DIGITALCOMMONS
— @WAYNESTATE —

Journal of Modern Applied Statistical Methods

Volume 17 | Issue 1 Article 23

9-5-2018

Comparison of Multiple Imputation Methods for Categorical Survey Items with High Missing Rates: Application to the Family Life, Activity, Sun, Health and Eating (FLASHE) Study

Benmei Liu
National Cancer Institute, liub2@mail.nih.gov

Erin Hennessy
Tufts University, erin.hennessy@tufts.edu

April Oh
National Cancer Institute, ohay@mail.nih.gov

Laura A. Dwyer
National Cancer Institute, laura.dwyer@nih.gov

Linda Nebeling
National Cancer Institute, nebelinl@mail.nih.gov

Recommended Citation

Liu, Benmei; Hennessy, Erin; Oh, April; Dwyer, Laura A.; and Nebeling, Linda (2018) "Comparison of Multiple Imputation Methods for Categorical Survey Items with High Missing Rates: Application to the Family Life, Activity, Sun, Health and Eating (FLASHE) Study," *Journal of Modern Applied Statistical Methods*: Vol. 17 : Iss. 1 , Article 23.
DOI: 10.22237/jmasm/1536146540
Available at: https://digitalcommons.wayne.edu/jmasm/vol17/iss1/23

Comparison of Multiple Imputation Methods for Categorical Survey Items with High Missing Rates: Application to the Family Life, Activity, Sun, Health and Eating (FLASHE) Study

Cover Page Footnote

The FLASHE study was funded by contract from the National Cancer Institute (NCI) to Westat, Inc., HHSN261201200039I. This project has been funded in whole or in part with federal funds from the NCI, National Institutes of Health, under Contract No. HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Journal of Modern Applied Statistical Methods
May 2018, Vol. 17, No. 1, eP2511
doi: 10.22237/jmasm/1536146540

בס"ד
Copyright © 2018 JMASM, Inc.
ISSN 1538 − 9472

Comparison of Multiple Imputation Methods for Categorical Survey Items with High Missing Rates: Application to the Family Life, Activity, Sun, Health and Eating (FLASHE) Study

Benmei Liu
National Cancer Institute
Rockville, MD

Erin Hennessy
Tufts University
Medford, MA

April Oh
National Cancer Institute
Rockville, MD

Laura A. Dwyer
National Cancer Institute
Rockville, MD

Linda Nebeling
National Cancer Institute
Rockville, MD

Two multiple imputation methods, the Sequential Regression Multivariate Imputation Algorithm and the Cox-Lannacchione Weighted Sequential Hotdeck, were examined and compared to impute highly missing categorical variables from the Family Life, Activity, Sun, Health and Eating (FLASHE) study. This paper describes the imputation approaches and results from the study.

Keywords: Perceptional categorical data, high missing rates, multiple imputation

Introduction

The Family Life, Activity, Sun, Health and Eating (FLASHE) study, sponsored by the National Cancer Institute (NCI), examines psychosocial, generational (parent-adolescent), and environmental correlates of cancer-preventive behaviors. The objective of this web-based survey was to examine cancer preventive lifestyle behaviors, mainly diet and physical activity, as well as sleep, sun-safety, and tobacco use. Data were collected in 2014 from dyads of caregivers and their adolescent children aged 12-17. After the data collection period ended, eight

doi: 10.22237/jmasm/1536146540 | Accepted: August 14, 2017; Published: September 5, 2018.
Correspondence: Benmei Liu, liub2@mail.nih.gov

variables related to physical limitations to physical activity and life goals about teens were found to be missing data from approximately half of the sample for the parent physical activity survey because of a system programming error in the web-based data collection tool. This missing scenario is rare but not unique. A considerable fraction of the accelerometer steps data from the 2003-2004 National Health and Nutrition Examination Survey (NHANES) were missing due to a device initialization error and had to be imputed (Liu, Yu, Graubard, Troiano, & Schenker, 2016).

When the missing data are not missing completely at random (MCAR), i.e., the missingness is not independent from the characteristics of the individuals surveyed, analyzing only the cases with non-missing data (called complete-case analyses) is known to produce biased estimates and also leads to reduced efficiency for many situations, especially when drawing inferences for subpopulations (Little & Rubin, 2002). Imputation is a flexible and commonly-used technique for handling missing-data problems. Both single imputation and multiple imputations have been commonly used in survey practice. Single imputation is used to fill in only one value for each missing value and then treat the imputed values as if they were true values in post-imputation analyses. Rubin (1987) described two attractive features of single imputation: First, standard complete-data methods of analysis can be used on the imputed data set. Second, in the context of public-use databases, the possibly substantial effort required to create sensible imputations needs to be carried out only once, by the data producer, and these imputations can incorporate the data collector’s knowledge.

Rubin (1987) noted one major disadvantage of single imputation is that the single value being imputed does not reflect either the sampling variability about the actual value when one model for nonresponse is being considered, or additional uncertainty when more than one model is involved in the imputation procedure. Multiple imputation, however, repeats the same imputation mechanism multiple times and creates multiple sets of imputed values, say *M* sets. These multiple values are used to empirically estimate both the variability from the sampling and imputation model.

Multiple imputation retains the virtues of single imputation and provides correct variance estimation. The superiority of multiple imputation over single imputation is magnified when the amount of missing data is large. With multiply imputed data, data users just need to analyze each completed data set one by one, and then combine the *M* analyses via simple formulas (Rubin, 1987). Many statistical packages contain routines for creating and/or analyzing multiply imputed data sets under selected models (Harel & Zhou, 2007). Given the high missing rate

(53%) for the eight variables, we chose to use multiple imputation to capture the additional variability due to imputation.

The sequential regression multivariate imputation (SRMI) algorithm (Raghunathan, Lepkowski, Van Hoewyk, & Solenberger, 2001), also called multiple imputation by chained equations (MICE) or fully conditional specification, was widely used in the literature to impute missing continuous or categorical survey items (e.g., Schenker, Raghunathan, et al., 2006; Schenker, Borrud, et al., 2010; Stuart, Azur, Frangakis, & Leaf, 2009). The SRMI approach uses a sequence of regression models and a Gibbs sampling style iterative algorithm to impute all variables with missing values in a sequential order. The algorithm has been implemented in several statistical software programs including IVEware, SAS, R, and Stata. Zhu and Raghunathan (2015) provided a detailed review on the SRMI approach.

The SRMI approach has two major practical advantages over other model-based imputation methods including: 1) It enables handling of complex data structures by focusing a set of regression models with a univariate outcome; and 2) The flexible selection of regression models enables better prediction of the missing values based on other variables. A theoretical limitation of this approach is that the specifications of conditional distributions for a set of variables do not guarantee the existence of a joint distribution. Therefore, it is not clear whether the iterative algorithm will achieve any stability. The convergence results established for the standard Gibbs sampling algorithms or its variations may not be applicable.

Another commonly used imputation approach for handling missing data is called hot deck imputation. Hot deck imputation replaces missing values of one or more variables for a nonrespondent (called the recipient) with observed values from a respondent (the donor) that is similar to the non-respondent with respect to characteristics observed for both cases (Little & Rubin, 2002; Andridge & Little, 2010). Andridge and Little (2010) reviewed different forms of the hot deck and existing research on its statistical properties. Among those, the Cox-Lannaccione Weighted Sequential Hot Deck (WSHD) (Cox, 1980; Cox & Folsom, 1981) was motivated by two issues: the unweighted sequential hot deck is potentially biased if the weights are related to the imputed variable, and respondent values can be used several times as donors if the sorting of the file results in multiple non-respondents occurring in a row, leading to estimates with excessive variance.

Implemented in SUDAAN version 10+, the WSHD provided another practical tool to multiply impute missing survey data. In the 2002 National Survey on Drug Use and Health (NSDUH) survey it was used sparingly. Grau, Frechtel, Odom, and Painter (2004) compared the WSHD approach with the Predictive Mean

Neighborhoods (PMN) procedure through a simple simulation and found significant but not substantial difference between the two methods. Through simulation Andridge and Little (2009) found the WSHD approach doesn't correct for bias when the outcome is related to the sampling weight and the response propensity. This condition doesn't apply to the FLASHE data.

Therefore, the purpose of this study is to consider the WSHD approach as the alternative approach to SRMI. There are few published, empirical studies with comparisons of the performance of the two algorithms.

Methodology

The FLASHE Study Sample and Missing Data

Parents with adolescent children between the ages of 12 and 17 years were recruited from the Ipsos Consumer Opinion Panel (Ipos) which includes over 700,000 active members. Balancing and quota sampling techniques were used (Lohr, 1999). The sample was intended to match the U.S. population on several key demographic characteristics as closely as possible. A screening instrument based on the FLASHE eligibility criteria was administered via the web to determine the panel member's eligibility for FLASHE. A panel member was deemed eligible for FLASHE if they: were at least 18 years of age; lived with at least one child between the ages of 12 and 17.5 for at least 50% of the time; and agreed to be contacted for participation in FLASHE. During the screening process, information on the eligible adolescents in the household was collected via a full household roster and one eligible adolescent was randomly selected until the quota for that age range was full.

Ipsos intended to provide a balanced sample of 4,500 eligible parents using only their panel. However, the size of the Ipsos panel unexpectedly did not provide an adequate number of respondents for the male adults and African-American adults in the balancing process. Ipsos therefore subsequently requested additional samples from four other panel companies: Global Marketing Insite, Inc; ROI Rocket; Clearvoice Research; Toluna. After an initial delivery of 4,527 eligible dyads (denoted as the main dyad sample, including the recruitment goal of 4,500 dyads plus additional dyads to allow for potential cases which might be unusable), an additional 500 eligible dyads with male parents were provided to improve the parent gender balance of the FLASHE sample, though the recruitment from the additional panels still did not meet the recruitment goal for African-American adults.

The main dyad sample and the additional male dyads were collected and delivered separately. FLASHE data collection materials and procedures were reviewed and approved by the U.S. Government's Office of Management and Budget (OMB), NCI's Special Studies Institutional Review Board (SSIRB), and Westat's Institutional Review Board (IRB). Detailed information on the FLASHE study design and recruitment was given elsewhere (Oh et al., 2017; Westat, 2015).

A total of 5,027 eligible dyads were invited to enroll into FLASHE in April of 2014. Among them, 1,945 dyads were fully enrolled (both parent and adolescent provided consent/assent, enrollment rate = 38.7%). Four surveys on physical activity-related behaviors and diet-related behaviors were administered to the dyads via the web with a cash incentive mailed to each participant upon completion: Parent Physical Activity Survey, Parent Diet Survey, Adolescent Physical Activity Survey, and Adolescent Diet Survey. The final response rates (RRs) out of the number of dyads invited to participate varied by survey. A participant was deemed a respondent to a specific survey if at least 80% of the questions were answered. Having an 80% threshold allowed for minor skips of questions. The final numbers of respondents were 1,802 (RR = 35.8%) for the Parent Physical Activity survey (which contained the eight items with a high percent of missing data), 1,754 for the Parent Diet Survey (RR = 34.9%), 1,670 for the Adolescent Physical Activity Survey (RR = 33.2%), and 1,667 for the Adolescent Diet Survey (RR = 33.2%). Participants received an incentive if they clicked "submit" on the survey, regardless of how many questions they skipped.

Prior to enrollment, a random half of the main dyad sample were selected to receive the Diet survey first and the other half were selected to receive the Physical Activity Survey first. Specifically, a random number was generated from a uniform distribution with a range of 0 to 1 for each dyad. Then, the dyads with a value lower than 0.5 was assigned to one group and the other half of the dyads was assigned to the other group. In addition, a random subsample (n = 1,690) of the main dyad sample was also invited to participate in a Motion study during which adolescents wore an accelerometer to assess physical activity, and among those 693 dyads fully enrolled into the motion study. Given the late delivery of the sampling frames and the complicated scheduling due to the inclusion of the motion study component, the additional all-male (n = 500) sample was assigned to the Diet survey first group for simplicity.

Table 1. Variables to be imputed

Variable name	Survey question
PPFEELLOVE	When my teenager is an adult, he/she will feel that there are people who really love him/her
PPOTHBETTER	The things my teenager will do as an adult will make other people's lives better
PPGETGDGRAD	My teenager will get good grades in school
PPATTRACTV	People will often comment about how attractive my teenager looks as an adult
PPJOBPAYWL	When my teenager is an adult, he/she will have a job that pays well
PPHCPALIMIT	Has a doctor or other healthcare professional ever told you that teen has any condition that could limit his/her ability to exercise, such as obesity, asthma, diabetes, high blood pressure, etc.
PPHCPASPORT	Do medical, behavioral or other health conditions interfere with teen's ability to participate in sports, clubs or other organized physical activities
PPHCPAOUT	Do medical, behavioral or other health conditions interfere with teen's ability to go on things such as the park, library, zoo, shopping, church, restaurants or family gatherings

Among the 1,802 final respondents in the parent physical activity survey, 951 respondents (53%) had eight variables all missing due to a system error. This missingness occurred only in the group of parents that had received the physical activity survey second, after completing the diet survey (Westat, 2015). Some of those with missing data completed the physical activity survey on an earlier date than some of the respondents who were assigned the physical activity survey first, just based on how fast people responded to their sets of surveys. Even if the physical activity survey was “first” for a given participant, he or she might have waited some time to complete it. The identified system error did not enable those parents to access the eight questions. Twenty-five respondents had one or more, but not all, of the eight variables missing. The remaining 826 respondents did not have any missing data for the eight variables in question. The variable names and corresponding questions are listed in Table 1.

A five-point Likert Scale was used for the first five variables in Table 1, which focused on parent-reported life goals for their adolescent child (PPFEELLOVE, PPOTHBETTER, PPGETGDGRAD, PPATTRACTV, PPJOBPAYWL): Not at all important to me (1), A little important to me (2), Somewhat important to me (3), Very important to me (4), Extremely important to me (5). Yes (1) or No (2) choices were used for the last three variables, which focused on physical limitations of the adolescent (PPHCPALIMIT, PPHCPASPORT, PPHCPAOUT).

Statistical Analysis

To determine if the missing data were MCAR, a series of cross-tabs and chi-square tests of the missing-skip (missing or not for all of the eight variables) and 18 parent socio-demographics were conducted. The socio-demographics included parent age, gender, education, marital status, health status, consistency of health insurance coverage, race/ethnicity, nativity, home ownership, housing security, work status, household income, language usually spoken at home, language used for media, health literacy, number of kids in home, BMI, and adolescent health insurance coverage. Those variables are either binary or categorical. The definition of each variable and associated categories are given in Table A1 in the appendix.

Because all eight variables with missing values were categorical or binary data, both SRMI and the WSHD multiple imputation approaches were considered to impute the missing data. The SRMI approach was implemented using IVEware (http://www.src.isr.umich.edu/software/). The specific imputation models are multiple linear regressions for continuous variables, logistic regressions for binary variables, and polynomial regressions for categorical variables. To create multiple imputations, it is recommended to include a large number of predictors in the imputation model, especially variables that will be used in subsequent analyses of the multiply imputed data, for congenial purpose. That is, to be accurate, the imputation model should be congenial with the analysis model. The two models don't have to be identical, but they cannot have major inconsistencies (Meng, 1994). Hence, for the SRMI approach, all of the 18 parent socio-demographic variables shown in Table A1 were included as the predictors. The module IMPUTE in the software package IVEware was implemented to simultaneously impute any missing values in the eight target variables and in the predictor variables. All the variables involved in the imputation were specified as categorical variables so logistic regression models were automatically picked for imputing binary variables and polynomial regressions were automatically picked for imputing variables with more than two categories.

The WSHD approach was implemented using PROC HOTDECK in SUDAAN. It requires defining a set of categorical variables that determine the imputation classes. It is advantageous to select classes of variables with a strong association with the imputation variables. Imputation is performed within each of the classes where both missing data and donor data are found. The objective was to find the best imputation model for each of the eight variables respectively. When too many predictors were included for the imputation model the software ran out of donors in one or more imputation cells, because each donor can only be used limited times depending on the sample weights. Thus, it failed to impute all the missing values for a targeted variable. To overcome this, the following stepwise procedures were used:

Step 1: Run chi-square tests of the eight variables and all the 18 parent socio-demographic characteristics and then pick the significant predictors (*p*-value < 0.05) to be included in the initial imputation model for each variable.

Step 2: Run each imputation model. If the model couldn't run properly due to too many predictors, remove the least significant predictor from the model and rerun the imputation. Repeat step 2 until each imputation model runs successfully.

All the remaining socio-demographic variables with *p*-values < 0.1 from the chi-square tests in step 1 were included in the ICSORT statement to allow for greater control over the sort order of observations within imputation classes. With WSHD, the assignment of a selection probability to a potential donor, or item respondent, depends both on the donor's weight and on the weights of nearby item nonrespondents. In other words, both the weights and the sort order of observations within an imputation class play a role in the selection of donors for imputation in the hot deck algorithm. Reordering item respondents and nonrespondents within an imputation class can yield different imputation results (Research Triangle Institute, 2012). Provided in Table 2 are the predictors included in the final WSHD imputation model for each variable to be imputed.

To further evaluate the two imputation methods and decide on the final imputation approach, a simulation study using the 826 respondents with observed data was conducted. Before imputation, there were 826 respondents with none of the eight variables missing, 951 respondents with all eight variables missing, and the remaining 25 respondents had one or more but not all of the eight variables missing. The 25 respondents were excluded from this evaluation study because the missingness was not caused by the system error. The missing rates for the eight variables by gender were calculated from the remaining sample containing the 826 respondents without, and the 951 respondents with the eight variables missing (*n* = 1,777). Among males, 64.7% had the eight variables missing due to the system error. Among females, 48.3% had the eight variables missing due to the same system error.

Table 2. Predictors included in the final WSHD imputation model

Variable name	Predictors included
PPFEELLOVE	Parent gender; Adolescent health coverage; Consistency of parent health coverage
PPOTHBETTER	Number of kids living in home; Parent health status; Home ownership
PPGETGDGRAD	Parent race/ethnicity; Adolescent health coverage
PPATTRACTV	Parent race/ethnicity; Parent BMI; Parent language used for media
PPJOBPAYWL	Parent race/ethnicity; Home ownership
PPHCPALIMIT	Housing security; Parent BMI
PPHCPASPORT	Parent health literacy; Housing security; Parent BMI
PPHCPAOUT	Parent health literacy; Housing security

The evaluation study sample was based on the 826 respondents (161 males and 665 females) with none of the eight variables missing. One hundred and four respondents were randomly chosen from the 161 males (64.7%) and 321 respondents from the 665 females (48.3%) and then set their values for the eight variables to be missing, to mimic the same missing pattern as the original data. Thus, in the simulated data, 425 persons had the eight variables all missing and 401 persons had none of the eight variables missing.

This simulation experiment was repeated 100 times by randomly resampling 104 males and 321 females from the 826 respondents and setting their values for the eight variables to missing. Each time different samples may be selected, thus the final 100 simulated data sets were different by simulation. The multiple imputation procedures in consideration were then applied to each of the 100 datasets to impute the missing values. For a fairer comparison between SRMI and WSHD, in this evaluation study, an alternative SRMI model was added by including the same set of predictors as in the WSHD method for each variable to be imputed. The alternative SRMI approach was denoted as SRMI2. Relative biases of point estimates and coverage of confidence intervals for the target quantities of interest based on the imputed data were then obtained because the observed values for the 425 persons are known.

Results

MCAR Assumption Tests

The chi-square tests of the missing-skip (missing or not for all of the eight variables) and the 18 parent socio-demographics (data not shown) showed that the missing-skip was independent ($p > 0.05$) from all the demographic variables except parent gender and work status, indicating that the missing scenario does not belong to MCAR, but may depend on parent gender and work status. This finding about parent gender is likely supported by the fact that the additional male sample ($n = 500$ males) was assigned to take the Physical Activity survey second and therefore did not have the opportunity to respond to the eight questions due to the system error.

The significant association between the missing-skip and parent work status may be due to its significant correlation with gender. To verify this, another chi-square test was conducted between the missing-skip and parent work status stratified by gender, which confirmed the missing-skip and parent work status were significantly associated only for males, but not for females.

Multiple Imputation Results for the Full FLASHE Sample (*n* = 1,802)

Recent research (e.g., Graham, Olchowski, & Gilreath, 2007) suggested the use of greater than the traditional number of five or fewer sets of imputed data, especially if the fractions of missing information for various analyses are high. After doing some sensitivity analyses based on 10, 20, and 50 sets of multiply imputed data (results not shown), it was decided to create 20 sets of multiply imputed data using each of the two multiple imputation methods (SRMI and WSHD).

Let θ denote the percentage of people that fall into one given category of a categorical variable (one of the eight variables, e.g., PPFEELLOVE = 4). Let θ_i and U_i denote the weighted percentage and associated variance computed from the i^{th} multiply imputed data, $i = 1,\ldots, M$. Then the point estimate for θ from the multiple imputations is the average of the M complete-data estimates:

$$\bar{\theta} = \frac{1}{M}\sum_{i=1}^{M}\theta_i \qquad (1)$$

Let $\bar{U}$ be the within imputation variance for the estimate, which is the average of the M complete-data estimates:

$$\bar{U} = \frac{1}{M}\sum_{i=1}^{M}U_i$$

Let B be the between-imputation variance:

$$B = \frac{1}{M-1}\sum_{i=1}^{M}\left(\theta_i - \bar{\theta}\right)^2$$

Then, the variance associated with $\bar{\theta}$ is the total variance:

$$T = \bar{U} + \left(1 + \frac{1}{M}\right) B \qquad (2)$$

The standard error of $\bar{\theta}$ is $\sqrt{T}$, and the 95% confidence interval bounds for $\bar{\theta}$ is

$$\bar{\theta} \pm 1.96 * \sqrt{T} \qquad (3)$$

With the same number of multiply imputed sets, to compare the performance of the two multiple imputation methods with the complete-case analyses, the averages of weighted percentage estimates along with their standard errors for each category of the eight variables computed using formulas (1) and (2), as well as the following standard measures for the multiple imputation methods are reported (Table 3). The number of missing respondents differed a little by variable due to the 25 respondents who had 1 to 7 variables missing at random and not because of the system error.

- Relative increase in variance due to imputation: $RIV = \left(1 + \frac{1}{M}\right) B / \bar{U}$
- Fraction of missing information: $FMI \approx \frac{RIV + 2}{(v_M + 3)(RIV + 1)}$, where v_M is adjusted degrees of freedom in multiple imputation variance.
- Relative efficiency of using finite M imputations: $RE = \left(1 + \frac{FMI}{M}\right)^{-1}$.

With 20 multiply imputed sets, the WSHD approach resulted in lower RIV and FMI, and higher RE than the SRMI approach did, meaning better performance for WSHD compared to SRMI with the same number of multiply imputed data. The post-imputation standard errors of the WSHD percentages are generally smaller than the complete-case standard errors indicating the achievement of efficiency using multiple imputations. Unexpectedly the post-imputation standard errors based on SRMI are larger than those of the complete-case results. The larger standard errors from the SRMI approach may indicate poor fitting of the sequential models or a joint distribution of the variables may not exist and thus stability was not achieved. To investigate this in the evaluation study, as we mentioned earlier, we added an alternative SRMI model by using the same set of variables as in the WSHD method which was denoted as SRMI2.

Table 3. Before and after imputation using two multiple imputation methods for the full data (*n* = 1,802)*

	Original data			**SRMI† (multiple = 20)**					**WSHD‡ (multiple = 20)**				
Variable	***n***	**Pct**	**SE**	**Pct**	**SE**	**RIV**	**FMI**	**RE**	**Pct**	**SE**	**RIV**	**FMI**	**RE**
PPFEELLOVE	miss = 953												
1, 2, 3	31	3.80	0.80	4.70	1.50	5.07	0.85	0.96	3.80	0.70	0.38	0.28	0.99
4	188	23.20	1.80	24.60	2.00	1.47	0.61	0.97	23.60	1.40	0.22	0.19	0.99
5	630	73.00	1.90	70.70	2.50	2.42	0.72	0.97	72.60	1.50	0.26	0.21	0.99
PPOTHBETTER	miss = 955												
1, 2, 3	127	14.00	1.40	15.00	1.70	1.76	0.65	0.97	14.00	1.10	0.17	0.15	0.99
4	322	38.30	2.00	38.50	2.20	1.56	0.62	0.97	38.50	1.60	0.25	0.20	0.99
5	398	47.70	2.10	46.60	2.30	1.58	0.63	0.97	47.50	1.60	0.21	0.18	0.99
PPGETGDGRAD	miss = 953												
1, 2, 3	101	11.50	1.30	12.70	1.90	2.81	0.75	0.96	11.80	0.90	0.10	0.09	1.00
4	279	32.50	1.90	33.70	2.00	1.26	0.57	0.97	32.90	1.50	0.18	0.16	0.99
5	469	56.00	2.10	53.60	2.10	1.19	0.56	0.97	55.30	1.50	0.13	0.12	0.99
PPATTRACTV	miss = 954												
1	125	12.80	1.20	12.70	1.10	0.50	0.34	0.98	12.90	0.90	0.11	0.10	0.99
2	164	16.50	1.50	17.30	1.60	1.58	0.63	0.97	16.30	1.00	0.09	0.09	1.00
3	271	32.80	1.90	32.10	2.00	1.30	0.58	0.97	32.70	1.50	0.18	0.15	0.99
4	130	19.60	1.80	20.30	1.70	0.82	0.46	0.98	19.70	1.30	0.11	0.10	0.99
5	158	18.30	1.60	17.50	2.10	2.71	0.74	0.96	18.40	1.20	0.06	0.06	1.00
PPJOBPAYWL	miss = 953												
1, 2	40	3.70	0.70	4.90	0.90	1.88	0.67	0.97	3.80	0.60	0.21	0.18	0.99
3	176	20.60	1.70	21.30	1.90	1.45	0.61	0.97	21.10	1.30	0.17	0.15	0.99
4	304	36.10	2.00	36.60	2.10	1.26	0.57	0.97	36.00	1.40	0.06	0.06	1.00
5	329	39.70	2.10	37.20	2.50	2.18	0.70	0.97	39.00	1.50	0.15	0.13	0.99
PPHCPALIMIT	miss = 966												
1	88	9.70	1.10	11.60	2.00	3.65	0.80	0.96	9.60	0.90	0.30	0.23	0.99
2	748	90.30	1.10	88.40	2.00	3.65	0.80	0.96	90.40	0.90	0.30	0.23	0.99
PPHCPASPORT	miss = 952												
1	102	12.40	1.40	12.90	1.90	2.58	0.74	0.96	11.70	1.00	0.19	0.16	0.99
2	748	87.70	1.40	87.10	1.90	2.58	0.74	0.96	88.30	1.00	0.19	0.16	0.99
PPHCPAOUT	miss = 953												
1	67	8.70	1.20	9.50	1.70	2.36	0.72	0.97	8.50	0.90	0.13	0.12	0.99
2	782	91.30	1.20	90.50	1.70	2.36	0.72	0.97	91.50	0.90	0.13	0.12	0.99

Note: * SE: Standard error; RIV: Relative Increase in Variance due to imputation; FMI: Fraction of Missing information; and RE: Relative Efficiency.

† The multiple imputation was run for all the eight variables together for the full data using all the 18 parent socio-demographic variables as predictors through IVEware.

‡ The multiple imputation was run for each of the eight variables separately for the full data using predictors specified in Table 2 through WSHD procedure in SUDAAN.

Evaluation Results Using Simulated Data

For each of the 100 simulated samples, 20 multiply imputed data sets were created using each of the three multiple imputation approaches (SRMI, SRMI2, WSHD) to impute the missing data, respectively. Let θ denote the parameter for the outcome of interest, i.e. the percentage of people that fall into one given category of a categorical variable, and let $\bar{\theta}^{\text{imp}}$ denote the estimate of θ based on multiply imputed data. Let $\bar{\theta}_j^{\text{imp}}$ and θ^{orig} denote the estimate based on multiply imputed data sets from the j^{th} ($j = 1,\ldots, K$) simulation and the original data from the evaluation study sample, respectively. The relative bias of $\bar{\theta}_j^{\text{imp}}$ (RELBIAS) was computed by averaging the relative differences between $\bar{\theta}_j^{\text{imp}}$ and θ^{orig}, $j = 1,\ldots, K$, across the K simulations, i.e.

$$\text{RELBIAS} = \frac{1}{K}\sum_{j=1}^{K}\frac{\left(\bar{\theta}_j^{\text{imp}} - \theta^{\text{orig}}\right)}{\theta^{\text{orig}}}$$

where $K = 100$ for our case and $\bar{\theta}_j^{\text{imp}}$ was computed using formula (1) for each j, $j = 1,\ldots, K$. The corresponding standard Monte Carlo simulation error for the relative bias of $\bar{\theta}_j^{\text{imp}}$ (SE_RELBIAS) was computed as

$$\text{SE\_RELBIAS} = \sqrt{\frac{1}{K\left(K-1\right)}\sum_{j=1}^{K}\left[\frac{\left(\bar{\theta}_j^{\text{imp}} - \theta^{\text{orig}}\right)}{\theta^{\text{orig}}} - \text{RELBIAS}\right]^2}$$

The coverage rate was computed as the proportion of confidence intervals that covers the original estimate (the truth) among the K simulations. The confidence interval for $\bar{\theta}_j^{\text{imp}}$ was computed using formula (3). The nominal coverage rate is 0.95.

Presented in Table 4 are the estimates of the percentage of people falling into each category of each outcome variable, relative bias, and standard Monto Carlo simulation errors of $\bar{\theta}_j^{\text{imp}}$ based on the three imputation approaches. The coverage rates of the associated confidence intervals are also reported. The absolute values of the estimated relative bias of $\bar{\theta}_j^{\text{imp}}$ based on WSHD approach varied from 0% (SE = 0.6%) to 7.2% (SE = 1.1%) across the 24 rows in Table 4, while the range is 0.9% (SE = 0.7%) to 274.1% (SE = 18.6%) based on the SRMI approach. The absolute values of 19 out of 24 of the estimated relative biases based on WSHD approach are less than 3%, while the absolute values of 20 out of 24 of the estimated relative biases based on SRMI are bigger than 3%.

Table 4. Percent relative bias (and standard Monte Carlo simulation errors) and associated 95% confidence interval coverage rate of imputed data based on 100 sets of simulated data (*n* = 826), with 20 sets of multiply imputed data for each simulation

	Original data		**Relative bias**			**Converge rate**		
Variable	***n***	**Pct**	**SRMI†**	**SRMI2‡**	**WSHD**	**SRMI†**	**SRMI2‡**	**WSHD**
PPFEELLOVE								
3	29	3.70	185.9 (19.3)	15.3 (2.6)	1.9 (2.6)	0.77	0.99	0.95
4	178	22.30	5.4 (1.2)	3.3 (1.0)	3.3 (0.9)	0.95	0.99	0.93
5	619	74.00	-10.9 (1.0)	-1.7 (0.3)	-1.1 (0.3)	0.73	0.99	0.93
PPOTHBETTER								
3	123	13.90	29.3 (2.9)	5.1 (1.2)	3.7 (1.2)	0.85	1.00	0.91
4	311	37.80	-1.8 (0.7)	-0.3 (0.6)	-0.7 (0.6)	0.96	0.99	0.95
5	392	48.30	-7.0 (0.8)	-1.2 (0.5)	-0.5 (0.4)	0.90	0.98	0.98
PPGETGDGRAD								
3	97	11.50	39.1 (3.1)	4.1 (1.2)	0.5 (1.1)	0.85	0.99	0.97
4	271	32.60	-0.9 (0.7)	2.0 (0.6)	0.0 (0.6)	0.99	1.00	0.98
5	458	55.90	-7.6 (0.6)	-2.0 (0.4)	-0.1 (0.4)	0.85	0.98	0.98
PPATTRACTV								
1	124	13.00	7.7 (1.2)	6.1 (1.2)	7.2 (1.1)	0.99	0.98	0.91
2	159	16.20	4.3 (1.1)	1.9 (1.0)	2.8 (1.0)	0.99	0.99	0.93
3	263	33.20	-6.6 (0.6)	-3.1 (0.6)	-0.5 (0.6)	0.95	0.99	0.96
4	124	19.30	-3.8 (1.1)	-3.0 (1.0)	-7.1 (1.0)	0.96	0.96	0.82
5	156	18.30	6.7 (1.1)	2.8 (1.1)	0.8 (1.1)	0.98	0.98	0.90
PPJOBPAYWL								
2	39	3.70	274.1 (18.6)	10.1 (2.2)	5.3 (2.0)	0.53	1.00	0.96
3	173	20.60	5.7 (1.9)	1.8 (0.9)	0.6 (0.9)	0.91	1.00	0.95
4	290	35.50	-7.5 (1.2)	0.6 (0.6)	-1.0 (0.6)	0.85	0.99	0.96
5	324	40.20	-21.6 (1.1)	-2.4 (0.6)	0.1 (0.6)	0.41	0.98	0.96
PPHCPALIMIT								
1	87	9.80	16.9 (1.9)	6.4 (1.4)	2.1 (1.3)	0.97	1.00	0.96
2	739	90.30	-1.8 (0.2)	-0.7 (0.1)	-0.2 (0.1)	0.97	1.00	0.96
PPHCPASPORT								
1	98	12.00	15.6 (1.4)	6.4 (1.3)	-0.6 (1.3)	0.96	1.00	0.93
2	728	88.00	-2.1 (0.2)	-0.9 (0.2)	0.1 (0.2)	0.96	1.00	0.93
PPHCPAOUT								
1	64	8.30	49.6 (5.7)	7.7 (1.6)	-0.9 (1.5)	0.82	1.00	0.95
2	762	91.70	-4.5 (0.5)	-0.7 (0.1)	0.1 (0.1)	0.82	1.00	0.95

Note: † The multiple imputation was run for all the eight variables together for each simulated data using all the 18 parent socio-demographic variables as predictors through IVEware.
‡ The multiple imputation was run for each of the eight variables separately for each simulated data using the same predictors as those used in WSHD (specified in Table 2) through IVEware.

Eight of the estimated relative biases based on SRMI are bigger than 15%. Among those, six are for categories with smaller sample sizes (*n* < 100). The range of the coverage rates based on the WSHD approach varied from 0.82 to 0.98 with many of the coverage rates being close to the 0.95 nominal value. The coverage rate range based on the SRMI approach varied from 0.41 to 0.99, with only a few coverage rates being close to the nominal value. With the SRMI2 approach, both the relative bias and coverage rates improved much compared to the SRMI model, but are still not as good as those from WSHD method. The range of the absolute values of the estimated bias becomes 0.3% (SE = 0.6%) to 15.3% (SE = 2.6%) and the number of absolute relative bias that are bigger than 3% reduces from 20 to 10. The coverage rates based on SRMI2 range from 0.96 to 1.00 which are too conservative. Those comparison results clearly show that WSHD is the winner among the three approaches/models compared in terms of relative bias and coverage rates.

Conclusion

Methods used to impute missing values were described for eight variables in the FLASHE Parent Physical Activity Survey due to a system error. Due to the large missing rates (around 53%), the focus was on multiple imputation methods and the results were compared between two commonly used approaches including SRMI and WSHD. An evaluation study through simulated data was conducted to fully evaluate the two different approaches.

For WSHD, with the FLASHE data, it was found including too many predictors may cause the software to fail to impute all the missing values for the target variable due to insufficient donors within one or more imputation cells. The number of times a donor is used is limited, depending on the donor’s sample weight (Research Triangle Institute, 2012). Thus, predictors were carefully selected for each individual variable to be imputed. Even though SRMI can incorporate all the target variables to be imputed and all the potential predictors into one model specification, which is desirable for congeniality purpose (Meng, 1994), and run through the imputation smoothly, the different analyses and evaluations showed that, without carefully choosing the predictors, the performance of SRMI could be poor in terms of bias and coverage rate for estimation of population quantity (e.g., percentage). Even with the same set of carefully chosen predictors, WSHD still out-performed SRMI in terms of estimation of percentages for categorical data. This evaluation focused on percentage estimation of those variables because that’s the focus of potential analyses. Different conclusions maybe drawn if the parameters of interest are associations (e.g., correlation coefficients).

Given that the respondents with the eight variables missing were all from the Diet survey first group, potential mode (group) effects on the responses to the eight items may impact the imputation of the FLASHE data. However, because respondents were randomly split into either receiving the Diet survey first or the Physical Activity Survey first groups, with the exception of the 500 all-male sample that was assigned to the Diet survey first group (due to late delivery of the sampling frame), it was expected that mode effect would be ignorable. To confirm this a mode effect analysis was performed on 22 variables appearing under the same section as those eight variables in the Parent Physical Activity Questionnaire (National Cancer Institute, 2015). The distributions of the 22 variables in the two groups were compared and logistic regressions were conducted using group indicator as the predictor. No significant model effect was detected for 21 out of the 22 variables being studied. For the one variable with mode effect being significant, the *p*-value was just 0.03. Therefore, we didn't consider model effect in the imputation for this paper.

Generally, imputing missing data has the potential to reduce bias that can occur with complete-case analysis and other methods by incorporating predictors observed for both complete and incomplete cases in the imputation model. Using multiple imputations instead of single imputation reflects the extra uncertainty in estimates that is due to imputation. Although there was not clear evidence of such bias in these analyses regardless of the unusual mechanism of missingness, there was greater efficiency from the imputation by utilizing data on the eight variables and other predictors in the imputation models when the WSHD approach was used. Even though this is a case study of missing data problem for a specific application, the methods applied in this study can extend to other applications as the methods have quite general applicability.

Acknowledgements

The FLASHE study was funded by contract from the National Cancer Institute (NCI) to Westat, Inc., HHSN261201200039I. This project has been funded in whole or in part with federal funds from the NCI, National Institutes of Health, under Contract No. HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. The authors would like to thank two referees for their valuable comments and suggestions, which led to significant improvements of the paper.

References

Andridge, R., & Little, R. J. A. (2009). The use of sample weights in hot deck imputation. *Journal of Official Statistics*, *25*(1), 21-36.

Andridge, R., & Little, R. J. A. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, *78*(1), 40-64. doi: 10.1111/j.1751-5823.2010.00103.x

Cox, B. G. (1980). The weighted sequential hot deck imputation procedure. In *Proceedings of the section on survey research methods* (pp. 721-726). Washington, D.C.: American Statistical Association.

Cox, B. G., & Folsom, R. E. (1981). An evaluation of weighted hot deck imputation for unreported health care visits. In *Proceedings of the section on survey research methods* (pp. 412-417). Washington, D.C.: American Statistical Association.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, *8*(3), 206-213. doi: 10.1007/s11121-007-0070-9

Grau, E. A., Frechtel, P. A., Odom, D. M., & Painter, D. (2004). A simple evaluation of the imputation procedures used in NSDUH. In *Proceedings of the section on survey research methods* (pp. 3588-3595). Washington, D.C.: American Statistical Association.

Harel, O., & Zhou, X-H. (2007). Multiple imputation: Review of theory, implementation and software. *Statistics in Medicine; 26*(16), 3057-3077. doi: 10.1002/sim.2787

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

Liu, B., Yu, M., Graubard, B. I., Troiano, R. P., & Schenker, N. (2016). Multiple imputation of completely missing repeated measures data within persons from a complex sample: Application to accelerometer data in the National Health and Nutrition Examination Survey. *Statistics in Medicine, 35*(28), 5170-5188. doi: 10.1002/sim.7049

Lohr, S. L. (1999). *Sampling: Design and analysis*. Boston, MA: Brooks/Cole.

Meng, X-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science, 9*(4), 538-573. doi: 10.1214/ss/1177010269

National Cancer Institute. (2015). *FLASHE – Annotated parent physical activity survey instrument*. Retrieved from https://cancercontrol.cancer.gov/brp/hbrb/docs/Parent_PA_PUF_Instrument.pdf

Oh, A. Y., Davis, T., Dwyer, L. A., Hennessy, E., Li, T., Yaroch, A. L., & Nebeling, L. C. (2017). Recruitment, enrollment, and response of parent-adolescent dyads in the FLASHE study. *American Journal of Preventive Medicine, 52*(6), 849-855. doi: 10.1016/j.amepre.2016.11.028

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

Research Triangle Institute. (2012). *SUDAAN language manual* (Vols. 1-2, Release 11). Research Triangle Park, NC: Research Triangle Institute.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley. doi: 10.1002/9780470316696

Schenker, N., Borrud, L.G., Burt, V.L., Curtin, L.R., Flegal, K. M., Hughes, J.,… Mirel, L. (2010). Multiple imputation of missing dual-energy X-ray absorptiometry data in the National Health and Nutrition Examination Survey. *Statistics in Medicine, 30*(3), 260-276. doi: 10.1002/sim.4080

Schenker, N., Raghunathan, T. E., Chiu, P., Makuc, D. M., Zhang, G., & Cohen, A. J. (2006). Multiple imputation of missing income data in the National Health Interview Survey. *Journal of the American Statistical Association, 101*(475), 924-933. doi: 10.1198/016214505000001375

Stuart, E. A., Azur, M., Frangakis, C., & Leaf, P. (2009). Multiple imputation with large data sets: A case study of the children's mental health initiative. *American Journal of Epidemiology, 169*(9), 1133-1139. doi: 10.1093/aje/kwp026

Westat. (2015). *Family, life, activity, sun, health, and eating (FLASHE) study methodology report*. Retrieved from https://cancercontrol.cancer.gov/brp/hbrb/docs/FLASHE_Methods_Report.pdf

Zhu, J., & Raghunathan, T. E. (2015). Convergence properties of a sequential regression multiple imputation algorithm. *Journal of the American*

Statistical Association, *110*(511), 1112-1124. doi: 10.1080/01621459.2014.948117

Appendix

Table A1. Definition of the parent socio-demographics variables

Socio-demographics variables	Categories
1. Parent age	1: 18-34; 2: 35-44; 3: 45-59; 4: 60+
2. Parent gender	1: male; 2: female
3. Parent highest education level	1: High school or less; 3: Some college; 4: 4-year college degree or higher
4. Parent marital status	1: Married; 2: Other
5. Parent health status	1: Excellent; 2: Very good; 3: Good; 4: Fair/Poor
6. Consistency of parent health insurance coverage	1: Currently uninsured or periods of no coverage during past 12 months; 2: Consistently insured during the past 12 months
7. Parent race/ethnicity	1: Hispanic; 2: Black only; 3: White only; 4: Other
8. Parent nativity	1: Born in the US; 2: Not born in the US
9. Home ownership	1: Currently own the home; 2; Not own the home
10: Housing security (How often in the past 12 months would you say you were worried or stressed about having enough money to pay for your rent or mortgage?	1: Never; Almost; 3: Sometimes; 4: Fairly often 5; Very often
11: Parent work status	1: Employed for wages; 2: Self-employed; 3: homemaker; 4: Out of work/student/retired/Other
12. Household income	1: $0 to $99,999; 2: $100,000 or more
13. Language usually spoken at home by parents	1: English only; 2: Not English only
14. Language used for media (In what languages are the TV shows, radio stations or newspapers that you usually watch, listen to or read?)	1. Only English; 2: English and/or other languages
15. Parent health literacy (How often do you need to have someone help you read written material from your doctor or pharmacy?)	1: Never; 2: sometimes to very often
16. Number of kids living in home	1: 1 Kid in home; 2: 2 kids in home; 3: 3 or more kids in home
17. Parent BMI	1; Under or normal weight (BMI < 25); 2: Over weight (25 ≤ BMI < 30); 3: Obesity (BMI ≥ 30)
18. Adolescent health insurance coverage (During the past 12 months, was there any time when teen had health care coverage?)	1: Yes; 2: No