Survey Methodology, December 2006
Vol. 32, No. 2, pp. 143-149
Statistics Canada, Catalogue No. 12-001

The Importance of Modeling the Sampling Design in Multiple Imputation for Missing Data

Jerome P. Reiter, Trivellore E. Raghunathan and Satkartar K. Kinney [1]

Abstract

The theory of multiple imputation for missing data requires that imputations be made conditional on the sampling design. However, most standard software packages for performing model-based multiple imputation assume simple random samples, leading many practitioners not to account for complex sample design features, such as stratification and clustering, in their imputations. Theory predicts that analyses of such multiply-imputed data sets can yield biased estimates from the design-based perspective. In this article, we illustrate through simulation that (i) the bias can be severe when the design features are related to the survey variables of interest, and (ii) the bias can be reduced by controlling for the design features in the imputation models. The simulations also illustrate that conditioning on irrelevant design features in the imputation models can yield conservative inferences, provided that the models include other relevant predictors. These results suggest a prescription for imputers: the safest course of action is to include design variables in the specification of imputation models. Using real data, we demonstrate a simple approach for incorporating complex design features that can be used with some of the standard software packages for creating multiple imputations.

Key Words: Complex sampling design; Multiple imputation; Nonresponse; Surveys.

1. Introduction

Typically in large surveys, less than 100% of the sampled units respond fully to the survey. Some units do not respond at all, and others respond only to certain items. One approach to handle such nonresponse is multiple imputation of missing data (Rubin 1987). It has been used in, for example, the Fatality Analysis Reporting System (Heitjan and Little 1991), the Consumer Expenditures Survey (Raghunathan and Paulin 1998), the National Health and Nutrition Examination Survey (Schafer, Ezzati-Rice, Johnson, Khare, Little and Rubin 1998), the Survey of Consumer Finances (Kennickell 1998) and the National Health Interview Survey (Schenker, Raghunathan, Chiu, Makuc, Zhang and Cohen 2005). Multiple imputation also has been suggested to protect confidentiality of public-release data (Rubin 1993; Little 1993; Raghunathan, Reiter and Rubin 2003; Reiter 2003, 2004, 2005). See Rubin (1996) and Barnard and Meng (1999) for a review of other applications.

Multiple imputation, in theory, conditions on the sampling design when deriving methods for obtaining inferences from multiply-imputed datasets (Rubin 1987). However, imputers seldom account for complex sampling design features, such as stratification and clustering, when using available software packages to construct imputation models. They instead use multivariate normal or general location models (*e.g.*, the software NORM written by Joe Schafer), or use sequential regression models (Raghunathan, Lepkowshi, van Hoewyk and Solenberger 2001). These methods can be modified to incorporate design features, but this is infrequently done.

This paper has two objectives. First, we illustrate the bias that can arise when imputers fail to account for complex design features in imputation models. To do so, we simulate multiple imputation in two-stage, stratified-cluster samples. The simulations indicate these biases can be severe, even when using design-based estimators in multiply-imputed data sets with moderate amounts of missing data. Second, we suggest two simple approaches to account for design features in imputation models. The first approach, which is relatively easy to implement, includes dummy variables for stratum or cluster effects in the imputation models. The second approach, which is computationally more complex than the first, uses hierarchical models where (i) the effects of clustering are incorporated using random effects, and (ii) the effects of stratification are incorporated using fixed effects. The simulations show that accounting for the design in these ways can reduce the bias. They also illustrate that controlling for design features that are unrelated to the survey variables can result in inefficient, but conservative, inferences relative to those from models that do not condition on such features, provided that the models include the predictors required to make the missing at random assumption (Rubin 1976) plausible. We demonstrate the first approach to incorporating the design features by imputing missing data from the National Health and Nutrition Examination Survey using a sequential regression approach.

1. Jerome P. Reiter and Satkartar K. Kinney, Institute of Statistics and Decision Sciences, Box 90251, Duke University, Durham, NC 27708, U.S.A.; Trivellore E. Raghunathan, Department of Biostatistics and Institute for Social Research, University of Michigan, Ann Arbor, MI 48106, U.S.A.

2. Inferences from Multiply-Imputed Data Sets

To describe construction of and inferences from multiply-imputed data sets, we use the notation of Rubin (1987). For a finite population of size *N*, let $I_j = 1$ if unit *j* is selected in the original survey, and $I_j = 0$ otherwise, where $j = 1, 2, \ldots, N$. Let $I = (I_1, \ldots, I_N)$. Let *n* be the size of the sample obtained using a complex design. To simplify notation, assume only one variable in the survey is subject to nonresponse. Let $R_j = 1$ if unit *j* responds to the original survey, and $R_j = 0$ otherwise. Let $R = (R_1, \ldots, R_N)$. The notation can be extended to handle multivariate item nonresponse, but such complication is not necessary for our purposes.

Let *Y* be the $N \times p$ matrix of survey data for all units in the population. Let $Y_{\text{inc}} = (Y_{\text{obs}}, Y_{\text{mis}})$ be the $n \times p$ matrix of survey data for units with $I_j = 1$; Y_{obs} is the portion of Y_{inc} that is observed, and Y_{mis} is the portion of Y_{inc} that is missing due to nonresponse. Let *Z* be the $N \times d$ matrix of design variables for all *N* units in the population, *e.g.*, stratum or cluster indicators or size measures. We assume that such design information is known at least approximately, for example from census records or the sampling frames.

Values for Y_{mis} are usually constructed from draws from some approximation to the Bayesian posterior predictive distribution of $(Y_{\text{mis}} | Z, Y_{\text{obs}}, I, R)$. These draws are repeated independently $l = 1, \ldots, M$ times to obtain *M* completed data sets, $D^{(l)} = (Z, Y_{\text{obs}}, Y_{\text{mis}}^{(l)}, I, R)$.

From these multiply-imputed data sets, some user of the data seeks inferences about some estimand $Q = Q(Z, Y)$. For example, *Q* could be a population mean or a population regression coefficient. In each imputed data set $D^{(l)}$, the analyst estimates *Q* with some estimator *q* and the variance of *q* with some estimator *u*. We assume that the analyst specifies *q* and *u* by acting as if each $D^{(l)}$ was in fact collected data from a random sample of (Z, Y) based on the original sampling design *I*, *i.e.*, *q* and *u* are complete-data estimators.

For $l = 1, \ldots, M$, let $q^{(l)}$ and $u^{(l)}$ be respectively the values of *q* and *u* in data set $D^{(l)}$. Under assumptions described in (Rubin 1987), the analyst can obtain valid inferences for scalar *Q* by combining the $q^{(l)}$ and $u^{(l)}$. Specifically, the following quantities are needed for inferences:

$$\bar{q}_M = \sum_{l=1}^{M} q^{(l)} / M \tag{1}$$

$$b_M = \sum_{l=1}^{M} (q^{(l)} - \bar{q}_M)^2 / (M - 1) \tag{2}$$

$$\bar{u}_M = \sum_{l=1}^{M} u^{(l)} / M. \tag{3}$$

The analyst then can use $\bar{q}_M$ to estimate *Q* and $T_M = (1 + \frac{1}{M}) b_M + \bar{u}_M$ to estimate the variance of $\bar{q}_M$. When *n* and *M* are large, inferences for scalar *Q* can be based on normal distributions, so that a $(1-\alpha)\%$ confidence interval for *Q* is $\bar{q}_M \pm z(\alpha/2)\sqrt{T_M}$. For moderate *M*, inferences can be based on *t*–distributions with degrees of freedom $v_M = (M-1)(1 + r_M^{-1})^2$, where $r_M = (1 + M^{-1}) b_M / \bar{u}_M$, so that a $(1-\alpha)\%$ confidence interval for *Q* is $\bar{q}_M \pm t_{vM}(\alpha/2)\sqrt{T_M}$. Refinements of these basic combining rules have been proposed by several authors, including Li, Raghunathan and Rubin (1991a), Li, Meng and Rubin (1991b), Raghunathan and Siscovick (1996), and Barnard and Rubin (1999).

3. Illustrative Simulations

In this section, we use simulations to illustrate the biases/inefficiencies associated with incorporating design features in imputation models. We simulate three target populations of $N = 100,000$ units that are stratified and clustered within strata. In the first population, *Y* depends on both stratum and cluster effects. In the second population, *Y* depends on strata but not on cluster effects. In the third population, *Y* is unrelated to the stratum and cluster indicators. The first population is used to demonstrate the importance of including all relevant design variables, and the second and third populations are used to examine the effect of including irrelevant design variables. The simulated populations are stylized to illustrate the importance of modeling the survey design; hence, the magnitudes of biases/inefficiencies may not be generalizable to other settings.

Each population is divided into five equally-sized strata comprised of $N_h = 200$ clusters, for $h = 1, \ldots, 5$. Each cluster *c* in stratum *h* is comprised of N_{hc} units. In each stratum, ten clusters have $N_{hc} = 300$, twenty clusters have $N_{hc} = 200$, sixty clusters have $N_{hc} = 100$, sixty clusters have $N_{hc} = 75$, and fifty clusters have $N_{hc} = 50$. Cluster sizes are varied to magnify design effects when taking multi-stage cluster samples. For each target population, there are two survey variables, *X* and *Y*. In all three populations, for simplicity we generate each X_{hcj}, where *j* indexes a unit within stratum and cluster *hc*, from $X_{hcj} \sim N(0, 10^2)$. To generate *Y*, we use different methods for each population, as shall be described in subsequent sections.

We randomly sample units from each population using multi-stage cluster sampling. First, we take a simple random sample of $n_1 = 40$ clusters from stratum 1, $n_2 = 20$ clusters from stratum 2, $n_3 = 30$ clusters from stratum 3, $n_4 = 10$ clusters from stratum 4, and $n_5 = 15$ clusters from stratum 5. The cluster sample sizes differ across strata to magnify

design effects relative to equal sampling. We then take a simple random sample of twenty units from each sampled cluster. Hence, there are 2,300 units with $I_{hcj} = 1$.

The estimands of interest in each population are $Q = \bar{Y}$, the population mean of Y, and the coefficients for the population regression of Y on X. The complete-data estimator of $\bar{Y}$ is the usual, unbiased design-based estimator,

$$q = \frac{1}{100,000}\left(\sum_{h=1}^{5} \frac{200}{n_h}\sum_{c\in h} \hat{y}_{hc}\right),$$

where $\hat{y}_{hc} = N_{hc}\bar{y}_{hc}$ is the estimated total in cluster *hc*. The complete-data estimator of the variance of q is,

$$u = \frac{1}{100,000^2}\left(\begin{array}{l}\sum_{h=1}^{5} 200^2\left(1-\frac{n_h}{200}\right)s_h^2/n_h \\ +\sum_{h=1}^{5} \frac{200}{n_h}\sum_{c\in h} N_{hc}^2\left(1-\frac{20}{N_{hc}}\right)s_{hc}^2/20\end{array}\right),$$

where s_h^2 is the sample variance of the $\hat{y}_{hc}$ and s_{hc}^2 is the sample variance of Y within cluster *hc*. The estimators of the coefficients in the regression of Y on X are the usual approximately unbiased, design-based estimators, which are computed using the "survey" routines (Lumley 2004) in the software package R. These routines estimate variances using Taylor series linearizations. These estimators are used for all multiply-imputed data sets in all simulations.

For each sample, we let X be fully observed, and let Y be missing for about 30% of the sampled units.

Each unit's binary response variable, R_{hcj}, is drawn from a Bernoulli distribution:

$$\Pr(R_{hcj} = 1) = \frac{\exp(-0.847 - 0.1X_{hcj})}{1+\exp(-0.847 - 0.1X_{hcj})} \qquad (4)$$

Here, $R_{hcj} = 1$ means that the unit's value of Y is missing. Equation 4 implies that Y_{mis} is missing at random (Rubin 1976). We can ignore the missing data mechanism provided that imputations for missing data are conditional on X. We purposefully do not allow missingness to depend on stratum or cluster membership to illustrate that bias can arise from failing to account for the survey design even when the ignorable missing data mechanism does not depend on the sampling design. Of course, if the sampling design is related to missingness, as it is in many real datasets, one must condition on the sampling design to make the missing data mechanism ignorable.

We examine three strategies to impute Y_{mis} that make different use of the design information. These strategies are summarized in Table 1. The first strategy, labeled SRS, completely disregards the sampling design. The second strategy, FX, incorporates the stratification and the clustering by using fixed effects for each cluster within stratum. The third strategy, HM, uses normal random effects models that incorporate the stratification and clustering. For SRS, one model is fit to the entire data set. For FX and HM, models are fit separately in each stratum. All three strategies regress on X because it is part of the missing data mechanism; not conditioning on X would violate ignorability and cause bias.

Table 1
Imputation Strategies

Label	Imputation model for missing Y_{hcj}
SRS	$N(\beta_0 + \beta_1 X_{hcj}, \sigma^2)$
FX	$N(\beta_{0h} + \beta_{1h} X_{hcj} + \omega_{hc}, \sigma_h^2)$
HM	$N(\beta_{0h} + \beta_{1h} X_{hcl} + \omega_{hc}, \sigma_h^2)$, $\omega_{hc} \sim N(0, \tau^2)$

All imputations are draws from the appropriate Bayesian posterior predictive distributions. First, we draw parameters of the imputation models from their posterior distributions given the components of the observed data, $(Z, X, Y_{\text{obs}}, I, R)$, that are included in the models. Second, we draw values of the missing data from the distributions given in Table 1. Diffuse priors are used for all parameters. For strategy HM, we draw values of the parameters using a Gibbs sampler (Gelfand and Smith 1990). We run the sampler for a burn-in period to get approximate convergence, then we use every tenth draw for imputations. Finally, we use $M = 5$ independently drawn imputations in each data set for each strategy.

3.1 Simulation A: Illustration of Disregarding Relevant Design Features

In this simulation, we generate a population in which the distributions of Y differ across strata and clusters. We call this "Population 1". Specifically, for unit j in stratum h and cluster c, we construct the population value of Y_{hcj} from

$$Y_{hcj} = 10X_{hcj} + \beta_{0h} + \omega_{hc} + \in_{hcj} \qquad (5)$$

where β_{0h} is a scalar constant for stratum h, the ω_{hc} is a scalar constant for cluster *hc*, and $\in_{hcj}$ is a random error term drawn from $N(0, 200^2)$. The values of the stratum effects are $\beta_{01} = 500, \beta_{02} = -250, \beta_{03} = 0, \beta_{04} = 250,$ and $\beta_{05} = -500$. The values of the ω_{hc} are obtained by drawing five sets of $N_h = 200$ values from independent $N(0, 70^2)$. The stratum and cluster effects are widely dispersed to magnify design effects relative to simple random sampling, which in turn magnifies the effects of disregarding the design in imputations. We then sample from this population using the stratified cluster sampling scheme outlined previously. We create the missing data indicator R using equation 4.

Table 2 shows the results of 1,000 replications of the three imputation strategies outlined in Table 1. The additional row labeled "Complete data" shows the results using the data for all sampled units, *i.e.*, assuming no units with $I_{hcj} = 1$ have $R_{hcj} = 0$. The column labeled "95% CI cov." contains the percentage of the 1,000 simulated confidence intervals that contain the population parameter. The column labeled "Pt. Est." contains the averages of the 1,000 point estimates of Q. The column labeled "Var" contains the variances of the 1,000 point estimates of Q. The column labeled "Est. Var" contains the averages across the 1,000 replications of the estimated variances of the point estimates. The columns labeled "Var(Est.Var)" and "MSE(Est.Var)" give the variance and mean squared error of the 1,000 estimated variances.

Imputations based on method SRS lead to severely biased estimates and very poor confidence interval coverage in this population. These problems exist even though there is not much missing information and despite the fact that we use design-unbiased estimators for inferences. Both FX and HM have point estimates that approximately match the complete-data point estimates, and both have coverage rates that approximately match the rates for the complete data inferences. FX and HM have similar profiles because the fixed effect models and the hierarchical models produce similar estimates of the parameters in equation 5.

When estimating the population mean, the variance associated with FX or HM is only slightly larger than the variance associated with the complete-data estimator. This is because of the large cluster effects, which makes the within-imputation variance a dominant factor relative to the between-imputation variance. That is, the fraction of missing information due to missing data is relatively small when compared to the effect of clustering.

3.2 Simulation B: Illustration of including irrelevant predictors

Modeling the design features is essential when the features are related to the survey variables of interest. How does modeling irrelevant design features affect inferences? In this section, we present the results of two simulation studies that explore this question.

First, we generate "Population 2" in which the distribution of Y differs across strata but does not depend on the clusters. To do so, we use the same generation method as in Equation 5, setting the ω_{hc} equal to zero. The $\in_{hcj}$ are drawn from $N(0, 100^2)$. We sample from Population 2 and generate missing data using the schemes outlined previously. The results for 1,000 replications are displayed in Table 3.

SRS continues to have severe bias and poor confidence interval coverage because it ignores the stratification. For FX and HM, the averages of their point estimates are within simulation error of the average of the point estimates for the complete data. Additionally, their confidence interval coverage rates approximately match the coverage rate for the complete-data intervals. This indicates that FX and HM are reasonable for these populations, even though the irrelevant cluster features are included in their imputation models.

We next generate "Population" 3 in which the distribution of Y is independent of the strata and cluster membership indicators. Specifically, to generate Y, we subtract the β_{0h} from the values of Y generated in Population 2. We then sample from Population 3 using the stratified cluster sampling scheme and create missing data using the methods outlined previously. The results for 1,000 replications are displayed in Table 4.

Table 2
Performance of Imputation Procedures when the Design Features are Related to the Survey Variable of Interest. The Population Mean Equals 3.2 and the Population Regression Coefficients Equal 3.0 and 10.1

	Method	95% CI cov.	Pt. Est.	Var	Est. Var	Var(Est. Var)	MSE (Est. Var)
Mean Y	Complete data	94.2	2.0	544.91	527.31	31,626.19	31,936.07
	SRS	38.0	45.8	327.79	360.74	11,927.97	13,013.35
	FX	94.8	2.4	554.09	579.92	37,474.82	38,141.70
	HM	94.5	2.3	551.02	553.16	34,056.39	34,060.99
Intercept	Complete data	93.0	2.4	529.51	499.73	18,543.13	19,430.21
	SRS	39.5	46.8	340.09	365.50	9,351.15	9,996.99
	FX	94.5	2.8	539.19	551.68	21,529.16	21,685.33
	HM	93.9	2.7	536.82	524.82	19,256.24	19,400.11
Slope	Complete data	93.3	10.1	1.24	1.15	0.14	0.15
	SRS	64.8	7.6	2.10	2.20	0.55	0.56
	FX	94.5	10.1	1.45	1.44	0.18	0.18
	HM	95.7	10.1	1.53	1.65	0.29	0.30

Table 3
Performance of Imputation Procedures when the Population has Stratum Effects but no Cluster Effects. The Population Mean Equals 0.34 and the Population Regression Coefficients Equal 0.14 and 10.13

	Method	95% CI cov.	Pt. Est.	Var	Est. Var	Var(Est. Var)	MSE (Est. Var)
Mean Y	Complete data	93.6	0.37	468.97	461.88	29,301.77	29,352.04
	SRS	31.1	42.90	259.46	303.46	10,228.40	12,164.74
	FX	93.7	0.32	473.86	474.21	30,408.95	30,409.07
	HM	93.4	0.34	476.03	465.53	29,406.61	29,516.85
Intercept	Complete data	93.0	0.72	451.46	432.74	14,955.20	15,305.73
	SRS	31.5	43.10	275.22	311.36	8,134.04	9,440.57
	FX	93.2	0.66	456.08	444.88	15,539.21	15,664.64
	HM	92.3	0.68	457.48	436.25	14,941.00	15,391.75
Slope	Complete data	93.1	10.09	0.99	0.91	0.09	0.10
	SRS	59.0	7.72	1.67	1.77	0.35	0.36
	FX	93.4	10.10	1.03	0.98	0.10	0.10
	HM	93.3	10.10	1.03	0.96	0.10	0.10

Table 4
Performance of Imputation Procedures when the Design Variables are Completely Unrelated to the Survey Variable of Interest. The Population Mean Equals 0.34 and the Population Regression Coefficients Equal 0.14 and 10.04

	Method	95% CI cov.	Pt. Est.	Var	Est. Var	Var(Est. Var)	MSE (Est. Var)
Mean Y	Complete data	94.7	0.35	14.61	14.73	32.65	32.66
	SRS	95.7	0.12	16.45	19.22	40.65	48.31
	FX	97.8	0.40	19.64	28.29	97.66	172.38
	HM	95.1	0.26	18.77	19.16	47.29	47.44
Intercept	Complete data	93.7	0.12	7.13	7.20	5.31	5.32
	SRS	96.8	–0.10	8.97	11.72	13.59	21.10
	FX	98.6	0.17	12.23	20.62	39.84	110.24
	HM	96.2	0.03	10.45	11.61	15.09	16.45
Slope	Complete data	94.5	10.04	0.07	0.07	0.001	0.001
	SRS	96.4	10.07	0.10	0.13	0.002	0.003
	FX	96.4	10.04	0.12	0.15	0.003	0.004
	HM	95.2	10.05	0.11	0.12	0.002	0.002

SRS finally produces point estimates whose averages are within simulation error of the complete data average point estimate. This is because the imputations in SRS reflect the population structure reasonably well. This suggests that disregarding the design in imputation models may provide acceptable inferences when the design variables are only weakly correlated with the survey outcomes. As in the previous simulations, FX and HM continue to have average point estimates within simulation error of the complete-data average point estimate. When comparing the three imputation strategies, we see that FX and HM are inefficient relative to SRS. This is because the imputation models for FX and HM estimate parameters that equal approximately zero in the population, whereas SRS sets them equal to zero. HM has smaller variance than FX does, because the hierarchical imputation model smoothes the estimated cluster effects towards zero.

For FX, the percentage of confidence intervals that cover Q is larger than the percentages for the complete-data intervals and HM intervals. This is because the estimated variance for FX tends to be larger than its actual variance. This apparent upward bias in T_M also exists for SRS, resulting in a larger coverage percentage than those for the complete-data and HM.

4. Real Data Example

We next examine the effect of accounting for stratification and clustering when imputing missing data in a genuine dataset. The data are taken from the public use file for the 1999–2002 National Health and Nutrition Examination Surveys. Individuals are grouped in 56 clusters divided among 28 strata. Many variables have 5% to 10% missing data.

We imputed missing data using two strategies: one ignoring design variables (like SRS) and one incorporating the design variables using fixed effects for cluster indicators (like FX). In the imputation model, we included 27 dummy variables to represent 28 strata and one dummy variable within-each stratum to represent the two clusters nested within each stratum. That is, a total of 55 dummy variables

were included as predictors. We used a stepwise variable selection procedure to identify statistically significant interactions between these dummy variables and survey variables, and we included these interactions as predictors in the imputation model as well. The values were imputed using the sequential regression approach implemented in the software package IVEWARE (www.isr.umich.edu/src/smp/ive). We generate $M = 10$ data sets for each strategy.

We consider three estimands. The first is the population percentage of people who have ever had their blood cholesterol level checked (BPQ060). This variable has about 15% missing values. The second and third are the population regression coefficients in a logistic regression of BPQ060 on family poverty income ratio (INDFMPIR), a continuous variable that has about 12% missing values. These estimands are estimated using design-based methods computed with the "survey" routines in the software package R.

Table 5 displays the results for both imputation strategies. The two sets of estimates for all analyses are very similar. In this case, incorporating the design variables into the imputation model hardly impacts the results. This is due in part to the small fractions of missing information and the relative unimportance of stratum and cluster effects. However, there is minimal penalty for including the design features in the imputation model. In light of the results of the simulations in section 3, we would incorporate the design features in this imputation model.

Table 5
Comparison of Real Data Results when Design Features are Included in Imputation Model and when Design Features are Ignored

	Pt. Est.	S.E.	95% CI
Mean BPQ060			
design	0.319	0.010	(0.299, 0.339)
no design	0.319	0.011	(0.296, 0.341)
Intercept: Logistic Regression			
design	0.362	0.054	(0.256, 0.467)
no design	0.352	0.052	(0.251, 0.454)
Slope: Logistic Regression			
design	–0.409	0.020	(–0.449, –0.369)
no design	–0.407	0.019	(–0.444, –0.371)

5. Concluding Remarks

The simulation studies, though limited, suggest disregarding the sampling design in multiple imputation can be a risky practice. When the design variables are related to the survey variables, as in our Simulation A, failing to include the design variables can lead to severe bias. On the other hand, including irrelevant design variables, as in our Simulation B and the NHANES example, leads at worst to inefficient and conservative inferences when the imputation models are otherwise properly specified.

Including dummy variables for cluster effects greatly reduced the bias relative to disregarding the design completely. However, blindly including dummy variables is not an automatic solution. When the regression slopes or variances differ across clusters, using FX or HM may result in biased estimates, since important design features are disregarded. Imputers suspecting such relationships should include appropriate interactions with the dummy variables for the design features, as we did in the NHANES example. In some surveys the design may be so complicated that it is impractical to include dummy variables for every cluster. In these cases, imputers can simplify the model for the design variables, for example collapsing cluster categories or including proxy variables (*e.g.*, cluster size) that are related to the outcome of interest.

The simulations suggest that there can be payoffs to using hierarchical models for imputation of missing data relative to using fixed effects models, particularly when cluster effects are similar. However, hierarchical models are more difficult to fit than fixed effect models. For example, it is daunting to fit hierarchical models in complex designs when data are missing for several continuous and categorical variables. It may be possible to fit sequential hierarchical models in a spirit similar to the sequential regression imputations of Raghunathan *et al.* (2001). This is an area for future research. A further disadvantage of hierarchical models is that they are easier to mis-specify than fixed effects models. For example, if the cluster effects follow a non-normal distribution, the hierarchical normal model used in this paper could provide implausible imputations.

With multiple imputation, the key to success is specifying an imputation model that reasonably describes the conditional distribution of the missing values given the observed values. Design features frequently are related to survey variables, so that including them in the imputation models reduces the risks of model mis-specification. We believe that in many cases the potential biases resulting from excluding important design variables, or other variables related to the missing data mechanism, outweigh the potential inefficiencies from estimating small coefficients. This reinforces the general advice provided by many on multiple imputation: include all variables that are related to the missing data in imputation models to make the missing data mechanism ignorable (*e.g.*, Meng 1994; Little and Raghunathan 1997; Schafer 1997, and Collins, Schafer and Kam 2001).

Acknowledgements

This research was funded by the National Science Foundation grant ITR-0427889. The authors thank the associate editor and reviewers for their comments and suggestions.

References

Barnard, J., and Meng, X. (1999). Applications of multiple imputation in medical studies: From AIDS to NHANES. *Statistical Methods in Medical Research*, 8, 17-36.

Barnard, J., and Rubin, D.B. (1999). Small-sample degrees of freedom with multiple-imputation. *Biometrika*, 86, 948-955.

Collins, L.M., Schafer, J.L. and Kam, C.K. (2001). A comparison of inclusive and restrictive strategies in modern missing-data procedures. *Psychological Methods*, 6, 330-351.

Gelfand, A.E., and Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.

Heitjan, D.F., and Little, R.J.A. (1991). Multiple imputation for the Fatal Accident Reporting System. *Applied Statistics*, 40, 13-29.

Kennickell, A.B. (1998). Multiple imputation in survey of consumer finances. In *Proceedings of the Section on Business and Economic Statistics*, American Statistical Association, 11-20.

Li, K.H., Raghunathan, T.E. and Rubin, D.B. (1991a). Large-sample significance levels from multiply-imputed data using moment-based statistics and an F reference distribution. *Journal of the American Statistical Association*, 86, 1065-1073.

Li, K.H., R.T.E., Meng, X.L. and Rubin, D.B. (1991b). Significance levels from repeated $p-$values with multiply-imputed data. *Statistica Sinica*, 1, 65-92.

Little, R.J.A. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, 9, 407-426.

Little, R.J.A., and Raghunathan, T.E. (1997). Should imputation of missing data condition on all observed variables? In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 617-622.

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, 9, 8.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input (disc: P558-573). *Statistical Science*, 9, 538-558.

Raghunathan, T.E., Lepkowski, J.M., van Hoewyk, J. and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a series of regression models. *Survey Methodology*, 27, 85-96.

Raghunathan, T.E., and Paulin, G.S. (1998). Multiple imputation of income in the Consumer Expenditure Survey: Evaluation of statistical inference. In *Proceedings of the Section on Business and Economic Statistics*, American Statistical Association, 1-10.

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

Raghunathan, T.E., and Siscovick, D.S. (1996). A multiple-imputation analysis of a case-control study of the risk of primary cardiac arrest among pharmacologically treated hypertensives. *Applied Statistics*, 45, 335-352.

Reiter, J.P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29, 181-189.

Reiter, J.P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology*, 30, 235-242.

Reiter, J.P. (2005). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society*, Series A, 168, 185-205.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-590.

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

Rubin, D.B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, 9, 462-468.

Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473-489.

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

Schafer, J.L., Ezzati-Rice, T.M., Johnson, W., Khare, M., Little, R.J.A. and Rubin, D.B. (1998). The NHANES III multiple imputation project. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 28-37.

Schenker, N., Raghunathan, T.E., Chiu, P.-L., Makuc, D.M., Zhang, G. and Cohen, A.J. Multiple imputation of missing income data in the National Health Interview Survey. *Journal of the American Statistical Association*, forthcoming.