A Semi-Parametric Approach to Account for Complex Designs in Multiple Imputation

Hanzhi Zhou, Trivellore E. Raghunathan and Michael R. Elliott

Program in Survey Methodology, University of Michigan

Department of Biostatistics, University of Michigan

zhouhanz@umich.edu, teraghu@umich.edu, mrelliot@isr.umich.edu

Multiple imputation (MI) has become one of leading approaches in dealing with missing data in survey research. However, existing software packages and procedures typically do not incorporate complex sample design features in the imputation process. Researcher has demonstrated that implementation of MI based on simple random sampling (SRS) assumption can cause severe bias in estimation and hence invalid inferences, especially when the design features are highly related to survey variables of interest (Reiter et al. 2006). Recent work to accommodate complex sample designs in imputation has focused on model-based methods which directly model the complex design features in the formulation of the imputation model. In this paper, we propose a semi-parametric procedure as an alternative approach to incorporate complex sample designs in MI. Specifically, we divide the imputation process into two stages: the complex feature of the survey design (including weights and clusters) is fully accounted for at the first stage, which is accomplished by applying a nonparametric method to generate a series of synthetic datasets; we then perform conventional parametric MI for missing data at the second stage using readily available imputation software designed for an SRS sample. A new combining rule for the point and variance estimates is derived to make valid inferences based on the two-stage procedure. Using health survey data from the Behavior Risk Factor Surveillance System, we evaluate the proposed method with a simulation study and compare it with the model-based method with respect to complete data analysis. Results show that the proposed method yields smaller bias and is more efficient than the model-based method.

Keywords: missing data, complex sample design, multiple imputation, Bayesian Bootstrap, synthetic data

1. Introduction and Research Question

Multiple imputation (MI) is a principled method in dealing with missing data in survey research and has been adopted by federal statistical agencies in recent years. A very important point underlying the MI theory is that the method was designed for complex sample surveys hence requires the imputation to be made conditional on sample designs. The purpose is to make the missing data mechanism ignorable: since design features are usually related to survey variables of interest in real survey data, severe bias on the estimates can be avoided when they are properly accounted for in the process (Reiter et al. 2006). However, survey practitioners usually assume simple random sampling (SRS) when they're performing MI, largely due to the inadequacy of standard software packages in handling complex sample designs. A typical example is the Sequential Regression Multivariate

Imputation procedure (SRMI) using IVEware (Raghunathan et al.2001), which has been gaining increasing popularity in handling multivariate missing data in large scale surveys. Applications such as MI for missing income data in NHIS (Schenker et al. 2006) focused on the strategies of modeling complicated data structure and failed to recognize the importance of incorporating complex sample design features, reflecting an inconsistency between theory and practice.

The question is then: how do we fully incorporate the sample designs in MI to achieve valid statistical inferences? Reiter et al. (2006) proposed a fixed effect modeling method in addressing the problem where they included design variables as predictors in the imputation model and it outperforms SRS scenario in terms of correcting the bias. Their conclusion thus supports the general advice of including complex sample designs in MI procedure. However, they did not look at survey weight as another important design variable. Treating weight as scalar summary of the design information in the imputation model may not work well for inference beyond means or totals, since interactions between the probabilities of selection and the population parameters of interest will not be accounted for (Elliott 2007). Besides, their results for real data application did not show significant gains of incorporating designs over ignoring designs as in their simulation study. Little following work has been done to replicate their results with real survey datasets or to investigate other potential methods in this regard, say, is there a way to incorporate design information within the MI framework other than including them as covariates in the imputation model? The goal of this paper is to propose a semi-parametric two-step method as an alternative to the existing fully model-based methods. Specifically, we divide the imputation process into two steps: the complex feature of the survey design (including weights and clusters) is fully accounted for at the first step, which is accomplished by applying a nonparametric method to generate a series of synthetic datasets; we then perform conventional parametric MI for missing data at the second step using readily available imputation software designed for an SRS sample. A new combining rule for the point and variance estimates is derived to make valid inferences based on the two-step procedure.

The rest of this paper is structured as follows: Section 2 describes the proposed method in detail. We first lay out the conceptual idea and then demonstrate the theoretical results. Section 3 provides the results from a simulation study in a PPS sampling design setting to empirically evaluate the new method. Section 4 concludes with discussion and directions for future research.

2. Two-step MI Procedure

In this section, we propose a two-step procedure to perform multiple imputation with which accounting for complex designs and multiply imputing missing data are divided into two separate steps. The basic idea is to “uncomplex” the complex designs before implementing standard MI procedure. By “uncomplex”, we mean a statistical procedure which makes the design features irrelevant at the analysis stage, i.e. turning a dataset into one that is self-weighting, so that we can treat the uncomplexed populations as if they were simple random samples from a superpopulation or the true population. Thus, simple estimation formulae for self-weighting samples directly apply. Specifically in our case, this is achieved by generating synthetic populations through a nonparametric resampling procedure adapted from Bayesian Bootstrap (Rubin 1981) i.e. the Finite Population Bayesian Bootstrap (FPBB) which will be introduced in section 2.2, such that complication of modeling those designs in the imputation can be avoided. The standard practice of MI assuming SRS can then apply directly to the generated populations which are free of sampling designs.

Two major assumptions are made for implementing the proposed method: 1) The method is proposed to treat item nonresponse problem under MI framework, rather than unit nonresponse. Therefore all input weights are assumed to be final weights after unit nonresponse adjustment and calibration. 2) Missing at random (MAR) as a more practical missing data mechanism is assumed for all types of analysis under study. We do not consider MCAR (Missing Completely At Random) which is too ideal for real world surveys, neither do we consider NMAR (Not Missing At Random) which is another scope of research topic.

2.1. Conceptual Idea

Figure 1 and Figure 2 together illustrate the conceptual idea of the procedure.

Figure 1 shows the proposed procedure to account for complex designs in MI. Denote Q as the population parameter we are interested in. Denote the actual sample survey data as $D = (Y_{mis}, Y_{obs})$, where Y_{mis} represents the portion of item missing data and Y_{obs} represents the portion of observed data. The first step of the synthetic data generation approach creates B FPBB synthetic populations $D^B = \{D^{(1)}, D^{(2)}, ..., D^{(B)}\}$, where $D^{(b)} = (Y_{mis}^{(b)}, Y_{obs}^{(b)})$, $b = 1, 2, ..., B.$ The second step of multiple imputation creates M imputed datasets for each of the FPBB synthetic population generated from the first step, $D_M^{(b)} = \{D_1^{(b)}, D_2^{(b)}, ..., D_M^{(b)}\}$, for $b = 1, 2, ..., B.$ Thus we end up with $D_M^B = \{D_1^{(1)}, D_2^{(1)}, ..., D_M^{(1)}, D_1^{(2)}, D_2^{(2)}, ..., D_M^{(2)}, ..., D_1^{(B)}, D_2^{(B)}, ..., D_M^{(B)}\}$ which contains all the imputed synthetic population datasets generated by the two-step procedure.

Figure 2 (a)-(c) shows the evolution of data structure in the proposed procedure. Suppose we have three survey variables of interest, Y1, Y2 and Y3, and a survey sample of size n was drawn from of a target population of size N through some type of complex sampling design. The character ‘m’ in pink square denotes the missing part for each survey variable in both sample data and synthetic population data. At the first step, we can think of the unobserved elements of the population as missing by design and we treat missing values as a separate category for each variable. By applying the adapted FPBB method, a synthetic population is created, ideally a plausible reflection of the target population. Note that in this process, the missing data are also brought up to the population level. At the second step, conventional SRMI assuming SRS is applied to fill in those item missing data and we end up with a complete dataset which we call imputed synthetic population. The whole process will be replicated for B by M times.

Figure1. Proposed Procedure to Account for Complex Designs in MI

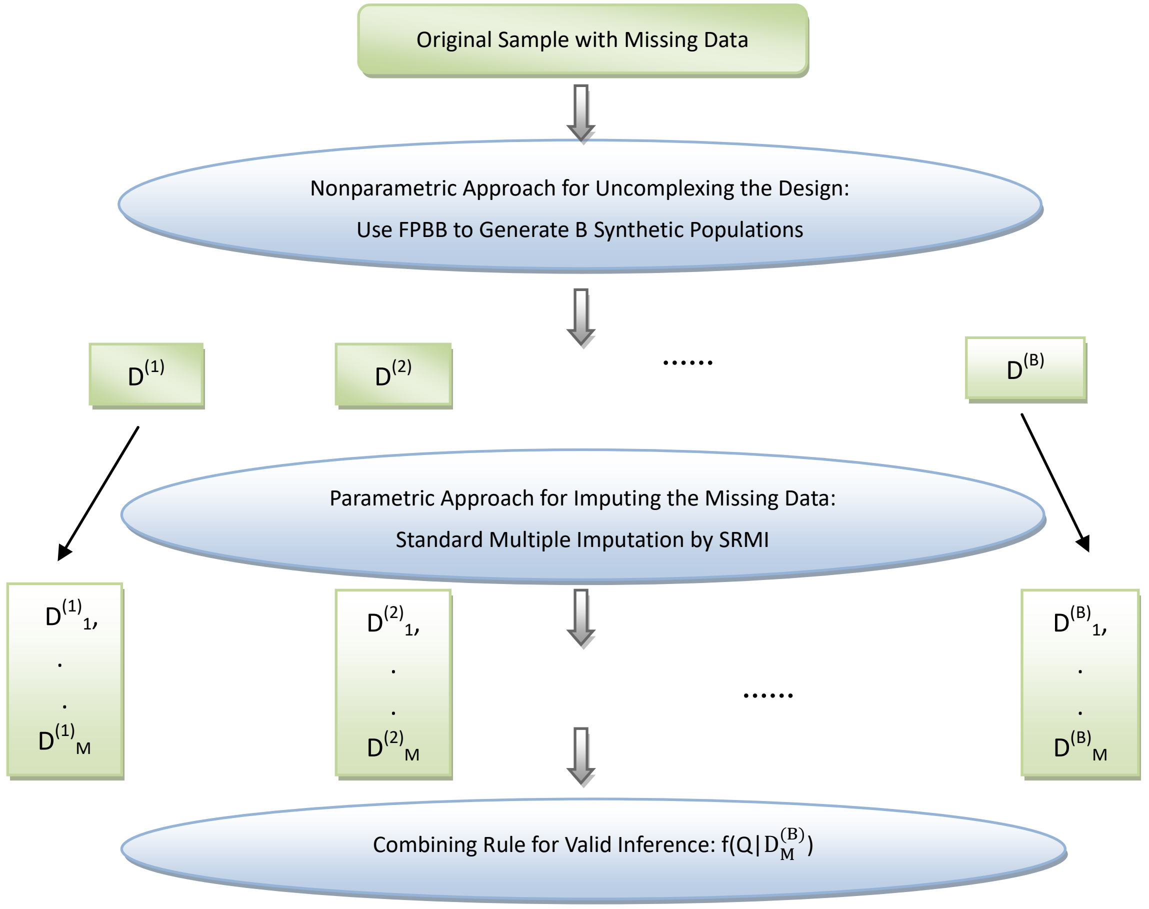
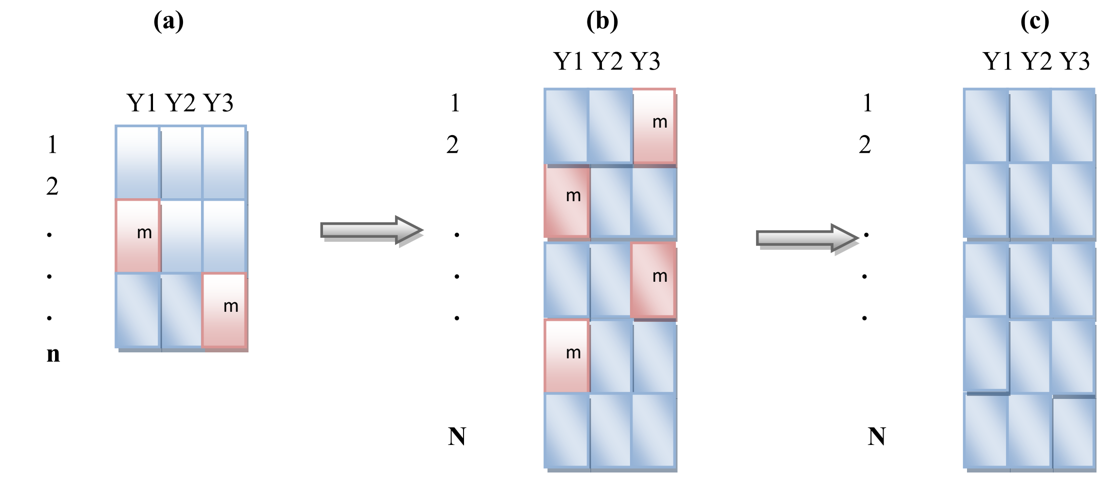

Original Sample with Missing Data
Nonparametric Approach for Uncomplexing the Design:
Use FPBB to Generate B Synthetic Populations
$D^{(1)}$
$D^{(2)}$
……
$D^{(B)}$
Parametric Approach for Imputing the Missing Data:
Standard Multiple Imputation by SRMI
$D^{(1)}{}_{1}$, . . $D^{(1)}{}_{M}$
$D^{(2)}{}_{1}$, . . $D^{(2)}{}_{M}$
……
$D^{(B)}{}_{1}$, . . $D^{(B)}{}_{M}$
Combining Rule for Valid Inference: $f(Q|D_M^{(B)})$

Figure2. Data Structure Evolution



Actual Sample Dataset **Synthetic Dataset** **Imputed Synthetic Dataset**

2.2. Methods

Now we demonstrate how the adapted-FPBB method for unequal probability sampling (Cohen 1997) can be applied to uncomplex weight as one design feature and how this method can be adapted to uncomplex cluster as another, hence more complicated designs such as stratified multistage cluster sampling can be handled. SRMI is illustrated as one option to perform the conventional MI once the complex designs have been appropriately accounted for in the first place.

2.2.1. First step---Nonparametric approach to Uncomplex Designs

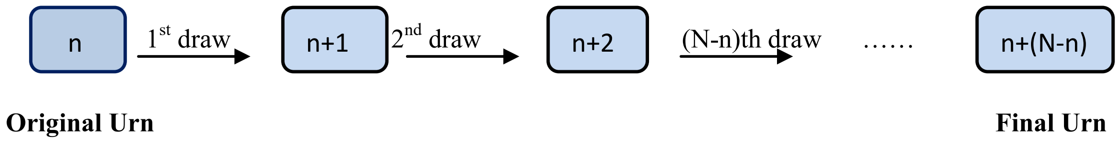
Per the limitations of fully model-based methods stated in section1, we propose using a nonparametric method, i.e. the adapted-FPBB by Cohen (1997) to account for complex sample designs in MI. Specifically, the adapted-FPBB serves as a procedure to restore the existing complex survey sample back to some SRS-type/self-weighting data structure. This will be realized by generating populations from the complex sample repeatedly in a spirit similar to synthetic population generation method in the context of MI for disclosure risk limitation (Raghunathan et al. 2003). Resampling technique is used in order to fully capture the uncertainty in the original sample. The nonparametric approach has minimum assumption of the distributional form of random effects and is robust to model misspecification that usually poses problems to the model-based methods. Additional to that, FPBB is a method developed from the conventional bootstrap whose Bayesian nature makes it fit to the MI framework well.

2.2.1.1. Finite population Bayesian Bootstrap:

Pólya's Urn Scheme:

Denote an urn containing finite number of balls as $\{y_1, y_2, \ldots, y_n\}$. A ball is randomly drawn from the urn and a same ball from outside of the urn is added back to the urn along with the originally picked one. Repeat such selection process until m balls have been selected as a sample, call this sample 'Pólya sample of size m'. Figure 3 is a flow chart of how a Pólya Sample of size m=N-n is drawn.

Figure3. A Flow Chart of Drawing a Pólya Sample of size m=N-n.



Adapted-FPBB method:

Based on the Pólya's urn scheme described above, the adapted-FPBB populations can be generated following the three-step procedure:

Step1: Take a Pólya sample of size N-n, denoted by $y_1^*, y_2^*, \ldots, y_{(N-n)}^*$ from the urn $\{y_1, y_2, \ldots, y_n\}$.

In this process, each y_i in the urn is selected with probability:

$$\frac{\mathrm{w_i - 1 + l_{i,k-1}} \times \left(\frac{\mathrm{N-n}}{\mathrm{n}}\right)}{\mathrm{N - n + (k-1)} \times \left(\frac{\mathrm{N-n}}{\mathrm{n}}\right)}, \qquad (1)$$

where w_i is the case weight for the *i*th unit and $l_{i,k-1}$ is the number of bootstrap selections of unit *i* up to (*k*-1)th selection, setting $l_{i,0} = 0$.

Step2: Form the FPBB population $y_1, y_2, \ldots, y_{\mathrm{n}}, y_1^*, y_2^*, \ldots, y_{N-n}^*$ so that the FPBB population has exact size N.

Step3: Repeat the previous steps a large number of times, say B times, to obtain B FPBB populations.

2.2.1.2. Relating FPBB to Multiple Imputation for Item Missing Data:

If we think of the sampling process and responding process as one combined process, in other words, treating the responding process as another level of sampling of responding units given the original sampled units, then it's easy to see the connection between FPBB and standard MI. FPBB tries to bootstrap the sample to the entire population while MI tries to 'bootstrap' the responding pool to the complete sample. A typical example is the approximate Bayesian bootstrap (ABB) suggested by Rubin & Schenker (1986) as a way of generating MI when the original sample can be regarded as IID and the response mechanism is ignorable. Now that Cohen's FPBB extends to unequal probability selection, we may well think of the unsampled part of population as missing and multiply impute this part using the adapted-FPBB procedure described above. Hence the essential of our approach---to carry out two levels of multiple imputation: 1) MI for unit nonresponse by Pólyaing up a complex sample onto a population where missing values are treated as a separate category for each variable with missing data, and 2) MI for item nonresponse by SRMI for the entire population.

2.2.1.3. Uncomplex Weights:

The adapted-FPBB can be applied directly to weights in a sampling design such as probability proportional to size (PPS) sampling. Note that in practice, the input weight in Formula (1) should be the final weight or poststratified weight after all types of adjustment including unit nonresponse adjustment and calibration for undercoverage, etc. The reason is obvious: if we use the base/design weight instead in applying the adapted FPBB for generating synthetic populations, the potential problems of undercoverage or unit nonresponse existing in the original complex sample would be brought up to the population level without being adjusted at all. Since ultimate analyses are based on such imputed synthetic populations, a direct consequence would be biased inference even if the procedure by itself is efficient.

Intuitive interpretation by assuming an SRS sample design:

Formula (1) for adjusting bootstrap selection probability based on case weight is interpreted as follows: Let $k = 1, 2, \ldots, N-n+1,\ \ i = 1, 2, \ldots, n,$ before making any bootstrap selection of unobserved units in the population from the observed original complex sample $y_1, y_2, \ldots, y_{\mathrm{n}}$, i.e. when $k = 1$ and $l_{i,k-1} = l_{i,0} = 0$, the probability of selecting unit i with sampling weight w_i is $(w_i - 1)/(N-n)$. To make it simpler to understand, suppose we have a simple random sampling of n units in the first place, then $w_i = N / n$ for all sampled units, each representing N / n units in the population, then the probability of that any one unit from the SRS sample is selected before any bootstrap selection is $(\frac{N}{n}-1)/(N-n)$, which is the selection probability of any units among all the rest $N-n$ units in the population, and this exactly equals $1/n$. As we proceed with the bootstrap selection, we adjust this selection probability according to the number of times each unit among $y_1, y_2, \ldots, y_{\mathrm{n}}$ was selected during the FPBB procedure, each unit now represents $(N-n)/n$ among the $N-n$ units to be selected during one bootstrap whenever it is selected once. After each selection, the denominator of the prior probability function needs to be inflated to reflect the total units being represented during all the bootstrap selections so far, while the numerator also needs to be inflated to reflect the total units represented by unit i in the process. Therefore we obtain the probability as in formula (1).

2.2.1.4. Uncomplex Sampling Error Codes (i.e. stratum and clusters):

Now we will show how the adapted FPBB also works when clusters are involved in the sample design. Suppose we have a stratified two-stage clustering design, with the probability of selection for each primary sampling unit (PSU)/cluster being proportional to its population size (PPS) within strata. As usual we treat each stratum independently and apply the method separately within each. Now we have two layers of bootstrap selection---one at the cluster level and the other at the element level. Suppose there are C_h clusters in h^{th} stratum in the population, among which c_h were sampled, denote them as z_{hj} where $j = 1, 2, ..., c_h$. Treating each cluster as the sampling unit, we can apply the same procedure as in previous section where only elements are involved. That is, we want to bootstrap selecting $C_h - c_h$ clusters out of $z_{h1}, z_{h2}, ..., z_{hc_h}$ to form a population of clusters: $z_{h1}, z_{h2}, ..., z_{hc_h}, \overset{*}{z}_{h1}, \overset{*}{z}_{h2}, ..., \overset{*}{z}_{h(C_h - c_h)}$. Accordingly, we need to change the corresponding terms in formula (1) to make it a cluster-level selection probability. Let $w_{hj}, j = 1, 2, ..., c_h$ be the sampling weight for j^{th} cluster, formula (1) thus can be adapted as formulae (2) and $(2)'$, corresponding to the cluster-level selection and element-level selection, respectively.

$$\frac{\mathrm{w}_{hj} - 1 + \mathrm{l}_{\mathrm{i,k-1}} \times \left(\frac{C_h - c_h}{c_h} \right)}{C_h - c_h + (\mathrm{k}-1) \times \left(\frac{C_h - c_h}{c_h} \right)}, \ (2) \quad \text{and} \quad \frac{\mathrm{w}_{ci} - 1 + \mathrm{l}_{\mathrm{i,k-1}} \times \left(\frac{N_{ch} - n_{ch}}{n_{ch}} \right)}{N_{hc} - n_{ch} + (\mathrm{k}-1) \times \left(\frac{N_{ch} - n_{ch}}{n_{ch}} \right)}, \ (2)'$$

Once the clusters have been selected by adapted FPBB procedure using formula (2), we can further select elements using the same method within each selected cluster using formula $(2)'$. Notice that at the element level of selection, each selected cluster should now be treated as a population therefore the sampling weight for elements in the original sample cannot be used anymore, instead we need to derive a new set of weights for elements conditional on cluster being selected. First, we need to obtain the conditional probability of selection for element given cluster (formula (3)), then inverse it to get the corresponding weight w_{ci} in formula $(2)'$.

$$\begin{aligned}&\Pr(ith \text{ element selected}|jth \text{ cluster selected})\\&=\frac{\Pr(ith \text{ element selected } \& \, jth \text{ cluster selected})}{\Pr(jth \text{ cluster selected})}\\&=\frac{\Pr(ith \text{ element selected in the original sample})}{\Pr(jth \text{ cluster selected})}=\frac{p_{hji}}{p_{hj}}, \quad (3)\end{aligned}$$

Where p_{hj} is the selection probability of cluster j among all clusters in stratum h, p_{hji} is the selection probability of unit i among all units in stratum h.

2.2.2. Second Step---MI for Missing Data Using SRMI:

Now that we have uncomplexed the sampling designs, we are in a good position to proceed with performing conventional multiple imputation. SRMI as a popular technique for complex survey data structure is one option. Without the need to include design variables in the imputation model due to a self-weighting FPBB population generated from previous step, our task should now be concentrated on correctly modeling the covariate variables as well as interactions among them whenever necessary.

2.3. Theoretical Results

Rubin (1987)'s standard MI rule for combining point and variance estimation does not fit the two-step MI procedure. A new combining rule is developed accordingly, which accommodates two sources of variability due to synthesizing populations by a nonparametric method at the first step and multiply imputing missing data by SRMI at the second step. The validity of the new combining rule is to be justified both from a Bayesian perspective and from a repeated sampling perspective.

<u>Combining rule for the point estimate:</u>

$$\bar{q}=\frac{1}{B}\sum_{b=1}^{B}\frac{1}{M}\sum_{m=1}^{M}q_m^{(b)}=\frac{1}{BM}\sum_{b=1}^{B}\sum_{m=1}^{M}q_m^{(b)}, \quad (4)$$

Combining rule for the variance:

$$T=(1+\frac{1}{B})\frac{1}{B-1}\sum_{b=1}^{B}(q^{(b)}-\overline{q^{B}})^{2}+\frac{1}{B}\sum_{b=1}^{B}U^{(b)}$$

$$=(1+\frac{1}{B})\frac{1}{B-1}\sum_{b=1}^{B}(\frac{1}{M}\sum_{m=1}^{M}q_{m}^{(b)}-\frac{1}{B}\sum_{b=1}^{B}\frac{1}{M}\sum_{m=1}^{M}q_{m}^{(b)})^{2}+\frac{1}{B}\sum_{b=1}^{B}(1+\frac{1}{M})\frac{1}{M-1}\sum_{m=1}^{M}(q_{m}^{(b)}-\overline{q_{M}^{(b)}})^{2}$$

$$=(1+\frac{1}{B})\frac{1}{B-1}\sum_{b=1}^{B}(\frac{1}{M}\sum_{m=1}^{M}q_{m}^{(b)}-\frac{1}{B}\sum_{b=1}^{B}\frac{1}{M}\sum_{m=1}^{M}q_{m}^{(b)})^{2}+\frac{1}{B}\sum_{b=1}^{B}(1+\frac{1}{M})\frac{1}{M-1}\sum_{m=1}^{M}(q_{m}^{(b)}-\frac{1}{M}\sum_{m=1}^{M}q_{m}^{(b)})^{2}$$

$$=(1+\frac{1}{B})V^{B}+\frac{1}{B}\sum_{b=1}^{B}(1+\frac{1}{M})V_{M}$$

$$=T_{B}+\frac{1}{B}T_{M}, \qquad (5)$$

When *n*, *M* and *B* are large, the inference can be approximated by normal distribution, thus the 95% confidence interval can be computed as $[\bar{q}-z_{0.975}\sqrt{T},\bar{q}+z_{0.975}\sqrt{T}]$. (See Appendix A for detailed derivation and notation)

Bayesian Proof of the New Combining Rules for Inference:

With $\mathrm{B}\times\mathrm{M}$ imputed FPBB synthetic datasets generated from the proposed two-step procedure, we need to find a way to combine inferences from both steps. Now we show how the combining rules for normal approximation is derived, assuming large samples.

Let the complete data be $D=(D_{obs},D_{mis})$, where $D_{obs}=(X,Y_{obs},R_{inc},I)$, X is covariate matrix, Y_{obs} is the observed part of survey variable with missing data, R_{inc} is the response indicator for all sampled units, and I is the sampling indicator.

Let the FPBB synthetic population be $D^{(b)}=(D_{obs}^{(b)},D_{mis}^{(b)})$, b=1,2,...,B.

Let the imputed synthetic population be $D_{(m)}^{(b)}=(D_{obs}^{(b)},D_{mis(m)}^{(b)})$, m=1,2,...,M and b=1,2,...,B.

The posterior mean and variance of Q are immediate using the rules for finding unconditional moments from conditional moments (according to Result 3.2 of Rubin 1987). In our case, we shall condition on two layers of observed data due to the two-step MI procedure:

$$\text{Posterior Mean: }E(Q\mid D_{obs})$$

$$=E\{E[E(Q\mid D_{(m)}^{(b)})\mid D_{obs}^{(b)}]\mid D_{obs}\}$$

$$=E\{\overline{Q}_{b\,\infty}\mid D_{obs}\}$$

$$=\overline{Q}_{\infty\infty}, \qquad (10)$$

Where $\overline{Q}_{b\,\infty}=\lim_{M\to\infty}\sum_{m=1}^{M}\frac{\hat{Q}_{m}^{(b)}}{M}=E(\hat{Q}_{B}\mid D_{obs}^{(b)})$, note that $\hat{Q}_{B}$ is the complete data statistic for the synthetic population and $\hat{Q}_m^{(b)} = \{\hat{Q}_1^{(b)},...,\hat{Q}_M^{(b)}\}$ is M repeated values of the posterior distribution of $\hat{Q}_B$. $\overline{Q}_{\infty\infty} = \lim_{B\to\infty}\lim_{M\to\infty}\sum_{b=1}^{B}\sum_{m=1}^{M}\frac{\hat{Q}_m^b}{BM} = E(\hat{Q}\mid D_{obs})$, note that $\hat{Q}$ is the complete data statistic for the original population and $\hat{Q}_m^b = \{\hat{Q}_1^1,...,\hat{Q}_M^1,...,\hat{Q}_1^B,...,\hat{Q}_M^B\}$.

$$\begin{aligned}
&\mathit{Posterior}\ \text{Variance: } V(Q\mid D_{obs})\\
&= V\{E[E(Q\mid D_{(m)}^{(b)})\mid D_{obs}^{(b)}]\mid D_{obs}\} + E\{V[E(Q\mid D_{(m)}^{(b)})\mid D_{obs}^{(b)}]\mid D_{obs}\}\\
&= V\{\overline{Q}_{b\,\infty}\mid D_{obs}\} + E\{T_{*M}\mid D_{obs}\}\\
&= T_B + \frac{1}{B}T_M, \qquad (11)
\end{aligned}$$

Where

$$T_B = \lim_{B\to\infty}(1+\frac{1}{B})*\frac{1}{B-1}\sum_{b=1}^{B}(\overline{Q}_{b\,\infty} - \overline{Q}_{\infty\infty})^2, T_M = \lim_{M\to\infty}(1+\frac{1}{M})*\frac{1}{M-1}\sum_{m=1}^{M}(\hat{Q}_m^{(b)} - \overline{Q}_{b\,\infty})^2.$$

3. Simulation Study

A simulation study was designed to investigate the properties of inference based on the proposed method. In particular, we are interested to see how the two-step MI procedure performs in comparison with the existing alternative methods including: (1) complete case analysis, (2) ignore designs in the imputation model, and (3) include designs as fixed effect in the imputation model.

3.1. Data:

Data from real surveys were manipulated to serve as our population. Specifically, BRFSS 2009 in the state of Michigan used a disproportionate stratified sampling design (DSS) with no PSU (cluster level) involved therefore is suitable for our purpose of looking at ***weight*** as a single design variable in multiple imputation. There are four strata in Michigan and for simplicity we only chose one stratum as the basis of our simulation study. Eight categorical variables were selected which we thought would be potentially correlated with survey weight. Table 1 shows the recoded variables we've chosen for analysis. After some data cleaning, we ended up with a complete dataset of N=1323 cases which will be serving as our population.

Table1. Survey Variable under Analysis

Survey Variable	Coding		
Race	1: Whites	2: Non-Whites	
Whether or Not Have Health Plan	1: Yes	2: No	
Income Level	1: Low	2: Medium	3: High
Employment Status	1: Employed	2: Unemployed	3: Other
Marital Status	1: Married	2: Unmarried	
Education	1: Lower Than High School	2: High School and Higher	
Whether or Not Have Diabete	1: Yes	2: No	
Whether or Not Have Asthma	1: Yes	2: No	

3.2. Design:

Our strategies of evaluating the performance of proposed method in comparison with alternative methods consist of the following steps:

Step1: Make design variable related to survey variables:

Since we want to examine the new method assuming the designs are relevant to missingness on survey variables, we achieved the assumption by regressing weights as the dependent variable on all other survey variables as predictors and obtained the predicted values of weight to be used for simulation. The purpose is to make sure that weight as a design variable is at least moderately related to survey variables.

Step2: Draw 100 samples/replicates with probability proportionate to the inverse of predicted weights, each of size 200 (we will call it before-deletion samples):

In this way, we obtained the probability proportional to size (PPS) sampling weights ($\hat{\boldsymbol{w}}_{\boldsymbol{2i}}$) which are directly related to the predicted weights ($\hat{\boldsymbol{w}}_{\boldsymbol{1i}}$) from the previous step, since we were implicitly treating the predicted weights as kind of a measure of size. Now $\hat{\boldsymbol{w}}_{\boldsymbol{2i}}$ becomes our target design variable to be examined with the new method.

Step3: Impose Missingness on the Complete Data under MAR Mechanism

we used a deletion function taking the form $\Pr(M_i = 1|y_{i1}, y_{i2}, \dots, y_{i5}, x; \varphi) = \Pr(M_i = 1|x; \varphi) = f(x)$,where M_i is a binary indicator for missingness, x represents 'Race', $y_{i1}, y_{i2}, \dots, y_{i5}$ represents unit's values of other survey variables (except 'marital status' and 'education') of which we purposefully delete values. Thus we obtained 100 *after-deletion samples*. Table 2 shows the fractions of missing by race for each survey variable of interest. The fractions of missingness by race are in a range of 15%~40%, where the missingness for income was made to be generally higher than all other variables.

Table2. Fractions of Missingness on Survey Variables of Interest by Race

Race	Employment Status	Health Plan	Diabete	Asthma	Income
Whites	15%	15%	15%	15%	20%
Non-Whites	25%	25%	25%	25%	40%

Step4: Generate B=100 FPBB synthetic populations for each replicate sample, with the PPS sampling weights as input weights, using Cohen (1997)'s adapted FPBB method. In the process, we made the population size five multiples of the original PPS sample size thus each FPBB population is of size 1000.

Step5: Create M=5 multiply imputed datasets by SRMI procedure for each FPBB population with each replicate sample, thus I obtained 5*100*100=50000 imputed datasets each of size 1000.

Step6: Obtain multiply imputed synthetic population estimates for the mean of each survey variable of interest. For each replicate sample, we use formulae (4) and (5) as our new combining rules to combine the estimated means and variances from 500 imputed synthetic datasets.

3.3. Results:

Three critical statistics were examined for comparison across the four methods. They are absolute relative bias, root mean square error, and empirical nominal 95% confidence interval coverage rate. All are calculated based on 100 replicate samples. Also note that except for the new method all estimates under the other three methods as well as the actual samples before deletion are design-based.

Figure4 displays the Q-Q plot matrix of 100 pairs of estimated proportions from the actual samples before deletion versus that from the corresponding imputed synthetic populations under proposed method, for each survey variable by level. The plots demonstrate a nearly perfect 45-degree straight line for all the variable levels. This indicates that the distributions of the imputed synthetic populations practically match the actual samples before deletion.

Table3 gives the detailed results from the simulation study. For simplicity, we only display the results for three variables, employment status and health plan, which has higher correlations with the design variable weight, and income, whose fraction of missingness is highest among all. According to table3, the new method has much smaller bias than all its competitors. In terms of RMSE, although the gain is not as substantial as that in the case of absolute relative bias, the new method performs generally the best. Nominal 95% CI coverage of the point estimate under the proposed method is in a range of 86%-96%, considering categorical nature of all the survey variables, this is a reasonable result. We can see that except for complete case analysis which has lowest CI coverage in general, there seems no much difference between the new method and the other two model-based MI methods.

Figure4. Q-Q Plot Matrix for Estimated Proportions: Actual Sample Before Deletion versus FPBB+SRMI

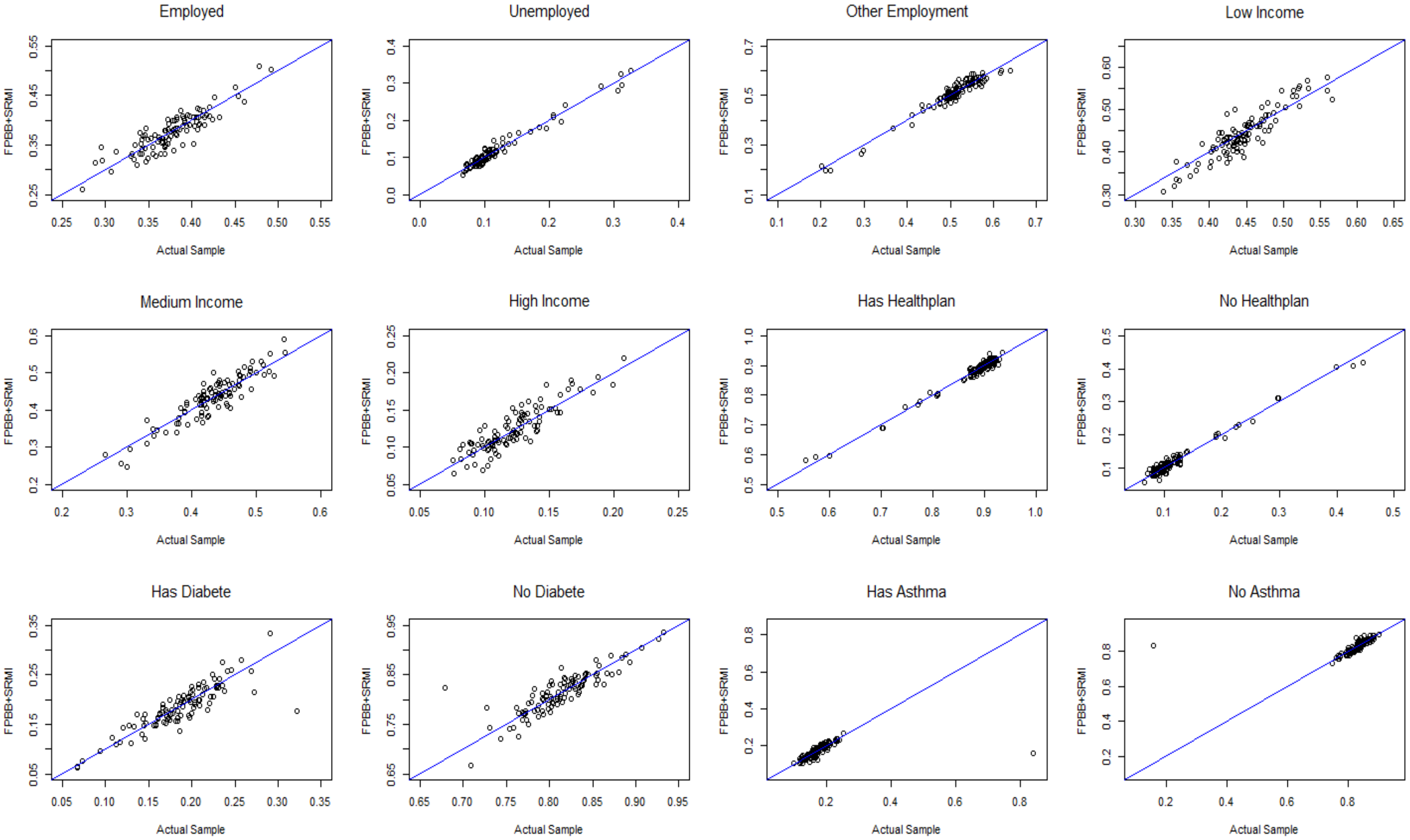


Table3. Absolute Relative Bias, RMSE and 95% CI Coverage Rates Compared across Four Methods

Variables	FPBB+SRMI			Include Weights in the Imputation Model			Do Not Include Weights in the Imputation Model			Complete Case Analysis		
EMPLOY_M	Relbias	RMSE	95% CI cov.	Relbias	RMSE	95% CI cov.	Relbias	RMSE	95% CI cov.	Relbias	RMSE	95% CI cov.
1	0.30%	3.99E-02	96%	1.03%	3.59E-02	95%	2.10%	3.93E-02	96%	3.23%	7.34E-02	95%
2	0.17%	5.41E-02	87%	19.15%	5.76E-02	86%	8.73%	5.67E-02	86%	5.98%	5.97E-02	79%
3	0.17%	7.69E-02	90%	3.66%	6.80E-02	90%	3.58%	7.69E-02	90%	1.02%	8.43E-02	92%
INCOME_M												
1	0.84%	5.54E-02	88%	1.38%	5.27E-02	91%	1.22%	5.63E-02	89%	6.77%	8.94E-02	88%
2	0.33%	6.16E-02	88%	0.25%	5.99E-02	85%	0.63%	6.20E-02	87%	3.17%	8.39E-02	93%
3	1.91%	3.02E-02	96%	5.93%	3.33E-02	95%	6.68%	3.35E-02	94%	13.53%	5.48E-02	90%
HLTHPLAN_M												
1	0.03%	6.68E-02	86%	4.59%	7.88E-02	87%	1.51%	6.96E-02	85%	0.80%	7.11E-02	80%
2	0.19%	6.68E-02	86%	32.70%	7.88E-02	87%	10.80%	6.96E-02	85%	2.40%	1.12E-01	80%

4. Discussions

Our primary goal was to propose a new method to account for complex sample design features in multiple imputation for item missing data and to evaluate the performance of the new method in a PPS sample setting. There are several advantages of the proposed two-step MI procedure: first, it relaxes the usually strong distributional assumptions of random effects in parametric models; second, it potentially protects against model misspecification, for example, wrongly inclusion or exclusion of interactions between design variables and other covariates in the imputation model. Meanwhile, it retains the nice features of SRMI in handling complex data structure and various types of missing variable. Another advantage is that both steps of the procedure are of a Bayesian flavor and implies a proper imputation method. This fits into the standard MI paradigm which requires Bayesian derived imputation method to attain randomization validity. A further advantage lies in that unlike the fully model-based methods which include designs in the imputation model and still require complex survey packages to analyze the imputed datasets, the new method fully accounts for the designs by uncomplexing them and restoring a population in a separate step, therefore only simple, unweighted complete-data analysis techniques are needed for inferences with the newly developed combining rules. This reduces a lot burden on data users. Our findings in the simulation study suggest that the new method can bring about significant gains in bias relative to the existing model-based methods without losing any efficiency. Therefore, even for categorical variables, the nominal 95% confidence interval coverage rates under the new method are quite reasonable.

A direct application of the proposed approach is in the context of MI for disclosure risk limitation context. The approach of using MI framework to generate synthetic populations has become popular in protecting the confidentiality of respondents in the survey world, yet most are model-based or based on nonparametric method alone (i.e. Bayesian Bootstrap not accounting for designs), the semi-parametric approach as a combining use of the two serves as a good alternative, especially for complex survey designs.

Our current work focuses on more comprehensive simulation studies to assess the general performance of the proposed method, including its robustness to different degrees of model misspecification. We also aim to extend the application of our method in more complex sample design settings where both unequal probabilities of selection and clustering are involved.

<u>Appendix A.</u> Direct Derivation for the Two-step Combining Rules

This is achieved by constructing an approximate posterior distribution of Q given D_M^B in analogy with the standard theory of multiple imputation for missing data. The conceptual framework of the proposed procedure suggests that we have the following decomposition:

$$f(Q \mid D_M^B) = \int f(Q \mid D_M^B, D^B, V_\infty, V) \bullet f(D^B, V_\infty \mid D_M^B, V) \bullet f(V \mid D_M^B) dD^B dV_\infty dV$$

$$= \int f(Q \mid D^B, V_\infty) \bullet f(D^B, V_\infty \mid D_M^B, V) \bullet f(V \mid D_M^B) dD^B dV_\infty dV$$

Where V_∞ is variance of the posterior mean $q^{(b)}$ for each $D^{(b)}$ obtained when $B = \infty$, let $V^{(b)}$ be the variance of the $q_m^{(b)}$ obtained when $M = \infty$. Then V is the average of the $V^{(b)}$ obtained when $B = \infty$.

Step1: Combining rule from Synthesizing data by adapted FPBB: $f(Q \mid D^B, V_\infty)$

Let $Q = f(Y)$ be a scalar population quantity we are interested in. For example, it can be the population mean of survey variable Y, or it can also be a regression coefficient. Let q and u be the point and variance estimates for Q based on the actual sample data, $q^{(b)} = \{q^{(1)}, q^{(2)}, ..., q^{(B)}\}$ denotes the point estimates from all B FPBB populations. Adapted from the combining rules developed by Raghunathan et al.(2003) for fully synthetic data, the posterior mean and variance can be estimated as:

$$\overline{q^B} = \frac{1}{B}\sum_{b=1}^{B} q^{(b)} \qquad (6)$$

$T^B =$ Between synthetic variance + Within synthetic variance

$$= (1+\frac{1}{B})V^B + \frac{1}{B}\sum_{b=1}^{B} U^{(b)} = (1+\frac{1}{B})\frac{1}{B-1}\sum_{b=1}^{B}(q^{(b)} - \overline{q^B})^2 + \frac{1}{B}\sum_{b=1}^{B} U^{(b)} \qquad (7)$$

Where $U^{(b)} = T_M^{(b)}$ which is the between imputation variance for FPBB population b, and will be defined later in step 2.

Step2: Combining rules from Multiply imputing missing data: $f(D^B, V_\infty \mid D_M^B, V) \bullet f(V \mid D_M^B)$ or $f(Q \mid D_M^{(b)})$

Let $q_M^{(b)} = \{q_1^{(b)}, q_2^{(b)}, ..., q_M^{(b)}\}$ denote the point estimates from all multiple imputed datasets for FPBB b. Adapted from Rubin (1987)'s conventional combining rules for multiple imputation, the posterior mean and variance estimates can be written as:

$$\overline{q_M^{(b)}} = \frac{1}{M}\sum_{m=1}^{M} q_m^{(b)} \qquad (8)$$

$$T_M^{(b)} = U^{(b)} = \text{Between imputation variance} + \text{Within imputation variance}$$

$$= (1+\frac{1}{M})V_M + \frac{1}{M}\sum_{m=1}^{M} U_m = (1+\frac{1}{M})V_M = (1+\frac{1}{M})\frac{1}{M-1}\sum_{m=1}^{M}(q_m^{(b)} - \overline{q_M^{(b)}})^2 \qquad (9)$$

The within imputation variance disappears because at this step we are actually multiply imputing missing data for a population (FPBB is treated as a population) therefore no sampling variance is involved here, i.e. $U_m = 0$.

Two-step combining rules for inference: $f(Q \mid D_M^B)$

When we were deriving $f(Q \mid D^B)$ at the first step, $q^{(b)}$ was viewed as the sufficient summaries of FPBB b, after constructing the unbiased estimator $\overline{q_M^{(b)}}$ for each $q^{(b)}$ from generating M multiply imputed datasets nested within each FPBB population, we need to approximate $f(Q \mid D^B)$ with $f(Q \mid D_M^B)$ by substituting $q^{(b)}$ with their estimators $\overline{q_M^{(b)}}$ to obtain the posterior mean and variance for the two-step procedure as in formulae (4) and (5). Where V^B and V_M represents the between synthetic population variance and between multiple imputation variance, respectively. Notice that the within variance component at the first step is actually represented by the between variance from the second step.

References:

1. Cohen, Michael P. (1997). The Bayesian bootstrap and multiple imputation for unequal probability sample designs. ASA Proceedings of the Section on Survey Research Methods, 635-638.
2. Dong, Qi. (2011). Combining Information from Multiple Complex Surveys. University of Michigan. Unpublished Dissertation.
3. Efron, B. (1979). Bootstrap methods: another look at the jackknife. Annals of Statistics, 7, 1-26.
4. Elliott, M.R. (2007). Bayesian Weight Trimming for Generalized Linear Regression Models. Survey Methodology 33(1):23-34.
5. Gross, S. (1980). Median estimation in sample surveys, presented at the 1980 Joint Statistical Meetings.
6. Kim, J.K. (2004). Finite Sample Properties of Multiple Imputation Estimator. Ann. Statist. 32(2):766-783.
7. Kim, Jae Kwang, Michael Brick, J., Fuller, Wayne A. and Kalton, Graham (2006). On the bias of the multiple-imputation variance estimator in survey sampling. Journal of the Royal Statistical Society, Series B: Statistical Methodology, 68, 509-521.
8. Little, R.A. and Rubin, D.B. (2002). Statistical Analysis with Missing Data (Second Edition), New

York: J Wiley & Sons, New York.

9. Lo, Albert Y. (1988). A Bayesian Bootstrap for a Finite Population. The Annals of Statistics 16(4):1684-1695.
10. Meng, Xiao-Li. (1994). Multipl-Imputation Inferences with Uncongenial Sources of Imput. Statistical Science 9(4):538-558.
11. 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.
12. Raghunathan, T.E. Reiter, J.P. and Rubin, D.B. (2003). “Multiple Imputation for Statistical Disclosure Limitation.” Journal of Official Statistics 19(1):1-16.
13. Rao, J.N.K. and Wu, C.F.J. (1988) Resampling Inference with Complex Survey Data. Journal of the American Statistical Association 83:231-241.
14. Reiter, J.P. Raghunathan, T.E. and Kinney, Satkartar K. (2006). “The Importance of Modeling the Sampling Design in Multiple Imputation for Missing Data.” Survey Methodology 32(2):143-149.
15. Reiter, J.P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. Survey Methodology 30(2):235-242.
16. Rubin, D. (1981). The Bayesian bootstrap. Annals of Statistics, 9, 130-134.
17. Rubin, D.B. and Schenker, N. (1986). Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse. Journal of the American Statistical Association, 81(394):366-374.
18. Rubin, D. (1987). Multiple Imputation for Nonresponse in Surveys. New York: Wiley.
19. Rubin, D. (1996). Multiple Imputation After 18+ Years. Journal of the American Statistical Association, 91(434):473-489.
20. Schafer, J.L. (1999). Multiple Imputation: A Primer. Statistical Methods in Medical Research 8:3-15.
21. Schafer, J.L. Ezzati-Rice, T.M. Johnson, W. Khare, M. Little, R.J.A. and Rubin, D.B. (1997). The Nhanes III Multiple Imputation Project.
22. Schenker, Nathaniel, Raghunathan, T.E. Chiu, Pei-Lu, Makuc, D.M. Zhang, Guangyu and 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.
23. Yu, Mandi (2008). Disclosure Risk Assessments and Control. Dissertation. The University of Michigan.