DE GRUYTER OPEN

Journal of Official Statistics, Vol. 32, No. 1, 2016, pp. 231–256, http://dx.doi.org/10.1515/JOS-2016-0011

Synthetic Multiple-Imputation Procedure for Multistage Complex Samples

Hanzhi Zhou[1], *Michael R. Elliott*[2], *and Trivellore E. Raghunathan*[3]

Multiple imputation (MI) is commonly used when item-level missing data are present. However, MI requires that survey design information be built into the imputation models. For multistage stratified clustered designs, this requires dummy variables to represent strata as well as primary sampling units (PSUs) nested within each stratum in the imputation model. Such a modeling strategy is not only operationally burdensome but also inferentially inefficient when there are many strata in the sample design. Complexity only increases when sampling weights need to be modeled. This article develops a general-purpose analytic strategy for population inference from complex sample designs with item-level missingness. In a simulation study, the proposed procedures demonstrate efficient estimation and good coverage properties. We also consider an application to accommodate missing body mass index (BMI) data in the analysis of BMI percentiles using National Health and Nutrition Examination Survey (NHANES) III data. We argue that the proposed methods offer an easy-to-implement solution to problems that are not well-handled by current MI techniques. Note that, while the proposed method borrows from the MI framework to develop its inferential methods, it is *not* designed as an alternative strategy to release multiply imputed datasets for complex sample design data, but rather as an analytic strategy in and of itself.

Key words: Finite population Bayesian bootstrap; Haldane prior; stratified sample; clustered sample; sample weights.

1. Introduction

Stratified multistage sampling is the most common type of sample design for large-scale surveys conducted by the U.S. federal statistical agencies. This type of sample design combines the advantages of both stratification (for statistical efficiency) and cluster sampling (for cost and logistical efficiency). Under this design, the primary sampling units (PSUs) are stratified in such a way that they are homogeneous with respect to a stratum-level aggregate of the variable(s) of interest. To permit a maximum degree of stratification

[1] Mathematica Policy Research, Princeton, NJ 08543, USA. Email: zhouhanz@umich.edu.
[2] Dept. of Biostatistics, University of Michigan School of Public Health, 1415 Washington Heights, Ann Arbor, MI USA 48109; Survey Methodology Program, Institute for Social Research, University of Michigan, 426 Thompson St., Ann Arbor, MI 48109, USA. Email: mrelliot@umich.edu.
[3] Dept. of Biostatistics, University of Michigan School of Public Health, 1415 Washington Heights, Ann Arbor, MI USA 48109; Survey Methodology Program, Institute for Social Research, University of Michigan, 426 Thompson St., Ann Arbor, MI 48109, USA. Email: teraghu@umich.edu.
Acknowledgments: This work was supported in part by Grant Number R01CA129101 from the National Cancer Institute. The authors would like to thank Rod Little, Brady West, and Richard Valliant, along with the Associate Editor and three reviewers, for their review and helpful comments.

and thus variance reduction, it is common practice to define a large number of strata where only a small number of PSUs are selected in each stratum.

Multiple imputation (MI) (Rubin 1976, 1987) is a method commonly used when item-level missing data are present. However, MI requires that survey design information be built into the imputation models. Reiter et al. (2006) demonstrated the importance of simultaneously accounting for stratum effects and clustering effects in multiple imputation. They showed that when design features were ignored in the imputation model, biases would occur on the estimated parameter, even if a design-based analysis method was applied to the imputed data. Current MI methods typically include dummy variables to represent strata as well as PSUs nested within each stratum in the imputation model. When necessary, they also identify statistically significant interactions between these dummies with other covariates through routine variable selection procedures such as stepwise regression (Reiter et al. 2006; Schenker et al. 2006). Such a modeling strategy is not only operationally burdensome but also inferentially inefficient when there are hundreds of strata in the sample design and the sample in each stratum consequently becomes sparse. For example, the Census Bureau's Current Population Survey design groups 1,768 nonself-representing PSUs into 220 strata.

Possibly a better strategy is to consider clusters as random effects while treating strata as either fixed (using dummies) or random effects. However, many of the popular software packages that implement multiple imputation (e.g., SAS MI procedure, R packages *mice* or *mi*, and IVEware) cannot be adapted easily to such an approach. While a few recent software modules (such as R package *pan* and MLwiN module *REALCOM-IMPUTE*) have started to incorporate mixed effects or multilevel modeling for imputation purposes, they typically assume normal or latent normal distribution for variables with missing data. Their performances for missing categorical variables (binary in particular) are unclear. Moreover, there has been only little research that formally investigates their use in incorporating strata as well as clusters.

To circumvent these problems with fully parametric model-based imputation techniques, we develop a two-step semiparametric MI method. The idea is to separate the need to account for complex sample designs from the treatment of missing data. In the first step, sample designs are "reversed" through synthetic population data generation using a weighted finite population Bayesian bootstrap (FPBB) (Cohen 1997; Little and Zheng 2007; Dong et al. 2014). In the second step, missing values are imputed in the created synthetic population based on a parametric model that assumes identically independently distributed (IID) data elements. To account for stratum effects, we combine a replication variance estimation method (Efron 1979; Kovar et al. 1988; Rao and Wu 1988; Rao et al.1992; Rust and Rao 1996) with the weighted FPBB. Under a standard missing at random (MAR) assumption (Little and Rubin 2002), our method requires neither complicated modeling of strata and clusters nor design-based analyses of the imputed data. Note that while the proposed method borrows from the multiple-imputation framework to develop its inferential methods, it is *not* designed as an alternative strategy to release multiply imputed datasets for complex sample design data. Rather, it is intended an alternative analytic strategy for population inference from complex sample design data with item-level missingness.

In this article, we focus on the estimation of two quantities: quantile estimates for a continuous variable, and estimates of rare proportions and their associated logistic regression estimates. We consider a stratified two-stage sample design and investigate a full range of quantiles including tail behaviors. While design-based methods for quantile estimation from complex survey data have been developed (Francisco and Fuller 1991; Woodruff 1952), quantile estimation after imputation is less commonly addressed in the literature. (A recent exception that considers nonparametric fractional imputation outside of the complex sample design setting is Yang et al. 2013.) This is the case despite the rapid development and increasing popularity of MI. We also consider MI for incomplete binary variables, with a focus on rare outcomes. It is well known that maximum-likelihood estimation of logistic regression models typically suffers from small sample bias, the degree of which is strongly dependent on the number of sample cases in the less frequent of the two categories (King and Zeng 2001). Thus when the dependent binary variable represents the occurrence of rare events, the logistic regression coefficients can be substantially biased even with a simple IID data structure. Random effects logistic models are commonly used for fitting clustered binary data; however, these models rely heavily on asymptotic theory assumptions, which may not be met in sparse samples. All these issues might extend naturally to the missing-data context. As shown by Zhao and Yucel (2009), sequential MI for binary data missing completely at random in a multilevel setting suffers from severe bias and poor coverage in estimating probabilities that are close to 0 or 1, particularly when the intraclass correlation is high.

The objectives of this article are: i) to develop a two-step synthetic MI method as a way to simultaneously account for stratification, clustering, and unequal inclusion probability; and ii) to demonstrate the effectiveness of the new method with respect to quantile estimation and logistic regression for binary rare events data as compared with existing fully parametric imputation strategies. Section 2 discusses the imputation strategies under three different models: simple random sample, fixed effects for clusters/strata, and random effects for cluster/strata. Section 3 introduces the newly proposed procedure and the MI inference rules for quantile estimation under this method. Section 4 presents a Monte Carlo simulation study as the validation tool to assess the repeated sampling properties of MI under the various approaches. Section 5 applies different MI procedures to the analysis of body mass index on youth data from the third National Health and Nutrition Examination Survey (NHANES III). Some concluding remarks follow in Section 6. We focus on the two-PSU-per-stratum design in this chapter, although the methods we develop can accommodate any number of PSUs per stratum.

1.1. Fully Parametric Imputation Methods for the Two-PSU-per-Stratum Design

Here we briefly describe fully parametric multiple-imputation techniques with complex sample design features incorporated to different degrees. We assume the missing data Y_i is a member of the exponential family, and that there are fully observed covariates X_i (a $(p + 1)$-dimension vector) such that $g(E(Y_i|X_i)) = X_i\beta$ for a known link function $g(\cdot)$ (e.g., $g(u) = \log(u/(1 - u))$ for binary outcomes (logistic regression), $g(u) = \log(u)$ for count outcomes (Poisson regression), or $g(u) = u$ for continuous outcomes (Gaussian regression)).

1.1.1. Standard Regression Model Assuming SRS

Based on the maximum-likelihood estimates $\hat{\beta}$ and the associated asymptotic covariance matrix $\hat{V}(\hat{\beta})$ for the generalized linear model $g(E(Y_i|X_i)) = X_i\beta$, the posterior predictive distribution of the parameters can be constructed, which is then used to impute the missing values (Rubin 1987, 169–170). Point and variance estimates of the regression parameters can then be obtained using the usual MI combining rules (Rubin 1987, 76). For the p^{th} component of the regression parameter:

$$\hat{\beta}_p = \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}_p^{(m)}, \tag{1}$$

$$\hat{V}(\hat{\beta}_p) = \frac{1}{M}\sum_{m=1}^{M}\hat{V}\left(\hat{\beta}_p^{(m)}\right) + \frac{M+1}{M(M-1)}\sum_{m=1}^{M}\left(\hat{\beta}_p^{(m)} - \hat{\beta}_p\right)^2 \tag{2}$$

and

$$\frac{(\hat{\beta}_p - \beta_p)}{\sqrt{\hat{V}(\hat{\beta}_p)}} \dot{\sim} t_\nu,\ \nu = (M-1)\left(1 + \frac{\sum_{m=1}^{M}\hat{\beta}_p^{(m)}}{\frac{(M+1)}{(M-1)}\sum_{m=1}^{M}\left(\hat{\beta}_p^{(m)} - \hat{\beta}_p\right)^2}\right)^2 \tag{3}$$

where $m = 1, \ldots, M$ imputations are taken from draws widely separated to practically eliminate autocorrelation. Multivariate combining rules for the joint distribution of $\hat{\beta}$ are available as well (Schafer 1997, 112–118).

1.1.2. Fixed-Effects Model (FX_APR)

Compared to the predictive model using standard generalized linear regression, we can add dummy variables indicating stratum and cluster memberships to account for stratification and clustering effects. Note that we also need to include the log transformation of sampling weight as a predictor if the missing-data mechanism depends on weights to make the imputation model truly appropriate. The model takes the following form:

$$g\left(E(Y_i|X_i)\right) = X_i\beta + D_i\gamma + E_i\eta + [\zeta \log(w_i)], \tag{4}$$

where D_i is a $1 \times (H-1)$ row vector of dummies representing the H strata, and E_i is a $1 \times Q$ row vector of dummies representing the clusters nested within each stratum. Note that $Q = \sum_h Q_h - H$, where Q_h is the number of clusters in each stratum; in the case of the two-PSU-per-stratum case, $Q = H$. The parameter space under this model is expanded as $\theta = (\beta, \gamma, \eta, \zeta)$, and the steps for imputation are similar to those in the SRS setting.

1.1.3. Mixed-Effects Model (RE_APR)

As there are only two PSUs selected from each stratum, it is not feasible to model clusters as random effects separately within each stratum. Here we pool all $Q + H$ clusters in the sample and model them using a single random-effect term. The imputation model is

specified as follows:

$$g\big(E(Y_j|X_j)\big) = X_j\beta + D_j\gamma + u_i + [\zeta \log(w_j)], \tag{5}$$

where $u_i \sim N\big(0, \sigma_u^2\big)$ is a random intercept term representing cluster effects, for $i = 1, \ldots, (Q + H)$, and σ_u^2 denotes the between cluster variance. Other terms are as previously defined. (In the two-PSU-per-stratum case, $Q + H = 2H$.)

2. Synthetic MI Using the Weighted FPBB for Stratified Samples

In this section, we develop the two-step multiple-imputation methodology for a stratified two-stage sample design where a combination of complex sampling techniques are considered, namely, stratification, clustering, and unequal inclusion probability. We develop methods for an unrestricted number of clusters per stratum, but for our simulations and application we focus on the special case of two PSUs selected per stratum, which mimics the form of a public-use dataset that is commonly released for analyses.

2.1. Synthetic Data Generation to Account for Complex Sample Designs

Consider a finite population P, which is stratified into H strata with N_h PSUs in the h^{th} stratum, and hence the population size of PSUs is $\sum_{h=1}^{H} N_h = N$. For the h^{th} stratum, select n_h PSUs with/without replacement from some probability sampling plan, independently across strata, and hence the total sample size of PSUs is $\sum_{h=1}^{H} n_h = n$. Subsampling of m_{hi} elements (treated as the ultimate sampling units in this example) from a total of M_{hi} is then conducted within the i^{th} sampled PSU of the h^{th} stratum for $i = 1, \ldots, n_h, h = 1,2, \ldots, H$. Hence the overall sample size and population size of elements are $\sum_{h=1}^{H}\sum_{i=1}^{n_h} m_{hi} = \sum_{h=1}^{H} m_h = m$ and $\sum_{h=1}^{H}\sum_{i=1}^{N_h} M_{hi} = \sum_{h=1}^{H} M_h = M$, respectively, where m_h and M_h are the sample size and population size of elements for the h^{th} stratum, respectively. The population consists of four types of survey variables: a single outcome Y, a single covariate X, a design matrix $Z = [S,C,w]$ including the stratum indicators (S), the cluster indicator (C) and the sample weight (w), and the response indicator R. Let $D = (D_s, D_{ns}) = \{(Y_{hij}, X_{hij}, Z_{hij}, R_{hij}), h = 1, \ldots, H, i = 1, \ldots, N_h, j = 1, \ldots, M_{hi}\}$ denote the population of values measured on the survey variables, which is divided into the sampled component (D_s) and the nonsampled (D_{ns}) component.

We generate synthetic populations using a two-stage procedure. The first stage accommodates stratification and clustering and the second weighting. We have two broad approaches. The first, which we term SYN1, assumes that first-stage (cluster-level) and second-stage (element-level) sample weights are available for the analysis and implements a weighted FPBB at each level to generate the synthetic population. The second, which we term SYN2, assumes that only final weights are available for the analysis; it uses a Bayesian bootstrap to account for stratification and clustering at the first stage and the weighted FPBB to account for the final weight at the second stage.

2.1.1. Double-Weighted Finite Population Bayesian Bootstrap (SYN1)

For the h^{th} stratum, let $t_{s,h}$ and $t_{ns,h}$ index the sampled and nonsampled clusters, respectively, and $\{b^1, \ldots, b^q, \ldots, b^{r_h}, q = 1, \ldots, r_h\}$ be the r_h $(1 \le r_h \le N_h)$ distinct

matrices of real numbers each of dimension $|b^q_{row}| \times |b^q_{col}|$ with no row vectors in common. Each cluster in the stratum can take the form of one of b^qs. Let $t_{hi} = q$ when the i^{th} cluster takes on the values of b^q, for $i = 1, \ldots, N_h$. Assume $n_h = r_h$ and $m_{hi} = \|b^{t_{s,hi}}\|$ (the number of distinct row vectors in $b^{t_{s,hi}}$) for convenience of exposition. Let $w_{t_{s,h}}(i)$ be the sample weight of the i^{th} sampled cluster in the h^{th} stratum which equals b^q, for $i = 1, \ldots, n_h$. Also let $w_{t_{s,hi}, D_{s,h}}(j)$ be the sample weight of the j^{th} sampled element in the i^{th} sampled cluster which equals $b_k^{t_{s,hi}}$, for $j = 1, \ldots, m_{hi}$. Finally, let $c_{t_{s,h}}(q)$ and $c_{t_{ns,h}}(q)$ be the number of sampled and nonsampled clusters that equal b^q, and $c^{hi}_{t_h, D_{s,h}}(k)$ and $c^{hi}_{t_h, D_{ns,h}}(k)$ be the number of sampled and nonsampled elements that equal $b_k^{t_{s,hi}}$.

It can be shown (cf. Zhou 2014) that, *within a stratum h*, the Polya posterior for the counts of distinct unobserved elements $D_{ns,h}$ is given by

$$p\left(D_{ns,h}|D_{s,h}\right) = \frac{\left\{\prod_{q=1}^{r_h}\left\{\Gamma(w_{t'_h}(q))/\Gamma(w_{t_{s,h}}(q))\right\}\right\}}{\left\{\Gamma(N_h)/\Gamma(n_h)\right\}} \times \frac{\left\{\prod_{k=1}^{m_h}\left\{\Gamma(w_{t'_h, D_{ns,h}}(k))/\Gamma(w_{t_{s,h}, D_{s,h}}(k))\right\}\right\}}{\left\{\Gamma(M_h)/\Gamma(m_h)\right\}}, \tag{6}$$

where $w_{t'_h}(q) = w_{t_{s,h}}(q) + c_{t_{ns,h}}(q)$ and $w_{t'_h, D_{ns,h}}(k) = w_{t_{s,h}, D_{s,h}}(k) + c^{hi}_{t_h, D_{ns,h}}(k)$, for $m_h = \sum_{k=1}^{m_h} c^{hi}_{t_h, D_{s,h}}(k)$ and $m'_h = M_h - m_h = \sum_{k=1}^{m_h} c^{hi}_{t_h, D_{ns,h}}(k)$. The *full posterior* is then given by the product of the posteriors within each stratum, since these strata are independent and all strata in the population are in the sample:

$$p\left(D_{ns}|D_s\right) = \prod_{h=1}^{H} p\left(D_{ns,h}|D_{s,h}\right). \tag{7}$$

A Monte Carlo procedure to simulate from this posterior distribution is then given as follows:

(i) Draw the $N_h - n_h$ nonsampled clusters in the population based on the Polya posterior distribution independently for each stratum. Each of the sampled clusters is resampled with probability

$$s_{hi} = \frac{w_{t_{s,h}}(i) - 1 + l_{hi,k-1} \times \left(\frac{N_h - n_h}{n_h}\right)}{N_h - n_h + (k-1) \times \left(\frac{N_h - n_h}{n_h}\right)}, k = 1, \ldots, N_h - n_h + 1, \tag{8}$$

where $l_{hi,k-1}$ is the number of times that the i^{th} cluster in the h^{th} stratum has been resampled at the $(k-1)^{th}$ resampling, and $w_{t_{s,h}}(i)$ is the weight for the i^{th} sampled cluster in the h^{th} stratum which is normalized to sum up to the total number of clusters, that is, $\sum_{i=1}^{n_h} w_{t_{s,h}}(i) = N_h$.

(ii) From Step 1, form a population of clusters $\{c_{11}, c_{12}, \ldots, c_{1n_1}, \overset{*}{c}_{11}, \overset{*}{c}_{12}, \ldots, \overset{*}{c}_{1N_1-n_1}, \ldots, c_{H1}, c_{H2}, \ldots, c_{Hn_H}, \overset{*}{c}_{H1}, \overset{*}{c}_{H2}, \ldots, \overset{*}{c}_{HN_H-n_H}\}$. Record the number of times each of the clusters from the original sample appears in the FPBB population of clusters, denoted by $\tau_{hi}, i = 1, \ldots, n_h, h = 1, \ldots, H.$, and $\sum_{h=1}^{H}\sum_{i=1}^{n_h} \tau_{hi} = N$. Then update the within cluster *element-level conditional weights* as follows: $\overset{*}{w}_{j|hi} = w_{j|hi} \times \tau_{hi}$,

$i = 1, \ldots, n_h, h = 1, \ldots, H$, where $w_{j|hi}$ is the inverse of the conditional probability that element j is selected given cluster i in stratum h is selected. Now pool all elements from these clusters together and treat them as a single *FPBB sample* (i.e., as if they have no stratum or cluster boundaries). Note that this FPBB sample has the same sample size $m = \sum_{h=1}^{H}\sum_{i=1}^{n_h} m_{hi}$ as the original sample, but different sampling weights. We then once more apply the weighted FPBB to these pooled elements to generate $M - m$ units from the m units in the FPBB sample. We resample from each of the resampled clusters $M - m$ elements, cycling through $M - m$ times and resampling with probability

$$\lambda_{j|hi} = \frac{w^{*}_{j|hi} - 1 + l_{hij,k-1} \times \left(\frac{M-m}{m}\right)}{M - m + (k-1) \times \left(\frac{M-m}{m}\right)}, \quad k = 1, \ldots, (M - m + 1), \tag{9}$$

where $l_{hij,k}$ is the number of times that the j^{th} element in the i^{th} cluster in the h^{th} stratum has been resampled at the k^{th} resampling, and $w_{j|hi}$ is the updated conditional weight for the j^{th} element in the i^{th} cluster in the h^{th} stratum. Again, they are normalized to sum up to the total number of units in the entire population, that is, $\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{j|hi} = M$. Thus we create a single synthetic population. Repeat Step 2 B times to obtain B FPBB synthetic populations.

(iii) Repeat Steps 1-2 L times to obtain L bootstrap samples, yielding $L \times B$ FPBB populations $P^{syn}_{(lb)} = \left(P^{syn}_{(lb)obs}, P^{syn}_{(lb)mis}\right)$, $l = 1, \ldots, L$, $b = 1, \ldots B$, each of which consists of both responding elements and nonresponding elements on a vector of variables $\{Y,X,Z,R\}$.

2.1.2. Bootstrap — Weighted Finite Population Bayesian Bootstrap (SYN2)

Because we often do not know the first- and second-stage weights in public-use datasets, we consider an alternative to the procedure proposed in Subsection 2.1.1. Rather than obtaining a sample of clusters from a draw from a Polya posterior, we use replication methods (Rust and Rao 1996) to capture the cluster-level sampling variance. The final sampling weights instead of the adjusted element-level conditional weights are then used directly as input in the second-stage weighted FPBB. We use Rao and Wu's (1988) rescaling bootstrap, which is a generalized extension of McCarthy and Snowden's (1985) "with replacement bootstrap". Once the PSUs have been sampled, we continue with the weighted FPBB approach to complete the synthetic population data generation. The proposed procedure is as follows:

(i) Select a sample of $n^{*}_{h} = n_h - 1$ PSUs from the parent sample in each stratum via SRSWR sampling;

(ii) Apply the "ultimate cluster principle" (Wolter 2007), that is, once a PSU is taken into the bootstrap replicate, all elements in that PSU are taken into the replicate also. Thus we obtain our first bootstrap sample;

(iii) Repeat the previous steps L times to obtain L bootstrap samples $\{Boot\_l, l = 1, \ldots, L\}$;

(iv) Within each bootstrap sample, update the element-level sampling weights as:

$$w^{*}_{hij} = w_{hij} \times \left(\tau_{hi}\frac{n_h}{n^{*}_h}\right) = \begin{cases} = \frac{n_h}{n_h - 1} w_{hij}, \text{ if the } i^{th} \text{ PSU selected in the bootstrap sample} \\ = 0, \text{ if the } i^{th} \text{ PSU not selected in the bootstrap sample} \end{cases}$$

As w^{*}_{hij} itself implicitly carries over the strata and PSU information in addition to unequal inclusion probability, we can drop the subscripts *hi* henceforth by pooling all elements in the bootstrap sample regardless of which stratum and PSU they originally came from. Normalize w^{*}_{j}s to sum up to m^{*}: $\sum_{j=1}^{m^{*}} w^{*}_{j} = m^{*}$, where m^{*} is the bootstrap sample size.

(v) For the l^{th} bootstrap sample, $l = 1, \ldots, L$, apply the weighted FPBB algorithm to create an entire population $D = \left(D_{ns}, D^{*}_{s}\right)$ based on the posterior predictive distribution of elements in the nonsampled population $D_{ns} = \left\{(Y_j, X_j, Z_j, R_j), j = m^{*} + 1, \ldots, M\right\}$ given the elements in the bootstrap sample $D^{*}_{s} = \left\{(Y_j, X_j, Z_j, R_j), j = 1, \ldots, m^{*}\right\}$.

Operationally, we draw a Polya sample of size $M^{*} = M - m^{*}$ from $mult\left(M^{*}; \lambda_1, \ldots, \lambda_K\right)$ where the selection probability $\lambda_k, k = 1, \ldots, K$ is a function of w^{*}_{j}:

$$\lambda_k = \frac{w^{*}_{j} - 1 + 1_{j,k-1} \times \left(\frac{M^{*}}{m^{*}}\right)}{M^{*} + (k-1) \times \left(\frac{M^{*}}{m^{*}}\right)}, k = 1, \ldots, M^{*} + 1, \tag{10}$$

Repeat Step (v) for B times to obtain $L \times B$ FPBB populations.

2.2. Imputation of the Synthesized Populations

Once the set of FPBB synthetic populations $P^{syn} = \left\{P^{(l)}_{(b)}, l = 1, \ldots, L, b = 1, \ldots, B\right\}$, where $P^{(l)}_{(b)} = \left(Y^{(l)}_{(b)mis}, P^{(l)}_{(b)obs}\right)$ are created using either the SYN1 method or the SYN2 method, we generate imputations $P^{imp} = \left\{P^{(l)}_{(ba)}, l = 1, \ldots, L, b = 1, \ldots, B, a = 1, \ldots, A\right\}$ from the posterior predictive distribution $p\left(Y^{(l)}_{(b)mis} | P^{(l)}_{(b)obs}\right)$ based on a parametric model that does not condition on sample design features, that is, a model taking a form similar to the SRS model given in Subsection 2.1. We consider imputations based on the covariate (X) only (SYN1_srs or SYN2_srs) or imputations that include the log of the sample weights in the linear predictors (SYN1_lwt or SYN2_lwt).

To obtain the MI inference, denote the observed set of synthetic populations by $P_R = \left\{P^{(l)}_{(b)obs}, b = 1, \ldots, B, l = 1, \ldots, L\right\}$ and the imputed set of synthetic populations by $P_{\bar{R}} = \left\{Y^{(l)}_{(ba)mis}, l = 1, \ldots, L, b = 1, \ldots, B, a = 1, \ldots, A\right\}$. The MI point estimator for the population statistic of interest Q (mean, regression estimator, quantile) is then given by the mean of the lba^{th} point estimators:

$$\hat{Q}_{MI} = \frac{1}{LBA} \sum_{l} \sum_{b} \sum_{b} \hat{Q}_{lba}. \tag{11}$$

The MI variance estimator is:

$$\hat{V}_{MI} = (1 + L^{-1})V_L = (1 + L^{-1})\frac{1}{L-1}\sum_l \left(\hat{Q}_l - \hat{Q}_{MI}\right)^2, \text{ where} \tag{12}$$

$$\hat{Q}_l = \frac{1}{BA}\sum_b \sum_a \hat{Q}_{lba}.$$

We then construct the 95% interval estimate for quantiles based on t reference distribution with degrees of freedom equal to $\min\left\{v_{com} = \sum_h n_h - H, v_{syn} = L - 1\right\}$. These results arise from the fact that, by the standard Rubin (1987) MI combining rules, we have

$$Q|P^{imp} \dot{\sim} t_{L-1}\left(\bar{Q}_L, (1 + L^{-1})V_L\right), \tag{13}$$

where $\bar{Q}_L = \frac{1}{L}\sum_l \tilde{Q}^{(l)}$, $V_L = \frac{1}{L-1}\sum_l \left(\tilde{Q}^{(l)} - \bar{Q}_L\right)^2$, and $\tilde{Q}^{(l)} = \lim_{\substack{B\to\infty \\ A\to\infty}} \frac{1}{BA}\sum_b \sum_a \hat{Q}_{lba}$.

Replacing $\tilde{Q}^{(l)}$ with its finite simulation estimator $\hat{Q}_l$ replaces $\bar{Q}_L$ with $\hat{Q}_{MI}$ and gives the results above. A complete theoretical justification for (13) is provided in Dong et al. (2014) and Zhou (2014). Some intuition of the result can be gained by noting that the generation of the synthetic population sets the within imputation variance to 0 so that the posterior variance of Q can be obtained using the between-bootstrap variance only. Moreover, (11) assumes that $E(\hat{q}_{ba}) = Q$ – a result guaranteed by our Bayesian bootstrap estimator if the imputation model is also correct – as well as a sufficiently large sample size for the t approximation is reasonable.

Lo (1988) showed that the variance estimator for the FPBB mean in a simple random sample setting should be inflated by the factor $(\frac{n+1}{n-1})$. In the double-weighted FPBB (SYN1) setting, a small sample correction to the variance estimate thus needs to be used when the number of clusters per stratum is small. When $n_h = a$ is a constant across all strata, we use $\frac{n_h+1}{n_h-1}(1 + L^{-1})\, V_L$; otherwise we suggest $\frac{\bar{n}_h+1}{\bar{n}_h-1}(1 + L^{-1})\, V_L$, where $\bar{n}_h = H^{-1}\sum_h n_h$.

The Appendix provides the sample R code used to conduct the analyses in the application in Section 4 and can easily be adapted to other settings.

3. Simulation Study

We conducted a simulation study to investigate the performance of the proposed method for incorporating stratified cluster-sampling effects in multiple imputation. We targeted three population statistics: 1) population quantiles, 2) proportions of binary event data, and 3) logistic regression parameters relating the covariate to the binary data. The simulation is a 2 × 2 factorial design based on the following factors:

1) keeping the first-stage sampling plan constant, we let the subsampling rate f_2 of elements within sampled clusters be
   a) independent of or
   b) dependent on the stratum effects, and
2) assume that
   a) the missingness on the Y-variable (continuous or binary) depends only on the covariate (X) (MAR_X), or
   b) depends on both X and the final sampling weight W(MAR_X,W).

We focus on a two-PSU-per-stratum sample design, both because it is a common design, especially in public-use settings, and because it is a "limiting case" in terms of the number of PSUs per stratum. In addition to the two variants of our synthetic MI estimators, we consider standard parametric MI under the SRS, appropriate fixed-effect (FX_APR), and appropriate random-effect (RE_APR) models.

3.1. Data Generation

Let i be the index for strata, j be the index for clusters, and k be the index for elements. Suppose there are 50 strata in the population. First, the number of PSUs in each stratum is randomly determined according to a uniform distribution, that is, $C_i \sim Unif(2,54)$, $i = 1,\ldots,50$; second, the number of population elements within PSUs is randomly generated as $N_{ij} \sim Unif(20,80)$, $i = 1,\ldots,50, j = 1,\ldots,C_i$. Thus we obtain a population of size $N = 67385$. The complete data for four survey variables $Y = (Y_1,Y_2,Y_3,Y_4)^T$ are generated from a superpopulation model according to a two-step process, In the first step, Y_1 and Y_2 are randomly selected from a bivariate linear mixed-effects model; let $N_2(\cdot)$ denote a bivariate normal distribution function:

$$\begin{pmatrix} Y_{1ijk} \\ Y_{2ijk} \end{pmatrix} \sim N_2(\mu,\Sigma), \text{ where } \mu = \begin{bmatrix} \beta_1 + S_i + u_{1ij} + \varepsilon_{1ijk} \\ \beta_2 + u_{2ij} + \varepsilon_{2ijk} \end{bmatrix}, \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}. \quad (14)$$

Let $\beta_1 = \beta_2 = 15$ be the fixed covariate effects, $S_i = \frac{i}{5}$ be the fixed stratum effects, and let $\begin{bmatrix} u_{1ij} & u_{2ij} \end{bmatrix}^T$ and $\begin{bmatrix} \varepsilon_{1ijk} & \varepsilon_{2ijk} \end{bmatrix}^T$ be the random cluster effects and random error terms drawn from two independent bivariate normal distributions: $N_2(0,\Sigma_u)$ and $N_2(0,\Sigma_\varepsilon)$. Elements of Σ_u are set as: $\sigma^2_{u_1} = 4$, $\sigma^2_{u_2} = 1$, $\sigma_{u_1u_2} = 0.2$, and elements of Σ_ε are set as: $\sigma^2_{\varepsilon_1} = 4$, $\sigma^2_{\varepsilon_2} = 3$, $\sigma_{\varepsilon_1\varepsilon_2} = 1.732$. This results in conditional intraclass correlations (ICC) of Y_1 and Y_2 as $\rho_{Y_1} = 0.5$ and $\rho_{Y_2} = 0.25$ (note that the unconditional ICC for the two variables may be smaller than these values). In the second step, a random-effects logistic regression model (Anderson and Aitkin 1985; Stiratelli, et al. 1984) is used to simulate two binary outcome variables Y_3 and Y_4 as a function of Y_2. Under this model, a random effect is added to the linear part of the logistic regression model for each element in the cluster. The conditional mean of Y_{3ijk} and Y_{4ijk} is

$$\pi_{ijk} = E\left(Y_{\cdot ijk}|Y_{2ijk},u_{\cdot ij}\right) = \Pr\left(Y_{\cdot ijk} = 1|Y_{2ijk},u_{\cdot ij}\right) = \frac{e^{\alpha_0+\alpha_1 S_i+\alpha_2 Y_{2ijk}+u_{\cdot ij}}}{1+e^{\alpha_0+\alpha_1 S_i+\alpha_2 Y_{2ijk}+u_{\cdot ij}}}, \quad (15)$$

where $u_{3ij} \sim N(0,6^2)$, $u_{4ij} \sim N(0,10^2)$ and $\alpha = (\alpha_0,\alpha_1,\alpha_2)^T$ is the vector of fixed covariate effects. We fix $\alpha_2 = 1.5$ and vary α_0 and α_1 to obtain two different binary variables Y_{3ijk} and Y_{4ijk}, with either moderate $(\alpha_0 = -5,\alpha_1 = -1.5)$ or rare probabilities $(\alpha_0 = -8,\alpha_1 = -6)$. Given $u_{\cdot ij}$, the $Y_{\cdot ijk}s$ in the cluster are independent Bernoulli variables, that is, $Y_{\cdot ijk}|u_{\cdot ij} \sim Bern(\pi_{ijk})$.

Figure 1 shows the correlations between variables in the simulated population, with the different shades of grey representing different degrees of association between any of the two variables. The darker shades indicate higher correlation. All survey outcome variables (Y_1,Y_3,Y_4) have a moderate to strong (0.2∼0.8) stratum effect (H or *strID*) and clustering effect (U_1,U_3,U_4), indicating that accounting for these effects in the analysis of missing data is essential.

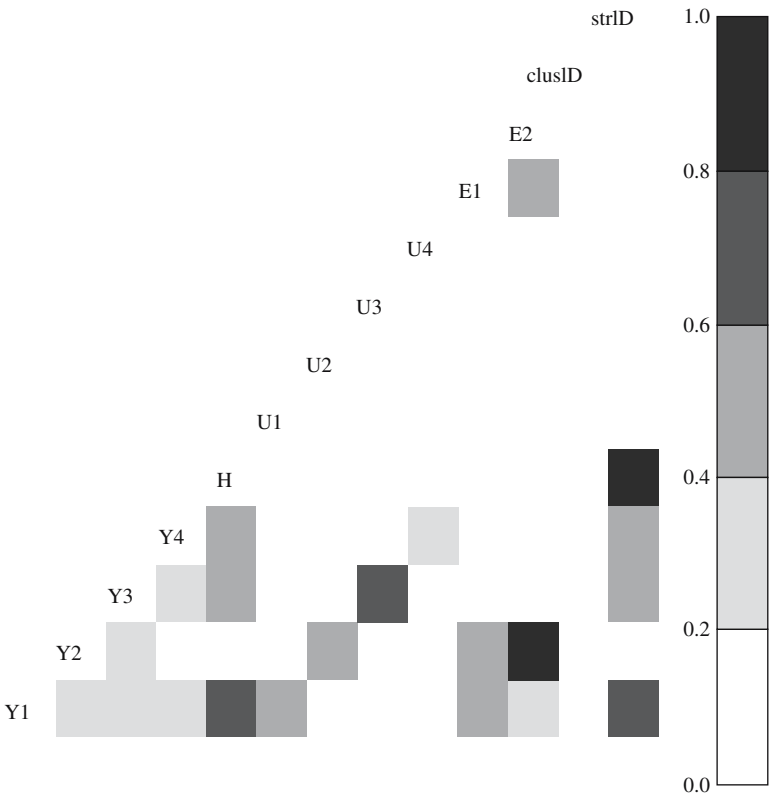

Fig. 1. Correlation between variables in the simulated population (darker shades = higher correlation)

3.2. Sample Design

Within each stratum, we draw a two-stage cluster sample according to the following procedure: first, we draw a sample of two PSUs without replacement with probability proportional to the cluster size $f_{1ij} = \frac{2^{*}N_{ij}}{\sum_{j} N_{ij}}$. Second, we sample elements from each sampled cluster using two different subsampling schemes:

1) sampling probability independent of S_i which is defined in (14): SRS with an equal sampling fraction of $f_{2k|ij} = 1/5$; and
2) sampling probability related to S_i: SRS with varying sampling fractions across strata, that is $f_{2k|ij} = \text{expit}(-0.8 - 0.12^{*}S_i)$, where $\text{expit}(x) = 1/(1 + \text{e}^{-1}(x))$.

An average of 1,122 elements are selected in each of the 200 simulation replications. The distributions of sampling weights are shown in Figure 2. The distributions of sampling weights under the two subsampling schemes are generally very similar with somewhat more skewness under subsampling scheme 2.

3.3. Imposing Missingness

Throughout the simulation study, we assume that Y_2 is always completely observed and we impose missing values on Y_1, Y_3, and Y_4 independently according to the following deletion

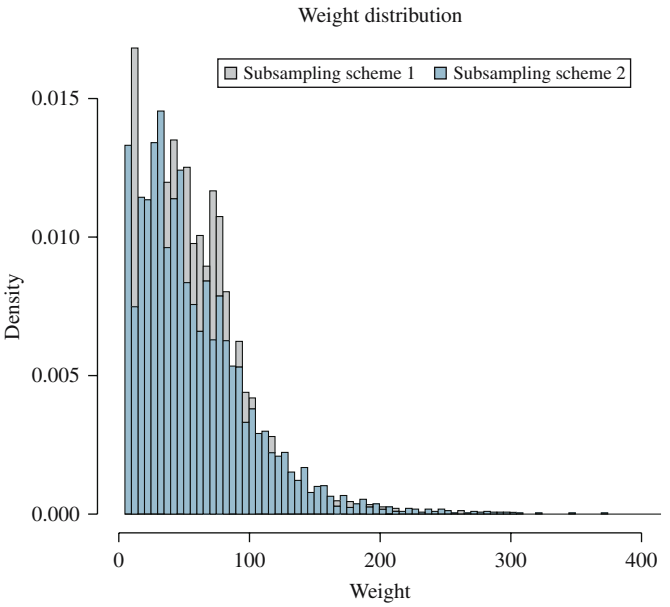

Fig. 2. Distribution of weights under the two subsampling schemes

function conditional on Y_2 and/or log transformation of the weight:

$$\Pr\big(R = 0|Y_2, W\big) = \frac{\exp\left(\lambda_0 + \lambda_1 * Y_2 + \lambda_2 * \log\left(W\right)\right)}{1 + \exp\left(\lambda_0 + \lambda_1 * Y_2 + \lambda_2 * \log\left(W\right)\right)}, \tag{16}$$

where R is the response indicator and W is the overall sample weight. Setting $\lambda_2 = 0$, we obtain the first MAR mechanism (i.e., MAR_X, note that we treat Y_2 as a covariate X here), under which we further set $\lambda_0 = 3.42$, $\lambda_1 = -0.2$ and $\lambda_0 = -2.58$, $\lambda_1 = 0.2$ for deleting values on Y_1 and Y_3, Y_4, respectively. Setting $\lambda_2 = -0.6$, we obtain the second MAR mechanism (i.e., MAR_X,W), under which we fix $\lambda_1 = 0.2$ and set two values on λ_0 ($= -0.274$ or -0.33) for deleting values independently on all three outcome variables under subsampling scheme 1 and subsampling scheme 2, respectively. All deletion functions result in approximately 40% missingness on each variable.

3.4. Parametric Multiple Imputation

Both simple random sample SRS (including SRS, SYN1_srs and SYN2_srs) and fixed-effects model FX_APR can be implemented in R (R Core Team 2013) using *mice* routines; for the logistic model associated with the binary outcome, the method '*logreg*' must be specified. We use the *pan* package in R for the mixed-effects imputation (RE_APR) for the missing continuous outcome; logistic mixed-effects imputation is programmed in SAS for the missing binary outcome, as there is no missing-data software package readily available for use.

3.5. Parameters of Interest and Inference

We focus on inference for the following population parameters: the mean of the continuous variable Y_1, the mean of the binary variables Y_3 and Y_4 (i.e., Bernoulli proportions), linear regression coefficients of Y_1 on Y_2, logistic regression coefficients of Y_3 (or Y_4) on Y_2, and the population percentiles of the continuous variable Y_1.

Weighted analyses and sandwich variance estimators accounting for strata and clusters are used to estimate smooth statistics (including proportions and regression parameters) under the three fully parametric MI methods. For estimating quantiles of the distribution of a continuous survey variable, we construct the sample-weighted point estimator with confidence intervals based on the test-inversion method (Francisco and Fuller 1991). We chose the test-inversion method instead of Woodruff's method (Woodruff 1952) despite the computational intensity, because the literature suggests that it may outperform Woodruff in heavily stratified samples or in small-to-moderate-sized samples (Kovar et al. 1988). Based on the a^{th} imputed dataset, the empirical distribution function can be written as

$$\hat{F}^{(a)}(y) = \frac{\left[\sum_{S_R} w_{hij} I\left(y_{hij}^{obs} < y\right) + \sum_{S_{\bar{R}}} w_{hij} I\left(y_{hij}^{(a)} < y\right)\right]}{\sum_{S} w_{hij}}, \tag{17}$$

where S_R and $S_{\bar{R}}$ are subsets of the sample data S, consisting of respondents and nonrespondents respectively. The estimator $\hat{F}(y)$ and its associated estimated variance $v(\hat{F}(y))$ can then be obtained using the variance estimator proposed by Francisco and Fuller (1991) together with standard Rubin combining rules as previously described. The sample γ^{th} quantile estimator thus is $\hat{q}_\gamma = (\hat{F})^{-1}(\gamma)$, with 95% asymptotic confidence interval (CI) given by

$$[L, U] = \left[[\hat{F}]^{-1}\left(\gamma - t_{0.025}\sqrt{\text{var}\left(\hat{F}(q_\gamma)\right)}\right), [\hat{F}]^{-1}\left(\gamma + t_{0.025}\sqrt{\text{var}\left(\hat{F}(q_\gamma)\right)}\right)\right]. \tag{18}$$

3.6. Results

Table 1 compares the average width $\times 10^{-2}$ and average coverage rates of the 95% CI of $q(\alpha)$, where $\alpha = 0.05, 0.10, 0.25, 0.50, 0.75, 0.90$, and 0.95, corresponding to seven selected population quantiles. Among all methods considered, the SRS imputation model yields the poorest coverage. This results from the compounding effects of biases and variance underestimation, due to ignoring stratum effects and clustering effects respectively. As we increase the dependence of both the sampling mechanism and response mechanism on stratum effects and sampling weights, the performance of SRS becomes even worse, as exhibited by the markedly increased RelBias and decreased coverage rates. In addition, ignoring stratum and/or weight effects that are highly relevant to either mechanism seems to impact the median and second and third quartiles more than the tail quantiles under SRS, as evident in the relatively lower coverage rates in the right part of Table 1.

Table 1. Comparison of average width $\times 10^{-2}$ and 95% CI coverage rates of $q(\alpha)$ for $\alpha = 0.05, 0.10, 0.25, 0.50, 0.75, 0.90,$ and 0.95.

Sampling scheme	Missingness mechanism	Methods	Average width of 95% CI x 10^{-2}							95% CI coverage						
			0.05	0.1	0.25	0.5	0.75	0.9	0.95	0.05	0.1	0.25	0.5	0.75	0.9	0.95
	Complete data	Actual	170	144	123	106	116	142	165	90.5%	92.5%	94.5%	93.5%	95.5%	91.5%	91.0%
		Syn1BD	172	144	126	105	117	146	165	90.5%	90.0%	95.0%	94.0%	95.0%	94.0%	89.0%
		Syn2BD	182	154	132	113	122	150	171	94.0%	95.0%	96.0%	94.5%	96.5%	96.0%	92.5%
$f_2 \propto$ const.	MAR_X	SRS	165	134	112	101	108	132	158	93.0%	91.5%	86.0%	82.5%	83.0%	89.0%	93.5%
		FX_APR	171	143	120	105	116	146	172	92.5%	90.5%	90.5%	92.5%	93.5%	94.0%	95.0%
		RE_APR	184	154	131	115	125	156	186	93.5%	94.0%	93.0%	97.5%	95.5%	95.5%	97.0%
		Syn1_srs	171	145	123	109	122	148	165	91.0%	89.5%	92.5%	95.0%	90.5%	89.5%	94.0%
		Syn2_srs	182	158	134	118	129	156	175	93.5%	93.0%	94.5%	96.5%	94.5%	94.5%	95.0%
	MAR_X,W	SRS	178	146	120	109	110	139	163	89.0%	81.0%	70.5%	69.0%	80.0%	90.0%	91.0%
		FX_APR	186	153	126	115	125	155	190	89.5%	92.5%	93.5%	95.5%	92.5%	92.5%	96.0%
		RE_APR	197	166	140	127	136	168	197	95.0%	97.0%	97.0%	98.0%	96.0%	95.0%	96.0%
		Syn1_srs	173	150	124	111	119	146	163	91.5%	92.0%	93.0%	91.5%	90.0%	94.0%	92.5%
		Syn2_srs	183	160	134	119	123	153	172	93.5%	95.5%	96.5%	92.5%	92.0%	93.0%	95.5%
		Syn1_lwt	174	151	126	115	124	148	166	90.0%	89.0%	93.0%	94.5%	90.5%	96.0%	94.0%
		Syn2_lwt	184	161	136	122	132	155	174	92.0%	93.0%	95.5%	96.0%	94.5%	96.0%	95.0%
	Complete data	Actual	170	143	120	110	121	148	169	92.5%	94.5%	95.0%	96.0%	92.5%	87.5%	87.5%
		Syn1BD	177	142	120	108	121	152	175	91.0%	92.5%	92.0%	94.5%	92.5%	87.5%	87.5%
		Syn2BD	182	152	128	116	126	154	178	95.0%	97.0%	96.0%	97.0%	94.5%	90.0%	90.5%
$f_2 \propto h(S_i)$	MAR_X	SRS	175	139	121	111	116	141	169	86.5%	73.0%	57.0%	48.5%	61.0%	72.0%	80.5%
		FX_APR	174	142	121	113	124	162	202	95.5%	95.0%	98.0%	95.5%	93.5%	92.5%	95.5%
		RE_APR	181	150	128	119	131	168	205	94.0%	96.5%	97.0%	96.5%	97.0%	94.0%	96.0%
		Syn1_srs	166	140	119	111	126	156	180	93.5%	94.0%	96.5%	92.5%	92.0%	91.0%	90.0%
		Syn2_srs	179	152	129	119	132	162	185	94.5%	95.5%	98.0%	96.5%	95.0%	93.5%	92.5%
	MAR_X,W	SRS	191	157	127	117	122	147	168	47.0%	31.5%	9.5%	8.0%	30.0%	60.0%	73.5%
		FX_APR	186	153	125	119	138	179	227	96.5%	97.0%	93.5%	96.5%	97.0%	95.0%	94.5%
		RE_APR	190	161	135	131	148	184	220	98.0%	99.5%	97.5%	98.0%	98.5%	96.5%	95.0%
		Syn1_srs	168	146	124	114	128	155	174	94.0%	92.5%	84.0%	73.0%	76.0%	87.5%	88.0%
		Syn2_srs	184	160	135	122	134	160	179	95.0%	95.5%	88.0%	77.5%	79.0%	87.0%	89.0%
		Syn1_lwt	168	143	121	113	130	158	176	92.5%	92.5%	94.5%	92.0%	89.0%	92.0%	91.5%
		Syn2_lwt	178	155	131	122	138	166	185	96.0%	95.5%	95.5%	92.5%	95.5%	95.0%	93.0%

The FX_APR model (Reiter et al. 2006; Rubin 1996; Schenker et al. 2006), generally performs fairly well in our simulation study with respect to the estimation of population quantiles. There is some modest underestimation of the small percentile quartiles with the second-stage sampling constant. The RE_APR model also performs well, with the exception of moderate to high overcoverage when the second-stage sampling probability is associated with the stratum mean and the missingness mechanism.

In contrast, our synthetic MI (SYN2 in particular) compares favorably with all of its competitors, and in most cases yields results comparable to the RE_APR, which is regarded as a "gold standard" as it is compatible with the data-generating mechanism (Meng 1994). There is some undercoverage when the stratified double-weighted FPBB estimator (SYN1) is used, perhaps due to the fact that the Lo small-sample adjustment is not as accurate when $n_h = 2$. However, use of a stratified bootstrap-weighted FPBB estimator (SYN2) generally eliminates this issue. Although an imputation model assuming SRS suffices for the synthetic MI method in most scenarios, we need to include the sampling weight as a predictor when the outcome Y and the response indicator R are strongly associated with each other through the sampling mechanism I, as is the case with the second subsampling scheme, when both the missingness indicator and the second-stage sampling rate are functions of the stratum mean.

Tables 2 and 3 compare the absolute relative bias $relbias = 100 \times \frac{|\hat{\theta} - \theta_{complete}|}{\theta_{complete}}\%$, RMSE and 95% nominal CI coverage for the estimated mean/proportions of Y_1, Y_3 and Y_4 and the slopes of the three outcome variables on Y_2, respectively. ($\theta_{complete}$ is the estimated parameter with complete data, and $\hat{\theta}$ is the estimated parameter under one of the different MI methods.) As in the estimation of the quantiles, the SRS imputation model is biased and has poor coverage as it ignores stratum and cluster effects. Again, dependence of subsampling on stratum effects and dependence of response on sampling weights damage the performance of SRS even further.

FX_APR generally performs well in estimating the mean of a continuous variable (Y_1) and a regular binary variable (Y_3) with moderate probability as well as the slopes. However, it fails for proportion estimation for rare events data (Y_4), yielding biased point estimates and less than nominal coverage throughout all scenarios. One interpretation might be that overfitting occurs when too many dummies are included to account for fixed strata and cluster effects, yielding dummy variables where all observed cases are 0 or 1. In this case, "complete separation" yields unstable coefficient estimates, damaging the predictive efficacy when the fitted model is used for drawing missing values. The problem is particularly prominent when the logistic fixed-effects imputation model is used along with the current sampling design, where an average of only ten elements are selected per PSU within each stratum; this results in even more substantial biases on $\bar{Y}_4$ than the SRS model. (Use of a Bayesian approach with an informative prior of the form $t_1(0,2.5)$ on the fixed-effect parameters using the *mi* function in R (Gelman et al. 2008) reduced but did not remove the impact of complete separation. A relative bias of 12–13% remained for the estimation of $\bar{Y}_4$ under the MAR_X missingness mechanism, with 95% nominal coverage of 89%, while a relative bias of 17–22% remained under the MAR_X,W mechanism, with nominal coverage of 84%.) The random-effects model RE_APR more effectively avoids the overfitting issue through shrinkage effects: note that under RE_APR, we pooled all PSUs from all

Table 2. Comparison of RelBias, RMSE and 95% CI coverage rates for the mean of Y1 and proportions of Y3 and Y4, Population true value: $\bar{Y}_1 = 20.4$, $P_{Y_3} = 0.608$, $P_{Y_4} = 0.117$

Sampling scheme	Missingness mechanism	Methods	RelBias			RMSE			95% CI coverage		
			$\bar{Y}_1$	P_{Y_3}	P_{Y_4}	$\bar{Y}_1$	P_{Y_3}	P_{Y_4}	$\bar{Y}_1$	P_{Y_3}	P_{Y_4}
	Complete data	Actual	–	–	–	0.220	0.042	0.024	95.0%	94.0%	90.5%
		Syn1BD	0.0%	0.0%	0.0%	0.221	0.042	0.024	94.5%	94.0%	91.5%
		Syn2BD	0.0%	0.0%	0.0%	0.222	0.043	0.024	95.0%	94.5%	93.0%
$f_2 \propto$ const.		SRS	0.8%	1.6%	10.8%	0.309	0.041	0.028	76.9%	90.0%	85.0%
		FX_APR	0.0%	1.3%	39.2%	0.243	0.040	0.054	91.0%	96.5%	72.5%
Actual samples BD:		RE_APR	0.0%	1.3%	15.1%	0.236	0.040	0.026	93.0%	93.5%	91.0%
$\bar{Y}_1 = 20.3$	MAR_X	Syn1_srs	0.0%	0.3%	0.4%	0.255	0.044	0.025	94.5%	93.5%	91.5%
$P_{Y_3} = 0.604$		Syn2_srs	0.0%	0.2%	0.4%	0.254	0.044	0.025	97.0%	95.0%	94.5%
$P_{Y_4} = 0.117$		SRS	1.4%	2.8%	19.4%	0.398	0.042	0.035	72.0%	85.5%	77.5%
		FX_APR	0.0%	2.7%	48.4%	0.260	0.042	0.065	91.5%	96.0%	60.0%
		RE_APR	0.1%	0.3%	6.8%	0.250	0.041	0.022	97.5%	95.5%	86.0%
		Syn1_srs	0.4%	1.4%	4.2%	0.285	0.043	0.026	92.0%	95.5%	91.5%
	MAR_X,W	Syn2_srs	0.5%	1.4%	4.4%	0.283	0.043	0.026	96.5%	95.0%	96.0%
		Syn1_lwt	0.0%	0.6%	0.3%	0.273	0.045	0.027	95.5%	93.5%	89.0%
		Syn2_lwt	0.0%	0.5%	0.0%	0.271	0.045	0.026	96.0%	96.0%	94.0%
	Complete data	Actual	–	–	–	0.218	0.037	0.023	96.0%	97.5%	92.0%
		Syn1BD	0.0%	0.0%	0.0%	0.220	0.037	0.023	93.5%	94.0%	92.0%
		Syn2BD	0.0%	0.0%	0.3%	0.219	0.038	0.023	96.0%	97.0%	94.0%
$f_2 \propto h(S_i)$		SRS	2.4%	4.7%	29.6%	0.540	0.048	0.045	42.0%	80.5%	62.5%
		FX_APR	0.0%	1.5%	42.0%	0.237	0.036	0.058	94.0%	97.0%	70.5%
Actual samples BD:		RE_APR	0.2%	1.6%	16.1%	0.230	0.039	0.025	96.5%	93.5%	91.5%
$\bar{Y}_1 = 20.4$	MAR_X	Syn1_srs	0.1%	0.0%	0.9%	0.266	0.042	0.025	92.5%	95.5%	91.5%
$P_{Y_3} = 0.609$		Syn2_srs	0.1%	0.1%	0.5%	0.266	0.042	0.025	94.0%	96.0%	93.5%
$P_{Y_4} = 0.117$		SRS	4.4%	9.2%	54.0%	0.912	0.067	0.071	6.5%	56.0%	34.5%
		FX_APR	0.1%	1.2%	55.3%	0.288	0.037	0.074	93.5%	95.5%	55.0%
		RE_APR	0.0%	0.7%	5.1%	0.239	0.038	0.022	97.5%	95.5%	87.0%
		Syn1_srs	1.5%	3.3%	15.0%	0.401	0.045	0.033	77.5%	91.5%	88.0%
	MAR_X,W	Syn2_srs	1.5%	3.2%	15.0%	0.400	0.045	0.033	82.0%	94.5%	91.5%
		Syn1_lwt	0.1%	0.2%	0.9%	0.281	0.042	0.025	89.5%	93.0%	91.0%
		Syn2_lwt	0.0%	0.1%	1.2%	0.278	0.043	0.025	93.5%	95.5%	92.5%

Table 3. Comparison of RelBias, RMSE and 95% CI coverage rates for the regression coefficients of Y1, Y3 and Y4 on Y2, Population true value: $\beta_{1,Y_1}|Y_2 = 0.488$, $\beta_{1,Y_3}|Y_2 = 0.227$, $\beta_{1,Y_4}|Y_2 = 0.083$

Sampling scheme	Missingness mechanism	Methods	RelBias			RMSE			95% CI coverage		
			$\beta_{1,Y_1\|Y_2}$	$\beta_{1,Y_3\|Y_2}$	$\beta_{1,Y_4\|Y_2}$	$\beta_{1,Y_1\|Y_2}$	$\beta_{1,Y_3\|Y_2}$	$\beta_{1,Y_4\|Y_2}$	$\beta_{1,Y_1\|Y_2}$	$\beta_{1,Y_3\|Y_2}$	$\beta_{1,Y_4\|Y_2}$
	Complete data	Actual	–	–	–	0.103	0.065	0.098	98.0%	96.0%	90.0%
		Syn1BD	0.4%	1.1%	1.9%	0.104	0.067	0.098	96.0%	93.5%	88.0%
		Syn2BD	0.2%	2.8%	5.0%	0.103	0.067	0.100	98.0%	97.5%	91.5%
$f_2 \propto$ const.		SRS	4.6%	4.6%	24.7%	0.110	0.071	0.100	93.0%	90.0%	91.0%
	MAR_X	FX_APR	0.2%	1.0%	44.7%	0.103	0.063	0.087	97.0%	97.0%	92.5%
Actual samples BD:		RE_APR	0.3%	2.1%	22.6%	0.100	0.056	0.068	98.0%	95.5%	95.0%
$\beta_{1,Y_1\|Y_2} = 0.481$		Syn1_srs	0.0%	0.5%	2.8%	0.114	0.079	0.111	95.5%	93.0%	88.0%
$\beta_{1,Y_3\|Y_2} = 0.232$		Syn2_srs	0.2%	3.0%	4.4%	0.115	0.082	0.111	96.5%	96.5%	94.5%
$\beta_{1,Y_4\|Y_2} = 0.086$	MAR_X,W	SRS	7.3%	7.5%	45.6%	0.121	0.070	0.100	93.0%	90.5%	87.0%
		FX_APR	0.4%	1.7%	53.5%	0.114	0.064	0.087	96.5%	96.0%	91.5%
		RE_APR	0.2%	6.5%	22.9%	0.105	0.054	0.073	97.5%	96.0%	96.0%
		Syn1_srs	3.6%	2.7%	9.7%	0.123	0.076	0.105	94.5%	91.5%	91.0%
		Syn2_srs	3.5%	0.5%	4.6%	0.121	0.076	0.107	96.5%	96.0%	93.0%
		Syn1_lwt	1.8%	1.4%	2.8%	0.121	0.075	0.104	95.5%	93.0%	90.0%
		Syn2_lwt	2.2%	1.5%	2.1%	0.120	0.075	0.106	96.5%	96.0%	96.5%
$f_2 \propto h(S_i)$	Complete data	Actual	–	–	–	0.108	0.066	0.088	95.0%	96.0%	95.0%
		Syn1BD	0.1%	0.6%	2.2%	0.109	0.068	0.089	95.0%	95.0%	93.0%
		Syn2BD	0.4%	2.9%	6.5%	0.109	0.069	0.090	95.0%	96.5%	96.0%
Actual samples BD:	MAR_X	SRS	12.8%	9.1%	52.0%	0.136	0.074	0.096	89.5%	90.0%	88.0%
$\beta_{1,Y_1\|Y_2} = 0.481$		FX_APR	0.5%	0.6%	43.5%	0.114	0.069	0.079	93.5%	95.0%	97.0%
$\beta_{1,Y_3\|Y_2} = 0.229$		RE_APR	0.8%	2.5%	19.0%	0.111	0.061	0.065	95.0%	95.5%	97.0%
$\beta_{1,Y_4\|Y_2} = 0.090$		Syn1_srs	0.4%	0.7%	5.6%	0.126	0.082	0.097	94.0%	92.0%	91.5%
		Syn2_srs	0.0%	3.5%	2.7%	0.124	0.082	0.098	95.0%	94.0%	96.5%
	MAR_X,W	SRS	17.6%	12.4%	69.5%	0.141	0.069	0.101	86.0%	94.0%	83.0%
		FX_APR	0.4%	5.7%	42.2%	0.118	0.066	0.082	93.5%	95.5%	55.0%
		RE_APR	1.7%	3.1%	30.1%	0.111	0.054	0.073	97.5%	98.0%	97.5%
		Syn1_srs	6.7%	3.1%	23.0%	0.136	0.073	0.093	93.0%	94.0%	94.5%
		Syn2_srs	7.4%	0.4%	19.0%	0.136	0.075	0.095	96.0%	97.5%	97.0%
		Syn1_lwt	0.9%	0.9%	6.0%	0.130	0.075	0.092	93.0%	95.5%	93.5%
		Syn2_lwt	1.7%	2.6%	3.3%	0.126	0.076	0.094	97.0%	98.0%	97.5%

strata as if there were no strata bounds, and the stratum effects can be thought as being implicitly modeled in the random intercept term ($u_j = I_h + u_{h(j)}$).

As in the quantile estimation setting, our synthetic MI compares favorably with all of its competitors, and in most cases yields comparable results to the RE_APR for estimation of means and logistic regression parameters. In the case of rare events data, our proposed new method increases the analytical size through generating synthetic population data thus is even superior to RE_APR, consistently yielding negligible biases and close to nominal coverage. The impact of ignoring the weights in the imputation (under MAR_X,W mechanism) is less than in the quantile estimation setting, with the exception of the estimation of the continuous mean $\bar{Y}_1$, where including the weight is required to obtain approximately correct coverage.

A disadvantage of the method lies in its relative inefficiency for estimating nonlinear parameters (regression coefficients) (e.g., the synthetic MI results in unbiased point estimates but a larger RMSE than the two model-based MI methods). This is typical in that nonparametric methods cannot typically compete with their fully parametric counterparts under the correct model, and is a tradeoff made to improve robustness to model misspecification.

4. Application to NHANES III

We apply our method to the National Health and Nutrition Examination Survey (NHANES) III (1988–1994), which is designed to provide national estimates of the health and nutritional status of the civilian noninstitutionalized population of the United States aged two months and older (National Center for Health Statistics 1996). The data are obtained from a stratified, multistage area probability sampling design with oversampling of certain age and ethnicity groups. For confidentiality and computational reasons, the public-use data provides two pseudo-PSUs per stratum. Another unique feature of NHANES is that data are collected through both interview and actual physical examinations of the sampled persons. Both unit- and item-level nonresponse occurs in both components of the survey, and there is a particularly high missing rate on the body mass index (BMI) measure for youth data in the physical examination component (30%). As a popular measure of overweight status and obesity, the percentiles of BMI for children and youths are of particular interest for public health reasons. The upper percentiles and the lower percentiles are also closely monitored for overweight and underweight status, respectively. As a result, we restrict our analysis sample to children and youths from two months to 16 years of age. The Appendix provides the sample R code used to conduct the analyses below.

We estimate population quantiles (from 0.05 to 0.95 with an increment of 0.05 along with two extreme percentiles: 0.03 and 0.97) of BMI for children and youths by gender. We also estimate the proportion of such a population being covered by health insurance, overall and by race. To assure congenial inference, we include the following variables that are either of primary interest in the substantive analysis or are important predictors for BMI measures in the imputation model: age, gender, race, education, mother's BMI, father's BMI and family income (Yuan and Little 2007). We compared three different methods in our treatment of the missing data:

1) complete case analysis (CC) with design-based estimation;
2) fully parametric model-based MI using design-based estimation, within which we apply both an imputation model assuming SRS and the appropriate model conditional on all three sample design features (i.e., dummy variables indicating cluster and stratum memberships as well as the log transformation of sampling weights); and
3) our proposed finite population Bayesian bootstrap method (using SYN2 since we do not have separate weights for the first and second stages of sampling), and including the log of the weight in the imputation model.

Estimates of the median BMI and the proportion of children with health insurance are given in Table 4. The CC method appears to overestimate the median of both the BMI measure and health-insurance coverage for the full sample and race domains relative to the MI approaches, and yields the widest confidence intervals or largest standard errors as a result of decreased sample size. Then again, the median of BMI obtained from synthetic MI is quite similar to that from the model-based MI, while demonstrating some advantages in efficiency by yielding shorter intervals. The generally lower health-insurance coverage estimates under the synthetic MI relative to model-based MI might be attributable to the fact that the synthetic MI are able to capture certain interactions between the sample design variables and the regular covariate matrix which are not explicitly modeled in the fully model-based MI.

Figure 3 displays a visual comparison of the percentile estimation for the three methods under consideration. We look at how those methods perform in three different percentile ranges by gender domains: the middle percentiles from 0.5 to 0.75, the upper percentiles from 0.90 to 0.97 and the lower percentiles from 0.03 to 0.1. We chose these percentile ranges because the extreme lower and upper percentiles of BMI are typically used to monitor under- and overweight for children and youths, and there is evidence that gender difference exists in these BMI percentile ranges (particularly when age is considered, i.e., growth patterns in BMI). In general, both MI methods result in very similar BMI estimates, and they are lower than those obtained from CC analysis. This makes sense since our comparison of the distributions of age for complete cases and for missing cases on the BMI measure revealed that younger children are more susceptible to missingness, and therefore CC analysis tends to overestimate BMI by excluding those younger missing cases. The inclusion of the age variable as a predictor in the imputation model corrects such an

Table 4. Alternative methods in estimating the median of BMI and the health-insurance coverage rate, for full sample and by gender and race, respectively

		Methods		
Variable	Domain	CC	Model-based MI	Synthetic MI
BMI	Overall	17.2 [17.1, 17.4]	17.1 [16.9, 17.3]	17.0 [16.9, 17.2]
	Male	17.2 [16.9, 17.4]	17.0 [16.7, 17.2]	17.0 [16.8, 17.2]
	Female	17.3 [17.0, 17.7]	17.1 [16.8, 17.4]	17.1 [16.8, 17.3]
Health insurance	Overall	0.785 (0.020)	0.778 (0.019)	0.761 (0.019)
	White	0.822 (0.018)	0.815 (0.017)	0.799 (0.016)
	Nonwhite	0.645 (0.036)	0.643 (0.033)	0.634 (0.036)

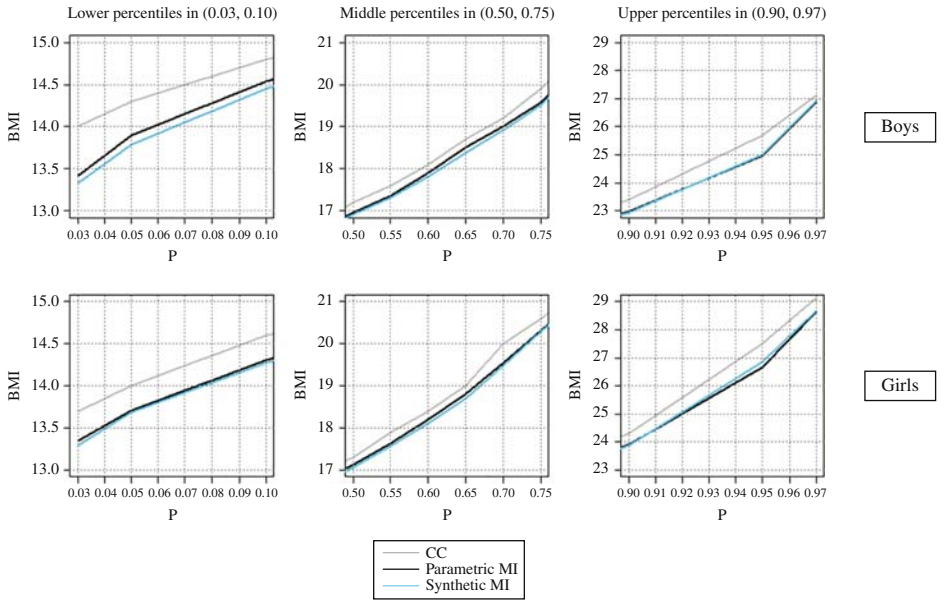

Fig. 3. Comparison of methods for quantile estimation of BMI, by gender

overestimation. The magnitude of this correction for boys is bigger than that for girls in estimating the lower percentiles (0.03, 0.05). When examining a report on BMI-for-age percentiles by gender released by the Center for Disease Control and Prevention (http://www.cdc.gov/nchs/data/series/sr_11/sr11_246.pdf), we find that baby boys (corresponding to the lower quantiles here) have a relatively higher BMI, which might be at least part of the explanation.

5. Discussion

While multiple imputation has become a popular option for the analysis of missing data, some issues remain unresolved in its practical application to complex sample survey data. The complex features of sampling compounded with nonresponse in survey data often result in a rather complicated data structure, which prevents the straightforward application of the standard MI techniques (such as a multivariate normal model assuming simple random sampling). In this article, we develop a general-purpose approach to account for various design features in a highly stratified two-stage sample using a two-step synthetic MI framework. We have focused on evaluating the performance of the new method compared with existing methods with respect to several missing-data issues frequently encountered in large population-based socioeconomic and epidemiological studies. These include: i) accommodating stratification and multistage sampling in the imputation process; ii) the employment of nonstandard or non-normal imputation models for estimating probabilities of rare events; and iii) the estimation of population quantiles with multiply imputed data. (For examples that consider alternative sample designs, such as independent unequal probability of selection designs, or cluster and weighted designs without stratification, as well as estimators of quantities such as means and linear regression parameters, see Zhou (2014).

Although multiple imputation is technically valid only for maximum-likelihood estimates (Kim et al. 2006), we demonstrate that the coverage properties of the proposed method are fairly good for nonsmooth statistics. Specifically, our stratified variations of the weighted Polya posterior exhibits robustness to the loss function for estimating the upper and lower tails of the distribution function where even the appropriate model-based method (i.e., FX_APR) fails. In contrast with existing fully parametric MI methods, most of which perform poorly when applied to rare outcome binary data, the proposed method yields quite stable parameter estimates regardless of the rarity of the outcome. An alternative approach for MI estimation of quantiles that relies on estimating the CDF using a smooth regression curve is given by Wei et al. (2012), and could be used at the second-stage imputation step after the weighted finite population Bayesian bootstrap has been implemented.

It is worth stressing that our method requires only the most straightforward form of imputation modeling and combining rules for inference. This is because the effects of the complex sample design and the effect of estimating the nuisance parameters in imputation (e.g., regression parameters when the main quantity of interest is a quantile of Y) are both correctly reflected in the replication variance estimation given the design-reversed and multiply imputed synthetic populations. Any higher-level and nonlinear interactions in the covariate data, including those with the weights, clusters, or strata, will automatically be captured in the synthesizing step. However, when the imputation is conducted parametrically, as it is here, such design-variable interactions will still need to be considered if they are associated with the missingness mechanism, although the impact of misspecification will generally be attenuated. Similarly, not-missing-at-random mechanisms that are dependent on the missing values are not accommodated in this framework. Finally, we note that assuming SRS for imputation results in correct inference only at the population level: correct inference for domain estimation requires that the domains be included in the imputation model. For example, if variables X and Y are positively correlated in stratum A but negatively correlated in stratum B, this interaction will be correctly averaged over for the population inference using weighted FPBB, but if this interaction is of direct interest, it will be attenuated unless incorporated in the imputation model for the synthetic population. Further, imputing under SRS does not absolve the imputer from correctly modeling the data. To give a trivial example, assume data are sampled from two strata denoted by $Z = \{1,2\}$, where $P(Z = 1) = P(Z = 2) = .5$ in the population, and $Y|Z = 1 \sim N(5,1)$ and $Y|Z = 2 \sim N(-5,1)$, and stratum 1 is oversampled with $P(I|Z = 1) \propto .8$ The method proposed here will correct the imbalance between the strata, and assuming a two-component normal mixture model will allow imputations of Y that maintain the correct marginal distribution of Y with equal-sized components. This will allow for correct estimation of percentiles, whereas simply assuming a unimodal normal distribution will only consistently estimate the mean. Correct estimation of percentiles *within* the strata will require also conditioning on the strata, as mentioned above. We note that one advantage of the proposed method is that, with design issues cleared out of the way, more focus can be given to developing missing-data models.

We also note that the method developed here does *not* allow for the release of a small number of multiply imputed datasets to be combined using the standard Rubin rules. It *would* be possible to publically release all $L \times B \times A$ multiply imputed datasets to be analyzed using the methods developed here, although this would typically involve

hundreds to thousands of datasets. Methods to allow a more modest release, with minimal impact on inference, are a topic for future research.

Future research will investigate the inferential properties of the proposed method in situations where auxiliary information on all population units is available, using a constrained version of the Polya posterior. Two other possible research directions include: (i) extending the two-step synthetic MI framework to deal with unit nonresponse problems, and (ii) extending it to deal with generating synthetic data for disclosure risk limitation.

Appendix: R Code for Using the Proposed Two-step MI Method on NHANES III

```
require(survey)
require(mice)
require(polyapost)
set.seed(seed #)

syn_bmi < -function(dt, N, Bt1, Bt2, Mt){

##Step 1: Generate synthetic populations with missing data;
#Stage 1: Create bootstrap samples from the parent sample;
        dsgn <- svydesign(ids = ~ predcl, strata = ~ pstrat, nest = TRUE, data =
        dat, weights = ~ predwt)
        dsgn.RW<-as.svrepdesign(design = dsgn, type = "subbootstrap", replicates
        = Bt1)
        dim(dsgn.RW$repweights)
        repwt<-as.matrix(dsgn.RW$repweights)
        repwt[repwt = =0]<-NA
        dim(repwt)

        #set up arrays to hold point estimates from bootstrap samples;
        btm<-matrix(0,nrow = Bt1,ncol = 3)
        btqt<-matrix(0,nrow = Bt1,ncol = 21)
        btqtm<-matrix(0,nrow = Bt1,ncol = 21)
        btqtf<-matrix(0,nrow = Bt1,ncol = 21)

        for (j in 1:Bt1){
                st.bb < -cbind(dat,repwt[,j])
                #delete those units with zero weights for each bootstrap sample;
                st.BB < -na.omit(st.bb)
                #recode those 999 back to NA so that the mice package can be used for
                imputation;
                st.BB$pybmi[st.BB$pybmi = = 999] < -NA

                #need to calculate the replicate weights;
                Samwt < -st.BB[,9]*st.BB[,13]
                #normalize again the adjusted weights;
                Samwts < -Samwt*N/sum(Samwt)
                np < -nrow(st.BB)
```

```
        ids < -seq(np)
        ns < -N-np

##Stage 2: Create unweighted synthetic populations within each bootstrap sample;
#Set up arrays to hold point estimates from imputed unweighted synthetic populations;
    fbm < -matrix(0,nrow = Bt2,ncol = 3)
    fbqt < -matrix(0,nrow = Bt2,ncol = 21)
    fbqtm < -matrix(0,nrow = Bt2,ncol = 21)
    fbqtf < -matrix(0,nrow = Bt2,ncol = 21)

    for(boott in 1:Bt2){
        l < -vector()
        smp < -wtpolyap(ids, Samwts, ns)
        #input the adjusted weights in the weighted Polya sampling algorithm;
        for (k in 1:np){
        l < -c(l,length(smp[smp = = k]))
        }
      #check if the vector of l sums up to the number of synthetic population size;
      sum(l);

      predY1 < -c(rep(st.BB[,1],l)) #bmi
      predY2 < -c(rep(st.BB[,2],l)) #race
      predY3 < -c(rep(st.BB[,3],l)) #gender
      predY4 < -c(rep(st.BB[,4],l)) #income
      predY5 < -c(rep(st.BB[,5],l)) #education
      predY6 < -c(rep(st.BB[,6],l)) #mother's bmi
      predY7 < -c(rep(st.BB[,7],l)) #father's bmi
      predY8 < -c(rep(st.BB[,8],l)) #age
      predwt1 < -c(rep(st.BB[,9],l))
      predlwt < -log(predwt1) #log of sample weight
      predCID < -c(rep(st.BB[,12],l)) #cluster ID
      predSTID < -c(rep(st.BB[,11],l)) #stratum ID

##Step 2: Multiple imputation of the unweighted synthetic populations;

    #use the imputation model including log of weight as a predictor (syn_lwt);
    temp1 < -data.frame(cbind(predY1, predY2, predY3, predY4, predY5, predY6,
    predY7, predY8, predlwt))
    temp1_imp < -mice(temp1,method = "norm", m = Mt)
    ml < -complete(temp1_imp, 'long')
    ml$bmit < -exp(ml$predY1) #back transform bmi to its normal scale
    mlmale < -subset(ml, predY3 = = 1)
    mlfem < -subset(ml, predY3 = = 2)
    multm < -cbind(as.vector(by(ml$bmit,ml$.imp,mean)),
    as.vector(by(mlmale$bmit,mlmale$.imp,mean)),
    as.vector(by(mlfem$bmit,mlfem$.imp,mean)))
```

```
        multqt < -sapply(with(ml,by(ml,.imp,function(x)quantile(x$bmit,
        c(0.03,seq(0.05,0.95,0.05),0.97)))),as.vector)
        multqtm < -sapply(with(mlmale,by(mlmale,.imp,function(x)quantile(x$bmit,
        c(0.03,seq(0.05,0.95,0.05),0.97)))),as.vector)
        multqtf < -sapply(with(mlfem,by(mlfem,.imp,function(x)quantile(x$bmit,
        c(0.03,seq(0.05,0.95,0.05),0.97)))),as.vector)
            fbm[boott,] < -t(apply(multm,2,mean))
            fbqt[boott,] < -t(apply(multqt,1,mean))
            fbqtm[boott,] < -t(apply(multqtm,1,mean))
            fbqtf[boott,] < -t(apply(multqtf,1,mean))
            print(boott)
            }

        btm[j,] < -t(apply(fbm,2,mean))
        btqt[j,] < -t(apply(fbqt,2,mean))
        btqtm[j,] < -t(apply(fbqtm,2,mean))
        btqtf[j,] < -t(apply(fbqtf,2,mean))
        print(j)
        }

        smpm < -apply(btm,2,mean)
        smpv < -(1 + 1/Bt1)*apply(btm,2,var)
        smpse < -sqrt(smpv)
        smpqt < -apply(btqt,2,mean)
        smpqtv < -(1 + 1/Bt1)*apply(btqt,2,var)
        smpqtse < -sqrt(smpqtv)
        smpqtm < - apply(btqtm,2,mean)
        smpqtvm < -(1 + 1/Bt1)*apply(btqtm,2,var)
        smpqtsem < -sqrt(smpqtvm)
        smpqtf < -apply(btqtf,2,mean)
        smpqtvf < -(1 + 1/Bt1)*apply(btqtf,2,var)
        smpqtsef < -sqrt(smpqtvf)

tt < -cbind(smpqt,smpqtm,smpqtf,smpqtse,smpqtsem,smpqtsef)
ss < -cbind(smpm,smpse)
write.table(tt,file = "D:/Dissertation/paper3/nhanes/synbmiqt_lwt.csv",row.
names = FALSE,sep = ",")
write.table(ss,file = "D:/Dissertation/paper3/nhanes/synbmimn_lwt.csv",
row.names = FALSE,sep = ",")
}

##Example##
syn_bmi(dt = dt, N = 100000, Bt1 = 50, Bt2 = 5, Mt = 5)
dt < -read.csv("D:/Dissertation/paper3/nhanes/synbmi.csv")
#Set the synthetic population size about 10 times the sample size;
N < -100000
```

```
#Normalize the weights to sum up to the assumed synthetic population size;
dt[,"predwt"] < -dt[,"predwt"]*N/sum(dt[,"predwt"])
sum(dt$predwt)
#Recode the missing values to 999;
dat[is.na(dat)] < -999
```

6. References

Anderson, D. and M. Aitkin. 1985. "Variance Component Models With Binary Response: Interviewer Variability." *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 47: 203–210.

Cohen, M. P. 1997. "The Bayesian Bootstrap and Multiple Imputation for Unequal Probability Sample Designs." In Proceedings of the Section on Survey Research Methods, American Statistical Association (ASA), Anaheim, CA, 1997, 635–638.

Dong, Q., M.R. Elliott, and T.E. Raghunathan. 2014. "A Nonparametric Method to Generate Synthetic Populations to Adjust for Complex Sample Design." Survey Methodology 40: 29–46

Efron, B. 1979. "Bootstrap Methods: Another Look at the Jackknife." *Annals of Statistics* 7: 1–26.

Francisco, C.A. and W.A. Fuller. 1991. "Quantile Estimation With a Complex Survey Design." *Annals of Statististics* 19: 454–469.

Kim, J.K., M.J. Brick, W.A. Fuller, and G. Kalton. 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. Doi: http://dx.doi.org/10.1111/j.1467-9868.2006.00546.x.

King, G. and L. Zeng. 2001. "Logistic Regression in Rare Events Data." *Political Analysis* 9: 137–163.

Kovar, J.G., J.N.K. Rao, and C.F.J. Wu. 1988. "Bootstrap and Other Methods to Measure Errors in Survey Estimates." *Canadian Journal of Statistics* 16: 25–45.

Little, R.J. and D.B. Rubin. 2002. *Statistical Analysis with Missing Data*, (2nd ed.). New York: Wiley and Sons, New York.

Little, R.J. and H. Zheng. 2007. "The Bayesian Approach to the Analysis of Finite Population Surveys." *Bayesian Statistics* 8: 283–302.

Lo, A.Y. 1988. "A Bayesian Bootstrap for a Finite Population." *The Annals of Statistics* 16: 1684–1695.

McCarthy, P.J., and C.B. Snowden. 1985. "*The Bootstrap and Finite Population Sampling. Vital and Health Statistics*." Data Evaluation and Methods Research, Series 2, No. 95. Public Health Service Publication 85–1369, U.S. Government Printing Office, Washington

Meng, X.L. 1994. "Multiple Imputation Inferences With Uncongenial Sources of Input." *Statistical Science* 9: 538–558. Doi: http://dx.doi.org/10.1214/ss/1177010269.

National Center for Health Statistics. 1996. *Analytic And Reporting Guidelines: The Third National Health and Nutrition Examination Survey, NHANES III (1988–94)*. National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville,

Maryland. Available at: http://www.cdc.gov/nchs/data/nhanes/nhanes3/nh3gui.pdf (accessed May 22, 2014)

Rao, J.N.K. and C.F.J. Wu. 1988. "Resampling Inference With Complex Survey Data." *Journal of the American Statistical Association* 83: 231–241. Doi: http://dx.doi.org/10.2307/2288945.

Rao, J.N.K.C.F., J. Wu, and K. Yue. 1992. "Some Recent Work on Resampling Methods for Complex Surveys." *Survey Methodology* 18: 209–217.

Reiter, J.P., T.E. Raghunathan, and S.K. Kinney. 2006. "The Importance of Modeling the Sampling Design in Multiple Imputation for Missing Data." *Survey Methodology* 32: 143–149.

Rubin, D.B. 1976. "Inference and Missing Data." *Biometrika* 63: 581–592.

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

Rubin, D.B. 1996. "Multiple Imputation After 18+Years." *Journal of the American Statistical Association* 91: 473–489. Doi: http://dx.doi.org/10.2307/2291635.

Rust, K. and J.N.K. Rao. 1996. "Variance Estimation for Complex Estimators in Sample Surveys." *Statistics in Medical Research* 5: 381–397.

Schafer, J.L. 1997. *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall.

Schenker, N., T.E. Raghunathan, P. Chiu, D.M. Makuc, G. Zhang, and A.J. Cohen. 2006. "Multiple Imputation of Missing Income Data in the National Health Interview Survey." *Journal of the American Statistical Association* 101: 924–933. Doi: http://dx.doi.org/10.1198/016214505000001375.

Stiratelli, R., N. Laird, and J. Ware. 1984. "Random-Effects Models for Serial Observations With Binary Response." *Biometrics* 40: 961–971. Doi: http://dx.doi.org/10.2307/2531147.

Wei, Y., Y. Ma, and R.J. Carroll. 2012. "Multiple Imputation in Quantile Regression." *Biometrika* 99: 423–438. Doi: http://dx.doi.org/10.1093/biomet/ass007.

Wolter, K.M. 2007. *Introduction to Variance Estimation*. New York: Springer.

Woodruff, R. 1952. "Confidence Interval for Medians and Other Position Measures." *Journal of the American Statistical Association* 47: 635–646. Doi: http://dx.doi.org/10.1080/01621459.1952.10483443.

Yang, S., J.K. Kim, and D.W. Shin. 2013. "Imputation Methods for Quantile Estimation under Missing at Random." *Statistics and Its Interface* 6: 369–377.

Yuan, Y. and R.J. Little. 2007. "Parametric and Semiparametric Model-Based Estimates of the Finite Population Mean for Two-Stage Cluster Samples With Item Nonresponse." *Biometrics* 63: 1172–1180. Doi: http://dx.doi.org/10.1111/j.1541-0420.2007.00816.x.

Zhao, E. and R.M. Yucel. 2009. "Performance of Sequential Imputation Method in Multilevel Applications." In Proceedings of the Section on Survey Research Methods, American Statistical Association ASA, August, Washington D.C., 2800–2810.

Zhou, H. 2014. "Accounting for Complex Sample Designs in Multiple Imputation Using the Finite Population Bayesian Bootstrap." Unpublished PhD Thesis

Received June 2014
Revised April 2015
Accepted April 2015