Biometrics 72, 242–252
March 2016

DOI: 10.1111/biom.12413

A Two-Step Semiparametric Method to Accommodate Sampling Weights in Multiple Imputation

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

[1]Mathematica Policy Research, Inc., Princeton, New Jersey
[2]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.
[3]Institute for Social Research, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.
*email: zhouhanz@umich.edu
**email: mrelliot@umich.edu
***email: teraghu@umich.edu

Summary. Multiple imputation (MI) is a well-established method to handle item-nonresponse in sample surveys. Survey data obtained from complex sampling designs often involve features that include unequal probability of selection. MI requires imputation to be congenial, that is, for the imputations to come from a Bayesian predictive distribution and for the observed and complete data estimator to equal the posterior mean given the observed or complete data, and similarly for the observed and complete variance estimator to equal the posterior variance given the observed or complete data; more colloquially, the analyst and imputer make similar modeling assumptions. Yet multiply imputed data sets from complex sample designs with unequal sampling weights are typically imputed under simple random sampling assumptions and then analyzed using methods that account for the sampling weights. This is a setting in which the analyst assumes more than the imputer, which can led to biased estimates and anti-conservative inference. Less commonly used alternatives such as including case weights as predictors in the imputation model typically require interaction terms for more complex estimators such as regression coefficients, and can be vulnerable to model misspecification and difficult to implement. We develop a simple two-step MI framework that accounts for sampling weights using a weighted finite population Bayesian bootstrap method to validly impute the whole population (including item nonresponse) from the observed data. In the second step, having generated posterior predictive distributions of the entire population, we use standard IID imputation to handle the item nonresponse. Simulation results show that the proposed method has good frequentist properties and is robust to model misspecification compared to alternative approaches. We apply the proposed method to accommodate missing data in the Behavioral Risk Factor Surveillance System when estimating means and parameters of regression models.

Key words: Bayesian bootstrap; Behavioral Risk Factor Surveillance System (BRFSS); Missing data; Polya posterior; Sampling design.

1. Introduction

Both item nonresponse and sampling weights are typical features of survey data obtained from complex sample designs. Item nonresponse occurs when some respondents do not answer all the items in a survey questionnaire, e.g., both "don't know" and refusal answers are considered as item nonresponse. Sampling weights arise as a correction factor to compensate for over- or under-representation of units in the target population due to unequal selection probabilities. The Behavior Risk Factor Surveillance System (BRFSS) has both a substantial proportion of missing data on income measures as well as survey weights that adjust for different sampling rates among states and oversampling of adults in smaller sized households, as well as for nonresponse bias by poststratifying and raking to known control totals for basic demographics.

When the proportion of item-level missing values is nontrivial and the data are not missing completely at random (MCAR), typical solutions for missing data like complete case analysis often lead to increased bias and reduced statistical power. Multiple imputation (MI) (Rubin, 1987) is a principled method for addressing item-level missing data. MI has a Bayesian conceptualization. The basic idea is to fill in missing data with M sets of plausible values. These are obtained as repeated draws from the posterior predictive distribution of the missing components of the sample Y_{mis} given its observed components Y_{obs}, i.e., $p(Y_{\text{mis}}|Y_{\text{obs}})$. The production of multiple "completed" data sets $\left\{(Y_{\text{obs}}, Y_{\text{mis}}^{(1)}), \ldots, (Y_{\text{obs}}, Y_{\text{mis}}^{(M)})\right\}$ is typically done by an "imputer" who has access to the data to develop reasonable models for generating the predictive distribution of Y_{mis}, allowing the "analyst" to then analyze each of the M imputed data sets and combine the point and variance estimates using the combining rules developed by Rubin (1987). Examples of this approach include imputation for blood alcohol concentration in the Fatal Accident Reporting System (FARS) (Heitjan and Little, 1991) and income imputation in the National Health Interview Survey (NHIS) (Schenker et al., 2006).

While the imputer/analyst distinction is convenient, Meng (1994) pointed out that this can lead to problems with inference when the imputer and analyst assume different data models ("uncongeniality"). Meng shows that, when the imputer assumes a richer model than the analyst, the resulting

MI analysis would typically be mildly conservative, whereas if the analyst assumed a richer model than the imputer, the resulting MI analysis could be either conservative or anticonservative. In settings where the observed data are obtained using an unequal probability sampling design, data are typically imputed using models in which variables are assumed to be independent and identically distributed (IID), and then analyzed using a weighted design-based approach that accounts for the unequal selection probability. This presents an example of the latter form of uncongeniality that can lead to biased point estimation and below-nominal confidence interval coverage. It is therefore important to incorporate unequal probabilities of selection/sampling weights in the imputation procedure.

A simple and seemingly straightforward way to incorporate sampling weights in MI is to let the imputer's model condition on a few key design variables that determine probabilities of inclusion, such as measure of size and stratification variables. However, not all design information is typically available in public use data due to disclosure risk concerns. Another option is to summarize the design information by using weights as a covariate in the imputation, perhaps after log transformation or categorization in "weight strata" and modeling them as dummy indicators. However, the modeling task may be complicated by attempting to include all interactions of weights (or weight-related design variables) with other covariates in the model (Meng, 1994; Kim et al., 2006). Moreover, this approach typically requires the functional form of the interaction to be modeled correctly, using a spline or other nonparametric form to be robust against model misspecification (Elliott and Little, 2000; Zheng and Little, 2005; Breidt, Claeskens, and Opsomer, 2005). In addition, Kim et al. and Seaman et al. (2011) show that, in the case of a target complete data estimator of a weighted total, the standard Rubin MI variance formula is no longer an unbiased even if the imputation model is correctly specified without the weights, since the weights induce a covariance between the MI point estimator and the (latent) complete data estimator, a quantity that is not accounted for in the Rubin MI variance formula; typically this covariance is negative, so that the standard MI variance estimators and associated confidence intervals are conservative, but complex adjustment must be made to regain nominal p-values and coverage.

This article develops a modified MI framework to account for sampling weights from single-stage designs. We propose a two-step MI procedure. In the first step, we develop and use a weighted finite population Bayesian bootstrap (weighted FPBB) to validly impute the whole population (including item nonresponse) from the observed data. In the second step, having generated posterior predictive distributions of the entire population, we use standard IID imputation to handle the item nonresponse. Our suggested procedure allows the parametric imputation model to no longer need to model interactions between weights and covariates in the imputation regression model to account for model misspecification. In addition, since we are imputing to a synthetic population, all weights are constant and equal to 1, so no covariance between the MI point estimator and the complete data estimator is induced.

The rest of this article is organized as follows. Section 2 provides a detailed overview of the proposed two-step semiparametric multiple imputation procedure to accommodate weighted data. (We term it "semiparametric" because the design features, in particular the weights, are accommodated nonparametrically, whereas the actual imputation is conducted under a standard parametric model.) We focus on the setting where the selection probabilities are obtained from a probability proportional to size (PPS) sample design, although the methods we develop can be used with any selection weights. Section 2 then discusses point estimation and inference using the MI data sets from the proposed procedure. Section 3 provides a simulation study in the context of a single-stage probability-proportional-to-size sample design to estimate population means and regression coefficients under a variety of settings where sampling weights are associated to differing degrees with both the outcome and the probability of nonresponse, and where failure to account for design in the imputation procedure has differing degrees of impact. We compare the performances of the proposed two-step MI and the fully parametric MI in terms of robustness to different degrees of model misspecification. Section 4 applies the proposed procedure to estimate means, linear, and log-linear regression models, describing marginal and joint distributions of income and health insurance accessibility, using data from the 2009 Behavioral Risk Factor Surveillance System (BRFSS). Section 5 concludes with a brief discussion of possible extensions.

2. A Two-Step Semiparametric MI Procedure

Bayesian finite population inference (Ericson, 1969) has been proposed as a means to harmonize design and model-based approaches for sample survey inference (Little, 2004, 2011). Under this approach, we focus on the posterior predictive distribution of our finite population quantity of interest (e.g., population mean, population regression parameter) obtained from the posterior predictive distribution for the nonsampled elements of the population. To make matters more concrete, consider the setting in the absence of missing data where we have a scalar outcome Y, sampling weight w based on a single stage PPS design, and no missing data. Our complete data consists of the vector of sampling indicators I for the population, sampled Y_s for which $I = 1$, the nonsampled Y_{ns} for which $I = 0$, and similarly w_s and w_{ns}. Given the sampling weights, the sampling mechanism generating I is assumed to be independent of Y ($p(I|Y, w) = p(I|w)$), and thus ignorable in the modeling. Assuming a model for the outcome given the sampling weights $p(Y|\theta, w)$ parameterized by θ with prior $p(\theta)$, the posterior predictive distribution for the nonsampled elements of the population Y_{ns} is given by

$$p(Y_{ns}|Y_s, w_s) \propto \int p(Y_{ns}|Y_s, \theta, w)\, p(\theta|Y_s, w)\, p(w_{ns}|w_s)\, d\theta\, dw_{ns} \tag{1}$$

Previous work has tackled estimation of this predictive distribution in a variety of ways. Zheng and Little (2004, 2005) and Chen, Little, and Elliott (2010) assumed that the sam-

pling weights were known for all subjects, so that $w_s = w$, reducing (1) to $p(Y_{ns}|Y_s, w) \propto \int p(Y_{ns}|Y_s, \theta, w)p(\theta|Y_s, w)d\theta$; these authors then obtained draws from the posterior predictive distribution under fairly weak modeling assumptions (parametric regression model for $p(Y|\theta, w)$ based on penalized splines). Little and Zheng (2007) and Zangeneh, Keener, and Little (2011) considered the situation in which weights are observed only for the sample (as in a public use data setting), and obtained predictive draws for $p(w_{ns}|w_s)$ under a Dirichlet model with a noninformative (Haldane) prior; the resulting predictive draw of the population of weights was then used as in Zheng and Little to obtain posterior predictive draws of Y_{ns}. Dong, Elliott, and Raghunathan (2014) consider a different factorization of (1):

$$p(Y_{ns}|Y_s, w_s) \propto \int p(Y_{ns}, w_{ns}|Y_s, w_s)p(Y_s, w_s)dw_{ns}, \qquad (2)$$

The parameter θ is dropped because the draws of Y_s, w_s are made directly from the posterior of the empirical CDF of Y_s, w_s using a Bayesian bootstrap (BB) procedure (Rubin, 1981). Draws from $Y_{ns}, w_{ns}|Y_s, w_s$ are then made using a weighted finite population Bayesian boostrap (FPBB) procedure described in Cohen (1997).

Here, we extend the approach of Dong, Elliott, and Raghunathan to accommodate missing data due to item-level nonresponse. We assume that, had we taken a census of the entire population, we could have observed a vector of response indicators $R = (R_s, R_{ns})$, where R_s corresponds to the response indicators observed in the sample, and R_{ns} to the response indicators associated with the nonsampled elements. We then divide the sampled $Y_s = (Y_{s,\text{obs}}, Y_{s,\text{mis}})$ into the fully observed and missing elements, corresponding to the sampled Y values associated with $R_s = 1$ and $R_s = 0$, respectively, and similarly the nonsampled $Y_{ns} = (Y_{ns,\text{obs}}, Y_{ns,\text{mis}})$ into those that would have been observed had they been sampled ($R_{ns} = 1$), and those that would have had missing values ($R_{ns} = 0$). We also assume a fully observable covariate $X = (X_s, X_{ns})$ consisting of the sampled and nonsampled elements, respectively. Note that we can combine the observed from the sampled and nonsampled parts of the population to obtain the potentially "observable" $Y_{\text{obs}} = (Y_{s,\text{obs}}, Y_{ns,\text{obs}})$, and similarly those missing $Y_{\text{mis}} = (Y_{s,\text{mis}}, Y_{ns,\text{mis}})$. We assume ignorable missingness, so that $p(R|Y, w) = p(R|Y_{\text{obs}}, w)$, allowing R to be ignored in the model along with I. Extending (1) to incorporate item-level missingness then yields

$$p(Y_{ns,\text{obs}}, X_{ns}|Y_{s,\text{obs}}, X_s, w_s) = \int p(Y_{ns,\text{obs}}, X_{ns}, Y_{\text{mis}}|Y_{s,\text{obs}}, X_s, w_s)dY_{\text{mis}}$$

We can generate from $p(Y_{ns,\text{obs}}, X_{ns}, Y_{\text{mis}}|Y_{s,\text{obs}}, X_s, w_s)$ by simply allowing the missing values in Y to be generated along with the observed values for Y and X using the weighted FPBB procedure. We then integrate out with respect to Y_{mis} by assuming a parametric model for $Y|X$:

$$\begin{aligned}
&\int p(Y_{ns,\text{obs}}, X_{ns}, Y_{\text{mis}}|Y_{s,\text{obs}}, X_s, w_s)dY_{\text{mis}} = \\
&\int p(Y_{\text{mis}}|Y_{ns,\text{obs}}, X_{ns}, Y_{s,\text{obs}}, X_s, w_s)p(Y_{ns,\text{obs}}, X_{ns}|Y_{s,\text{obs}}, X_s, w_s) \\
&\quad \times dY_{\text{mis}} = \int\int p(Y_{\text{mis}}|Y_{ns,\text{obs}}, X_{ns}, Y_{s,\text{obs}}, X_s, w_s, \theta) \\
&\quad \times p(Y_{ns,\text{obs}}, X_{ns}|Y_{s,\text{obs}}, X_s, w_s, \theta)p(\theta|Y_{s,\text{obs}}, X_s, w_s)d\theta dY_{\text{mis}} \propto \\
&\int\int p(Y_{\text{mis}}|Y_{ns,\text{obs}}, X_{ns}, Y_{s,\text{obs}}, X_s, w_s, \theta) \\
&\quad \times p(Y_{ns,\text{obs}}, X_{ns}|Y_{s,\text{obs}}, X_s, w_s, \theta)p(Y_{s,\text{obs}}, X_s, w_s|\theta)p(\theta)d\theta dY_{\text{mis}}
\end{aligned} \qquad (3)$$

We can implement the integration in (3) by use of a standard Gibbs sampler for multiple imputation that iterates between draws of

$$p(\theta|Y_{ns,\text{obs}}, X_{ns}, Y_{s,\text{obs}}, X_s, w_s, Y_{\text{mis}}) = p(\theta|Y, X, w_s) = p(\theta|Y, X) \qquad (4)$$

and

$$p(Y_{\text{mis}}|Y_{ns,\text{obs}}, X_{ns}, Y_{s,\text{obs}}, X_s, w_s, \theta) \qquad (5)$$

Note that (4) follows from the fact that, conditional on the entire population, the observed weights are superfluous for the draws of θ, so that it is sufficient to develop a parametric model for Y that does not involve the weights together with a prior for θ (possibly conditional on X): $p(\theta|Y, X, w_s) = p(\theta|Y, X) \propto p(Y|\theta, X)p(\theta|X)$. The presence w_s in (5) indicates that the observed weights may still be important in the imputation of the missing elements of Y if missingness itself is a function of the probability of selection, as we note below.

2.1. *Step 1: Undo Sampling Weights through Nonparametric Synthetic Data Generation*

Here, we briefly review the work of Dong, Elliott, and Raghunathan (2014) to obtain draws from a posterior predictive distribution of the population that is free of the effects of unequal probability of selection. This work builds on the work of Ghosh and Meeden (1983), Lo (1988), and Cohen (1997), where details of the derivations of the results can be found.

2.1.1. The weighted Pólya posterior. The purpose of developing the weighted Polya posterior is to be able to draw from a posterior predictive distribution of a finite population based on an unequal probability-of-selection sampling design without making any parametric assumptions about the probability mechanism that generated the data. We begin by describing the Polya posterior developed by Ghosh and Meeden (1983) in the simple random sampling setting. Assume that a simple random sample of size n is drawn from a finite population of size N, denoted by $y_s = \{y_1, ..., y_n\}$. Let $\Gamma(\bullet)$ denote the gamma function, $\{d_1, d_2, \ldots, d_K\}$ denote the set of K distinct values in the sample and $\lambda = \{\lambda_1, \lambda_2, ..., \lambda_K\}$

denote the vector of probabilities that $\Pr(y_i = d_k|\lambda) = \lambda_k$, for $i = 1, 2, \ldots, n$, with $\sum_{j=1}^{K} \lambda_j = 1$. Let n_j and u_j be the number of units taking value d_j in the sample and in the nonsampled part of the population, respectively, for $j = 1, 2, ..., K$, and $\sum_{j=1}^{K} n_j = n, \sum_{j=1}^{K} u_j = N - n$. Assuming a noninformative Haldane prior of λ, $\lambda \sim \text{Dir}(0, \ldots, 0)$, together with a multinomial distribution for the counts of sample data, $n_1, .., n_K|\lambda \sim \text{Mult}(n; \lambda)$, Ghosh and Meeden show that predictive distribution of counts in the nonsampled data is given by the following:

$$p(u_1, \ldots, u_K|n_1, \ldots, n_K) = \frac{\prod_{j=1}^{K} \Gamma(n_j + u_j)/\Gamma(n_j)}{\Gamma(N)/\Gamma(n)}. \quad (6)$$

Cohen (1997) generalized (6) to the case where the sample is selected with unequal probabilities. We now assume that we have a sample of size n consisting of $(Y_s, X_s, w_s, R_s) = \{(Y_i, X_i, w_i, R_i), i = 1, ..., n.\}$, where R is a response indicator for Y, so that $Y_i = Y_{i,\text{obs}}$ if $R_i = 1$ and $Y_i = Y_{i,\text{mis}}$ if $R_i = 0$, X consists of fully observed covariates, and w_i denotes the sampling weight for the ith unit in the sample, which is normalized to sum up to N, i.e., $\sum_{i=1}^{n} w_i = N$. Let $\{\tilde{d}_1, \tilde{d}_2, \ldots, \tilde{d}_K\}$ denote the set of K distinct vectors of (Y_i, X_i, w_i, R_i) in the sample and $\varsigma = \{\varsigma_1, \varsigma_2, \ldots, \varsigma_K\}$ denote the vector of probabilities that $\Pr\left((Y_i, X_i, w_i, R_i) = \tilde{d}_k|\varsigma\right) = \varsigma_k$, for $i = 1, 2, \ldots, n, k = 1, \ldots, K$, and $\sum_{j=1}^{K} \varsigma_i = 1$. Let n_j and u_j be the number of units taking vector $\tilde{d}_j$ in the sample and in the nonsampled part of the population, respectively, for $j = 1, 2, \ldots, K$, and $\sum_{j=1}^{K} n_j = n, \sum_{j=1}^{K} u_j = N - n$. Again assuming a noninformative Haldane prior of ς: $\varsigma \sim \text{Dir}(0, \ldots, 0)$ together with multinomially distributed weighted counts in the data $p(w_1, \ldots, w_K|\varsigma) \propto \prod_{j=1}^{K} \varsigma_j^{w_j}$, Cohen (1997) posits and Dong, Elliott, and Raghunathan (2014) prove that the posterior predictive distribution of counts in the nonsampled data is given by the following:

$$p(u_1, \ldots, u_K|w_1, \ldots, w_K) = \frac{\prod_{j=1}^{K} \Gamma(w_j + u_j)/\Gamma(w_j)}{\Gamma(2N - n)/\Gamma(N)}. \quad (7)$$

2.1.2. The adapted-weighted FPBB method. The adapted-weighted FPBB (Dong, Elliott, and Raghunathan, 2014) consists of two stages. The first stage resamples the original sample using the standard Bayesian bootstrap assuming IID, and the second stage reverses/undoes the sampling weights using the weighted FPBB. This two-stage algorithm is analogous to the fully parametric Bayesian method, where the first stage is equivalent to drawing values of the parameter (ς) from its posterior distribution given the counts in sampled data $(n_1, \ldots, n_K)$ and the second stage draws the predicted counts in the nonsampled data $(u_1, \ldots, u_K)$ given the drawn parameter. The method is described as follows:

• *Resampling Using the Standard Bayesian Bootstrap (BB)*

The standard Bayesian Bootstrap of Rubin (1981) assuming IID is used to generate L replicate BB samples each of size n, i.e., $\left\{(Y_s^{(l)}, X_s^{(l)}, w_s^{(l)}, R_s^{(l)}), l = 1, \ldots, L.\right\}$. This essentially generates the posterior joint distribution (denoted by f) of all the variables in the population given their realized values in the sample data set. Or equivalently, the posterior distribution of the parameter vector ς is drawn given the sample, i.e.,

$$\begin{aligned} f^{(l)}(Y, X, w, R)|(Y_s, X_s, w_s, R_s) &\Leftrightarrow \left(\varsigma^{(l)}|Y_s, X_s, w_s, R_s\right) \\ &\sim \text{Dir}(n_1, \ldots, n_K), \quad \text{for} \quad l = 1, \ldots, L., \\ &\text{where } \varsigma^{(l)} = \left(\varsigma_1^{(l)}, \ldots, \varsigma_K^{(l)}\right). \end{aligned} \quad (8)$$

This stage captures the sampling variability. The uncertainty in the posterior draws of the parameter $\varsigma^{(l)}$ is reflected in the varying counts of distinct units in the original sample being selected in different replicate BB samples. Let $t_l(i)$ denote the number of times unit i is selected in the lth replicate BB sample, for $l = 1, \ldots, L$. We incorporate this source of uncertainty in computing "the lth bootstrap weight for unit i", i.e., $w_i^{(l)*} = w_i \cdot t_l(i)$, where w_i denotes the original sampling weight for unit i. The bootstrap weights are carried forward as input weights to the next stage.

• *Undo Sampling Weight Using the Weighted Polya Posterior/Weighted FPBB*

To capture the variability due to "imputing" the nonsampled units, the weighted Polya posterior in equation (7) is used to create S synthetic populations for each of the L BB sample obtained from the previous stage, i.e., $\left\{(Y_s^{(l)}, X_s^{(l)}, R_s^{(l)}), (Y_{ns}^{(ls)}, X_{ns}^{(ls)}, R_{ns}^{(ls)})\right\}$, for $s = 1, \ldots, S, l = 1, \ldots, L$. The distribution in equation (7) does not lend itself to direct calculation; however, draws from (7) can be obtained using Monte Carlo simulation. Specifically, we apply a procedure suggested by Cohen (1997), who extended the algorithm developed by Lo (1998) in the simple random sampling setting to a weighted sampling setting

(i) Take a Pólya sample of size $N - n$, denoted by $(Y_s^{(ls)}, X_s^{(ls)}, R_s^{(ls)})$ from the urn $(Y_s^{(l)}, X_s^{(l)}, R_s^{(l)})$ by selecting each element in the urn with probability

$$\frac{w_i^{(l)*} - 1 + l_{i,k-1} \times (\frac{N-n}{n})}{N - n + (k-1) \times (\frac{N-n}{n})}, \quad k = 1, 2, \ldots, N - n + 1. \quad (9)$$

where $w_i^{(l)*}$ is the bootstrap weight for the ith unit in the lth replicate BB sample, and $l_{i,k-1}$ is the number of selections of unit i up to (k-1)th selection, setting $l_{i,0} = 0$.

(ii) Form the weighted FPBB synthetic population $P_{(s)}^{(l)} = \left\{(Y_s^{(l)}, X_s^{(l)}, R_s^{(l)}), (Y_{ns}^{(ls)}, X_{ns}^{(ls)}, R_{ns}^{(ls)})\right\}$ so that it has exact size N.

This results in the "unweighted" synthetic populations $P_{(s)}^{(l)} = (Y^{(ls)}, X^{(ls)}, R^{(ls)}) = (P_{(s)\text{obs}}^{(l)}, Y_{\text{mis}}^{(ls)}), s = 1, \ldots, S, l = 1, 2, \ldots, L$, where L and S are the numbers of data sets generated from first- and second-stage, respectively, and $P_{(s)\text{obs}}^{(l)} = ((Y_{s,\text{obs}}^{(l)}, X_s^{(l)}, R_s^{(l)}), (Y_{ns,\text{obs}}^{(ls)}, X_{ns}^{(ls)}, R_{ns}^{(ls)}))$ and $Y_{\text{mis}}^{(ls)} = (Y_{s,\text{mis}}^{(l)}, Y_{ns,\text{mis}}^{(ls)})$ consist of the observed and unobserved data in the lsth FPBB synthetic population data set, respectively.

2.2. *Step 2: Multiply Impute Missing Data through Parametric Models*

Now that we have effectively "undone" the sampling design, we are ready to perform conventional MI under an IID assumption. Following the standard MI procedure or approximations such as SRMI (Raghunathan, Lepkowski, Van Hoewyk, and Solenberger, 2001), we obtain draws from the posterior predictive distribution $p(Y_{\text{mis}}^{(ls)}|P_{(s)\text{obs}}^{(l)})$. Without the need to include weights in the imputation model due to a self-weighting FPBB population generated from previous step, our task can now be concentrated on correctly modeling the covariate variables. Note that the elimination of the weights from the self-weighting FPBB population does not obviate the need to account for the weights in the imputation process, if the probability of selection (I) and nonresponse (R) is associated with each other (i.e., $p(R|Y_{\text{obs}}, w) \neq p(R|Y_{\text{obs}})$). This step results in M imputed synthetic data sets for each of the $L \times S$ FPBB synthetic populations generated from the first step, $P_{sM}^{l} = (P_{(s)1}^{(l)}, P_{(s)2}^{(l)}, \ldots, P_{(s)M}^{(l)})$, *for* s = $1, 2, \ldots, S, l = 1, 2, \ldots, L$.

2.3. *Point and Variance Estimates for the Two-Step MI Procedure*

Conditional on $P^{\text{imp}} = \{P_{(11)}^{(1)}, \ldots, P_{(1\text{M})}^{(1)}, \ldots, P_{(\text{S1})}^{(1)}, \ldots, P_{(\text{SM})}^{(1)}, \ldots, P_{(\text{SM})}^{(\text{L})}\}$, the posterior predictive distribution of a scalar population statistic $Q(Y) \equiv Q$ is given by

$$Q|P^{\text{imp}} \overset{\bullet}{\sim} t_{L-1}(\bar{Q}_L, (1 + L^{-1})V_L) \tag{10}$$

where $\bar{Q}_L = \frac{1}{L}\sum_l \tilde{Q}^{(l)}$ and $V_L = \frac{1}{L-1}\sum_l (\tilde{Q}^{(l)} - \bar{Q}_L)^2$, where $\tilde{Q}^{(l)} = \lim_{\substack{S\to\infty \\ M\to\infty}} \frac{1}{SM}\sum_s \sum_m q^{(lsm)}$, where $q^{(lsm)}$ is an estimate of Q obtained from the mth imputation of the sth synthetic population within the lth Bayesian Bootstrap sample; in practice we estimate $\tilde{Q}^{(l)}$ by $\hat{Q}^{(l)} = \frac{1}{SM}\sum_s \sum_m q^{(lsm)}$. The result follows immediately from Section 4.1 of Raghunathan, Reiter, and Rubin (2003), and is based on the standard Rubin (1987) multiple imputation combining rules, where (Y_{ns}, X_{ns}, R_{ns}) and $Y_{s,\text{mis}}$ are missing data and $(Y_{s,\text{obs}}, X_s, R_s)$ is observed. The average "within" imputation variance is zero, since the entire population is being synthesized; hence the posterior variance of Q is entirely a function of the between-imputation variance, and the degrees of freedom is simply given by the number of BB samples. This result requires $E(q^{(lsm)}) = Q$, which implies that our imputation model for Y_{mis} is correctly specified, as well as the standard sufficiently large sample size for the t approximation to be reasonable. In addition, since we are imputing under the synthesized population, all weights are constant and equal to 1, so no covariance between the MI point estimator and the complete data estimator is induced (Kim et al., 2006; Seamen et al., 2012).

These results assume $S \to \infty$ and $M \to \infty$; in practice, we have found that relatively modest values of S and M are needed for the imputation approximations to hold. In particular, we show below that $S = 20$ and $M = 5$ yield reasonable results in simulation studies, results that are also consistent with in Dong et al. (2014). In addition, in settings where N is very large, generating a synthetic population large enough to have a relatively trivial sampling fraction (e.g., $N^* = 10n$) will generally be sufficient.

3. Simulation Study

A simulation study was designed to investigate the inferential properties of the proposed method. In particular, we are interested to see how the two-step MI procedure performs in comparison with the existing fully parametric methods under four simulation designs defined by crossing the following two factors:

(1) Associations of the probabilities of selection with the mechanism generating the data. We call the design "outcome relevant" if the probabilities of selection are correlated with the outcome variable Y, otherwise we term it an "outcome irrelevant" design.
(2) Associations of the probabilities of selection with the mechanism generating the missing values. We use "MAR_X" (weight-independent missingness) and "MAR_X, W" (weight-dependent missingness), respectively to denote respective situations where the missing data mechanism is dependent on fully observed covariates only and where it depends on probabilities of selection as well as other fully observed covariates.

We first generate a population of three variables: the outcome variable Y, a covariate X, and a variable Z based on which probability proportionate to size without replacement (PPSWOR) sampling is conducted. The joint distribution of Z, X, and Y is given by the following:

$$\begin{aligned}
\log Z &\sim N(2, 1) \\
X|Z &\sim N(0.1 * \log Z, \sigma_x^2) \\
Y_1|X, Z &\sim N(0.1 * X + 0.5 * \log Z + 0.6 * X * \log Z, \sigma_{y_1}^2) \\
Y_2|X, Z &\sim N(0.2 * X, \sigma_{y_2}^2)
\end{aligned}$$

Thus (Y_1, X, Z) constitutes the "relevant design" population and (Y_2, X, Z) constitutes the "irrelevant design" population. Both populations have size $N = 4000$. For each population, we drew 500 independent samples of size $n = 200$ without replacement, with inclusion probability for the ith unit $\pi_i = nZ_i / \sum_{j=1}^{N} Z_j$. We call the 500 PPSWOR samples "before deletion (BD) samples."

Next, probit models were used as deletion functions to create missing data in the outcome variable Y for each of the 100 simulations. Both X and Z are assumed to be completely observed.

Table 1
Before deletion study of the effects of the number of generated FPBB populations (S) on variance estimate

Parameters Of interest	Performance criteria	Weighted FPBB method with S synthetic populations created								Actual sample
		$S=1$	$S=5$	$S=10$	$S=15$	$S=20$	$S=25$	$S=30$	$S=40$	
Mean	Pt. est.	1.460	1.459	1.458	1.460	1.458	1.460	1.460	1.460	1.450
	Emp.Est.Var	0.048	0.036	0.034	0.034	0.033	0.033	0.033	0.033	0.033
	Emp.Var	0.031	0.031	0.032	0.031	0.032	0.031	0.031	0.031	0.032
	RMSE	0.176	0.178	0.177	0.177	0.178	0.175	0.177	0.177	0.178
	95% CI cov.	99%	97%	96%	97%	96%	96%	95%	96%	96%
Intercept	Pt. est.	1.251	1.250	1.249	1.250	1.249	1.250	1.250	1.250	1.241
	Emp.Est.Var	0.028	0.022	0.022	0.022	0.021	0.021	0.021	0.021	0.021
	Emp.Var	0.021	0.021	0.021	0.011	0.021	0.021	0.021	0.021	0.022
	RMSE	0.145	0.145	0.146	0.145	0.145	0.144	0.144	0.146	0.150
	95% CI cov.	96%	94%	94%	94%	94%	95%	94%	94%	94%
Slope	Pt. est.	1.280	1.281	1.280	1.281	1.280	1.281	1.280	1.280	1.264
	Emp.Est.Var	0.036	0.029	0.028	0.027	0.027	0.027	0.027	0.027	0.027
	Emp.Var	0.030	0.029	0.029	0.029	0.029	0.029	0.030	0.030	0.033
	RMSE	0.172	0.171	0.172	0.171	0.171	0.171	0.173	0.173	0.181
	95% CI cov.	95%	92%	91%	91%	91%	90%	91%	90%	89%

We generate $T_1 = -0.635 + 0.4X + e$ and $T_2 = -0.55 + 0.4X - 0.5\log Z + 0.4X * \log Z + e$, where $e \overset{iid}{\sim} N(0,1)$, corresponding to the MAR_X condition and MAR_X,W condition, respectively. The outcome is then missing if $T_j > 0$ (i.e., $P(M = 1|T_j) = \Phi(E(T_j))$, $j = 1, 2.$, where $\Phi(x)$ corresponds to the standard normal CDF). This yields a missingness fraction of approximately 30% in all four scenarios.

For each of the four simulation designs, we analyze the data using five imputation models. Model 1 ignores weights altogether in the imputation process, a procedure typically adopted. Model 2 includes log(Z) in the imputation model (Schenker et al., 2006). Model 3 includes both the log(Z) and its interactions with other covariates in the imputation model. Model 4 and Model 5 are equivalent to Model 2 and Model 3, except that log(Z) is replaced with 1/Z corresponding to the weight, as suggested in Kim et al. (2006)and Seaman et al. (2011). All five imputation models will be tested with both the fully parametric MI method and the proposed two-step synthetic MI procedure. The only difference is that we perform design-based analyses on the imputed data from the former, while with the new method we perform simple unweighted analyses instead. We implement the MI using the MICE package (R Core Team, 2013).

Finally, we focus on estimating the population mean of Y (i.e., $\bar{Y}$) and the population regression coefficients of Y on X: $Y = \beta_0 + \beta_1 X$. We used five quantities to evaluate the performance of the various methods under comparison: bias, empirical root mean square error (RMSE), empirical interval coverage, empirical variance, and the mean of the estimated variance (to compare with the empirical variance). For the standard parametric analysis, we use Rubin's combining rules (Rubin, 1987), using weighted point estimates and Taylor Series approximations (Binder, 1983) to account for the weights in the variance estimation of the filled-in data sets. Population means and regression parameters are used to compute bias and mean square error.

3.1. *Simulation Results*

In deciding how many synthetic populations S are needed, we conducted a preliminary study based on the before deletion (BD) data. (We let $L = 100$.) Simulation results are shown in Table 1. We observe that as we increase S, the variance estimate decreases, and stabilizes close to the actual sample variance when $S \geq 20$. This is consistent with a similar result in Dong et al. (2014), which found that 20 synthetic populations were sufficient to yield appropriate coverage intervals in a complete data setting. Therefore, we use $S = 20$ along with $L = 100$ and $M = 5$ in the after deletion (AD) simulation.

Tables 2 and 3 present the results from our simulation study. Each table is divided into two parts, containing the results from MAR_X scenario and MAR_X,W scenario, respectively. Within each scenario, we compare our new method with the fully parametric method, with the columns indicated by "X," "X,log(Z)," "X*log(Z)," "X,W," and "X*W" each corresponding to the estimates under the five imputation models described above.

When the design is relevant to the outcome variable Y yet uncorrelated with missingness (Table 2: MAR_X scenario), obvious advantages can be observed for the synthetic methods over the fully parametric method. For the fully parametric method to work properly under this condition, the imputation model has to be correctly specified, otherwise all inferences based on this method are invalid—not only is there substantial bias attached to all three parameter estimates, but there is a corresponding disruption in coverage rates as well, which is particularly poor when the design is completely ignored in the model. In contrast, our proposed method results in nearly unbiased estimates and actual coverage that is closer to the nominal level under all five models, regardless of the misspecification. Substantial gains in terms of RMSE over the model-based method were also consistently observed in all scenarios considered. This indicates that the "unweighting" procedure has actually played *dual roles* in the process: its effect is not

Table 2

*Performance of the proposed method in contrast to the fully parametric method under the relevant design condition (*true model *italicized)*

		MAR_X										MAR_X,W									
		Standard parametric MI					Semi-parametric MI					Standard parametric MI					Semi-parametric MI				
Actual parameters	Performance criteria	X	X, logZ	X*logZ	X,W	X*W	X	X, logZ	X*logZ	X,W	X*W	X	X, logZ	X*logZ	X,W	X*W	X	X, logZ	X*logZ	X,W	X*W
Mean=1.450	Bias	0.248	0.084	*0.003*	0.063	−0.078	0.019	0.012	*0.005*	−0.011	−0.056	0.211	0.032	*0.000*	0.001	−0.098	0.061	0.022	*0.006*	−0.001	−0.048
	MeanEst.Var	0.057	0.050	*0.041*	0.080	0.133	0.047	0.043	*0.040*	0.059	0.089	0.062	0.065	*0.049*	0.080	0.123	0.040	0.039	*0.037*	0.057	0.078
	Emp.Var	0.047	0.042	*0.035*	0.065	0.125	0.044	0.039	*0.036*	0.051	0.067	0.062	0.066	*0.045*	0.074	0.125	0.036	0.034	*0.032*	0.048	0.065
	RMSE	0.329	0.222	*0.188*	0.262	0.362	0.211	0.196	*0.190*	0.224	0.265	0.327	0.258	*0.211*	0.273	0.367	0.199	0.186	*0.179*	0.219	0.259
	95%Cov	84%	95%	*96%*	95%	98%	95%	95%	*95%*	96%	97%	84%	92%	*96%*	96%	97%	95%	95%	*96%*	97%	98%
Intercept=1.241	Bias	0.208	0.056	*0.004*	0.035	−0.064	0.019	0.010	*0.007*	−0.013	−0.043	0.181	0.003	*−0.001*	−0.019	−0.081	0.051	0.015	*0.007*	−0.009	−0.038
	MeanEst.Var	0.032	0.030	*0.028*	0.058	0.090	0.029	0.028	*0.027*	0.045	0.062	0.032	0.033	*0.032*	0.058	0.087	0.023	0.024	*0.024*	0.043	0.053
	Emp.Var	0.027	0.030	*0.026*	0.065	0.092	0.029	0.027	*0.026*	0.043	0.051	0.027	0.031	*0.026*	0.067	0.097	0.022	0.023	*0.022*	0.039	0.048
	RMSE	0.266	0.183	*0.162*	0.257	0.309	0.171	0.165	*0.160*	0.205	0.227	0.243	0.174	*0.160*	0.258	0.321	0.155	0.151	*0.147*	0.195	0.220
	95% Cov	76%	92%	*94%*	90%	95%	93%	93%	*94%*	95%	95.%	76%	92%	*96%*	94%	96%	92%	94%	*95%*	94%	95%
Slope=1.264	Bias	0.208	0.144	*−0.009*	0.156	−0.064	0.003	0.013	*0.008*	0.019	−0.041	0.155	0.165	*−0.003*	0.117	−0.083	0.060	0.047	*0.015*	0.053	−0.026
	MeanEst.Var	0.035	0.031	*0.034*	0.046	0.128	0.035	0.032	*0.034*	0.039	0.085	0.038	0.036	*0.039*	0.049	0.117	0.028	0.027	*0.030*	0.038	0.073
	Emp.Var	0.034	0.032	*0.037*	0.048	0.155	0.037	0.033	*0.035*	0.036	0.072	0.054	0.046	*0.037*	0.052	0.138	0.030	0.029	*0.031*	0.037	0.066
	RMSE	0.277	0.229	*0.192*	0.268	0.398	0.191	0.181	*0.186*	0.191	0.269	0.278	0.269	*0.191*	0.257	0.381	0.182	0.176	*0.177*	0.198	0.256
	95% Cov	75%	82%	*94%*	82%	94%	93%	93%	*93%*	94%	95%	78%	82%	*97%*	85%	93%	88%	89%	*91%*	90%	92%

Table 3

*Performance of the proposed method in contrast to the fully parametric method under the irrelevant design condition (*true model *italicized)*

		MAR_X										MAR_X,W									
		Standard parametric MI					Semi-parametric MI					Standard parametric MI					Semi-parametric MI				
Actual parameters	Performance criteria	X	X, logZ	X*logZ	X,W	X*W	X	X, logZ	X*logZ	X,W	X*W	X	X, logZ	X*logZ	X,W	X*W	X	X, logZ	X*logZ	X,W	X*W
Mean = −0.0191	Bias	*−0.025*	−0.034	−0.037	0.004	−0.006	*−0.000*	−0.000	0.002	0.001	−0.006	*−0.006*	−0.025	−0.026	0.007	−0.001	*0.007*	0.006	0.004	0.009	0.003
	MeanEst.Var	*0.020*	0.020	0.021	0.025	0.032	*0.018*	0.019	0.020	0.025	0.034	*0.024*	0.029	0.029	0.021	0.027	*0.015*	0.016	0.016	0.023	0.030
	Emp.Var	*0.014*	0.019	0.018	0.023	0.026	*0.018*	0.018	0.019	0.021	0.024	*0.013*	0.021	0.022	0.019	0.022	*0.015*	0.016	0.016	0.019	0.022
	RMSE	*0.122*	0.141	0.140	0.154	0.168	*0.133*	0.134	0.136	0.146	0.155	*0.114*	0.147	0.151	0.143	0.154	*0.121*	0.127	0.128	0.137	0.148
	95%Cov	*97%*	94%	93%	96%	94%	*94%*	95%	96%	94%	95%	*97%*	94%	95%	95%	94%	*95%*	95%	95%	95%	95%
Intercept = −0.0569	Bias	*−0.102*	−0.109	−0.113	0.003	−0.003	*−0.003*	−0.003	−0.004	0.001	−0.002	*−0.001*	−0.021	−0.023	0.006	0.001	*0.007*	0.005	0.004	0.008	0.003
	MeanEst.Var	*0.019*	0.019	0.020	0.023	0.026	*0.017*	0.018	0.019	0.023	0.028	*0.022*	0.027	0.025	0.020	0.023	*0.015*	0.016	0.016	0.022	0.026
	Emp.Var	*0.014*	0.018	0.018	0.022	0.024	*0.018*	0.018	0.018	0.020	0.021	*0.012*	0.023	0.024	0.020	0.021	*0.015*	0.016	0.016	0.019	0.021
	RMSE	*0.157*	0.174	0.174	0.185	0.193	*0.132*	0.132	0.133	0.140	0.145	*0.110*	0.152	0.157	0.176	0.185	*0.121*	0.126	0.127	0.136	0.143
	95% Cov	*88%*	85%	86%	86%	85%	*95%*	95%	95%	94%	95%	*98%*	95%	97%	93%	94%	*94%*	95%	94%	95%	96%
Slope = 0.222	Bias	*0.006*	0.003	0.003	0.005	−0.012	*0.008*	0.008	0.006	0.001	−0.008	*−0.023*	−0.022	−0.008	0.004	−0.014	*0.000*	−0.000	−0.003	0.002	−0.010
	MeanEst.Var	*0.017*	0.017	0.019	0.019	0.030	*0.017*	0.017	0.019	0.019	0.033	*0.023*	0.020	0.025	0.016	0.026	*0.013*	0.013	0.015	0.015	0.028
	Emp.Var	*0.014*	0.014	0.017	0.014	0.027	*0.018*	0.017	0.019	0.017	0.024	*0.011*	0.014	0.025	0.013	0.025	*0.013*	0.013	0.015	0.014	0.021
	RMSE	*0.118*	0.119	0.132	0.119	0.164	*0.133*	0.131	0.136	0.128	0.154	*0.108*	0.120	0.158	0.113	0.157	*0.115*	0.115	0.121	0.116	0.143
	95% Cov	*95%*	95%	94%	97%	95%	*94%*	95%	95%	94%	95%	*98%*	98%	92%	96%	95%	*93%*	94%	94%	94%	95%

limited to untying the unequal probability selection and saving the effort of design-based analyses afterward, but it also captures much of the interactions between the design and the survey variable of interest so that ignoring the design in the imputation model does little harm. Incorporating the probability of selection in the imputation model unnecessarily has a modest impact, with some greater increase in variability and MSE when weights versus log MOS are included, since the weights are more variable.

With a relevant design that is also a correlate of missingness (Table 2: MAR_X,W scenario), the imputation models require use of the design variable (here the weight) to maintain an ignorable missing data mechanism. The model-based method behaves similarly to the case where the design is associated only with Y: failure to include the weight in the imputation model substantially biases all of the estimators considered, while including the weight as a covariate corrects for bias in the mean and intercept estimator but not in the slope. The synthetic model partially corrects for these biases by providing a correct estimate of the population distribution in the presence of missing data; however, unless the imputation model is correctly specified, some biases remain. Nevertheless, the synthetic model still has substantially reduced RMSE relative to the fully parametric approach for the mean and intercept estimator when the weight is ignored in the imputation model, and reduced RMSE when estimating the slope when the weight is included as a covariate but the interaction between the slope and the probability of selection is ignored. The synthetic model also has nearly exact to slightly conservative coverage properties, in contrast to the anti-conservative coverage of the fully parametric estimator when the model is misspecified for the estimator of interest. Misspecifying the functional form of the probability of selection in the imputation model (using weights instead of the log MOS) generally increases bias and MSE for both the fully parametric and semi-parametric approach, although the increased bias is not sufficient to reduce nominal coverage over the correctly specified functional form.

With an outcome irrelevant design (Table 3), there are very slight effects on the estimates when compared across methods and models. Including the irrelevant design variable in the imputation model results in negligible biases and introduces some modest inefficiencies, consistent with the findings in Reiter et al. (2006). The only impact of using weights rather than log MOS in the imputation is to modestly increase MSE. It is also worth noting that the MI variance/standard errors under the new method are consistently lower than the fully parametric method, in addition to their better coverage properties. This is observed for all 12 scenarios considered.

4. Application to the Behavioral Risk Factor Surveillance System (BRFSS)

We next examine the effect of incorporating the survey weight in MI using data from one design stratum ($n = 388$) of the 2009 Michigan BRFSS. This design stratum contains sampled households that belong to the medium-density (unlisted) telephone numbers group. The BRFSS is a telephone survey conducted with a random sample of adults living in telephone-equipped households in the US. An independent sample of telephone numbers are used as the sampling frame; thus case weights are constructed to account for the fact that the probability of selection is proportional to the number of telephone lines and inversely proportional to the number of adults in a household; in addition, poststratification weights are used to adjust age–sex–race/ethnic distributions to Census totals. A mix of categorical and continuous variables is selected for analysis. These include health insurance coverage (yes/no), body-mass index (BMI) in kg/m^2, high blood pressure (yes/no), and five demographic variables (age (in years), race (White versus Nonwhite), annual household income (low = <\$25, 000, medium = \$25, 000–75, 000, high > \$75, 000), and gender and employment status (yes/no/other)). All survey variables except gender have certain degrees of missing data: income has the highest missing rate (16.5%), while others are missing 0–6%.

4.1. *Imputation Method*

We compare results from the conventional fully parametric MI method with the proposed two-step semi-parametric MI method, with two imputation modeling strategies applied with each method: (1) assuming SRS, and (2) including the log of weights as a predictor in the model. We also include the weighted complete case analysis. Both imputation models used all available substantive covariates (health insurance, BMI, high blood pressure status, age, race, income, gender, and employment status). For the standard parametric analysis, as in the simulation study, we use Rubin's combining rules (Rubin, 1987) with weighted point estimates and Taylor Series approximations to account for the weights in the variance estimation. For the new method, we generated $L = 100$ Bayesian bootstrap (BB) samples and created $S = 30$ FPBB populations within each BB sample, with $M = 5$ multiple imputations performed for each FPBB population. Since we do not know the population size in advance and the individual final weights sum up to nearly 200,000 cases which is unrealistic to generate, we assume that $N = 4500$ is large enough to be treated as a synthetic population (corresponding to a sampling fraction of less than 10%). Since the degrees of freedom is $L - 1 = 99$, a normal distribution was used for inference.

4.2. *Analyses*

We consider three different analyses: (1) the marginal distribution of income and health insurance accessibility (Table 4); (2) a linear regression model of BMI on key demographic variables (Table 4); and (3) a log-linear model of a four-way contingency table defined by four categorical variables with no second-or-higher-order interactions (Table 5). We consider an analysis using the full data set, as well as a stratified analysis restricted to subjects identifying as white ("white domain"). Multivariate imputation by chained equations (MICE) in R was used to impute the missing data under both MI methods.

4.3. *Results*

Since the poststratification adjustment factor constitutes an important component of the final weight in BRFSS data set, we presume that including the variables used to construct poststratification cells (age, race, and gender in this case) in the imputation model should help in predicting the missing Y variable. A linear regression of final weights on age, sex, and

Table 4

Estimation of marginal distributions for income and health insurance, and linear regression coefficients for the regression of BMI (dependent variable) on income, age, and gender (independent variables). (Complete case analysis presented is weighted.)

			Methods									
					Parametric MI ($M = 5$)				Synthetic MI ($L = 100, S = 30, M = 5$)			
			Complete case		Exclude weights		Include log (weights)		Exclude weights		Include log (weights)	
Sample	Estimation	Variable	Pt.est.	SE	Pt.est.	SE	Pt.est.	SE	Pt.est.	SE	Pt.est.	SE
Full sample	Marginal	Low income	0.50	0.04	0.50	0.04	0.52	0.05	0.52	0.04	0.51	0.04
		Medium income	0.38	0.04	0.36	0.04	0.36	0.04	0.36	0.04	0.36	0.04
		High income	0.12	0.03	0.14	0.03	0.12	0.03	0.13	0.03	0.13	0.03
		No insurance	0.22	0.04	0.24	0.04	0.24	0.04	0.24	0.04	0.24	0.04
	Regression	Intercept	27.0	2.8	26.1	2.0	25.8	2.0	26.3	2.3	26.2	2.3
		Medium income	0.47	1.40	0.35	1.21	0.39	1.19	0.37	0.94	0.37	0.95
		High income	0.27	1.43	−0.47	1.32	−0.33	1.37	−0.36	1.40	−0.31	1.40
		Age	0.02	0.04	0.03	0.03	0.04	0.03	0.03	0.03	0.03	0.03
		Female	2.29	1.30	2.72	1.06	2.56	1.07	2.57	1.06	2.55	1.05
Whites domain	Marginal	Low income	0.30	0.07	0.36	0.07	0.35	0.06	0.34	0.06	0.34	0.06
		Medium income	0.53	0.08	0.48	0.07	0.50	0.07	0.49	0.06	0.49	0.06
		High income	0.17	0.06	0.16	0.06	0.15	0.05	0.17	0.06	0.17	0.06
		No insurance	0.24	0.07	0.21	0.06	0.21	0.06	0.19	0.06	0.19	0.06
	Regression	Intercept	31.1	3.9	32.4	4.7	31.0	4.1	31.0	4.2	31.0	4.1
		Medium income	−1.6	3.25	−2.8	2.83	−2.1	2.72	−1.8	2.96	−1.7	2.97
		High income	−3.1	3.60	−3.5	3.42	−3.2	3.18	−3.1	3.65	−3.0	3.62
		Age	0.02	0.06	−0.01	0.06	0.02	0.06	0.02	0.06	0.02	0.05
		Female	−1.7	2.39	−0.13	2.13	−0.68	2.11	−0.80	2.17	−0.75	2.17

race shows that these covariates explain 40% of the variance of the weights, suggesting that there are other design variables that contribute to the survey weights unknown to us. Thus, we conclude that imputation approaches that condition only on the available design variables will be insufficient to account for the sampling weights.

Table 4 shows that under the fully parametric MI method, including survey weights in the imputation model has a large

Table 5

Estimation of log-linear model for four categorical variables (collapse categories for medium and high income): 2009 Michigan BRFSS

		Methods									
				Parametric MI		Parametric MI		Synthetic MI		Synthetic MI	
		Complete case		exclude weights		include log(weights)		exclude weights		include log(weights)	
Estimation	Variable level	Coef.	SE	Coef.	SE	Coef.	SE	Coef.	SE	Coef.	SE
Main effects	Low income	−0.01	0.12	0.08	0.13	0.04	0.12	0.04	0.13	0.02	0.13
	Has insurance	0.61	0.12	0.64	0.12	0.62	0.11	0.69	0.12	0.68	0.12
	White	−0.94	0.11	−1.00	0.10	−1.00	0.11	−1.10	0.12	−1.10	0.12
	Male	−0.11	0.12	−0.09	0.10	−0.07	0.10	−0.07	0.11	−0.07	0.11
	Low income × has insurance	−0.36	0.12	−0.37	0.12	−0.33	0.13	−0.31	0.11	−0.30	0.11
Two-way interactions	Low income × white	−0.28	0.10	−0.20	0.09	−0.20	0.09	−0.22	0.09	−0.22	0.09
	Low income × male	−0.03	0.09	−0.03	0.09	−0.07	0.09	−0.05	0.08	−0.05	0.08
	Has insurance × white	−0.13	0.12	−0.02	0.12	−0.02	0.12	0.02	0.13	0.03	0.13
	Has insurance × male	−0.01	0.13	−0.12	0.10	−0.14	0.10	−0.15	0.10	−0.15	0.10
	White × male	−0.08	0.09	−0.11	0.08	−0.11	0.09	−0.11	0.08	−0.10	0.08

impact on the estimated proportions of income levels and the regression coefficients of BMI on income and gender. In fact, these differences are particularly significant for the whites-only analysis. Under the new method, however, all estimates are similar to those from the model-based method with weights accounted for. Moreover, there is essentially no difference whether or not we incorporate weights into the imputation model once the sample data are synthesized, indicating that, as we expect, the new method can adjust for the weight effects at the synthesizing step without the need to model survey weights at the imputation step. Similar results are obtained in Table 5 with respect to the log-linear model.

5. Discussion

We propose using weighted finite population Bayesian Bootstrap to account for one-stage sampling weights in MI for item missing data in the Behavioral Risk Factor Surveillance Survey. We also evaluate the performance of this method in a simulation setting: our findings suggest that it can bring significant reductions in bias relative to the existing model-based methods with little loss in efficiency. Meanwhile, the weighted FPBB method potentially protects against model misspecification, for example, wrongly including or excluding interactions between design variables and other covariates in the imputation model, while also maintaining population-level multivariate relationships. A further advantage lies in that, unlike the fully parametric methods which include designs in the imputation model and still require complex survey packages to analyze the imputed data sets, the new method fully accounts for the unequal selection probabilities by unweighting them and restoring a population in a separate step; therefore, only simple, unweighted complete-data analysis techniques are needed for inferences with the combining rules. This potentially allows a much wider variety of models to be considered using existing software, which, despite recent improvements, often does not have straightforward methods for accounting for complex sample designs.

A limitation of the proposed method is the need for the weights to be included in the imputation model if the probability of item response is a function of selection probability. However, by separating the modeling of the weights in the complete data by use of a relatively easy-to-implement nonparametric algorithm from the modeling of the weights in the missingness mechanism, it (i) reduces the impact of misspecified missingness mechanisms (as noted in Table 2, where the RMSEs and coverage of the misspecified models are greatly improved over the standard parametric approaches), and (ii) allows more careful inspection and modeling of the missingness mechanism as a function of the weights. In particular, this suggests that the imputation model be developed using the weighted FPBB data sets, to include appropriate functions of and interactions with the design weights.

The proposed two-stage semi-parametric multiple imputation approach has a number of possible extensions. First, while we have imputed the missing data in our second step using a model-based approach, a fully nonparametric approach using a Bayesian bootstrap (Rubin and Schenker, 1986) can be used instead. Second, while our approach has focused on sampling weights, extensions that incorporate unit nonresponse into the synthetic population generation and multiple imputation to propagate uncertainty in unit-nonresponse weighting adjustments are possible. (However, when only final weights incorporating nonresponse adjustments are provided, treating the final weight as a sampling weight as we did in the BRFSS application may be the only practical alternative.) Third, while we made a missing at random assumption with a single missing outcome in our simulation study and application, it is certainly possible at the imputation stage to accommodate missingness in multiple covariates via sequential regression multiple imputation (Raghunathan et al., 2001), or even to consider not missing at random mechanisms (Little, 2008). Fourth, we could extend the method to incorporate unit nonresponse by generating Bayesian bootstraps of the entire sample including the unit nonresponders, applying standard unit nonresponse adjustments to the base weights to obtain the nonresponse-adjusted weights, and then applying the weighted Polya posterior with the nonresponse-adjusted weights as the input weight in the algorithm to create synthetic populations. Finally, our method developed here is for a one-stage design; extensions to account for multi-stage designs with clustering and stratification as part of the finite population Bayesian bootstrap are required as well, and are the focus of current research efforts.

Acknowledgements

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 for their review and helpful comments, as well as the Editor, Associate Editor, and two reviewers, whose careful review and comments substantially improved the article.

References

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review* **51**, 279–292.

Breidt, J., Claeskens, G., and Opsomer, J. (2005). Model-assisted estimation for complex surveys using penalized splines. *Biometrika* **92**, 831–846.

Chen, Q., Elliott, M. R., and Little, R. J. A. (2010). Bayesian penalized spline model-based inference for finite population proportions in unequal probability sampling. *Survey Methodology* **36**, 22–34.

Cohen, M. P. (1997). The Bayesian bootstrap and multiple imputation for unequal probability sample designs. *ASA Proceedings of the Section on Survey Research Methods*, 635–638.

Dong, Q., Elliott, M. R., and Raghunathan, T. E. (2014). A nonparametric method to generate synthetic populations to adjust for complex sample designs. *Survey Methodology* **40**, 29–46.

Elliott, M. R. and Little, R. J. (2000). Model-based alternatives to trimming survey weights. *Journal of Official Statistics* **16**, 191–209.

Ericson, W. A. (1969). Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society* **B31**, 195–233.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, Vol. I. New York, NY: Wiley.

Ghosh, M. and Meeden, G. (1983). Estimation of the variance in finite population sampling. *Sankhyā: The Indian Journal of Statistics* **B45**, 362–375.

Heitjan, D. F. and Little, R. J. A. (1991). Multiple imputation for the fatal accident reporting system. *Applied Statistics* **40**, 13–29.

Kim, J. K., Brick, J. M., Fuller, W. A., and Kalton, G. (2006). On the bias of the multiple-imputation variance estimator in survey sampling. *Journal of the Royal Statistical Society* **B68**, 509–521.

Little, R. J. A. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association* **99**, 546–556.

Little, R. J. A. (2008). Selection and Pattern Mixture Models, in *Longitudinal Data Analysis*, G. Fitzmaurice, M. Davidian, G. Verbeke, G. Molenberghs (eds), 409–430. Boca Raton Florida: Chapman & Hall/CRC Press.

Little, R. J. A. (2011). Calibrated Bayes, for statistics in general, and missing data in particular (with discussion and rejoinder). *Statistical Science* **26**, 162–186.

Little, R. J. A. and Zheng, H. (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. *Annals of Statistics* **16**, 1684–1695.

Meng, X. L. (1994). Multiple imputation inferences with uncongenial sources of input. *Statistical Science* **9**, 538–558.

Olkin, I. and Tate, R. F. (1961). Multivariate correlation models with mixed discrete and continuous variables. *Annals of Mathematical Statistics* **32**, 448–465.

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

Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics* **19**, 1–16.

Reiter, J. P., Raghunathan, T. E., and Kinney, S. K. (2006). The importance of modeling the sampling design in multiple imputation for missing data. *Survey Methodology* **32**, 143–149.

Rubin, D. B. (1981). The Bayesian bootstrap. *Annals of Statistics* **9**, 130–134.

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

Rubin, D. and Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association* **81**, 366–374.

Schafer, J. L. (1997a). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

Schafer, J. L. (1997b). Imputation of missing covariates under a multivariate linear mixed model. *Technical Report 97-04, Dept. of Statistics, The Pennsylvania State University, http://www.stat.psu.edu/reports/1997/tr9704.pdf*.

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

Seaman, S. R., White, I. R., Copas, A. J., and Li, L. (2012). Combining multiple imputation and inverse-probability weighting. *Biometrics* **68**, 129–137.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software* **45**, 1–67.

Zangeneh, S. Z., Keener, R. W., and Little, R. J. (2011). Bayesian nonparametric estimation of finite population quantities in absence of design information on nonsampled units. *ASA Proceedings of the Section on Survey Research Methods*, 3429–3440.

Zheng, H. and Little, R. J. (2004). Penalized spline nonparametric mixed models for inference about a finite population mean from two-stage samples. *Survey Methodology* **30**, 209–218.

Zheng, H. and Little, R. J. (2005). Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model. *Journal of Official Statistics* **21**, 1–20.

Received March 2014. *Revised July* 2015. *Accepted August* 2015.