DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 33, No. 4, 2017, pp. 1005–1019, http://dx.doi.org/10.1515/JOS-2017-0047

Multiply-Imputed Synthetic Data: Advice to the Imputer

Bronwyn Loong[1] *and Donald B. Rubin*[2]

Several statistical agencies have started to use multiply-imputed synthetic microdata to create public-use data in major surveys. The purpose of doing this is to protect the confidentiality of respondents' identities and sensitive attributes, while allowing standard complete-data analyses of microdata. A key challenge, faced by advocates of synthetic data, is demonstrating that valid statistical inferences can be obtained from such synthetic data for non-confidential questions. Large discrepancies between observed-data and synthetic-data analytic results for such questions may arise because of uncongeniality; that is, differences in the types of inputs available to the imputer, who has access to the actual data, and to the analyst, who has access only to the synthetic data. Here, we discuss a simple, but possibly canonical, example of uncongeniality when using multiple imputation to create synthetic data, which specifically addresses the choices made by the imputer. An initial, unanticipated but not surprising, conclusion is that non-confidential design information used to impute synthetic data should be released with the confidential synthetic data to allow users of synthetic data to avoid possible grossly conservative inferences.

Key words: Data confidentiality; data utility; multiple imputation.

1. Introduction and Review of Issues

Releasing synthetic microdata rather than actual microdata is a method of statistical disclosure control in the context of public dissemination of survey data. The goal is to limit the risk of disclosure of survey respondents' identities or sensitive attributes, while simultaneously retaining enough detail in the synthetic data to preserve valid conclusions drawn about the target population for many non-confidential population-level estimates. The idea was first proposed by Rubin (1993) based on the theory of *multiple imputation* (Rubin 1987). Synthetic microdata sets are created by the imputer, often a federal agency, ideally using samples drawn from the posterior predictive distribution of the target population under a proper imputation model. Using (i) an acceptable imputation model that correctly captures basic relationships among survey variables, and (ii) estimation methods based on the principles of multiple imputation, analysts can make valid inferences about non-confidential attributes in the target population using standard complete-data statistical methods, without accessing the actual confidential microdata. If all observed values are replaced and no true values are released, this process is known in the literature as creating *fully synthetic* data. A *partially synthetic* data set consists of a mix of multiply-imputed and actual data values.

[1] Australian National University – Research School of Finance, Actuarial Studies and Statistics, College of Business and Economics Building 26C The Australian National University Canberra, Canberra, Australian Capital Territory 2601, Australia. Email: bronwyn.loong@anu.edu.au
[2] Harvard University – Department of Statistics, Cambridge, MA 02138-2901, U.S.A. Email: rubin@stat.harvard.edu

Some basic inferential methods for the analysis of fully synthetic data were derived in Raghunathan et al. (2003). Simulated and empirical data examples of fully synthetic data can be found in Reiter (2002), Raghunathan et al. (2003) and Reiter (2005a). Since then, the basic fully synthetic data framework has been adapted to meet other disclosure control criteria. Some key developments requiring new inferential methods include inference from partially synthetic data (Reiter 2003); releasing multiply-imputed synthetic data in two stages (Reiter and Drechsler 2010), which allows agencies to release different numbers of imputations for different variables (see also the related approach, called nested multiple imputation, in Shen 2000; Harel and Schafer 2003; Rubin 2003); and sampling with synthesis (Drechsler and Reiter 2010), which combines the disclosure control benefits of partially synthetic data and random sampling, so that potential intruders no longer know whether their target units are in the released data.

Prior to public release, the disclosure risk and data utility of any synthetic data set should be assessed. Abowd and Vilhuber (2008) proposed some disclosure risk measures for fully synthetic data based on the ideas of differential privacy from the computer science literature. Methods to estimate risks of identification disclosure with partially synthetic data were developed by Reiter and Mitra (2009), based on the earlier approach in Duncan and Lambert (1989) which computes estimated probabilities of unit identification conditional on the released data. Reiter et al. (2014) proposed a Bayesian framework to estimate the disclosure risks in fully and partially synthetic data based on the distribution of the original values given the released synthetic data and information on the data generation mechanism. Data-utility measures attempt to characterize the quality of inferences about the target population using the synthetic data, relative to those using the actual data set. Such comparisons can be tailored to specific analyses (Karr et al. 2006), or can be broadened to reflect global differences in distributions (Woo et al. 2009).

Reiter (2009) summarized some research challenges in multiple imputation for disclosure limitation. One challenge is ‘confidence in synthetic data’, which requires demonstrating to the public that useful statistical conclusions can be drawn about the target population using synthetic data. This task is difficult because the imputer and any analyst comprise distinct entities, and it is generally impossible for the imputer to foresee and incorporate in its imputation models all estimands of importance to all future analysts.

Meng (1994, 539) defined *imputation input* to include the imputer’s model assumptions; purpose of imputation; available information and data from the data collection phase, as well as any other potentially related resources (e.g., past similar surveys). Imputation input is summarized by an *imputation model*. *Analysis input* is more vague but includes the analyst’s purpose of investigation; data made available for analysis; information about the imputation model that is made available by the imputer; and the analyst’s computational skills. Analysis input is summarized by an *analysis procedure*. Meng (1994, 539) coined the term *uncongeniality* of the analysis procedure with the imputation model when using multiple imputation to fill in missing data, expressed as follows:

> ‘Uncongeniality essentially means that the analysis procedure does not correspond to the imputation model. The uncongeniality arises when the analyst and the imputer have access to different amounts of information and have different assessments.’

In other words, imputers and analysts have access to, and use, different types of inputs, and this uncongeniality can lead to large discrepancies between observed-data analytic results and synthetic-data analytic results. A common example occurs when a simple imputation model does not capture some relationships of the observed data that are important to the analyst. Uncongeniality is a challenging theoretical topic, with, at present, only limited analytic results available. Furthermore, we cannot directly apply the technical work on congeniality in Meng (1994) to synthetic data imputation and analysis, because with synthetic data, at least part of the observed data set used for creating imputations is confidential and is not available to analysts.

The aim of this article is to provide advice to the imputer regarding what information and advice should be communicated to analysts for valid and efficient inference. If analysts can ignore complex sampling designs and methods of imputation when analysing synthetic data, that is, if they can assume that the multiply-imputed synthetic data come from simple random samples, then the analysis burden is reduced. The utility of including survey design information in synthetic data sets has been discussed briefly in Reiter (2002), Drechsler et al. (2008), and Reiter and Drechsler (2010). We present a simple simulation experiment to advance the discussion.

In Section 2 we review the creation of fully synthetic data. Section 3 describes the analysis of fully synthetic data from Raghunathan et al. (2003). We describe our simulation experiment in Section 4, and Section 5 discusses its results. Section 6 provides concluding remarks.

2. Creation of Fully Synthetic Data

Suppose the data collector, conducts a survey to collect detailed information from a sample of n units from a target population of $N >> n$ units. The original sampling mechanism (denoted $\mathscr{I}_{\text{inc}}$, *inc* for "included") that generates the data uses r background covariates (e.g., stratum indicators), which are known to the data collector for all N units in the population and comprise the $(N \times r)$ population matrix $\mathbf{X}$. For the n units included in the survey, information is collected on p variables of interest, whose values are unknown prior to the survey; the values of these survey variables in the population comprise the $(N \times p)$ matrix $\mathbf{Y}$. For simplicity, assume all survey variables in $\mathbf{Y}$ are confidential, but none of the variables in $\mathbf{X}$ are confidential.

Let $\mathbf{Y}_{\text{inc}}$ be the $(n \times p)$ matrix representing the portion of $\mathbf{Y}$ corresponding to the sampled (that is, *included* in the sample) units; $\mathbf{Y}_{\text{inc}}$ is observed by the data collector. Let $\mathbf{Y}_{\text{exc}}$ be the $((N - n) \times p)$ matrix representing the portion of $\mathbf{Y}$ corresponding to not-sampled (that is, *excluded* from the sample) units. Similarly let $\mathbf{X}_{\text{inc}}$ be the values of $\mathbf{X}$ corresponding to the included units, and let $\mathbf{X}_{\text{exc}}$ be the values of $\mathbf{X}$ corresponding to the excluded units. Define $Z = \{\mathbf{X}, \mathbf{Y}_{\text{inc}}\}$ to be the set of known and observed (to the data collector) microdata (that is, data at the unit level), which includes the values of the survey design variables, $\mathbf{X}$, for *all* units in the population.

If there were no confidentiality concerns, the data collector would release Z. Even if Z were released, some users would discard, and hence not use, $\mathbf{X}_{\text{exc}}$ if it is viewed as too much of a burden to store and to attempt to use this information. The task for multiply-imputed synthetic microdata is to create $m > 1$ synthetic microdata sets (denoted $Z^{(1)}_{\text{syn}}, \ldots, Z^{(\text{m})}_{\text{syn}}$) for release instead of Z, where each $Z^{(l)}_{\text{syn}} (l = 1, \ldots, m)$ consists of $\mathbf{X}$ and synthetic (i.e., imputed) values for n_{syn} units in the excluded part of the population, that is, n_{syn} rows of $\mathbf{Y}_{\text{exc}}$. Multiply-imputed data sets are created using the following steps:

- **Step 1**: Draw m values of the excluded values $\mathbf{Y}_{\text{exc}}$ from their joint posterior predictive distribution $g(\mathbf{Y}_{\text{exc}}|Z)$, say $\mathbf{Y}_{\text{exc}}^{(l)}$, $l = 1, \ldots, m$, which is conditional on all information in the population known to the data collector. At the conclusion of Step 1, we obtain m versions of the complete-data population, $\left(\mathbf{X}, \mathbf{Y}_{\text{pop}}^{(l)}\right)$ $(l = 1, \ldots, m)$.
- **Step 2**: Draw n_{syn} units (rows) from each $\mathbf{Y}_{\text{pop}}^{(l)}$ using sampling mechanism $\mathscr{I}_{\text{syn}}$, thereby producing $Z_{\text{syn}}^{(l)} = \left\{\mathbf{X}, \mathbf{Y}_{\text{syn}}^{(l)}\right\}$ for release, where $\mathbf{Y}_{\text{syn}}^{(l)}$ is the $(n_{\text{syn}} \times p)$ matrix of synthetic survey values of $\mathbf{Y}$; because $N >> n$, and we assume $N >> n_{\text{syn}}$, for simplicity we assume no actual observed values are included in any $\mathbf{Y}_{\text{syn}}^{(l)}$. Let $\underset{\sim}{\mathcal{Z}}_{\text{syn}} = \left\{Z_{\text{syn}}^{(l)}, l = 1, 2, \ldots m\right\}$ be the set of m synthetic microdata sets.

Step 2 can be computationally merged into Step 1 by only imputing synthetic values for units drawn by $\mathscr{I}_{\text{syn}}$. Various methods can be used to generate approximate draws of $\mathbf{Y}_{\text{exc}}$ from their joint posterior predictive distribution: joint modelling (Schafer 1997); sequential regression multivariate imputation (SRMI) (Van Buuren and Oudshoorn 2000; Raghunathan et al. 2001), also known as multiple imputation by chained equations (MICE); as well as other related approaches (e.g., Rubin 2003; Li et al. 2014). Note that some of these methods (such as MICE) do not necessarily converge to the target joint posterior predictive distribution. This topic of convergence of such methods has been an active area of statistical research, but is beyond our topic here.

3. Analysis of Synthetic Data Sets

Let Q be the target population quantity of interest, that is, the estimand, and for simplicity of exposition, let Q be a scalar function of $\mathbf{X}$ and $\mathbf{Y}$, $Q = Q(\mathbf{X},\mathbf{Y})$. Suppose that, given the original data set $Z = \{\mathbf{X},\mathbf{Y}_{\text{inc}}\}$, the analyst would use a point estimate q_{inc} of Q, and an associated measure of sampling variance, v_{inc}, and further suppose that using these, the analyst's inference in large samples would be valid; that is, $q_{\text{inc}} \pm 1.96\sqrt{v_{\text{inc}}}$ would be a valid 95% confidence interval in the standard Neyman (1934) sense of including Q at least 95% of the time in repeated samples drawn using $\mathscr{I}_{\text{inc}}$; both q_{inc} and v_{inc} are generally functions of $\mathbf{Y}_{\text{inc}}$ and $\mathbf{X} = (\mathbf{X}_{\text{inc}}, \mathbf{X}_{\text{exc}})$.

In place of Z, the collection of synthetic data sets $\underset{\sim}{\mathcal{Z}}_{\text{syn}}$ generated by the imputer is released to analysts. Let $\left(q_{\text{syn}}^{(l)}, v_{\text{syn}}^{(l)}\right)$ be the values of the statistics q_{inc} and v_{inc} computed from synthetic data set $Z_{\text{syn}}^{(l)}$ $(l = 1, \ldots, m)$. The analyst needs rules to combine the results from the analysis of the m synthetic data sets to draw inference about the target population quantity Q. These rules are passed on to data analysts so that they can obtain valid inferences in the standard frequentist sense.

Raghunathan et al. (2003) derived approximations to the first and second moments of the posterior distribution of Q, assuming that both $\mathscr{I}_{\text{inc}}$ and $\mathscr{I}_{\text{syn}}$ are simple random sampling mechanisms, and that the data are analyzed using standard estimators. Reiter (2002) showed in a simulation study that valid inferences can be obtained for fully synthetic data under sampling designs more complex than simple random sampling, but still assuming that $\mathscr{I}_{\text{inc}}$ and $\mathscr{I}_{\text{syn}}$ are the *same* sampling mechanism. Raghunathan et al. (2003) conjectured that $\mathscr{I}_{\text{syn}}$ and $\mathscr{I}_{\text{inc}}$ can be different sampling plans because each synthetic data set being released is created from the imputed complete-data population $\left(\mathbf{X}, \mathbf{Y}_{\text{pop}}^{(l)}\right)$, but this conclusion requires careful consideration, as we will see here.

For combining the m answers from the m synthetic data sets, the following three quantities are used for inference about Q, assuming for the moment that both $\mathscr{I}_{\text{inc}}$ and $\mathscr{I}_{\text{syn}}$ are simple random sampling:

$$\bar{q}_{\text{m}} = \sum_{l=1}^{\text{m}} \frac{q_{\text{syn}}^{(l)}}{m}, \tag{1}$$

$$b_m = \sum_{l=1}^{\text{m}} \frac{\left(q_{\text{syn}}^{(l)} - \bar{q}_{\text{m}}\right)^2}{(m-1)}, \tag{2}$$

and

$$\bar{v}_{\text{m}} = \sum_{l=1}^{\text{m}} \frac{v_{\text{syn}}^{(l)}}{m}. \tag{3}$$

The analyst uses $\bar{q}_{\text{m}}$ as the point estimate of Q. The sampling variance of $\bar{q}_{\text{m}}$ is estimated by

$$\hat{V}_{\text{syn}} = \left(1 + \frac{1}{m}\right) b_m - \bar{v}_{\text{m}}. \tag{4}$$

Raghunathan et al. (2003) denoted the variance estimator in (4) as $\hat{T}_{\text{m}}$, whereas we have changed the notation to $\hat{V}_{\text{syn}}$ to avoid confusion with the standard sampling variance estimator when using multiple imputation for missing data in $\mathbf{Y}_{\text{inc}}$: $\hat{T}_{\text{m}} = \left(1 + \frac{1}{m}\right) b_{\text{m}} + \bar{v}_{\text{m}}$ (Rubin 1987); $\hat{V}_{\text{syn}}$ is not a sum (total) of two components, but rather, the mean within variance $\bar{v}_{\text{m}}$ is subtracted in (4) because there is sampling variability when creating the synthetic data that is included in the estimated between-imputation sampling variance estimator b_m. As with all method-of-moments estimators, (4) is not perfect; for example, it is not constrained to be in the parameter space; for instance, the estimated sampling variance, $\hat{V}_{\text{syn}}$, may be negative. When $\hat{V}_{\text{syn}}$ is negative, alternative variance estimates (see, for example, Reiter 2002; Drechsler and Reiter 2010) have been proposed that are always positive. When $\hat{V}_{\text{syn}} > 0$, inferences for scalar Q can be based on a t-distribution with degrees of freedom

$$df_{\text{syn}} = (m-1)\left(1 - \frac{1}{r_{\text{m}}}\right)^2 = (m-1)\left(1 - \frac{m}{m+1}\frac{\bar{v}_{\text{m}}}{b_{\text{m}}}\right)^2, \tag{5}$$

where $r_{\text{m}} = \frac{(1+\frac{1}{m})b_{\text{m}}}{\bar{v}_{\text{m}}}$, so that a nominal $100(1-\alpha)\%$ asymptotic confidence interval estimate for Q is

$$\bar{q}_{\text{m}} \pm t_{df_{\text{syn}},\alpha/2}\sqrt{\hat{V}_{\text{syn}}}. \tag{6}$$

The t-reference distribution was presented by Raghunathan and Rubin (2000) at the International Society for Bayesian Analysis conference–May 2000. For large m, inference can be based on a standard normal distribution.

Extensions for multivariate Q are presented in Reiter (2005b).

4. Simple Simulation Experiment

In our simple simulation experiment, each unit, indexed by $i = 1, \ldots, N = 1{,}000{,}000$, in the target population, belongs to one of two strata, each of size 500,000, indicated by $X_i = 1$ or $X_i = 2$, known to the data collector for all 1,000,000 units. The estimand, Q, is the population mean of a univariate outcome variable $\mathbf{Y}$. Population values for $\mathbf{Y}$ are generated by $(Y_i|X_i = 1) \sim N(\mu_1 = 100, \sigma_1 = 1)$, and $(Y_i|X_i = 2) \sim N(\mu_2 = 10, \sigma_2 = 1)$, where the simulation results are affinely invariant, that is, the identical statistical conclusions would be obtained for any two-strata population with normal distributions whose means are 90 standard deviations apart. The true population value of Q is clearly 55. The sample drawn by $\mathscr{I}_{\text{inc}}$ has $n = 5{,}000$ units, and thus the matrices $\mathbf{X}_{\text{inc}}$ and $\mathbf{Y}_{\text{inc}}$ have 5,000 rows.

We conducted a 2^4 factorial experiment to investigate the implications of using different methods at each of the four stages: data collection, imputation, sampling of the synthetic data, and analysis of the synthetic data. The factors in the experiment are listed in Table 1.

The first factor (A) is the Actual sampling plan, either simple random sampling (SRS) with 5,000 units or stratified random sampling using $\mathbf{X}$ (StRS) with 2,500 units in each stratum. We assume that factor A is beyond the control of the creator of synthetic data.

The second factor (I) is the Imputation method, either (a) conditional on $\mathbf{X}$ – which is the proper imputation model under StRS because $\mathbf{X}$ is used by StRS, or (b) not conditional on $\mathbf{X}$ – which is an ad-hoc imputation model because it does not condition on all information about the population that is known to the imputer and is improper under StRS. Since Rubin (1978), the advice has been that any variable used in the actual sampling plan must be included in the multiple imputation model for valid inferences to result.

The third factor (S) is the Synthetic-data sampling plan, $\mathscr{I}_{\text{syn}}$, either SRS (with 5,000 units) or StRS (with 2,500 units in each stratum); the number of such data sets is fixed at $m = 200$. We assume that both factors I and S are under the control of the imputer of the synthetic data.

Finally, the fourth factor (E) is the analysis (Estimation) procedure used by the analyst: the simple random sampling estimator, $\widehat{\text{SRS}}$, or the stratified random sampling estimator, $\widehat{\text{StRS}}$, and their associated standard sampling variance estimators. The $\widehat{\text{SRS}}$ procedure completely ignores $\mathbf{X}$, even if it is released. The $\widehat{\text{StRS}}$ requires knowledge of the exact proportions of units in the two strata in the population. The $\widehat{\text{StRS}}$ estimator is used to improve efficiency and this estimator is the appropriate estimator if the synthetic-data sampling plan is StRS, and it is the poststratification estimator if the synthetic-data sampling plan is SRS. Another idea is to estimate the sampling rates for each stratum from

Table 1. Simulation factors.

Factor	Description	Levels
A	Actual sampling plan	{SRS, StRS} each with $n = 5{,}000$
I	Imputation model	{Ad-hoc – not conditional on $\mathbf{X}$, Proper – conditional on $\mathbf{X}$}
S	Synthetic sampling plan	{SRS, StRS}
E	Estimation procedure	$\widehat{\text{SRS}}$,$\widehat{\text{StRS}}$

all the released synthetic samples. Because it is not clear how to use such information for the two tasks of point estimation and for sampling variance estimation, such methods were not studied here, but are considered methods worthy of future research. We ignored the finite population correction factors in the analysis.

As already stated, Factor A, the actual sampling plan, is under the control of the original data collector, and so is generally beyond the control of the creator of the synthetic data. The next two factors (I, the imputation model, and S, the synthetic sampling plan) are under the control of the imputer. However, the imputer cannot control what the analyst will do (that is, factor E, the estimation procedure), although the imputer can convey recommendations for the choice of E. The objective of the simulation study is to address what settings for the factors I and S should the imputer use, and do these influence advice for factor E, and how does this advice depend on factors A, I, and S.

For each combination of the 8 (=2 × 2 × 2) factor levels A, I, and S, a complete data set is created, and for it, $m = 200$ synthetic data sets are created and analysed as if each were a complete data set for each of the two estimation methods of factor E; analysis Equations (1) – (4) are then applied to obtain a point estimate and sampling variance estimate for Q, and thus an interval estimate for Q. This process is replicated 1,000 times at each combination of factors A, I, and S to obtain 1,000 point and 1,000 interval estimates of Q for each estimation procedure. The proportion of the 1,000 interval estimates that contain the true value Q is the "coverage rate" for that procedure in that $A \times I \times S$ cell. We also calculate the "average interval width" of the 1,000 interval estimates for that procedure, for each of the 2 × 2 × 2 combinations of factors A, I, and S. Simulations were conducted using the R software and computing environment.

5. Simulation Experiment – Results

Table 2 reports the coverage rates for observed data collected by SRS, for each combination of factor levels of the imputation model (I), the synthetic data sampling mechanism (S) and the estimation procedure for the synthetic data (E). Table 3 reports the corresponding average interval width values. Tables 4 and 5 are structured analogously, but for actual data collected by StRS.

Table 2 shows that when the actual sampling plan is SRS and the ad-hoc imputation model is used, valid coverage rates are obtained for both analysis procedures and both synthetic data sampling plans. That is, when neither the actual sampling plan nor the

Table 2. Coverage rates for actual data collected by SRS – (factor A, level 1).

		Estimation procedure (E)	
Imputation model (I)	Synthetic data sampling plan (S)	$\widehat{\text{SRS}}$	$\widehat{\text{StRS}}$
Ad-hoc – not conditional on **X**	SRS	94.8	94.9
	StRS	95.4	95.4
Proper – conditional on **X**	SRS	100	95.2
	StRS	100	95.1

Table 3. Average interval widths for actual data collected by SRS – (factor A, level 1).

Imputation model (I)	Synthetic data sampling plan (S)	Estimation procedure (E)	
		$\widehat{\text{SRS}}$	$\widehat{\text{StRS}}$
Ad-hoc – not conditional on **X**	SRS	2.558	2.564
	StRS	2.543	2.543
Proper – conditional on **X**	SRS	6.476	0.057
	StRS	2.495	0.057

imputation model uses **X**, valid confidence intervals are obtained whether the analyst uses **X**, or completely ignores **X** in the analysis; the extremely conservative coverage when using proper imputation and $\widehat{\text{SRS}}$ estimation is worrisome, however. If proper imputation is used, with an actual SRS, then when the synthetic sampling plan is either SRS or StRS, *accurate* coverage is obtained using the $\widehat{\text{StRS}}$ estimation procedure. Table 3 shows that the shortest interval widths are obtained when using proper imputation with either the SRS or StRS synthetic-data sampling plan, but only if $\widehat{\text{StRS}}$ estimation is used. That is, for valid and accurate results, the imputer should use proper imputation and tell the analyst to use the $\widehat{\text{StRS}}$ estimation procedure. The imputer can choose either the SRS or StRS synthetic sampling plan.

When the actual sampling plan is StRS and the imputation model is improper (not conditional on **X**, a method that has been recommended to be avoided for multiple imputation since its inception in the late 1970s), the results in Table 4 show that gross over-coverage will result, no matter what the analyst does (factor E) and no matter what the synthetic data sampling plan is (factor S). The imputer must use proper imputation for accurate coverage as recommended for decades, and the analyst must use the $\widehat{\text{StRS}}$ estimation procedure. Table 5 shows that to produce tight intervals, $\widehat{\text{StRS}}$ should be used for estimation when the imputation model is proper. Again we observe that the imputer can choose either the SRS or StRS synthetic sampling plan.

To emphasize that the user should use the $\widehat{\text{StRS}}$ estimation procedure, users should be told that **X** was used in the imputation model. Our results are compatible with the results in Reiter et al. (2006), which studied the importance of using the sampling design in multiple imputation for missing data using simulation studies, and concluded that the safest course of action is to include design variables in the specification of imputation models to avoid biased estimates. Our results also confirm previous discussions in Reiter (2002), Drechsler

Table 4. Coverage rates for actual data collected by StRS – (factor A, level 2).

Imputation model (I)	Synthetic data sampling plan (S)	Estimation procedure (E)	
		$\widehat{\text{SRS}}$	$\widehat{\text{StRS}}$
Ad-hoc – not conditional on **X**	SRS	100	100
	StRS	100	100
Proper – conditional on **X**	SRS	100	95.3
	StRS	100	95.5

Table 5. Average interval widths for actual data collected by StRS – (factor A, level 2).

Imputation model (I)	Synthetic data sampling plan (S)	Estimation procedure (E)	
		$\widehat{\text{SRS}}$	$\widehat{\text{StRS}}$
Ad-hoc – not conditional on **X**	SRS	2.558	2.559
	StRS	2.559	2.559
Proper – conditional on **X**	SRS	6.640	0.057
	StRS	2.496	0.057

et al. (2008), and Reiter and Drechsler (2010), on the utility of including survey design information in synthetic data sets, which agrees with the old advice from Rubin (1978).

We extended our simulation study to vary the variance between the two strata, to include more than two strata with varying sample sizes, and we also conducted the simulation for multivariate **Y**. These additional simulation results are in the Appendix. For all these extensions, the numerical results for the coverage rate are similar to those in Tables 2 and 4. The magnitude of the average interval width values do differ from those in Tables 3 and 5, but the settings of factors for which accurate interval widths are obtained do not change. Thus, we maintain the same advice to the imputer: use proper imputation, and recommend to analysts that they use estimation procedures that include the covariates used in the imputation model. This advice holds no matter which actual sampling plan was used or what synthetic sampling plan was used.

Xie and Meng (2014) proposed a general theoretical framework for multiple imputation allowing for uncongeniality. Xie and Meng (2014) used a general estimating-equation decomposition theorem to present multiple imputation inference as an integration of the knowledge of the imputer and the analyst. Three parties are involved in their framework: God (G), the Imputer (I), and the Analyst (A). In our simulation study, "God's model" is described by factor A (the actual sampling plan), the imputer's model is formed by factor I (the imputation method) and factor S (the synthetic sampling plan), and the analyst's estimation model is specified by factor E. They considered different scenarios for the relationship between G (God's model), I (the imputer's model) and A (the analyst's estimation model). Although the theoretical findings in Xie and Meng (2014) formally concern Rubin's standard combining rules for multiple imputation (Rubin 1987), and not the combining rules for synthetic data multiple imputation (Raghunathan et al. 2003), it may still be helpful to consider our simulation results in light of their work. We look forward to seeing future work that applies the perspective of Xie and Meng (2014) to the problem of synthetic data.

6. Conclusion

We have discussed a simple, possibly canonical, example of uncongeniality for synthetic data sets created by multiple imputation. Our simulation experiment identified important factors that affect coverage rates and average interval widths with choices at each of the four stages: data collection, imputation, sampling for creation of the synthetic data, and analysis of the synthetic data. A proper imputation method is critical for valid and efficient inference; that is, the imputation method must condition on covariates for accurate inferences to result. Furthermore, the imputer should inform the analyst that this information was used in the imputation, and must recommend to analysts to use this covariate information in their estimation procedures in order to achieve valid and efficient interval estimates. This advice holds whether the acutal sampling plan used this covariate information or not, and whether the synthetic sampling plan used this information or not. Given that our conclusions are drawn from a simple simulation experiment, how broadly this simple conclusion holds should be explored in further work.

Appendix

A1. Simulation Results – Different Variance Between the Two Strata

For the simulation results in Tables 6–9, $(Y_i|X_i = 1) \sim N(\mu_1 = 100, \sigma_1 = 5)$, and $(Y_i|X_i = 2) \sim N(\mu_2 = 10, \sigma_2 = 1)$.

Table 6. Coverage rates for actual data collected by SRS – (factor A, level 1).

		Analysis procedure (E)	
Imputation model (I)	Synthetic data sampling plan (S)	$\widehat{\text{SRS}}$	$\widehat{\text{StRS}}$
Ad-hoc – not conditional on **X**	SRS	95.5	95.5
	StRS	94.4	94.4
Proper – conditional on **X**	SRS	100	95.8
	StRS	100	95.6

Table 7. Coverage rates for actual data collected by StRS – (factor A, level 2).

		Analysis procedure (E)	
Imputation model (I)	Synthetic data sampling plan (S)	$\widehat{\text{SRS}}$	$\widehat{\text{StRS}}$
Ad-hoc – not conditional on **X**	SRS	100	100
	StRS	100	100
Proper – conditional on **X**	SRS	100	94.7
	StRS	100	94.5

Table 8. Average interval widths for actual data collected by SRS – (factor A, level 1).

		Analysis procedure (E)	
Imputation model (I)	Synthetic data sampling plan (S)	$\widehat{\text{SRS}}$	$\widehat{\text{StRS}}$
Ad-hoc – not conditional on **X**	SRS	2.574	2.574
	StRS	2.574	2.574
Proper – conditional on **X**	SRS	6.724	0.204
	StRS	2.503	0.204

Table 9. Average interval widths for actual data collected by StRS – (factor A, level 2).

		Analysis procedure (E)	
Imputation model (I)	Synthetic data sampling plan (S)	$\widehat{\text{SRS}}$	$\widehat{\text{StRS}}$
Ad-hoc – not conditional on **X**	SRS	2.575	2.575
	StRS	2.572	2.572
Proper – conditional on **X**	SRS	6.851	0.204
	StRS	2.504	0.204

A2. Simulation Results – Ten Strata with Varying Sample Sizes and Different Means, Constant Variance

For the simulation results in Tables 10–13, $N_j = j \times 10^j$, $\mu_j = j \times 10$, $n_j = 0.01N_j$ and $\sigma_j = 1$ $(j = 1,\ldots,10)$.

Table 10. Coverage rates for actual data collected by SRS – (factor A, level 1).

		Analysis procedure (E)	
Imputation model (I)	Synthetic data sampling plan (S)	$\widehat{\text{SRS}}$	$\widehat{\text{StRS}}$
Ad-hoc – not conditional on **X**	SRS	94.7	94.7
	StRS	95.8	94.4
Proper – conditional on **X**	SRS	100	93.8
	StRS	100	94.4

Table 11. Coverage rates for actual data collected by StRS – (factor A, level 2).

		Analysis procedure (E)	
Imputation model (I)	Synthetic data sampling plan (S)	$\widehat{\text{SRS}}$	$\widehat{\text{StRS}}$
Ad-hoc – not conditional on **X**	SRS	100	100
	StRS	100	100
Proper – conditional on **X**	SRS	100	95.9
	StRS	100	95.4

Table 12. Average interval widths for actual data collected by SRS – (factor A, level 1).

		Analysis procedure (E)	
Imputation model (I)	Synthetic data sampling plan (S)	$\widehat{\text{SRS}}$	$\widehat{\text{StRS}}$
Ad-hoc – not conditional on **X**	SRS	0.584	0.584
	StRS	0.629	0.594
Proper – conditional on **X**	SRS	1.505	0.024
	StRS	0.564	0.024

Table 13. Average interval widths for actual data collected by StRS – (factor A, level 2).

Imputation model (I)	Synthetic data sampling plan (S)	Analysis procedure (E)	
		$\widehat{\text{SRS}}$	$\widehat{\text{StRS}}$
Ad-hoc – not conditional on **X**	SRS	0.587	0.586
	StRS	0.631	0.592
Proper – conditional on **X**	SRS	1.515	0.027
	StRS	0.567	0.027

A3. *Simulation Results – Multivariate Y*

For the simulation results in Tables 14–17, $(\mathbf{Y}_i|X_i = 1) \sim MVN(\boldsymbol{\mu}_1 = (100, 200, 300), \Sigma_1)$ and $\mathbf{Y}_i|X_i = 2 \sim MVN(\boldsymbol{\mu}_2 = (10, 20, 30), \Sigma_2)$, where

$$\Sigma_1 = \Sigma_2 = \begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix}.$$

Table 14. Coverage rates for actual data collected by SRS – (factor A, level 1).

Imputation model (I)	Synthetic data sampling plan (S)	Analysis procedure (E)	
		$\widehat{\text{SRS}}$	$\widehat{\text{StRS}}$
Ad-hoc – not conditional on **X**	SRS	94.6	94.6
	StRS	95.1	95.1
Proper – conditional on **X**	SRS	100	94.8
	StRS	100	94.9

Table 15. Coverage rates for actual data collected by StRS – (factor A, level 2).

Imputation model (I)	Synthetic data sampling plan (S)	Analysis procedure (E)	
		$\widehat{\text{SRS}}$	$\widehat{\text{StRS}}$
Ad-hoc – not conditional on **X**	SRS	100	100
	StRS	100	100
Proper – conditional on **X**	SRS	100	94.6
	StRS	100	95.4

Table 16. Average interval widths for actual data collected by SRS – (factor A, level 1).

Imputation model (I)	Synthetic data sampling plan (S)	Analysis procedure (E)	
		$\widehat{\text{SRS}}$	$\widehat{\text{StRS}}$
Ad-hoc – not conditional on **X**	SRS	2.556	2.556
	StRS	2.562	2.563
Proper – conditional on **X**	SRS	6.828	0.057
	StRS	2.496	0.057

Table 17. Average interval widths for actual data collected by StRS – (factor A, level 2).

Imputation model (I)	Synthetic data sampling plan (S)	Analysis procedure (E)	
		$\widehat{\text{SRS}}$	$\widehat{\text{StRS}}$
Ad-hoc – not conditional on **X**	SRS	2.558	2.558
	StRS	2.565	2.561
Proper – conditional on **X**	SRS	6.911	0.057
	StRS	2.495	0.057

7. References

Abowd, J.M. and L. Vilhuber. 2008. "How Protective are Synthetic Data?" In *Privacy in Statistical Databases*, edited by J. Domingo-Ferrer and V. Yucel, 239–246. New York: Springer.

Drechsler, J., A. Dundler, S. Bender, S. Rässler, and T. Zwick. 2008. "A New Approach for Disclosure Control in the IAB Establishment Panel – Multiple Imputation for a Better Data Access." *Advances in Statistical Analysis* 92: 439–458. Doi: http://dx.doi.org/10.1007/s10182-008-0090-1.

Drechsler, J. and J.P. Reiter. 2010. "Sampling with Synthesis: a New Approach for Releasing Public Use Census Microdata." *Journal of the American Statistical Association* 105: 1347–1357. Doi: http://dx.doi.org/10.1198/jasa.2010.ap09480.

Duncan, G.T. and D. Lambert. 1989. "The Risk of Disclosure for Microdata." *Journal of Business and Economic Statistics* 7: 207–217. Doi: http://dx.doi.org/10.1080/07350015.1989.10509729.

Harel, O. and J.L. Schafer. 2003. "Multiple Imputation in Two Stages." In Proceedings of Federal Committee on Statistical Methodology 2003 Conference, November 17–19, 2003, Washington DC. Available at: http://fcsm.sites.usa.gov/files/2014/05/2003FCSM_Harel.pdf (accessed August 2017).

Karr, A.F., C.N. Kohnen, A. Oganian, J.P. Reiter, and A.P. Sanil. 2006. "A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality." *The American Statistician* 60: 224–232. Doi: http://dx.doi.org/10.1198/000313006X124640.

Li, F., M. Baccini, F. Mealli, E.Z. Zell, C.E. Frangakis, and D.B. Rubin. 2014. "Multiple Imputation by Ordered Monotone Blocks, with Applications to the Anthrax Vaccine

Adsorbed Trial." *Journal of Computational and Graphical Statistics* 23: 877–892. Doi: http://dx.doi.org/10.1080/10618600.2013.826583.

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

Neyman, J. 1934. "On Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection (with Discussion)." *Journal of the Royal Statistical Society* 97: 558–625.

Raghunathan, T.E. and D.B. Rubin. 2000. "Bayesian Multiple Imputation to Preserve Confidentiality in Public-Use Data Sets." In Proceedings of ISBA 2000 – The Sixth World Meeting of the International Society for Bayesian Analysis, Crete, May 2000.

Raghunathan, T.E., J.M. Lepkowski, J. van Hoewyk, and P. Solenberger. 2001. "A Multivariate Technique for Multiply Imputing Missing Values Using a Series of Regression Models." *Survey Methodology* 27: 85–96.

Raghunathan, T.E., J.P. Reiter, and D.B. Rubin. 2003. "Multiple Imputation for Statistical Disclosure Limitation." *Journal of Official Statistics* 19: 1–16.

Reiter, J.P. 2002. "Satisfying Disclosure Restrictions with Synthetic Datasets." *Journal of Official Statistics* 18: 531–543.

Reiter, J.P. 2003. "Inference for Partially Synthetic, Public Use Microdata Sets." *Survey Methodology* 29: 181–189.

Reiter, J.P. 2005a. "Releasing Multiply Imputed Synthetic Public Use Microdata: An Illustration and Empirical Study." *Journal of the Royal Statistical Society,* Series A 168: 185–205. Doi: http://dx.doi.org/10.1111/j.1467-985X.2004.00343.x.

Reiter, J.P. 2005b. "Significance Tests for Multi-Component Estimands from Multiply Imputed, Synthetic Microdata." *Journal of Statistical Planning and Inference* 131: 365–377. Doi: http://dx.doi.org/10.1016/j.jspi.2004.02.003.

Reiter, J.P. 2009. "Multiple Imputation for Disclosure Limitation: Future Research Challenges." *Journal of Privacy and Confidentiality* 1: 223–233.

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

Reiter, J.P. and R. Mitra. 2009. "Estimating Risks of Identification and Disclosure in Partially Synthetic Data." *Journal of Privacy and Confidentiality* 1: 99–110.

Reiter, J.P. and J. Drechsler. 2010. "Two Stage Multiple Imputation to Protect Confidentiality." *Statistica Sinica* 20: 405–422.

Reiter, J.P., Q. Wang, and B.E. Zhang. 2014. "Bayesian Estimation of Disclosure Risks for Multiply Imputed, Synthetic Data." *Journal of Privacy and Confidentiality* 6: 17–33.

Rubin, D.B. 1978. "Multiple Imputation in Sample Surveys." In Proceedings of the Survey Research Methods Section of the American Statistical Association, 20–34. Alexandria, VA: American Statistical Association, August 14-17, San Diego. Available at: https://ww2.amstat.org/sections/srms/Proceedings/papers/1978_004.pdf (accessed August 2017).

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: 461–468.

Rubin, D.B. 2003. “Nested Multiple Imputation of NMES via Partially Incompatible MCMC.” *Statistica Neerlandica* 57: 3–18. Doi: http://dx.doi.org/10.1111/1467-9574.00217.

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

Shen, Z. 2000. *Nested Multiple Imputation*. Ph.D. thesis, Harvard University, Dept. of Statistics: Cambridge, MA.

Van Buuren, S. and C.G.M. Oudshoorn. 2000. *Multivariate Imputation by Chained Equations: MICE v1.0 user’s manual*. Leiden: TNO. Available at: http://www.stefvanbuuren.nl/publications/mice%20v1.0%20manual%20tno00038%202000.pdf (accessed september 2017).

Woo, M.J., J.P. Reiter, A. Oganian, and A.F. Karr. 2009. “Global Measures of Data Utility for Microdata Masked for Disclosure Limitation.” *Journal of Privacy and Confidentiality* 1: 111–124.

Xie, X. and X.L. Meng. 2014. “Dissecting Multiple Imputation from a Multi-Phase Inference Perspective: What Happens When God’s, Imputer’s and Analyst’s Models are Uncongenial?” *Statistica Sinica*. Preprint. Doi: http://dx.doi.org/10.5705/ss.2014.067.

Received March 2016
Revised August 2017
Accepted September 2017