods. In no case do the substantive interpretations differ among the imputation methods, including the complete-case analysis.

4.1.1 Effect on Univariate Estimates

In univariate estimates (Table 3), differences across imputation methods tend to be in the first and second decimal place, and standard errors (in parentheses under means/proportions) seem to be only mildly affected by imputation method.

4.1.2 Effect on Regression Coefficients

We estimated a regression model which includes self-reported and MEC lab data, and for which we found a reasonable linear fit. In our model, Age^2, Sex (Female), Race (African American), Standing Height, Weight, and Cholesterol were all significantly associated with systolic blood pressure. The same substantive finding can be seen in each imputation approach (including no imputation). R^2 values varied slightly across approaches (.262 unimputed to .271 in IVEware), but would not likely lead an analyst to different decisions about model fit.

The most notable difference across imputation methods is in significance of individual coefficients (e.g., cholesterol is significant at the .005 level in the hot-deck imputation, but only the .05 level in all others). In no case does the variation in significance across imputation change substantive interpretation of significant coefficients at the .05 level (i.e., the same set of coefficients are significant at the .05 level or lower across all imputation models).

Table 3: Univariate Estimates (Interview Sample and MEC-only) under Different Imputation Methods

No Imputation	Interview	MEC
Married	63.25% (1.567)	63.73% (1.354)
Education HS or Less	45.26% (1.355)	45.52% (1.298)
Poverty Index	2.98 (.0747)	2.96 (.0756)
Self Report of Heart Attack	3.93% (.434)	4.03% (.432)
Self Report of Diabetes (incl. borderline)	8.92% (.705)	9.03% (.738)
Self Reported Height (in)	66.98 (.0895)	66.99 (.0989)
Self Reported Weight (lbs)	176.54 (.975)	176.92 (.993)
Standing Height (cm)	169.15 (.241)	169.11 (.241)
Body Weight at Exam (kg)	81.03 (.439)	80.96 (.433)
Systolic BP at Exam	122.57 (.531)	122.70 (.530)
Diastolic BP at Exam	71.26 (.349)	71.23 (.348)
Total Cholesterol (mg/dL)	201.69 (.728)	201.73 (.721)
Hemoglobin (g/dL)	14.55 (.0676)	14.55 (.0667)
Hematocrit (%)	42.99 (.181)	42.97 (.180)
Iron (ug/dL)	86.75 (.723)	86.72 (.715)

Hotdeck	Interview	MEC
Married	63.30% (1.554)	63.77% (1.340)
Education HS or Less	45.32% (1.367)	45.55% (1.304)
Poverty Index	2.96 (.0742)	2.95 (.0762)
Self Report of Heart Attack	3.95% (.44)	4.03% (.432)
Self Report of Diabetes (incl. borderline)	8.91% (.705)	9.03% (.738)
Self Reported Height (in)	66.96 (.0875)	66.97 (.0955)
Self Reported Weight (lbs)	176.39 (.994)	176.81 (1.02)
Standing Height (cm)	169.07 (.240)	169.09 (.248)
Body Weight at Exam (kg)	81.01 (.433)	80.96 (.448)
Systolic BP at Exam	122.85 (.535)	122.88 (.504)
Diastolic BP at Exam	71.12 (.353)	71.08 (.346)
Total Cholesterol (mg/dL)	201.84 (.746)	201.67 (.775)
Hemoglobin (g/dL)	14.55 (.0637)	14.54 (.0646)
Hematocrit (%)	42.97 (.169)	42.95 (.172)
Iron (ug/dL)	86.42 (.685)	86.60 (.723)

Mix R Package	Interview	MEC
Married	63.25% (1.567)	63.73% (1.353)
Education HS or Less	45.29% (1.356)	45.53% (1.296)
Poverty Index	2.96 (.071)	2.95 (.073)
Self Report of Heart Attack	3.93% (.434)	4.03% (.431)
Self Report of Diabetes (incl. borderline)	8.92% (.705)	9.04% (.738)
Self Reported Height (in)	66.94 (.0894)	66.95 (.0988)
Self Reported Weight (lbs)	176.68 (1.01)	177.08 (1.03)
Standing Height (cm)	169.04 (.232)	169.07 (.252)
Body Weight at Exam (kg)	80.95 (.448)	80.75 (.460)
Systolic BP at Exam	122.82 (.522)	122.81 (.528)
Diastolic BP at Exam	71.10 (.330)	71.16 (.346)
Total Cholesterol (mg/dL)	201.80 (.748)	201.81 (.714)
Hemoglobin (g/dL)	14.54 (.0651)	14.54 (.0659)
Hematocrit (%)	42.94 (.174)	42.95 (.177)
Iron (ug/dL)	86.63 (.746)	86.62 (.705)

IVEware	Interview	MEC
Married	63.27% (1.561)	63.74% (1.351)
Education HS or Less	45.29% (1.355)	45.55% (1.299)
Poverty Index	2.97 (.0714)	2.95 (.0726)
Self Report of Heart Attack	3.94% (.437)	4.04% (.432)
Self Report of Diabetes (incl. borderline)	8.92% (0.705)	9.03% (.738)
Self Reported Height (in)	66.94 (.0892)	66.96 (.0994)
Self Reported Weight (lbs)	176.66 (1.009)	177.09 (1.03)
Standing Height (cm)	169.05 (0.234)	169.06 (.251)
Body Weight at Exam (kg)	80.76 (0.449)	80.95 (.456)
Systolic BP at Exam	122.85 (0.531)	122.82 (.508)
Diastolic BP at Exam	71.07 (0.328)	71.11 (.337)
Total Cholesterol (mg/dL)	201.68 (0.745)	201.70 (.778)
Hemoglobin (g/dL)	14.54 (0.064)	14.54 (.0665)
Hematocrit (%)	42.95 (0.1713)	42.95 (.177)
Iron (ug/dL)	86.69 (0.729)	86.70 (.700)

Table 4: Regression Estimates Predicting Systolic Blood Pressure under Different Imputation Methods

No Imputation (R^2 .262)	Estimate	*SE*
Intercept	141.09**	*5.53*
Age	.0832	*.1001*
Age^2	.00412**	*.00101*
Female	-5.24**	*.691*
African American	3.94**	*.937*
Mexican American	-.962	*1.11*
Standing Height (MEC)	-.272**	*.0294*
Weight (MEC)	.143**	*.0120*
Cholesterol	.0212*	*.0087*
Not Married	.802	*.575*

Hot Deck (R^2 .268)	Estimate	*SE*
Intercept	137.96**	*5.37*
Age	.124	*.0953*
Age^2	.00391**	*.000914*
Female	-5.29**	*.590*
African American	4.08**	*.748*
Mexican American	-.0148	*1.26*
Standing Height (MEC)	-.253**	*.0282*
Weight (MEC)	.121**	*.0125*
Cholesterol	.0229**	*.00603*
Not Married	.924	*.637*

Mix R Package (R^2 .264)	Estimate	*SE*
Intercept	139.72**	*5.52*
Age	.175	*.0985*
Age^2	.00329**	*.000964*
Female	-5.35**	*.663*
African American	4.63**	*.853*
Mexican American	-.524	*1.18*
Standing Height (MEC)	-.271**	*.0304*
Weight (MEC)	.13**	*.0128*
Cholesterol	.0215*	*.0076*
Not Married	.864	*.597*

IVEware (R^2 .271)	Estimate	*SE*
Intercept	138.62**	*6.87*
Age	.166	*.0924*
Age^2	.00345*	*.000883*
Female	-4.99**	*.708*
African American	4.53**	*.801*
Mexican American	-.386	*1.10*
Standing Height (MEC)	-.267**	*.0362*
Weight (MEC)	.13**	*.0167*
Cholesterol	.0228*	*.00822*
Not Married	1.060	*.595*

5. Conclusions

For the variables we analyzed from the 2003-2004 NHANES data set, our four different approaches to missing data yield very similar results. In particular, substantive interpretations of our regression models and univariate estimates did not change across imputation methods, including no imputation, single imputation and multiple imputation. This suggests that for some data sets and certain missing data problems, the choice of treatment for incomplete data may not warrant much concern. The NHANES item missing data rates in our example were, however, relatively low so we cannot extend this conclusion to data sets with higher rates of missing data.

Since we do not know the true values of the parameters that we estimated, it is hard to evaluate which of the four methods was the best for our data set. Any imputation method would probably be preferable to complete-case analysis, as the benefits of having more analyzable cases may outweigh any minor changes in estimates. Further, some imputation methods were easier to implement than others. In our experience, imputation with IVEware was the easiest, except of course for complete-case analysis. The Hot-Deck imputation method was probably the next easiest to implement since it is based on very simple statistical arguments, and partial SAS code was provided by Stiller and Dalzell (1998). The Mix package for R implements a much more complicated statistical model, which in our experience can not be estimated easily for large data sets. It is up to the analyst to specify appropriate alternative models and monitor the convergence of the algorithm.

In light of these results, we believe that statistical "exactness" in missing data problems should be balanced against available resources and statistical expertise of the research team, as well as the potential benefit to estimates of interest.

Acknowledgements

We would like to thank the staff of the NHANES and NCHS for their diligent documentation of this complex survey, and for the development and maintenance of its clear and comprehensive online access pages. This specific analysis would not have been possible without NHANES data, and the quality of the online documentation and tutorials made the use of these data very easy. We would also like to thank the Summer Institute in Survey Research Techniques for making it possible for Ms. Charest to visit Ann Arbor for the summer and work on this project.

References

Ake, C. F. (April, 2005). Rounding after multiple imputation with non-binary categorical covariates. *SAS Conference Proceedings: SAS Users Group International 30*, Philadelphia, PA. Accessed at http://www.lexjansen.com/cgi-bin/xsl_transform.php?x=sugi30&s=sugi&c=sugi October 27, 2008.

Little, R. J. A. and Rubin, D. B. (2002). Statistical analysis with missing data, 2nd edition, New York: John Wiley

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, *27,* 1, 85-95.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581-592.

Schafer, J. L. (1996) *MIX: Multiple imputation for mixed continuous and categorical data*, software library for S-PLUS. Written in S-PLUS and Fortran-77.

Schafer, J. L. (1997) *Analysis of Incomplete Multivariate Data*. Chapter 9 Methods for Mixed Data, 333-347. London: Chapman & Hall.

Schafer, J. L., Ezatti-Rice, T.M., Johnson, W., Khare, M., Little, R.J.A., and Rubin, D.B.
(1996), The NHANES III multiple imputation project. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 696–701.

Stiller, J. G. and Dalzell, D. R. (1998). Hot-deck imputation with SAS arrays and macros for large surveys. *Proceedings of the Twenty-Third Annual AS Users Group International Conference*, 1378-1383.