Published in final edited form as:

J Surv Stat Methodol. 2016 June 1; 4(2): 139–170. doi:10.1093/jssam/smv031.

Multiple Imputation in Two-Stage Cluster Samples Using The Weighted Finite Population Bayesian Bootstrap

Hanzhi Zhou,

Statistician, Mathematica Policy Research, Princeton, NJ, USA

Michael R. Elliott* [Professor of Biostatistics], and

Department of Biostatistics, University of Michigan School of Public Health, and the Survey Methodology Program, Institute for Social Research, University of Michigan, Ann Arbor, MI, USA

Trivellore E. Raghunathan [Professor]

Department of Biostatistics, University of Michigan School of Public Health, and Director, Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI, USA

Abstract

Multistage sampling is often employed in survey samples for cost and convenience. However, accounting for clustering features when generating datasets for multiple imputation is a nontrivial task, particularly when, as is often the case, cluster sampling is accompanied by unequal probabilities of selection, necessitating case weights. Thus, multiple imputation often ignores complex sample designs and assumes simple random sampling when generating imputations, even though failing to account for complex sample design features is known to yield biased estimates and confidence intervals that have incorrect nominal coverage. In this article, we extend a recently developed, weighted, finite-population Bayesian bootstrap procedure to generate synthetic populations conditional on complex sample design data that can be treated as simple random samples at the imputation stage, obviating the need to directly model design features for imputation. We develop two forms of this method: one where the probabilities of selection are known at the first and second stages of the design, and the other, more common in public use files, where only the final weight based on the product of the two probabilities is known. We show that this method has advantages in terms of bias, mean square error, and coverage properties over methods where sample designs are ignored, with little loss in efficiency, even when compared with correct fully parametric models. An application is made using the National Automotive Sampling System Crashworthiness Data System, a multistage, unequal probability sample of U.S. passenger vehicle crashes, which suffers from a substantial amount of missing data in “Delta-V,” a key crash severity measure.

Keywords

Cluster sampling; Complex sample design; Missing data; National Automotive Sampling System Crashworthiness Data System (NASS-CDS)

For permissions, please journals.permissions@oup.com

*Address correspondence to Michael R. Elliott, Department of Biostatistics, University of Michigan School of Public Health, 1415 Washington Heights, Ann Arbor, MI, 48109, USA; mrelliot@umich.edu.

1. Introduction

Multistage sampling is often employed for cost and convenience reasons. For example, area probability sampling is a cost-effective way to select households because households are naturally clustered by geographic areas (e.g., primary sampling units [PSUs]/counties, blocks). It is often the case that units in the same cluster tend to be more alike than units in different clusters. This leads to clustering effects, typically in the form of intraclass correlation (ICC) (Kish 1965). Because a positive ICC implies an effective sample size less than the total sample, accounting for clustering effects is important.

Methods for doing this have been well developed for complete-data approaches using either robust sandwich estimators based on Taylor Series approximations (Binder 1983) or replication methods such as balanced repeated replication, jackknife, or bootstrap (Rust and Rao 1996). A model-based approach is multilevel modeling (Rabe-Hesketh and Skrondal 2006), which decomposes the overall variance estimates into components of variance due to different levels of sample selection.

Concurrent with the development of methods for clustered sample designs, multiple imputation (MI) (Rubin 1987) has become a widely used method for dealing with item-level missing data in complex sample surveys. The need to incorporate sample design information into the imputation process has been recognized; for example, Rubin (1996) in his MI review paper stated that any imputation model should minimally include major stratification and clustering indicators as well as design weights to make it proper. Schafer (1997) also asserted that, whether the estimation techniques ultimately applied to survey data are model based or design based, imputation models for missing values should incorporate important features of the sample design. However, as this can be difficult to do in practice, imputation is typically performed under an assumption of simple random sampling (SRS).

This issue is considered in detail by Reiter, Raghunathan, and Kinney (2006). They attempt to incorporate clustered sample designs in the model by including cluster membership indicators as fixed effects in the imputation, finding large increases in variance when the number of clusters is large. Andridge (2011) also considered fitting a fixed cluster effects imputation model to data from cluster randomized trials, finding an upward bias in the MI variance estimator through both analytical proof and simulation studies.

An alternative approach consistent with standard model-based approaches is to impute data using a random-effects model, where each cluster has its own random intercept, usually assumed to be sampled from a normal distribution with zero mean and unknown variance. Software packages allowing the use of mixed-effects models for MI include (i) the *R pan* package (Schafer and Yucel 2002) for imputing continuous variables in clustered or panel designs under a multivariate normal model and (ii) the *REALCOM-IMPUTE* module of the multilevel model fitting software MLwiN (Carpenter, Goldstein, and Kenward 2011). This also adopts a joint modeling approach and Markov Chain Monte Carlo (MCMC)-based solutions but assumes a multivariate latent normal model with random effects for mixed-type response variables subject to missingness. Although these imputation models achieve

congeniality (Meng 1994) with respect to the cluster effects and result in more efficient estimates than a fixed-effect approach, their use in practice is limited due to several shortcomings: the *pan* package, for example, assumes normality on binary variables, and the linear mixed-effects imputation model typically leads to implausible values that need rounding (to the nearest 0 or 1) after imputation. This can cause biased estimates of parameters (Horton, Lipsitz, and Parzen 2003). With *REALCOM-IMPUTE*, uncongeniality remains an issue in the sense that conditional relationships among variables in the imputation model are typically assumed to be linear while the models of interest usually include nonlinear relationships and/or interactions. Also, both procedures are confined to relatively simpler data structures with no more than two levels.

On another front, attempts to extend the variable-by-variable imputation methodology to clustered designs, for example, the sequential hierarchical regression imputation models (*SHRIMP*) for multivariate clustered data (Yucel and Raghunathan 2006; Zhao and Yucel 2009; Yucel 2011), demonstrate problems with convergence. There is evidence of poor finite-sample repeated sampling properties in logistic mixed-effect models with high ICC. Other problems with the fully parametric imputation methods include the following: (i) imputations drawn from these models may perform poorly if the random effects follow a non-normal distribution; and (ii) the accommodation of both sample weights and cluster effects simultaneously is particularly problematic, especially in high-dimensional settings with many missing covariates. Such an accommodation requires the estimation of interactions between clusters and weights, as well as between weights and other quantities of interest (e.g., regression parameters). This can be difficult to achieve using a standard parametric approach.

Thus this article focuses on developing an adaptation of the two-step semi-parametric MI procedure proposed in Zhou (2014) for disentangling clustering effects and sampling weight effects using a weighted finite population Bayesian bootstrap (FPBB) procedure developed in Dong, Elliott, and Raghunathan (2014). The new MI procedure is designed to produce draws from the posterior predictive distribution of the population that incorporate both clustering and weighting elements of the complex sample design. Item-level missingness is incorporated in these “uncomplexed” synthetic populations; these missing data can then be imputed under an independent and identically distributed (IID) assumption without explicit modeling of clustering and weight effects. We consider two-stage unbalanced cluster samples obtained from an unequal probability of selection design, and we assume a missing-at-random (MAR) missing data mechanism on a single survey outcome variable (Y). The parameters of interest we consider are the population mean of Y and population regression parameters of Y on a covariate X based on the multiply imputed data.

Section 2 discusses MI under SRS, together with standard fixed-effects and random-effects models to incorporate cluster effects. Section 3 develops the newly proposed weighted FPBB procedure. Section 4 conducts and discusses a series of simulation studies designed to assess the repeated sampling properties of MI under the various approaches discussed in sections 2 and 3. These simulation studies use different population models, differing numbers of clusters, differing degrees of ICC, and differing MAR mechanisms (dependent only on the fully observed covariate independent of sample weights or dependent on both

the covariate and sample weights). Section 5 applies the different MI procedures to the analysis of passenger vehicle injury data from the National Automotive Safety System Crashworthiness Data System (NASS-CDS) survey. Section 6 concludes with a discussion and suggestions for next research steps.

1.1 Fully Parametric Imputation Methods for Clustered Sample Designs

1.1.1 SRS model—Let Y_i be the survey outcome with missing values (assumed scalar for ease of exposition) and $\mathbf{X}_i$ be a p -variate vector of other survey variables whose values are known for all cases in the sample $i = 1, \dots, n$. We assume that Y is a member of the exponential family and that the known transformation of $E(Y_i|\mathbf{X}_i) = \mu_i$ can be modeled as a linear function of $\mathbf{X}$: $g(\mu_i) = \mathbf{X}_i^T \beta$, with $\text{var}(Y_i|\mathbf{X}_i) = \sigma_e V(\mu_i)$ for a known variance function V and a possibly estimated scale parameter σ_e . We assume independence of the Y_i and equal probability of selection. A Gibbs sampler can then be used to simulate the joint posterior of the unknown parameters $\theta = (\beta^T, \sigma_e^2)^T$ and the missing data $\mathbf{y}_{\text{mis}}$:

1. Draw model parameters from their posterior distributions conditional on the “filled-in data” (imputed plus observed) from the previous draw:

$$\theta^{(t+1)} \sim p(\theta | \mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}}^{(t)}).$$

2. Draw missing values in Y from its posterior given all other parameters drawn:

$$\mathbf{y}_{\text{mis}}^{(t+1)} \sim f(\mathbf{y}_{\text{mis}} | \mathbf{y}_{\text{obs}}, \theta^{(t+1)})$$

After a sufficient number of draws have been obtained to attain convergence of the Gibbs sampler (Gelman and Rubin 1992), M widely separated draws of the missing data are

combined with the observed data to form the completed data $\mathbf{y}_{\text{comp}}^{(m)} = (\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{imp}}^{(m)})$, $m = 1, \dots, M$. Inference about the population quantity of interest Q is estimated using the combining rules of Rubin (1987):

$$\begin{aligned} \hat{Q} &= M^{-1} \sum_{m=1}^M Q(\mathbf{y}_{\text{comp}}^{(m)}) \\ T^{1/2}(\hat{Q} - Q) &\sim t_{v_m} \\ T &= U + (1 + M^{-1})V_B \\ U &= M^{-1} \sum_{m=1}^M \hat{v}ar(Q(\mathbf{y}_{\text{comp}}^{(m)})) \\ V_B &= (M - 1)^{-1} \sum_{m=1}^M (\hat{Q} - Q(\mathbf{y}_{\text{comp}}^{(m)}))^2 \\ v_m &= (M - 1) \left[1 + \frac{U}{(1 + M^{-1})V_B} \right]^2, \end{aligned}$$

where $Q(\mathbf{y}_{\text{comp}}^{(m)})$ is the (presumed scalar) point estimate obtained from the m th completed dataset $\mathbf{y}_{\text{comp}}^{(m)}$, U is the within-imputation variance calculated as the average of variance

estimates based on the M completed datasets, V_B is the between-imputation variance, and v_m is the degrees of freedom associated with the t reference distribution for inference about Q .

1.1.2 Fixed cluster effects model—We expand the SRS model to $g(\mu_i) = \mathbf{X}_i^T \beta + \mathbf{Z}_i^T \gamma$, for $i = 1, \dots, n$, where $\mathbf{Z}_i$ is a $(c-1)$ -dimension vector of dummy variables for the c clusters selected in the sample. Alternatively, we can further expand to include interactions between the fully observed covariates and the cluster dummy variables:

$g(\mu_i) = \mathbf{X}_i^T \beta + \mathbf{Z}_i^T \gamma + \mathbf{Z}_i^T \otimes \mathbf{Z}_i^T \eta$. Imputation proceeds as in the SRS setting, with expanded parameter space $\theta = (\beta^T, \gamma^T, \eta^T, \sigma_e)^T$.

1.1.3 Random cluster effects model—Here we group the variables by cluster for the $i = 1, \dots, c$ clusters with $j = 1, \dots, m_i$ observations. Our imputation model then becomes

$g(\mu_{ij}) = \mathbf{X}_{ij}^T \beta + \omega_i$, where we assume that the random cluster effects are distributed as $\omega_i \sim N(0, \sigma_\omega)$ independently for $i = 1, \dots, c$. The unknown parameters in the model are now $\theta = (\beta^T, \sigma_\omega, \sigma_e)$ and the set of ω_i . A Gibbs sampler is then used to simulate the joint posterior of the unknowns:

$$\begin{aligned} \omega_i^{(t+1)} &\sim p(\omega_i | \mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}}^{(t)}, \theta^{(t)}), i = 1, \dots, c. \\ \theta^{(t+1)} &\sim p(\theta | \mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}}^{(t)}, \tilde{\omega}^{(t+1)}), \tilde{\omega} = (\omega_1, \dots, \omega_c)^T \\ \mathbf{y}_{\text{mis}}^{(t+1)} &\sim f(\mathbf{y}_{\text{mis}} | \mathbf{y}_{\text{obs}}, \tilde{\omega}^{(t+1)}, \theta^{(t+1)}). \end{aligned}$$

2. MI Using The Weighted FPBB in Clustered and Weighted Sample Designs

In this section, we develop a semiparametric two-step MI methodology to account for complex sampling designs with two-stage cluster samples. The method is simple to implement and follows a “generate” then “impute” procedure. In the first step, a weighted FPBB generates predictive draws of a population that capture the clustering and unequal probability of selection design features nonparametrically. If the probabilities of selection for the sampled clusters and the conditional probabilities of selection for the sampled elements within a sampled cluster are known, a weighted FPBB of the clusters is drawn using (4) and repeated L times; within each of these resampled clusters, a conditional weighted FPBB of the elements is drawn using (5) and repeated B times. If only the final weights (inverse of the product of the probabilities of selection for the clusters and the conditional probabilities of selection for the elements) are known, as is typically the case in public-use files, an unweighted Bayesian bootstrap of the clusters of size L is obtained, followed by a weighted FPBB of size B for the elements within each of the resampled clusters, as in (6). The second-step multiply imputes the missing data under a parametric model A times within each of the $L \times B$ populations. Inference can then be carried out by computing the mean of the statistic of interest $\hat{Q}^{(l)} = (1/BA) \sum_b \sum_a q^{(lba)}$ within each of the L first-stage bootstraps: the point estimate $\hat{Q} = L^{-1} \sum_{l=1}^L \hat{Q}^{(l)}$ is the mean of the $\hat{Q}^{(l)}$ s, and the variance is $(1 + L^{-1}) V_L$, where $V_L = (L-1)^{-1} \sum_{l=1}^L (\hat{Q}^{(l)} - \hat{Q})^2$ is the variance of the $\hat{Q}^{(l)}$ s,

which is asymptotically t distributed with $L - 1$ degrees of freedom. The remainder of the section provides detailed derivations of these steps.

2.1 Notation and Assumptions

Define $\mathbf{I}$ as a (fully observed) vector of sampling indicators for a target finite population of size N . (In multistage samples, sampling indicators below the first stage may only be fully observed in the units that were sampled at the higher stages [e.g., adults will only be fully listed within sampled households], but we ignore this in order to avoid making the notation overly complex.) To develop the population posterior predictive distribution generation for the first step, let $\mathbf{Y} = [\mathbf{Y}_s, \mathbf{Y}_{ns}]$ be the survey outcome of interest where missing data occur, which divides into the sampled part $\mathbf{Y}_s$ (for which $I_j = 1$) and the non-sampled part $\mathbf{Y}_{ns}$ (for which $I_j = 0$). Define separate response mechanisms for the sampled and the nonsampled parts of $\mathbf{Y}$: $\mathbf{R} = (\mathbf{R}_s, \mathbf{R}_{ns})$, such that $\mathbf{Y}_s = [\mathbf{Y}_{s,obs}, \mathbf{Y}_{s,mis}]$ further divides into the observed component $\mathbf{Y}_{s,obs}$ and the missing component $\mathbf{Y}_{s,mis}$, corresponding to the sampled $\mathbf{Y}$ values associated with $R_{s,i} = 1$ and $R_{s,i} = 0$, respectively, and similarly $\mathbf{Y}_{ns} = [\mathbf{Y}_{ns,obs}, \mathbf{Y}_{ns,mis}]$ divides into those that would have been observed had they been sampled ($R_{ns,i} = 1$) and those that would have had missing values ($R_{ns,i} = 0$). Note that we can recombine as $\mathbf{Y} = [\mathbf{Y}_{obs}, \mathbf{Y}_{mis}]$ where $\mathbf{Y}_{obs} = [\mathbf{Y}_{s,obs}, \mathbf{Y}_{ns,obs}]$ and $\mathbf{Y}_{mis} = [\mathbf{Y}_{s,mis}, \mathbf{Y}_{ns,mis}]$. In a similar fashion, let $\mathbf{X} = [\mathbf{X}_s, \mathbf{X}_{ns}]$ be the complete covariate, and $\mathbf{Z} = [\mathbf{Z}_s, \mathbf{Z}_{ns}] = [(\mathbf{w}_s, \mathbf{C}_s), (\mathbf{w}_{ns}, \mathbf{C}_{ns})]$ be the complete design matrix, which contains all of the essential sample design features (here we restrict to sampling weights $\mathbf{w}$ and cluster membership indicators $\mathbf{C}$). Note that $\mathbf{w}$ is a matrix consisting of rows $\{(w_{ji}), (w_{ji}, w_{ij}), i = 1, \dots, n, j = 1, \dots, m_i\}$, where w_{ji} is the cluster-level weight for element j in cluster i , w_{ji} is the element-level conditional weight for element j given cluster i is selected, and $w_{ij} = w_i w_{ji}$ is the overall sampling weight for element j .

Assuming ignorable sampling—that is, that the sampling indicators are at most dependent on observed data, $p(\mathbf{I}|\mathbf{Y}, \mathbf{X}, \mathbf{Z}, \mathbf{R}) \propto p(\mathbf{I}|\mathbf{Y}_s, \mathbf{X}_s, \mathbf{Z}_s, \mathbf{R}_s)$ —the joint predictive distribution of the nonsampled part of the population is given as (Rubin 1987):

$$p(\mathbf{Y}_{ns}, \mathbf{X}_{ns}, \mathbf{Z}_{ns}, \mathbf{R}_{ns} | \mathbf{Y}_s, \mathbf{X}_s, \mathbf{Z}_s, \mathbf{R}_s, \mathbf{I}) \propto p(\mathbf{Y}_{ns}, \mathbf{X}_{ns}, \mathbf{Z}_{ns}, \mathbf{R}_{ns} | \mathbf{Y}_s, \mathbf{X}_s, \mathbf{Z}_s, \mathbf{R}_s). \quad (1)$$

Explicit modeling for $\mathbf{I}$ could be avoided by incorporating design variables ($\mathbf{Z}$) into a model for $p(\mathbf{Y}_{ns}, \mathbf{X}_{ns}, \mathbf{Z}_{ns}, \mathbf{R}_{ns} | \mathbf{Y}_s, \mathbf{X}_s, \mathbf{Z}_s, \mathbf{R}_s)$ that conditionally eliminates dependence on $\mathbf{I}$ (Gelman 2007; Little 2011). For example, in a stratified sample design, a model for a population mean that assumes separate means within each stratum allows (1) to hold (Holt and Smith 1979; Elliott and Little 2000) because, within each stratum, sampling is ignorable ($p(\mathbf{I}|\mathbf{Y}, \mathbf{X}, \mathbf{Z}, \mathbf{R}) \propto p(\mathbf{I}|\mathbf{Z}) = p(\mathbf{I}|\mathbf{Z}_s)$) because the strata are fully observed. However, the practical implementation of this in a robust manner using standard parametric models can be daunting; for example, in Elliott and Little (2000), normality was assumed to allow for tractability. In more complex settings, for example, involving unequal probabilities of selection at an individual level (e.g., probability proportional to size) or involving cluster sampling, assumptions might have to be complex to be reasonable (e.g., using smoothing splines to relate probability of selection to means [Zheng and Little 2003]) or might have to

be strong to be tractable (e.g., normality for clusters [Reiter et al. 2006]). Matters grow even more complex if more complicated statistics such as population regression parameters are of interest, in which case, for (1) to hold in the presence of unequal probabilities of selection, interactions between the weights and the independent variables of interest must be specified (Elliott 2007). Hence we are motivated to avoid modeling to yield (1) and instead pursue a non-parametric approach using a weighted FPBB that combines a Polya sampling scheme developed by Ghosh and Meeden (1997) in the SRS setting and extended by Meeden (1999) into a simple balanced cluster design setting with a second extension by Cohen (1997) to accommodate weighted data. As long as correct selection weights and sampling clusters are specified for the sampled data, our nonparametric method will allow (1) to hold.

The second step requires a correct imputation model that generates the posterior predictive distribution of missing values in the entire population. In this article, we will assume an MAR mechanism so that we do not need an explicit model for R :

$$p(\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}, \mathbf{X}, \mathbf{Z}, \mathbf{R}) \propto \begin{cases} p(\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}, \mathbf{X}), & \text{if } p(\mathbf{R}|\mathbf{Y}, \mathbf{X}, \mathbf{Z}) \propto p(\mathbf{R}|\mathbf{Y}_{\text{obs}}, \mathbf{X}) \\ p(\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}, \mathbf{X}), & \text{if } p(\mathbf{R}|\mathbf{Y}, \mathbf{X}, \mathbf{Z}) \propto p(\mathbf{R}|\mathbf{Y}_{\text{obs}}, \mathbf{X}, \mathbf{Z}) \end{cases} \quad (2)$$

Here we will proceed with a standard parametric approach, modeling $\mathbf{Y}_{\text{mis}}$ conditional on the observed $\mathbf{X}$ and $\mathbf{Z}$. Note that the extent to which $\mathbf{Z}$ should be incorporated into the imputation model still depends on the specific form of the response mechanism, as shown in (2).

2.2 The Weighted FPBB in Clustered and Weighted Sample Designs

In a simple two-stage cluster sampling (balanced case) design with equal probabilities of selection at both cluster and element levels, Meeden (1999) proposed a “two-stage Polya posterior” approach to simulate draws that form an entire population of clusters and then an entire population of elements within each cluster. He showed that the posterior mean and variance are very close to standard design-based results. Cohen (1997) proposed a method to generate draws from a weighted Polya posterior using data obtained from weighted sample designs in a nonclustered setting. Here we extend the method of Meeden to a two-stage cluster sampling design with unequal cluster sizes (unbalanced case) and combine it with the approach of Cohen to further handle unequal probability of selection.

We propose two variations of a two-stage procedure to simulate draws from (1). The first approach assumes that we know the sampling probabilities at both sampling stages and uses a weighted FPBB at each stage to generate a population of clusters and then a population of elements (and hence is termed an “adapted two-stage Polya posterior” approach). The second approach requires only the final weights, as is typical in most public-use surveys, and uses a Bayesian bootstrap at the first stage to account for clustering effects and a weighted FPBB using the final weights to generate the population of elements. Details are given below, but conceptually this procedure generates a draw of the posterior predictive distribution of the population that is free of design effects due to sample weights at both cluster and element levels as well as design effects due to cluster sampling. A simple

illustrative example goes as follows: Suppose we obtain a sample of households (HHs) from a two-stage cluster sampling design, where all U.S. metropolitan statistical areas (MSAs) are treated as clusters to be sampled at the first stage. Suppose Detroit is among such sampled clusters. By using the proposed procedure, we first replicate the number of all of the “Detroit-like MSAs” based on the sample weight for Detroit and use it to update the conditional sample weights of the sampled HHs within Detroit; then we replicate the number of all of the nonsampled HHs in Detroit as well as that in the Detroit-like MSAs in the population. The same reasoning applies to other sampled MSAs that are different from Detroit (e.g., Los Angeles and Los Angeleslike MSAs). Once we generate the population of HHs, HH-level inference becomes straightforward without the need to consider effects due to sampling at the MSA level.

2.2.1 Double-weighted FPBB (SYN1)—We will call the first approach “double-weighted FPBB.” As indicated by the name, we propose to simulate the posterior predictive distribution in (1) by utilizing the weighted FPBB method at both the PSU (or cluster) level and the element level such that a synthetic population of clusters is created at first and then a synthetic population of elements is created. Our discussion below parallels that of Meeden (1999), extended to allow for multiple variables and sampling weights as in Cohen (1997).

Let $D = \{(Y_{ij}, \mathbf{X}_{ij}^T, \mathbf{Z}_{ij}^T, R_{ij})\}$, $i = 1, \dots, N$, $j = 1, \dots, M_i$ denote the population of values for element j within cluster i , where N and M_i are the number of clusters and the number of elements within the i th cluster in the population, respectively. Thus the population size is

$\sum_{i=1}^N M_i = M$. Let $\{\mathbf{b}^q\}$, $q = 1, \dots, r$, be $r(1 \leq r \leq N)$ distinct matrices of real numbers each of dimension $|b_{\text{row}}^q| \times |b_{\text{col}}^q|$ (each column vector corresponds to a survey variable) with no row vectors in common so that each matrix consists of one of the possible sets of values for all of the survey variables for a given cluster. Thus each cluster in the population can take the form of one of the $\mathbf{b}^q$ s; let $\mathbf{t}$ be a vector of length N , $t_i = q$ when the elements in the i th cluster take on the values of $\mathbf{b}^q$ for $i = 1, \dots, N$. Finally, let $c(q) =$ the number of t_i s that equal q , for $q = 1, \dots, r$, and $c_{t,D}^i(k) =$ the number of $(Y_{ij}, \mathbf{X}_{ij}^T, \mathbf{Z}_{ij}^T, R_{ij})$ s, for $k = 1, \dots, b_k^{t_i}$, $i = 1, \dots, N$.

Suppose a two-stage cluster sample is selected from the above finite population. Thus $\mathbf{D}$ divides into the sampled components $\mathbf{D}_s = (\mathbf{Y}_s, \mathbf{X}_s, \mathbf{Z}_s, \mathbf{R}_s)$ and the nonsampled components $\mathbf{D}_{ns} = (\mathbf{Y}_{ns}, \mathbf{X}_{ns}, \mathbf{Z}_{ns}, \mathbf{R}_{ns})$. Let h and $\bar{h}$ each represent the collection of the sampled and nonsampled clusters, and thus we can break $\mathbf{t}$ into $\mathbf{t} = (\mathbf{t}_h, \mathbf{t}_{\bar{h}})$. Let n denote the number of sampled clusters and m_i the number of sampled elements within the i th sampled cluster, and thus the total sample size is $m = \sum_{i=1}^n m_i$. For convenience of exposition, assume that each cluster has a unique set of observations, so that $n = r$, $m_i = \|\mathbf{b}^{t_{hi}}\|$ the number of distinct row vectors in $\mathbf{b}^{t_{hi}}$, and thus $m = \sum_{i=1}^n \|\mathbf{b}^{t_{hi}}\|$. Let $w_{t_{hi}}(i) =$ the first-stage sample weight of the sampled cluster $i \in h$ that equals $\mathbf{b}^q$ for $i \equiv q) = 1, \dots, n(\equiv r)$, normalized to sum to the number of clusters: $\sum_{i=1}^n w_{t_{hi}}(i) = n$, and $w_{t_{hi}, D_s}(j) =$ the second-stage sample weight of the j th

sampled element in the j th sampled cluster that equals $\mathbf{b}_k^{t_{h_i}}$, for $j \equiv k = 1, \dots, m_j \equiv \|\mathbf{b}^{t_{h_i}}\|$,

normalized to sum to the number of observations in a cluster: $\sum_{j=1}^{m_i} w_{t_{h_i}, D_s}(j) = m_i$.

Let the set of distinct clusters in the observed sample ($\mathbf{D}_s$) be $\{\mathbf{b}^1, \dots, \mathbf{b}^q, \dots, \mathbf{b}^r\}$, and $\boldsymbol{\varsigma} = \{\varsigma_1, \varsigma_2, \dots, \varsigma_r\}$ be the vector of probabilities that $\Pr(\mathbf{D}_{s,i} = \mathbf{b}^q | \boldsymbol{\varsigma}) = \varsigma_q$ for $i = 1, 2, \dots, n$, and $\sum_{q=1}^r \varsigma_q = 1$. Let $c_{t_h}(q)$ and $c_{t_h^-}(q)$ be the numbers of clusters equal to $\mathbf{b}^q$ in the sampled and

the nonsampled parts of the population, respectively, for $q = 1, 2, \dots, r$, $\sum_{q=1}^r c_{t_h}(q) = n$, and $\sum_{q=1}^r c_{t_h^-}(q) = N - n$. Let the set of distinct elements in the sampled clusters be $\{\mathbf{d}^1, \dots, \mathbf{d}^k, \dots, \mathbf{d}^l\}$, $\boldsymbol{\lambda} = \{\lambda_1, \lambda_2, \dots, \lambda_l\}$ be the vector of probabilities that $\Pr(\mathbf{D}_{s,j} = \mathbf{d}^k | \boldsymbol{\lambda}) = \lambda_k$ for $i =$

$1, 2, \dots, n, j = 1, \dots, m_j$ and $\sum_{k=1}^l \lambda_k = 1$, and $c_{t, D_s}^i(k)$ and $c_{t, D_{ns}}^i(k)$ be the numbers of elements equal to $\mathbf{d}^k$ in the sampled and the nonsampled parts of the population,

respectively, for $k = 1, 2, \dots, l$, and $\sum_{k=1}^l c_{t, D_s}^i(k) = m$ and $\sum_{k=1}^l c_{t, D_{ns}}^i(k) = M - m$.

Assuming a noninformative Haldane prior of $\boldsymbol{\varsigma}$: $\boldsymbol{\varsigma} \sim \text{Dir}(0, \dots, 0)$ together with multinomially distributed weighted counts of distinct clusters in the sample data,

$p(w_{t_h}(1), \dots, w_{t_h}(r) | \boldsymbol{\varsigma}) \propto (\prod_{q=1}^r \varsigma_q^{w_{t_h}(q)})$, yields a Dirichlet posterior distribution of $\boldsymbol{\varsigma}$: $\boldsymbol{\varsigma} | w_{t_h}(1), \dots, w_{t_h}(r) \sim \text{Dir}(w_{t_h}(1), \dots, w_{t_h}(r))$. Similarly, the multinomially distributed weighted counts of distinct observations within a sampled cluster

$p(w_{t_h, D_s}(1), \dots, w_{t_h, D_s}(l) | \boldsymbol{\lambda}) \propto \prod_{k=1}^l \lambda_k^{w_{t_h, D_s}(k)}$ yields a Dirichlet posterior distribution of $\boldsymbol{\lambda}$.

We then have the adapted two-stage Polya posterior derived as follows (for simplicity of exposition and by an abuse of notation, we use $\mathbf{t}_h$ and $\mathbf{t}_h^-$ to represent $c_{t_h}(1), \dots, c_{t_h}(r)$ and

$w_{t_h}(1), \dots, w_{t_h}(r)$, respectively; similarly, we use $\mathbf{D}_{ns}$ and $\mathbf{D}_s$ for $c_{t, D_{ns}}^i(1), \dots, c_{t, D_{ns}}^i(l)$ and $w_{t_h, D_s}(1), \dots, w_{t_h, D_{ns}}(l)$, respectively):

$$\begin{aligned}
p(\mathbf{D}_{ns}|\mathbf{D}_s) &= \frac{p(\mathbf{D})}{p(\mathbf{D}_s)} = \frac{p(\mathbf{t}) \cdot p(\mathbf{D}|\mathbf{t})}{p(\mathbf{t}_h) \cdot p(\mathbf{D}_s|\mathbf{t}_h)} \\
&= \frac{p(\mathbf{t}_h, \mathbf{t}_{\bar{h}}) \cdot p(\mathbf{D}_{ns}, \mathbf{D}_s|\mathbf{t}_h, \mathbf{t}_{\bar{h}})}{p(\mathbf{t}_h) \cdot p(\mathbf{D}_s|\mathbf{t}_h)} \\
&= \frac{p(\mathbf{t}_h, \mathbf{t}_{\bar{h}})}{p(\mathbf{t}_h)} \cdot \frac{p(\mathbf{D}_{ns}|\mathbf{D}_s, \mathbf{t}_h, \mathbf{t}_{\bar{h}}) p(\mathbf{D}_s|\mathbf{t}_h, \mathbf{t}_{\bar{h}})}{p(\mathbf{D}_s|\mathbf{t}_h)} \\
&= p(\mathbf{t}_{\bar{h}}|\mathbf{t}_h) \cdot p(\mathbf{D}_{ns}|\mathbf{D}_s, \mathbf{t}_h, \mathbf{t}_{\bar{h}}) \\
&= \frac{\int_0^1 \cdots \int_0^1 \prod_{q=1}^{r-1} \zeta_q^{w_{t'}(q)-1} \left(1 - \sum_{q=1}^{r-1} \zeta_q\right)^{w_{t'}(r-1)-1} d\zeta_1 \cdots d\zeta_{r-1}}{\int_0^1 \cdots \int_0^1 \prod_{q=1}^{r-1} \zeta_q^{w_{t'}(q)-1} \left(1 - \sum_{q=1}^{r-1} \zeta_q\right)^{w_{t_h}(r-1)-1} d\zeta_1 \cdots d\zeta_{r-1}} \\
&\times \frac{\int_0^1 \cdots \int_0^1 \prod_{k=1}^{l-1} \lambda_k^{w_{t', D_{ns}}(k)-1} \left(1 - \sum_{j=1}^{l-1} \lambda_j\right)^{w_{t', D_{ns}}(l-1)-1} d\lambda_1 \cdots d\lambda_{l-1}}{\int_0^1 \cdots \int_0^1 \prod_{k=1}^{l-1} \lambda_k^{w_{t_h, D_s}(k)-1} \left(1 - \sum_{j=1}^{l-1} \lambda_j\right)^{w_{t_h, D_s}(l-1)-1} d\lambda_1 \cdots d\lambda_{l-1}} \\
&= \frac{\left\{ \prod_{q=1}^r \frac{\Gamma(w_{t'}(q))}{\Gamma(w_{t_h}(q))} \right\}}{\left\{ \frac{\Gamma(N)}{\Gamma(n)} \right\}} \\
&\times \frac{\left\{ \prod_{k=1}^l \frac{\Gamma(w_{t', D_{ns}}(k))}{\Gamma(w_{t_h, D_s}(k))} \right\}}{\left\{ \frac{\Gamma(M)}{\Gamma(m)} \right\}} \quad (3)
\end{aligned}$$

where $w_t(q) = w_{th}(q) + c_{th}^i(q)$, $w_{th, D_s}(k)$ is the element-level weight and $w_{t', D_{ns}}(k) = w_{th, D_s}(k) + c_{t, D_{ns}}^i(k)$. We note that this derivation assumes a weighted multinomial likelihood, as in Cohen 1997 and Dong et al. 2014; a reviewer notes that, because this might not be considered a true likelihood, the resulting posterior and posterior predictive distributions might be considered as approximations. However, we find that its properties work well, both in derivation and simulation.

The posterior distribution in (3) does not lend itself to direct calculation, but draws can be obtained by a Monte Carlo simulation procedure that extends the simulation described in Lo (1998) into the unequal probability of selection setting (for details, see Dong et al. 2014, section 3.2):

Step 1: Apply the weighted FPBB to sampled clusters; that is, draw non-sampled clusters in the population based on the weighted Polya urn distribution. Denote the original sample of clusters by $\{c_1, c_2, \dots, c_n\}$. In generating $N - n$ clusters $\{c_1^*, c_2^*, \dots, c_{N-n}^*\}$ from the original n , we resample the sampled clusters with probability

$$\frac{w_{t_h}(i) - 1 + l_{i, k-1} \times (N - n/n)}{N - n + (k - 1) \times (N - n/n)}, k=1, \dots, (N - n + 1), i=1, \dots, n, \quad (4)$$

where $l_{i, k-1}$ is the number of times that the i th cluster has been sampled at the $(k - 1)$ th resampling and $w_{th}(i)$ is the weight for the i th cluster, which is normalized to sum up to the total number of clusters; that is, $\sum_{i=1}^n w_{th}(i) = N$. Repeat step 1 L times to obtain L FPBB synthetic populations of clusters.

Step 2: For each repetition of step 1, form a population of clusters $\{c_1, c_2, \dots, c_n, c_1^*, c_2^*, \dots, c_{N-n}^*\}$. Record the number of times each of the n clusters from the original sample appears in the FPBB population of clusters, denoted by $\tau_i, i = 1, \dots, n$ and $\sum_{i=1}^n \tau_i = N$. Then update the within-cluster element-level conditional weights to unconditional weights as follows: $w_{ij}^* = w_{j|i} \times \tau_i, i = 1, \dots, n$, where $w_{j|i}$ is the inverse of the conditional probability that element j is sampled given that cluster i is sampled. (Note w_{ij}^* is the same as $w_{th_i D_s}(j)$ previously defined.) Now each observed element in these n clusters is associated with an updated weight that represents all nonsampled elements in both the sampled cluster it belongs to and similar nonsampled clusters in the population. Note that the resulting sample has the same sample size of elements (m) but different survey weights than the original sample, and we term it an “FPBB sample.”

We then apply the weighted FPBB again to elements in the FPBB sample. In generating $M - m$ units from the m elements in the FPBB sample, the j th element in the i th cluster in the FPBB sample is resampled at the k th replication with probability

$$\frac{w_{ij}^* - 1 + l_{ij,k-1} \times (M - m/m)}{M - m + (k - 1) \times (M - m/m)}, k = 1, \dots, (M - m + 1), i = 1, \dots, n, j = 1, \dots, m_i, \quad (5)$$

where w_{ij}^* is the updated conditional weight for the j th element in the i th cluster. Note that we are resampling the clusters at the first stage so that the elements within the cluster are “pulled through” at that stage; the w_{ij}^* s inherit the information from the cluster-level selection proportion τ_i/N and are again normalized to sum up to the total number of elements in the entire population—that is, $\sum_i \sum_{j=1}^{m_i} w_{ij}^* = M$. Once this first stage of resampling is complete, the cluster IDs can be dropped and resampling completed at the second stage, thus creating a single synthetic population. Repeat step 2 B times to obtain B FPBB synthetic populations of elements for each of the L clusters.

2.2.2 Bayesian bootstrap weighted FPBB (SYN2)—A disadvantage of the procedure proposed in 2.1.1 is that, to implement, it requires knowledge of both the first- and second-stage selection probabilities, which are often not provided in public-use datasets. Here we propose an alternative method that uses a regular Bayesian bootstrap (BB) for each of the L synthetic populations of clusters to capture the cluster-level sampling variance. Instead of the adjusted element-level conditional weights (i.e., w_{ij}^*), the final sample weights (w_{ij}) are then directly used as input in the second-stage weighted FPBB. We will show that this procedure not only serves as a handy approximation to the double-weighted FPBB due to easier implementation in practice, but it can also deal with complex missing data problems as effectively as the double-weighted FPBB where fully parametric MI fails. The full procedure is described as follows:

Step 1: Apply BB to draw a sample of the clusters from the original sample of clusters:

1. Sample n clusters with replacement with the vector of probabilities as described in Rubin, and record the number of times (τ_i) each cluster is selected in the BB sample, $\sum_{i=1}^n \tau_i = n$.
2. Modify the initial case weight in the parent sample as follows:

$$w_{ij}^* = \begin{cases} \tau_i \cdot w_{ij}, & \text{if } i\text{th cluster is selected or duplicated;} \\ 0, & \text{if } i\text{th cluster is not selected.} \end{cases},$$

Normalize the w_{ij}^* s to sum up to the population size. Note that the bootstrap sample size (m^*) is different from the parent sample size (m). Repeat step 1 L times to obtain L FPBB synthetic populations of clusters.

Step 2: Generate nonsampled elements of the population accounting for inclusion weights: Select a FPBB Polya sample of size $M^* = M - m^*$ by resampling the j th element in the i th cluster in the FPBB sample at the k th replication with probability

$$\frac{w_{ij}^* - 1 + l_{ij,k-1} \times (M - m^*)/m^*}{M - m^* + (k-1) \times (M - m^*)/m^*}, k=1, \dots, (M - m^* + 1), i=1, \dots, n, j=1, \dots, m_i, \quad (6)$$

where w_{ij}^* is the updated conditional weight for the j th element in the i th cluster. Simulate many (B) copies of the nonsampled population for each of the L clusters.

2.3 Multiply Imputing Missing Data

Denote the $L \times B$ FPBB “unweighted” synthetic populations generated by either of the weighted FPBB procedures in section 2.2 by $\mathbf{P}^{\text{syn}} = \{\mathbf{P}_{(b)}^{(l)}, b=1, \dots, B, l=1, \dots, L, \}$, where $\mathbf{P}_{(b)}^{(l)} = (\mathbf{P}_{(b)\text{mis}}^{(l)}, \mathbf{P}_{(b)\text{obs}}^{(l)})$. We are now ready to perform conventional parametric MI under an IID assumption. Following the standard MI procedure or approximations such as sequential regression multiple imputation (Raghunathan, Lepkowski, Van Hoewyk, and Solenberger 2001), we obtain A draws for each of the synthetic populations from the posterior predictive distribution $p(\mathbf{P}_{(b)\text{mis}}^{(l)} | \mathbf{P}_{(b)\text{obs}}^{(l)})$, where $\mathbf{P}_{(b)\text{mis}}^{(l)}$ and $\mathbf{P}_{(b)\text{obs}}^{(l)}$ consist of the unobserved and observed data in the l th FPBB dataset, respectively.

Point and variance estimation then proceed as follows. Denote the “filled in” synthetic datasets as $\mathbf{P}^{\text{imp}} = \{\mathbf{P}_{(11)}^{(1)}, \dots, \mathbf{P}_{(1A)}^{(1)}, \dots, \mathbf{P}_{(B1)}^{(1)}, \dots, \mathbf{P}_{(BA)}^{(1)}, \dots, \mathbf{P}_{(BA)}^{(L)}\}$. By the standard Rubin (1987) MI combining rules, we have

$$Q | \mathbf{P}^{\text{imp}} \sim t_{L-1}(\bar{Q}_L, (1+L^{-1})V_L), \quad (7)$$

where $\bar{Q}_L = (1/L) \sum_l \bar{Q}^{(l)}$, $V_L = (L-1)^{-1} \sum_l (\bar{Q}^{(l)} - \bar{Q}_L)^2$, and

$\bar{Q}^{(l)} = \lim_{\substack{B \rightarrow \infty \\ A \rightarrow \infty}} (1/BA) \sum_b \sum_a q^{(lba)}$, where $q^{(lba)}$ is an estimate of Q obtained from the a th imputation of the b th synthetic population within the l th (finite population) BB sample, obtained as $\bar{Q}^{(l)} = (1/BA) \sum_b \sum_a q^{(lba)}$. Note 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 V_L only (see Dong et al. 2014 and Zhou 2014). In small samples, we suggest using the correction $\{(n+1)/(n-1)\} (1+L^{-1}) V_L$.

3. Simulation Study

The simulation study considers two different clustered population data structures. For the first structure, we impose only clustering effects in the data model while leaving the weight independent of the data-generating mechanism.

$$\text{Structure1: } Y_{ij} \sim \alpha_0 + \alpha_1 X_{ij} + u_i + \varepsilon_{ij}, \text{ where } u_i \sim N(0, \sigma_u^2), \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2).$$

For the second structure, the data model depends on both clustering effects and the measure of size (Z) for second-stage sampling as well as its interaction with the covariate (X). Under this model, the subsampling probability is correlated with the response variable even after conditioning on the covariate:

$$\text{Structure2: } Y_{ij} \sim \beta_0 + \beta_1 X_{ij} + \beta_2 Z_{ij} + \beta_3 X_{ij} Z_{ij} + v_i + \varepsilon_{ij}, \text{ where } v_i \sim N(0, \sigma_v^2), \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2).$$

We conduct a separate factorial design study for each population structure by considering the following simulation factors:

- residual intraclass correlation coefficient ($ICC_{Y|X}$): $ICC_{Y|X} = 0.1, 0.5$;
- the number of selected clusters in the sample $n = 10, 40$, given a fixed number of clusters (M) in the hypothetical population; and
- the response mechanism ($\mathbf{R}$) for $\mathbf{Y}$: missingness depends only on the covariate matrix $\mathbf{X}$ (MAR_X), missingness depends both upon $\mathbf{X}$ and the survey weights $W(\text{MAR}_X, W)$.

Based on the above stated factors, we conduct a full $2 \times 2 \times 2$ factorial design for each data structure. Because the value of accounting for the sample design is greatest when there are large design effects, our goal in the simulation study is to consider a range of moderate to large design effects (on the order of two to twenty) due to clustering and in a setting where the missingness fractions are nontrivial (on the order of 30 percent). A total of five MI methods are compared in the simulation study. They are described as follows:

- SRS: the linear model where clustering effects are ignored altogether;
- FX: the linear fixed cluster effects model;

- RE: the linear random cluster effects model;
- SYN1: the two-step MI method with the double-weighted FPBB as the first step; and
- SYN2: the two-step MI method with the BB-weighted FPBB as the first step.

The imputation tasks under SRS, FX, SYN1, and SYN2 are all implemented using the R *mice* package (van Buuren and Groothuis-Oudshoorn 2011). The imputation under RE is implemented using the R *pan* package. Following the recommendations in the packages, we selected noninformative priors for the regression parameters and diffuse inverse chi-square priors for the variance components ($p(\sigma_u^{-2}) \sim \chi_1^2(1)$, $p(\sigma_e^{-2}) \sim \chi_1^2(1)$, where $\chi_v^2(s)$ is a chi-square distribution with v degrees of freedom scaled by s). We imputed the missing data five times, taking one hundred iterations between imputations after a burn-in period of one thousand iterations. Visual inspection of a handful of MCMC chains suggested this thinning interval was easily sufficient to ensure approximate independence among the imputations.

Finally, for imputation, we consider three model specifications:

- X -only model: $y \sim \mathcal{N}(\beta_0 + \beta_1 * x + [b], \sigma^2)$;
- Main effect model: $y \sim \mathcal{N}(\beta_0 + \beta_1 * x + \beta_2 * w + [b], \sigma^2)$; and
- Interaction model: $y \sim \mathcal{N}(\beta_0 + \beta_1 * x + \beta_2 * w + \beta_3 * x * w + [b], \sigma^2)$,

where b is a cluster-level effect, either ignored in the imputation (SRS), treated as a fixed effect (FX), or treated as a random effect (RE). For the two variations of the synthetic MI methods, we will use the X -only model for imputation throughout all scenarios.

3.1 Description of the Design

3.1.1 Step 1: Population data generation—We fix the number of clusters in the population as $N = 400$ for all data structures. Varying cluster sizes are generated from $M_i \sim \text{Unif}(50, 250)$, $i = 1, \dots, 400$, which sum up to a population size of $M = 59,501$ for population structure 1 and $M = 59,230$ for population structure 2. For structure 1, we let $\alpha_0 = 1$, $\alpha_1 = 1.5$, and $X_{ij} \sim \mathcal{N}(0, 1)$. For structure 2, we let $\log(Z_{ij}) \sim \mathcal{N}(2, 1)$, $X_{ij} \sim \mathcal{N}(0.1 * Z, 1)$, $\beta_0 = 0$, $\beta_1 = 0.2$, $\beta_2 = 0.6$, $\beta_3 = 0.5$. The random cluster effects are given by $u_i, v_i \sim \mathcal{N}(0, 1)$. To generate hypothetical populations with varying ICC , we vary the random error terms as $e_{ij}^1 \sim \mathcal{N}(0, 1)$ and $e_{ij}^2 \sim \mathcal{N}(0, 9)$, resulting in $ICC_1 = 0.5$ and $ICC_2 = 0.1$, respectively.

3.1.2 Step 2: Drawing replications of samples—We draw clusters with probabilities proportional to the cluster size M_j , with first-stage sampling fraction $f_{1i} = n \cdot M_i / \sum_{k=1}^N M_k$. The subsampling methods within each sampled cluster are different for the two data structures: we select elements using simple random sampling without replacement with an equal second-stage sampling fraction of $f_{2ij} = 1/10$ for the first data structure, while using probability proportional to size without replacement with Z as the measure of size for the second data structure so that $f_{2ij} = m_i Z_{ij} / 10 \sum_{j=1}^{M_i} Z_{ij}$. The final weight $w_{ij} = \{f_{1i} f_{2ij}\}^{-1}$ used

in SYN2 is thus dominated by the cluster-level weight for the former and the element-level weight for the latter.

3.1.3 Step 3: Imposing missingness—For samples drawn from the first data structure: under MAR_X, we use a logistic function as the deletion function, $\Pr(R_{ij} = 1) = \text{expit}(\lambda_0 + \lambda_1 X_{ij})$, for $\lambda_0 = 0.8$, and $\lambda_1 = -0.2$; under MAR_{X,W}, we define a probit model $\Pr(R_{ij} = 0 | S_{ij}) = \Phi(S_{ij})$ to generate missingness indicators, where the latent variable is $S_{ij} = \eta_0 + \eta_1 X_{ij} + \eta_2 \log(\text{weight}_{ij}) + e_{ij}$, $e_{ij} \sim N(0, 1)$ and $\eta_0 = -3.8$, $\eta_1 = -0.8$, and $\eta_2 = 1$. Similar forms of the probit model are used for data structure 2 with a different latent variable: $S_{ij} = \xi_0 + \xi_1 X_{ij} + \xi_2 Z_{ij} + \xi_3 X_{ij} * Z_{ij} + e$, $e \sim N(0, 1)$, where we set $\xi_0 = -0.635$, $\xi_1 = 0.4$, and $\xi_2 = \xi_3 = 0$ under condition MAR_X, and $\xi_0 = -0.55$, $\xi_1 = 0.4$, $\xi_2 = -0.5$, and $\xi_3 = 0.4$ under condition MAR_{X,W}. All scenarios result in a missing rate of approximately 30 percent on Y .

3.1.4 Step 4: MI inference—For each scenario, two hundred simulations are obtained. The number of imputations under the fully parametric MI methods is $A = 5$, and we set $L = 100$, $B = 20$, and $A = 5$ for the synthetic MI methods. Our focus is on estimating the population mean of Y and regression coefficients of Y on X , where the analyst's model is assumed to be $Y = b_0 + b_1 X$.

Design-based analyses are performed for original samples before deletion (termed “BD”) and for fully parametric methods (SRS/FX/RE). Specifically, we use a Horvitz-Thompson-type estimator for point estimation and an ultimate cluster variance estimator for the within-imputation variance estimation. Because neither $n = 10$ nor $n = 40$ allows for a normal approximation, we use a t reference distribution instead to construct the 95 percent confidence intervals. Different degrees of freedom (df) are used for different methods. For the actual samples BD, complete data df are used; that is, $v_{\text{com}} = n - p$ where p is the number of parameters. For the fully parametric MI method, we use the small sample df derived in Barnard and Rubin (1999); that is, $\tilde{v}_m = ((1/v_m) + (1/\hat{v}_{\text{obs}}))^{-1} = v_{\text{com}}[\{\lambda(v_{\text{com}})(1 - \hat{\gamma}_m)\}^{-1} + v_{\text{com}}/v_m]^{-1}$, where v_m is the large-sample, finite-number imputation df and $\hat{\gamma}_m = (1 + A^{-1})V_B/T$ is approximately the Bayesian fraction of missing information (FMI). For the proposed synthetic MI method, we propose using $v_{\text{syn}} = \min\{L - 1, v_{\text{com}}\}$ as an approximation. The reason for doing this is that the FMI is not well defined in the context of synthetic data, and, hence, the df for constructing interval estimation for the small sample t approximation synthetic data cannot be constructed using FMI. This accounts, simultaneously, for finite number synthetic data generation ($L - 1$) and small sample size (v_{com}).

3.2 Results

Results pertaining to varying simulation conditions are summarized in tables 1–6. We display the following measures in the result tables: point estimate (Pt.est), relative bias (Relbias), standard error of the estimate (SE), empirical standard error of the estimate (Emp.SE), root mean squared error (RMSE), interval widths (Intv.wth), and coverage rate of the nominal 95 percent confidence interval (Cov.).

3.2.1 Population structure 1 (mean unrelated to probability of selection)—

Under population structure 1, the final sampling weight is dominated by cluster-level weights and is uncorrelated with the outcome variable, so the estimator of Y is unbiased even when we exclude weight as a predictor in the imputation model (see tables 1 and 2). Thus, our focus is on the clustering effects, which only impact the variance and hence coverage properties.

As expected, imputation based on SRS leads to underestimation of standard errors and consequently poor confidence interval coverage for the estimated mean. It also leads to overestimates of SEs and thus conservative confidence intervals when estimating the slope because the variability due to clustering is absorbed in the residual variance. On the other hand, the fixed-effects model tends to overestimate the SEs for the mean and thus results in intervals that are too conservative, especially when the ICC is low (Andridge 2011). The results hold regardless of whether the response indicator is a function of X alone (MAR_X) or both X and W (MAR_{X,W}).

The two newly proposed synthetic methods and the RE model have similar results when estimating the mean; their performances are generally better than FX and SRS (table 1). Our synthetic methods demonstrate some advantages over RE when estimating the slope in the small sample scenario ($n = 10$). While the use of a t -corrected inference for a small sample size under RE tends to overcorrect the coverage, SYN1 and SYN2 yield approximately nominal coverage (table 2). Again, the results are similar for both the MAR_X and MAR_{X,W} missingness mechanisms.

3.2.2 Population structure 2 (mean related to probability of selection)—

Under population structure 2, the final weight is dominated by the element-level weights and is associated with the outcome variable Y after conditioning on X . Therefore, the clustering effects on the inference for Y are confounded by the weight effects.

Tables 3 and 4 display the results for SYN1 and SYN2 in the complete data context (SYN1_BD, SYN2_BD) for estimation of the mean and slope, respectively. Both versions of the proposed method result in point and variance estimates very close to those obtained from the actual replication samples (BD). There are even slight gains in efficiency for the proposed methods over the design-based BD estimates.

Tables 5 and 6 display the results for different MI methods for estimating the mean and slope, respectively. When estimating the mean under the MAR_X condition, all three fully parametric methods yield sizable biases and large RMSEs in the point estimates under the X -only model, and the main effects model is required in this case for these biases to disappear. But if the missingness also relates to the weight (i.e., under the MAR_{X,W} mechanism), then only the interaction model gives approximately unbiased estimates. For the slope, however, the correct (interaction) model is always required for obtaining unbiased estimates, regardless of the missingness mechanisms. As before, the SRS method underestimates the variance of the mean and overestimates the variance of the slope because the clustering effects are ignored. The fixed-effects method with small ICC overestimates the

variance and hence the confidence interval coverage for MAR_X; this problem becomes more pronounced for MAR_{X,W}.

Our synthetic MI performs consistently well in all scenarios and is superior to the fully parametric methods in this combination of population structure and sampling design. Use of the simplest *X*-only imputation model not only yields RMSEs comparable to the FX and RE methods using the appropriate interaction model, but it also results in the smallest relative biases for point estimates. This is because, despite the operational and inferential similarity to the frequentist bootstrap (Efron 1979), the (finite population) BB simulates the posterior distribution of the true parameter (ϕ) instead of the sampling distribution of a statistic estimating that parameter ($\hat{\phi}$). As a result, the point estimates based on our method average closer to the true value than those obtained from fully parametric methods, which more resemble the statistic based on actual before-deletion samples. There are slight underestimations for the variance (i.e., SE. < Emp.SE.) with SYN1 relative to SYN2, as well as the appropriate fully parametric models (i.e., FX_{x,w,x*w} and RE_{x,w,x*w}), but this does not affect the overall conclusion.

Furthermore, although the correct imputation model is required for our semiparametric method to break associations between the sample design (W) and response mechanism (R) under a MAR_{X,W} condition, the correct structure in the observed synthetic population data greatly reduces the effect of this association in the ultimate inference, even if the design is not included in the second-step imputation.

4. Application to NASS-CDS Data

The NASS-CDS publishes microlevel crash record datasets that are obtained from a representative three-stage probability sample design: geographic region (county or groups of counties), police departments, and police-reported crashes. The design also has unequal probabilities of selection, with vehicles in more-severe crashes much more likely to be sampled than vehicles in less-severe crashes. The sample is selected annually from all police-reported crashes that resulted in at least one vehicle having to be towed from the scene for damage (<http://www.nhtsa.gov/NASS>).

A key feature of this car-crash data is the Delta-V measure (roughly defined as the “instantaneous” change in velocity). Unfortunately, a very high proportion of car-crash cases are missing this variable. Although detailed reasons for the missingness are not available, they are typically related to the inability to directly assess the vehicle or other vehicles in the crash, which is sometimes positively associated with crash severity (either because vehicles are unavailable to the crash investigators or because the damage severity is too great to obtaining meaningful measurements). This suggests that a missing completely at random assumption is unreasonable. However, because there exists a rich set of covariates that are potentially good predictors in the imputation for Delta-V—among others, basic demographics, road condition, type of vehicle, injury condition of the driver, and number of vehicles in the crash—an MAR assumption, if not entirely plausible, is almost certainly more reasonable than a complete case analysis. The survey variables selected for imputation and analysis are summarized in table 7. We estimate (i) the mean of Delta-V as a continuous

variable ($\bar{Y}_{\text{Delta-V}}$) and (ii) the odds ratio of having (a) any severe injury or (b) severe head injury given varying levels of Delta-V.

We analyze data for a sample of $n = 7,525$ drivers in 2011, where nearly half of the sample cases (45 percent) are missing the Delta-V measure. There are twenty-four PSUs selected in the first sampling stage (out of a total of 1,195 PSUs in the population, which were grouped into twelve strata based on geographic region and type, with at least two PSUs selected from each stratum). For illustration purposes, we treat stratum effects (at both the first and third stage) as being reflected in the unequal selection probabilities or sample weights and analyze the data using an ultimate-cluster approach with a focus on the PSU-level clustering effects.

As a preliminary analysis, we examined features of the design variables and the missing data pattern before imputation. The logistic regression coefficients of the response indicator on the available covariates are mostly significant, suggesting at least an MAR missing data mechanism. Therefore all covariates are included as predictors for imputation. A linear regression of final weights on all of the survey variables indicates that several covariates as well as the Delta-V measure are strong predictors of final weights. Yet all together they explain only 10 percent of the variance in weights; therefore incorporating weights in the imputation is essential. Because the estimated population size based on the sum of the weights is in excess of one million, we assume for operational simplicity that $N = 80,000$, which is large enough to eliminate the (inappropriate) finite population correction that would be present in a smaller synthetic population.

Table 8 displays the results of applying different MI methods to the NASS-CDS Delta-V measure. Different modeling strategies are used under each method according to the extent the design information is modeled. Fully parametric imputation that does not include the weight appears to overestimate the mean Delta-V, as would be expected, because more-severe crashes are over-represented in the unweighted sample. When weights are included in the imputation, the fixed-effect model FX_X,W tends to result in wider confidence intervals. The FPBB method (SYN2) mimics the performance of the random-effects model with weights in the imputation (RE_X,W). Results are broadly similar for estimates of the odds ratios of injury. Here the complete case model appears to underestimate the effect of Delta-V on injury risk, suggesting that injury cases are more likely to have high Delta-V measures missing. The SYN2 approach even enjoys some gains in precision when estimating the odds ratio of having head injury, where all imputation methods yield similar point estimates.

5. Discussion

Two-stage cluster sampling is a popular sampling scheme in survey research because it is cost effective and easy to administer. However, fully parametric MI does not work well with two-stage cluster sampling when the sampling units at both stages are selected with unequal probabilities unless the sampling probabilities are properly accounted for. We propose an alternative two-step MI approach, where the first step generates synthetic populations with missing data that account for weights and clusters, and the second step imputes under an IID assumption. We propose two alternative versions. The first (SYN1) uses a FPBB-FPBB approach, generating clusters from an FPBB and population elements from an FPBB from

those FPBB-generated clusters. Our procedure correctly restores the population configuration of an outcome variable (Y) that is built on complex relationships of the regular covariate (X) and the sample design variables (Z). Our second approach (SYN2) uses a BB-FPBB procedure, resampling the clusters using a BB, and then generating the population elements from an FPBB. Simulation results show that both approaches are insensitive to either sample size or population structure. However, because SYN1 requires knowing both cluster-level weights and within-cluster element-level conditional weights, whereas SYN2 only needs the final weights, we recommend SYN2 for general purposes.

We developed our approach in a standard MI missing data framework using Rubin's combining rules; however, as pointed out by a reviewer, draws from the posterior could in principle be treated in a straightforward Bayesian manner, treating the imputation as a data augmentation step and working directly with the empirical distributions. This approach was assessed in some detail in the complete data setting in Dong et al. (2014, section 6.1), where a simulation study showed very similar results from the direct draws when compared with an approach using a t approximation with Rubin's combining rules, as long as a sufficient number of bootstraps were used (typically a larger number than the t approximation required). Use of Rubin's combining rules provides a convenient means to reduce computational burden, although if time and resources permit, estimates can be obtained from the empirical distribution of the direct draws. As an example, we reran the NASS-CDS analysis using either $L = 50$ and $B = 5$ (EMP1) or $L = 250$ and $B = 1$ (EMP2) and a single imputation ($A = 1$) associated with each (see table 9). The EMP2 approach matched more closely the MI results; both approaches yielded narrower intervals, a feature that was also observed in Dong et al., with little impact on correct nominal coverage.

We also note that, although we made an MAR assumption in our simulation study and application, there is nothing intrinsic to our methods that requires this. Non-MAR models (pattern-mixture models that assume common parameters across missingness patterns, or selection models with nonidentified sensitivity parameters) can certainly be implemented at the imputation stage for item-level missing data.

In this article, we focus on the mean structure of the imputation model for Y and use design-based analysis to account for weight and clustering effects after fully parametric MI. There are many opportunities for extensions of our work. For example, we assume no population-level design information is available, as is usually the case in public-use datasets. However, if such data are available, constrained Polya posterior estimation (Lazar, Meeden, and Nelson 2008) may provide an opportunity to incorporate information such as known population totals into inference. Another natural extension of the method would be to adapt the two-step MI framework to incorporate unit nonresponse; for example, by developing a BB of the entire sample (including the unit non-responders), applying standard unit nonresponse adjustments to the base weights for the respondent set using available auxiliary variables for the entire sample 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. Other important extensions include consideration of the robustness of the proposed new method in comparison with other MI methods against potentially incompatible normal assumptions on random cluster effects and

random errors (Yucel and Demirtas 2010), as well as extensions to stratified, clustered, and unequal probability of selection sample designs.

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 two anonymous reviewers and the editor, for their review and helpful comments.

References

- Andridge RR. Quantifying the Impact of Fixed Effects Modeling of Clusters in Multiple Imputation for Cluster Randomized Trials. *Biometrical Journal*. 2011; 53:57–74. [PubMed: 21259309]
- Barnard D, Rubin D. Small-Sample Degrees of Freedom with Multiple Imputation. *Biometrika*. 1999; 86:948–955.
- Binder DA. On the Variances of Asymptotically Normal Estimators from Complex Surveys. *International Statistical Review*. 1983; 51:279–292.
- Carpenter JR, Goldstein H, Kenward MG. REALCOM-IMPUTE Software for Multilevel Multiple Imputation with Mixed Response Types. *Journal of Statistical Software*. 2011; 45:1–14.
- Cohen MP. The Bayesian Bootstrap and Multiple Imputation for Unequal Probability Sample Designs. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*. 1997:635–638.
- Dong Q, Elliott MR, Raghunathan TE. A Nonparametric Method to Generate Synthetic Populations to Adjust for Complex Sampling Design Features. *Survey Methodology*. 2014; 40:29–46. [PubMed: 29200608]
- Efron B. Bootstrap Methods: Another Look at the Jackknife. *Annals of Statistics*. 1979; 7:1–26.
- Elliott MR. Bayesian Weight Trimming for Generalized Linear Regression Models. *Survey Methodology*. 2007; 33:23–34.
- Elliott MR, Little RJA. Model-Based Approaches to Weight Trimming. *Journal of Official Statistics*. 2000; 16:191–210.
- Gelman A. Struggles with Survey Weighting and Regression Modeling. *Statistical Science*. 2007; 22:153–164.
- Gelman A, Rubin D. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*. 1992; 7:457–472.
- Ghosh, M., Meeden, G. *Bayesian Methods for Finite Population Sampling*. Chapman and Hall; London: 1997.
- Holt D, Smith TMF. Poststratification. *Journal of the Royal Statistical Society*. 1979; A142:33–46.
- Horton N, Lipsitz S, Parzen M. A Potential for Bias when Rounding in Multiple Imputation. *American Statistician*. 2003; 57:229–232.
- Kish, L. *Survey Sampling*. New York: Wiley; 1965.
- Lazar R, Meeden G, Nelson D. A Noninformative Bayesian Approach to Finite Population Sampling Using Auxiliary Variables. *Survey Methodology*. 2008; 34:51–64.
- Little RJ. Calibrated Bayes, for Statistics in General, and Missing Data in Particular (with Discussion and Rejoinder). *Statistical Science*. 2011; 26:162–186.
- Meeden G. A Non-Informative Bayesian Approach for Two-Stage Cluster Sampling. *Sankhya, Series B*. 1999; 61:133–144.
- Meng XL. Multiple Imputation Inferences with Uncongenial Sources of Input. *Statistical Science*. 1994; 9:538–558.
- Rabe-Hesketh S, Skrondal A. Multilevel Modeling of Complex Survey Data. *Journal of the Royal Statistical Society, Series A*. 2006; 169:805–827.
- Raghunathan TE, Lepkowski JM, Van Hoewyk J, Solenberger P. A Multivariate Technique for Multiply Imputing Missing Values Using A Sequence of Regression Models. *Survey Methodology*. 2001; 27:85–95.

- Reiter JP, Raghunathan TE, Kinney SK. The Importance of Modeling the Sampling Design in Multiple Imputation for Missing Data. *Survey Methodology*. 2006; 32:143–149.
- Rubin, DB. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley; 1987.
- Rubin DB. Multiple Imputation after 18+ Years. *Journal of the American Statistical Association*. 1996; 91:473–489.
- Rust K, Rao JNK. Variance Estimation for Complex Estimators in Sample Surveys. *Statistics in Medical Research*. 1996; 5:381–397.
- Schafer, JL. *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall; 1997.
- Schafer JL, Yucel RM. Computational Strategies for Multivariate Linear Mixed-Effects Models with Missing Values. *Journal of Computational and Graphical Statistics*. 2002; 11:421–442.
- Yucel RM. Random Covariances and Mixed-Effects Models for Imputing Multivariate Multilevel Continuous Data. *Statistical Modeling*. 2011; 11:351–370.
- Yucel RM, Demirtas H. Impact of Non-Normal Random Effects on Inference by Multiple Imputation: A Simulation Assessment. *Computational Statistics and Data Analysis*. 2010; 54:790–801. [PubMed: 20526424]
- Yucel RM, Raghunathan TE. Sequential Hierarchical Regression Imputation (SHRIMP). *Proceedings of the Health Policy Statistics Section of the American Statistical Association*. 2006
- Zhao E, Yucel RM. Performance of Sequential Imputation Method in Multilevel Applications. *American Statistical Association Proceedings of the Survey Research Methods Section*. 2009:2800–2810.
- Zheng H, Little RJA. Penalized Spline Model-Based Estimation of the Finite Population Total from Probability-Proportional-to-Size Samples. *Journal of Official Statistics*. 2003; 19:99–117.
- Zhou, H. unpublished PhD thesis, Survey Methodology Program. University of Michigan; 2014. Accounting for Complex Sample Designs in Multiple Imputation Using the Finite Population Bayesian Bootstrap.

Table 1
Performance of Alternative MI Methods for Estimating the Mean: Population 1, Unbalanced Two-Stage Sample Design

Simulation factor	Method	MAR_X					MAR_X,W						
		Pt.est.	SE.	Emp.SE	RMSE	Intv.wth	Cov. (%)	Pt.est.	SE.	Emp.SE	RMSE	Intv.wth	Cov. (%)
ICC = 0.5, $a = 10$	BD	1.02	0.34	0.36	0.36	1.54	95.50	1.02	0.34	0.36	0.36	1.54	95.5
	SRS	1.03	0.28	0.36	0.36	1.33	90.50	1.03	0.30	0.37	0.37	1.38	91.5
	FX	1.02	0.37	0.36	0.36	1.69	96.50	1.03	0.37	0.37	0.37	1.68	96.5
True value = 1.05	RE	1.02	0.36	0.36	0.36	1.65	96.50	1.02	0.36	0.37	0.37	1.64	95.5
	SYN1	1.03	0.36	0.38	0.38	1.59	95.50	1.02	0.36	0.38	0.38	1.56	95.0
	SYN2	1.02	0.37	0.37	0.37	1.65	95.50	1.02	0.36	0.38	0.38	1.62	95.0
ICC = 0.5, $a = 40$	BD	1.04	0.18	0.19	0.19	0.73	95.50	1.04	0.18	0.19	0.19	0.73	95.5
	SRS	1.04	0.15	0.19	0.19	0.59	86.50	1.04	0.15	0.19	0.19	0.61	86.5
	FX	1.04	0.19	0.19	0.19	0.76	95.50	1.04	0.19	0.19	0.19	0.76	97.0
True value = 1.05	RE	1.04	0.18	0.19	0.19	0.74	94.50	1.04	0.18	0.19	0.19	0.75	96.0
	SYN1	1.04	0.18	0.20	0.20	0.72	92.50	1.03	0.18	0.20	0.20	0.71	92.5
	SYN2	1.05	0.19	0.20	0.20	0.77	94.50	1.03	0.19	0.20	0.20	0.77	94.5
ICC = 0.1, $a = 10$	BD	1.06	0.42	0.43	0.43	1.89	95.00	1.06	0.42	0.43	0.43	1.89	95.0
	SRS	1.05	0.41	0.45	0.45	2.08	95.50	1.07	0.43	0.46	0.46	2.27	95.5
	FX	1.07	0.54	0.45	0.45	2.54	98.50	1.07	0.54	0.48	0.48	2.58	98.0
True value = 1.06	RE	1.07	0.47	0.45	0.45	2.30	98.50	1.08	0.49	0.47	0.47	2.40	98.0
	SYN1	1.05	0.47	0.46	0.46	2.06	95.50	1.07	0.48	0.47	0.47	2.10	96.0
	SYN2	1.07	0.48	0.46	0.46	2.19	97.00	1.07	0.48	0.48	0.48	2.19	97.0
ICC = 0.1, $a = 40$	BD	1.05	0.21	0.22	0.22	0.87	94.50	1.05	0.21	0.22	0.22	0.87	94.5
	SRS	1.05	0.20	0.24	0.24	0.86	92.50	1.05	0.21	0.24	0.24	0.87	90.5
	FX	1.05	0.27	0.24	0.24	1.09	99.00	1.05	0.27	0.24	0.24	1.08	98.5
True value = 1.06	RE	1.05	0.23	0.24	0.24	0.95	93.50	1.05	0.24	0.24	0.24	0.97	94.0
	SYN1	1.05	0.23	0.25	0.25	0.89	94.50	1.04	0.23	0.25	0.25	0.90	92.5
	SYN2	1.05	0.25	0.25	0.25	1.01	96.50	1.04	0.24	0.25	0.25	0.99	93.5

Table 2
Performance of Alternative MI Methods for the Slope: Population 1, Unbalanced Two-Stage Sample Design

Simulation factor	Method	MAR_X					MAR_X,W						
		Pt.est.	SE.	Emp.SE	RMSE	Intv.wth	Cov. (%)	Pt.est.	SE.	Emp.SE	RMSE	Intv.wth	Cov. (%)
ICC = 0.5, $a = 10$	BD	1.51	0.11	0.12	0.12	0.48	94.5	1.51	0.11	0.12	0.12	0.48	94.5
	SRS	1.51	0.14	0.14	0.14	0.85	98.5	1.53	0.17	0.16	0.16	1.04	97.5
	FX	1.52	0.13	0.14	0.14	0.73	98.0	1.52	0.14	0.14	0.14	0.83	98.0
True value = 1.51	RE	1.52	0.13	0.13	0.13	0.69	98.0	1.52	0.14	0.14	0.14	0.80	99.0
	SYN1	1.52	0.13	0.14	0.14	0.57	93.5	1.52	0.14	0.16	0.16	0.62	95.0
	SYN2	1.52	0.13	0.14	0.14	0.61	96.0	1.52	0.15	0.15	0.15	0.71	96.5
ICC = 0.5, $a = 40$	BD	1.51	0.06	0.06	0.06	0.23	95.5	1.51	0.06	0.06	0.06	0.23	95.5
	SRS	1.51	0.07	0.07	0.07	0.31	93.5	1.51	0.09	0.08	0.08	0.37	96.5
	FX	1.51	0.07	0.07	0.07	0.28	95.5	1.51	0.07	0.07	0.07	0.31	98.0
True value = 1.51	RE	1.51	0.07	0.07	0.07	0.27	95.5	1.51	0.07	0.07	0.07	0.30	96.5
	SYN1	1.51	0.07	0.07	0.07	0.26	93.5	1.51	0.07	0.07	0.07	0.30	94.0
	SYN2	1.51	0.07	0.07	0.07	0.29	95.0	1.51	0.08	0.07	0.07	0.33	97.5
ICC = 0.1, $a = 10$	BD	1.52	0.25	0.26	0.26	1.17	96.5	1.52	0.25	0.26	0.26	1.17	96.5
	SRS	1.51	0.33	0.32	0.31	2.00	98.5	1.53	0.37	0.34	0.34	2.77	100.0
	FX	1.50	0.33	0.32	0.32	2.01	98.0	1.52	0.37	0.34	0.34	2.60	99.0
True value = 1.52	RE	1.51	0.33	0.32	0.32	1.93	99.0	1.54	0.35	0.35	0.35	2.50	99.0
	SYN1	1.50	0.31	0.32	0.32	1.38	96.5	1.53	0.35	0.36	0.36	1.54	96.0
	SYN2	1.50	0.32	0.32	0.32	1.49	97.0	1.53	0.36	0.35	0.35	1.67	98.0
ICC = 0.1, $a = 40$	BD	1.52	0.13	0.13	0.13	0.53	95.5	1.52	0.13	0.13	0.13	0.53	95.5
	SRS	1.52	0.16	0.16	0.16	0.69	96.0	1.51	0.18	0.17	0.17	0.78	95.5
	FX	1.51	0.16	0.16	0.16	0.72	96.5	1.51	0.18	0.17	0.17	0.81	98.0
True value = 1.52	RE	1.51	0.16	0.16	0.16	0.68	97.5	1.51	0.17	0.16	0.16	0.76	96.5
	SYN1	1.51	0.15	0.16	0.16	0.60	94.0	1.51	0.17	0.17	0.17	0.66	94.5
	SYN2	1.51	0.16	0.16	0.16	0.65	96.0	1.51	0.18	0.17	0.17	0.75	97.5

Table 3
Performance of Alternative Methods for Estimating the Mean: Population 2, Unbalanced Two-Stage Sample Design, Before Deletion Results

Simulation factor	Method	Summary statistics					
		Ptest.	SE.	Emp.SE	RMSE	Intv.wth	Cov. (%)
ICC = 0.5, $a = 40$; True value = 1.542	BD	1.498	0.179	0.176	0.181	0.723	96.5
	SYN1_BD	1.507	0.176	0.179	0.182	0.712	95.0
	SYN2_BD	1.514	0.178	0.175	0.177	0.719	95.5
ICC = 0.1, $a = 40$; True value = 1.541	BD	1.493	0.233	0.232	0.237	0.944	94.5
	SYN1_BD	1.502	0.230	0.235	0.237	0.931	93.5
	SYN2_BD	1.509	0.231	0.232	0.234	0.936	94.5
ICC = 0.5, $a = 10$; True value = 1.542	BD	1.511	0.354	0.369	0.370	1.603	97.5
	SYN1_BD	1.512	0.362	0.377	0.377	1.640	98.0
	SYN2_BD	1.521	0.358	0.371	0.370	1.620	98.0
ICC = 0.1, $a = 10$; True value = 1.541	BD	1.507	0.474	0.517	0.517	2.145	95.5
	SYN1_BD	1.505	0.485	0.520	0.520	2.196	95.0
	SYN2_BD	1.527	0.470	0.511	0.510	2.126	95.0

Table 4
Performance of Alternative Methods for Estimating the Slope: Population 2, Unbalanced Two-Stage Sample Design, Before Deletion Results

Simulation factor	Method	Summary statistics					
		Ptest.	SE.	Emp.SE	RMSE	Intv.wth	Cov. (%)
ICC = 0.5, $a = 40$; True value = 1.265	BD	1.236	0.095	0.102	0.106	0.384	96.5
	SYN1_BD	1.246	0.089	0.097	0.099	0.361	94.0
	SYN2_BD	1.248	0.093	0.099	0.100	0.378	95.5
ICC = 0.1, $a = 40$; True value = 1.255	BD	1.226	0.174	0.170	0.172	0.706	94.0
	SYN1_BD	1.237	0.166	0.166	0.166	0.671	92.0
	SYN2_BD	1.241	0.173	0.168	0.168	0.700	94.5
ICC = 0.5, $a = 10$; True value = 1.265	BD	1.251	0.171	0.206	0.206	0.787	92.5
	SYN1_BD	1.267	0.169	0.193	0.193	0.780	91.0
	SYN2_BD	1.268	0.173	0.190	0.190	0.800	93.5
ICC = 0.1, $a = 10$; True value = 1.255	BD	1.253	0.330	0.357	0.356	1.520	95.5
	SYN1_BD	1.270	0.338	0.351	0.350	1.557	95.5
	SYN2_BD	1.274	0.335	0.343	0.342	1.545	95.0

Table 5
Performance of Alternative MI Methods for Estimating the Mean: Population Structure 2, Unbalanced Two-Stage Sampling Design

Simulation factor	Method	MAR_X					MAR_X,W						
		Relbias (%)	SE.	Emp. SE	RMSE	Intv. with	Cov. (%)	Relbias (%)	SE.	Emp. SE	RMSE	Intv. with	Cov. (%)
ICC = 0.5, $a = 40$ True value = 1.542	BD	-2.9	0.179	0.176	0.181	0.723	96.5	-2.9	0.179	0.176	0.181	0.723	96.5
	SRS_x	13.0	0.165	0.186	0.274	0.671	74.5	20.0	0.169	0.185	0.359	0.691	57.0
	SRS_x,w	1.8	0.161	0.187	0.188	0.655	89.5	6.5	0.167	0.191	0.215	0.683	86.0
	SRS_x,w,x*w	-2.9	0.156	0.181	0.186	0.635	90.0	-2.7	0.162	0.187	0.191	0.662	91.0
	FX_x	13.2	0.197	0.183	0.273	0.802	83.5	20.1	0.201	0.188	0.362	0.819	65.5
	FX_x,w	1.8	0.191	0.181	0.182	0.775	96.5	6.3	0.197	0.189	0.212	0.801	93.5
	FX_x,w,x*w	-2.9	0.187	0.179	0.184	0.759	96.0	-2.8	0.191	0.182	0.187	0.778	98.0
	RE_x	13.2	0.192	0.184	0.275	0.780	83.0	20.0	0.196	0.188	0.362	0.798	66.0
	RE_x,w	1.8	0.187	0.183	0.185	0.758	94.5	6.4	0.192	0.187	0.211	0.782	92.5
	RE_x,w,x*w	-3.0	0.184	0.178	0.184	0.745	95.5	-2.7	0.188	0.180	0.185	0.763	97.5
ICC = 0.1, $a = 40$ True value = 1.541	SYN1_x	-2.1	0.185	0.190	0.192	0.750	94.5	0.6	0.189	0.198	0.198	0.767	94.0
	SYN2_x	-1.9	0.188	0.183	0.185	0.761	95.5	0.8	0.193	0.192	0.192	0.779	95.5
	BD	-3.1	0.233	0.232	0.237	0.944	94.5	-3.1	0.233	0.232	0.237	0.944	94.5
	SRS_x	12.7	0.244	0.242	0.311	1.003	86.5	20.3	0.260	0.237	0.392	1.077	79.0
	SRS_x,w	1.3	0.246	0.258	0.258	1.009	93.5	6.7	0.265	0.265	0.284	1.103	94.0
	SRS_x,w,x*w	-3.2	0.244	0.252	0.257	1.005	94.0	-2.7	0.269	0.266	0.268	1.126	95.5
	FX_x	12.8	0.287	0.237	0.309	1.170	92.0	20.5	0.304	0.243	0.398	1.246	87.0
	FX_x,w	1.3	0.285	0.250	0.250	1.164	97.0	6.3	0.306	0.270	0.286	1.261	97.0
	FX_x,w,x*w	-3.3	0.283	0.254	0.259	1.155	97.0	-2.8	0.307	0.267	0.270	1.267	96.5
	RE_x	13.0	0.263	0.240	0.313	1.076	90.0	20.3	0.279	0.241	0.395	1.148	85.5
ICC = 0.5, $a = 10$ True value = 1.542	RE_x,w	1.3	0.261	0.259	0.259	1.067	95.5	6.6	0.280	0.261	0.280	1.152	95.0
	RE_x,w,x*w	-3.4	0.261	0.254	0.259	1.068	96.5	-2.6	0.283	0.258	0.260	1.170	97.5
	SYN1_x	-2.7	0.254	0.266	0.268	1.029	92.0	0.6	0.262	0.270	0.270	1.058	94.5
	SYN2_x	-2.4	0.262	0.261	0.263	1.059	94.0	0.6	0.267	0.265	0.264	1.080	93.5
	BD	-2.0	0.354	0.369	0.370	1.603	97.5	-2.0	0.354	0.369	0.370	1.603	97.5
	SRS_x	12.8	0.326	0.364	0.413	1.564	91.5	20.4	0.333	0.390	0.500	1.622	87.0
SRS_x,w	1.8	0.316	0.370	0.370	1.519	95.5	6.4	0.326	0.381	0.392	1.596	93.0	

Simulation factor	Method	MAR_X						MAR_X,W					
		Relbias (%)			Cov. (%)			Relbias (%)			Cov. (%)		
		SE.	Emp. SE	RMSE	Intv. wth	Intv. wth	SE.	SE.	Emp. SE	RMSE	Intv. wth	Intv. wth	Cov. (%)
ICC = 0.1, $a = 10$ True value = 1.541	SRS _{x,w,x*w}	-2.9	0.308	0.372	0.374	1.479	94.5	-2.6	0.320	0.373	0.374	1.591	98.0
	FX _x	13.8	0.391	0.366	0.423	1.847	95.5	20.4	0.397	0.397	0.506	1.887	93.0
	FX _{x,w}	2.5	0.377	0.367	0.368	1.779	98.0	6.4	0.388	0.397	0.408	1.845	97.0
	FX _{x,w,x*w}	-2.0	0.369	0.372	0.373	1.741	98.5	-2.7	0.381	0.379	0.380	1.812	98.5
	RE _x	13.7	0.381	0.364	0.420	1.801	95.0	20.2	0.387	0.394	0.502	1.841	93.0
	RE _{x,w}	2.3	0.371	0.366	0.366	1.752	97.5	6.4	0.382	0.392	0.403	1.821	96.5
	RE _{x,w,x*w}	-2.3	0.363	0.369	0.370	1.713	98.0	-2.8	0.374	0.374	0.376	1.787	98.0
	SYN1 _x	-1.6	0.377	0.381	0.381	1.705	98.0	1.6	0.383	0.399	0.398	1.734	95.5
	SYN2 _x	1.9	0.377	0.375	0.374	1.707	96.0	1.6	0.379	0.405	0.405	1.716	93.0
	BD	-2.2	0.474	0.517	0.517	2.145	95.5	-2.2	0.474	0.517	0.517	2.145	95.5
	SRS _x	12.6	0.490	0.492	0.528	2.428	96.5	19.3	0.522	0.514	0.593	2.698	97.0
	SRS _{x,w}	1.1	0.491	0.538	0.537	2.466	96.0	4.3	0.530	0.543	0.546	2.790	97.5
	SRS _{x,w,x*w}	-3.6	0.491	0.540	0.541	2.476	97.0	-4.2	0.544	0.552	0.555	2.983	98.0
	FX _x	14.1	0.575	0.488	0.533	2.788	98.0	18.9	0.606	0.519	0.595	3.011	98.5
	FX _{x,w}	2.1	0.570	0.532	0.532	2.791	98.5	4.2	0.613	0.578	0.580	3.084	99.0
	FX _{x,w,x*w}	-2.3	0.569	0.538	0.538	2.785	99.0	-4.4	0.622	0.567	0.570	3.187	99.5
	RE _x	13.5	0.532	0.485	0.527	2.591	97.0	18.6	0.561	0.514	0.587	2.818	97.5
	RE _{x,w}	1.2	0.536	0.529	0.528	2.652	98.0	4.3	0.581	0.558	0.561	3.013	98.5
	RE _{x,w,x*w}	-3.4	0.533	0.532	0.533	2.644	99.5	-4.3	0.588	0.555	0.558	3.096	99.0
	SYN1 _x	-2.3	0.524	0.563	0.562	2.373	96.0	-1.0	0.537	0.570	0.569	2.438	92.5
	SYN2 _x	-1.9	0.527	0.547	0.547	2.382	95.5	-0.4	0.536	0.571	0.570	2.425	92.0

Table 6
Performance of Alternative MI Methods for Estimating the Slope: Population Structure 2, Unbalanced Two-Stage Sampling Design

Simulation factor	Method	MAR_X						MAR_X,W					
		Relbias (%)	SE.	Emp. SE	RMSE	Intv. with	Cov. (%)	Relbias (%)	SE.	Emp. SE	RMSE	Intv. with	Cov. (%)
ICC = 0.5, $a = 40$	BD	-2.3	0.095	0.102	0.106	0.384	96.5	-2.3	0.095	0.102	0.106	0.384	96.5
	SRS_x	14.5	0.112	0.104	0.211	0.462	65.5	22.3	0.120	0.101	0.299	0.501	35.0
	SRS _{x,w}	8.8	0.108	0.100	0.149	0.443	83.5	13.8	0.117	0.100	0.201	0.488	74.0
	SRS _{x, w, x^{*w}}	-2.1	0.110	0.112	0.115	0.454	96.0	-2.5	0.129	0.124	0.128	0.539	98.5
	FX_x	14.4	0.109	0.103	0.209	0.445	62.5	22.3	0.107	0.102	0.300	0.444	29.5
	FX _{x,w}	8.5	0.103	0.101	0.148	0.420	81.5	13.5	0.104	0.101	0.198	0.432	65.0
True value = 1.265	FX _{x,w, x^{*w}}	-2.4	0.105	0.110	0.114	0.430	96.0	-2.5	0.113	0.119	0.123	0.467	97.5
	RE_x	14.6	0.107	0.105	0.213	0.440	59.5	22.1	0.106	0.101	0.297	0.441	28.0
	RE _{x,w}	8.6	0.101	0.100	0.148	0.415	81.5	13.6	0.103	0.098	0.198	0.424	64.0
	RE _{x,w, x^{*w}}	-2.5	0.105	0.110	0.114	0.427	96.0	-2.5	0.113	0.119	0.123	0.465	97.0
	SYN1_x	-1.8	0.107	0.116	0.120	0.434	93.0	-3.3	0.118	0.133	0.142	0.479	90.0
	SYN2_x	-2.4	0.113	0.114	0.118	0.459	93.5	-3.7	0.123	0.131	0.139	0.498	93.5
ICC = 0.1, $a = 40$	BD	-2.3	0.174	0.170	0.172	0.706	94.0	-2.3	0.174	0.170	0.172	0.706	94.0
	SRS_x	14.6	0.211	0.190	0.264	0.871	91.5	22.6	0.234	0.187	0.339	0.989	84.0
	SRS _{x,w}	8.8	0.210	0.186	0.216	0.869	98.0	14.1	0.233	0.190	0.259	0.984	92.5
	SRS _{x, w, x^{*w}}	-2.2	0.214	0.210	0.212	0.888	95.0	-2.7	0.254	0.237	0.239	1.083	98.0
	FX_x	14.5	0.214	0.184	0.258	0.883	92.0	22.8	0.233	0.197	0.347	0.981	80.5
	FX _{x,w}	8.5	0.212	0.186	0.214	0.977	97.5	13.5	0.235	0.196	0.259	0.988	93.5
True value = 1.255	FX _{x,w, x^{*w}}	-2.2	0.217	0.209	0.210	0.899	96.0	-2.5	0.250	0.238	0.239	1.060	98.5
	RE_x	15.1	0.209	0.190	0.268	0.863	87.0	22.2	0.229	0.193	0.339	0.963	82.5
	RE _{x,w}	8.7	0.207	0.186	0.215	0.854	97.5	13.7	0.227	0.187	0.254	0.955	92.5
	RE _{x,w, x^{*w}}	-2.5	0.213	0.207	0.209	0.881	96.0	-2.5	0.246	0.235	0.237	1.046	99.0
	SYN1_x	-3.5	0.207	0.219	0.221	0.837	93.0	-5.5	0.232	0.248	0.254	0.938	90.5
	SYN2_x	-2.5	0.214	0.218	0.222	0.866	93.0	-4.1	0.241	0.244	0.251	0.975	94.5
ICC = 0.5, $a = 10$	BD	-1.1	0.171	0.206	0.206	0.787	92.5	-1.1	0.171	0.206	0.206	0.787	92.5
	SRS_x	14.9	0.207	0.225	0.294	1.112	92.5	22.8	0.224	0.213	0.359	1.308	92.0
	SRS _{x,w}	9.4	0.199	0.219	0.249	1.066	94.5	13.6	0.217	0.213	0.274	1.261	96.5

Simulation factor	Method	MAR_X						MAR_X_W					
		Relbias (%)	SE.	Emp. SE	RMSE	Intv. wth	Cov. (%)	Relbias (%)	SE.	Emp. SE	RMSE	Intv. wth	Cov. (%)
ICC = 0.1, $a = 10$	SRS_x, w,x* _w	-0.9	0.208	0.237	0.236	1.111	97.5	-1.2	0.239	0.253	0.253	1.389	98.5
	FX_x	15.3	0.198	0.224	0.296	1.028	88.5	23.1	0.196	0.203	0.355	1.089	85.5
	FX_x,w	9.7	0.186	0.218	0.249	0.953	93.0	13.9	0.188	0.206	0.271	1.027	92.0
	FX_x,w,x* _w	-0.6	0.192	0.230	0.229	0.988	97.0	-1.3	0.206	0.243	0.243	1.142	99.5
	RE_x	15.3	0.196	0.222	0.294	1.018	87.5	23.1	0.193	0.201	0.354	1.082	85.5
	RE_x,w	9.7	0.188	0.218	0.250	0.967	92.5	13.8	0.189	0.205	0.270	1.054	94.0
	RE_x,w,x* _w	-0.9	0.192	0.225	0.225	0.985	96.5	-1.3	0.207	0.242	0.242	1.150	98.5
	SYN1_x	0.1	0.196	0.231	0.231	0.902	92.0	-1.1	0.206	0.254	0.254	0.950	91.5
	SYN2_x	-0.2	0.205	0.228	0.227	0.945	94.0	-1.5	0.216	0.251	0.251	0.998	92.5
	BD	-0.2	0.330	0.357	0.356	1.520	95.5	-0.2	0.330	0.357	0.356	1.520	95.5
True value = 1.255	SRS_x	16.6	0.404	0.401	0.450	2.219	97.0	23.6	0.447	0.412	0.506	2.769	98.5
	SRS_x,w	10.8	0.402	0.395	0.416	2.218	98.5	13.5	0.447	0.408	0.441	2.770	100.0
	SRS_x, w,x* _w	0.7	0.421	0.436	0.435	2.376	97.0	-0.7	0.484	0.476	0.475	3.205	99.5
	FX_x	17.1	0.407	0.404	0.456	2.215	98.0	23.7	0.442	0.403	0.500	2.711	99.5
	FX_x,w	11.2	0.403	0.398	0.421	2.190	98.5	14.0	0.445	0.404	0.440	2.698	99.5
	FX_x,w,x* _w	1.8	0.422	0.441	0.441	2.393	98.5	-0.6	0.485	0.473	0.472	3.232	99.5
	RE_x	17.0	0.400	0.395	0.448	2.172	98.5	23.5	0.431	0.400	0.496	2.661	98.5
	RE_x,w	11.3	0.404	0.399	0.422	2.235	98.5	13.9	0.444	0.402	0.437	2.786	100.0
	RE_x,w,x* _w	0.7	0.417	0.427	0.426	2.375	98.0	-0.6	0.483	0.472	0.471	3.286	100.0
	SYN1_x	-0.2	0.400	0.444	0.443	1.843	92.5	-3.7	0.441	0.500	0.500	2.033	94.0
	SYN2_x	1.2	0.411	0.440	0.439	1.894	94.0	-1.8	0.454	0.481	0.481	2.096	96.0

Table 7
Summary of Selected Survey Variables for Imputation

Selected variables	Values and labels
Delta V	Continuous
Driver's age	Continuous
Maximum injury	1: not injured/minor injury; 2: severe injury;
Direction of force	1: frontal; 2: right side; 3: left side; 4: rear; 5: other
Driver's gender	1: male; 2: female
Light condition	1: daylight; 2: dark; 3: dark, lighted; 4: dawn/dusk
Road surface condition	1: dry; 2: wet; 3: other
Vehicle type	1: passenger car; 2: truck
Model year	1: >2006; 2: 2006
Vehicle make	1: American; 2: Japanese; 3: Korean; 4: European and other
Body region	1: head; 2: thorax; 3: lower extremity; 4: other
Primary sampling unit	1, 2, ..., 24
Final survey weight	Continuous

Table 8
Estimating Mean Delta-V, Odds Ratio of Severe Injury Given Delta-V, and Odds Ratio of Head Injury Given Delta-V

Methods		CC			SRS			FX			RE			SYN2		
Models		CC	X	X,W	X	X,W	X	X,W	X	X,W	X	X,W	X	X,W	C	X,W,C
Estimates	$\bar{Y}_{\text{Delta-V}}$	26.7	29.8	27.9	29.6	27.3	29.6	27.9	29.6	27.9	28	27.6	28.1	27.4		
	SE	0.88	0.88	0.96	1.12	1.05	1.09	0.88	1.09	0.88	0.85	0.89	0.81	0.89		
	OR ^{Severe Injury} _{high vs. low DV}	5.03	5.78	6.23	5.43	6.38	5.53	6.71	5.53	6.71	6.55	7.49	6.23	7.29		
	95% CI	[2.53, 9.99]	[3.16, 10.60]	[3.37, 11.50]	[2.99, 9.86]	[3.45, 11.80]	[3.23, 9.48]	[3.74, 12.00]	[3.41, 12.60]	[3.85, 14.60]	[3.26, 11.90]	[3.70, 14.40]				
	OR ^{Severe Injury} _{medium vs. low DV}	1.28	1.58	1.52	1.61	1.77	1.56	1.77	1.81	1.86	1.76	1.80				
	95% CI	[0.612, 2.680]	[0.838, 2.980]	[0.862, 2.680]	[0.917, 2.830]	[0.994, 3.150]	[0.911, 2.670]	[1.01, 3.11]	[1.07, 3.06]	[1.05, 3.29]	[1.03, 3.00]	[0.992, 3.270]				
	OR ^{Head Injury} _{high vs. low DV}	1.05	1.55	1.45	1.69	1.52	1.58	1.36	1.62	1.60	1.72	1.56				
	95% CI	[0.581, 1.900]	[0.76, 3.16]	[0.751, 2.800]	[0.904, 3.160]	[0.765, 3.020]	[0.706, 3.540]	[0.919, 2.650]	[1.02, 2.56]	[1.09, 2.35]	[1.09, 2.71]	[1.09, 2.23]				

Note.—High Delta-V is in excess of 35 kph; medium Delta-V is 15–35 kph; and low delta-V is less than 15 kph. CC = complete case; FX = imputation using fixed cluster effects; RE = imputation using random cluster effects; SRS = imputation under simple random sampling assumption; SYN2 = results from two-stage finite-population Bayesian bootstrap model using Bayesian bootstrap–finite population Bayesian bootstrap for synthetic population generation.

Table 9
Estimating Mean Delta-V and Odds Ratio of Severe Injury Given Delta-V from Two-Stage Finite Population Bayesian Bootstrap Model Using Bayesian Bootstrap-Finite Population Bayesian Bootstrap for Synthetic Population Generation (SYN 2) and Imputation Using Both Covariates and Weights

Models	MI	EMP1	EMP2
$\bar{V}_{\text{Delta-V}}$	27.6	28.2	28.3
SE	0.89	0.74	0.94
$OR_{\text{Severe Injury high vs. low DV}}$	7.49	7.15	7.57
95% CI	[3.85, 14.60]	[3.22, 10.60]	[4.23, 14.50]
$OR_{\text{Severe Injury medium vs. low DV}}$	1.86	1.78	1.82
95% CI	[1.05, 3.29]	[1.11, 2.76]	[1.06, 2.82]