Multiple Imputation for Statistical Disclosure Limitation

T. E. Raghunathan
Department of Biostatistics and
Institute for Social Research
University of Michigan
Ann Arbor, MI 48106

J.P. Reiter
Institute of Statistics and Decision Sciences
Duke University
Durham, NC 27708-0251

D. B. Rubin
Department of Statistics
Harvard University
Cambridge, MA 02138

Abstract

This article evaluates the use of the multiple imputation framework to protect the confidentiality of respondents' answers in sample surveys. The basic proposal (Rubin, 1993) is to simulate multiple copies of the population from which these respondents have been selected and release a random sample from each of these synthetic populations. Users can analyze the synthetic sample data sets with standard complete-data software for simple random samples, then obtain valid inferences by combining the point and variance estimates using the methods in this article. Both parametric and nonparametric approaches for simulating these synthetic databases are discussed and evaluated. It is shown, using actual and simulated data sets in simple settings, that statistical inferences from these simulated research databases and the actual data sets are similar, at least for the class of analyses considered. Arguably, this class in practice will be large enough to satisfy many users of public-use data. Users with more detailed demands may have to apply for special access to the confidential data.

Keywords: Bayesian approach, Bayesian bootstrap, Combining rules, Confidentiality protection, Sample survey, Synthetic data sets

1 Introduction

The recent explosion in users' demands for microdata, especially when the data collection is paid for with public funds, has increased concerns about confidentiality protection. Such protection is often promised to potential survey respondents by data-collecting agencies. To minimize the chances of disclosures of respondents' data, some agencies alter or limit the variables in the public-release data, or they restrict users' access to data. Several data enclaves have been established where persons

wishing to use the microdata must perform analyses in these locations, either physically using the computers at these locations or through remote access. The latter option requires several passes through "checkpoints" to ensure that the output does not contain any potential identifying information or raw microdata. Even these data enclaves seldom provide unfettered access to data. Typically, a detailed proposal has to be submitted that provides a list of variables, rationale for the analysis and, in some instances, evidence of extramural funding for the proposed analysis. These proposals are reviewed by an appropriate committee and, upon approval, access is granted only for the requested variables. Requests for additional variables may need new, albeit expedited, administrative processes.

These solutions limit the potential utility of publicly collected data. Such data, when released to a broad spectrum of society, inform analyses that can have major positive social, medical and economic implications. The core of the issue, therefore, is how to achieve wider dissemination of data for analytical purposes and at the same time avoid accidental or malicious disclosures of respondents' data.

Our proposal, first described in Rubin (1993), is to release synthetic research databases constructed using multiple imputation. The heart of the proposal is to view all data from nonsampled units as missing data to be filled in by multiple imputation. The proposal is related to, but differs from, some other research efforts on masking data to preserve confidentiality (Cox (1980, 1994), Dalenius and Reiss (1989), Fienberg, Steele and Makov (1996), Fienberg, Makov and Steele (1998)). Valid analysis of masked data requires knowledge of the masking techniques used and special purpose statistical software tuned to those masking techniques (Duncan and Lambert (1989), Fienberg, Makov and Sanil (1997), Fuller (1993), Keller and Bethlehem (1992), and Little (1993)). In contrast, our approach preserves the user's ability to obtain valid statistical inferences using standard statistical methods and software. The approach is also related to the work of Kennickel (1999) and Abowd and Woodcock (2001). These works both use multiple imputation to replace sensitive values for the units originally sampled, which are then released instead.

Another advantage of this methodology is the potential for more efficient inferences by capitalizing on auxiliary variables in the generation of the multiply-imputed databases. For example, suppose that information from administrative sources is available on some or all of the sampled and non-sampled units, and the variables in the administrative sources are highly correlated with the variables in the survey. These correlations can be exploited in the generation of imputations to yield more efficient inferences. Such information is available when a large survey is used as a sampling frame for another survey. For instance, the National Survey of Family Growth and Medical Provider Expenditure Survey both use the National Health Interview Survey as the sampling frame (Ezzati-Rice et al. (1999)). The linked files cannot be released because that can significantly increase the chances of disclosures. However, the linked files can be used to simulate research databases. Other cases involve auxiliary data not available to the public. For example, in Clogg et al. (1999), imputations

were created using a special double-coded data base, thereby increasing the precision of the imputations.

A further advantage of our approach is the ability to simplify the users' analyses of the public-use data by releasing simulated simple random samples from the parent population rather than complex multi-stage samples typically used in practice. The breadth of software available to analyze simple random samples is substantially greater than for complex surveys.

Despite these obvious advantages, there has been relatively little progress until recently on the implemation of this proposal from a decade ago. There are several possible reasons for this. First, when it was originally made, although met with enthusiasm from some, it was met with disbelief by others: could we seriously propose spending time analyzing completely "fake" data? The simple analogies with the analysis of small surveys to learn about huge populations was not entirely convincing to some, even though accurate. Second, a decade ago, the use of multiple imputation, although supported in some quarters, had not yet been generally accepted as a method to address the problem of missing data. Since that time, not only has multiple imputation been much more broadly accepted and available in common software, but the entire battery of simulation tools, including MCMC algorithms, has become relatively common. A third reason for the dearth of work on the proposal may be the complexity of the objects of its application: real world major surveys. Multiple imputation has recently been successfully used in complex multi-stage surveys, as pointed out repeatedly (e.g., Rubin 1987, 1996; Reiter and Raghunathan, 2002) and illustrated in applications to the National Health and Nutritional Examination Survey (Schafer et al. (1996)) and the National Health Interview Survey (Schenker et al. (2002)).

This paper begins to address this sitution in simple, but not unrealistic, settings. In Section 2, procedures for combining inferences from multiply imputed, synthetic databases are provided. In Section 3, it is shown via simulation studies that inferences based on multiply-imputed research databases using complete-data models can be very similar to the corresponding inferences based on actual data sets. In the simulations, both a parametric approach and a nonparametric approach using the Bayesian bootstrap are used to create the synthetic data,. The basic idea in both approaches is to draw several sets of values from the posterior predictive distribution of the observations for the nonsampled units given the observations from the sampled units. Both artificial and genuine data sets are used in the simulations. The artificial data sets are generated under multivariate normal assumptions. The genuine data set is the 1994 Consumer Expenditure Survey.

Section 4 develops theory for inferences from synthetic samples. A cursory glance may suggest that the standard multiple imputation combining formulas (the repeated imputation rules, Rubin (1987)) are adequate. However, the correct Bayesian development for combining inferences from synthetic samples leads to a different formula for constructing variance estimates because of the extra sampling from the synthetic

populations to create synthetic samples, an issue first addressed by Raghunathan and Rubin (2000). Section 5 provides theoretical results concerning validity of the combining rules from the randomization perspective. Finally, Section 6 concludes with a discussion and directions for future research.

2 Synthetic Data

2.1 Creation of Synthetic Samples

Let the actual microdata be a sample of size n from a finite population $\mathcal{P} = (X, Y)$ of size N, with $X = (X_i, i = 1, 2, \ldots, N)$ representing background (including design and administrative records) variables available on all N units in the population and $Y = (Y_i; i = 1, 2, \ldots, N)$ representing survey variables of interest, observed only for sampled units. Without loss of generality, let $Y_{inc} = (Y_i, i = 1, 2, \ldots, n)$ be the observed portion of Y and $Y_{exc} = (Y_i, i = n+1, n+2, \ldots, N)$ be the unobserved portion of Y corresponding to nonsampled units. The observed microdata is $\mathcal{D} = \{X = (X_i, i = 1, 2, \ldots, N), Y_{inc} = (Y_i, i = 1, 2, \ldots, n)\}$. For simplicity, assume there are no item-missing data in the observed data set, though the framework developed in this article can be extended to handle this situation; this is a topic for future work.

The approach developed in this article conceptually involves two steps. First, construct multiple synthetic populations, $\mathcal{P}^{(l)} = \{(X^{(l)}, Y^{(l)}), l = 1, 2, \ldots, M\}$. Second, draw a sample, usually a simple random sample, from each synthetic population; release these samples.

More specifically for the first step, when there are no confidentiality constraints on releasing X, let $X^{(l)} = X$ and simulate $(Y_{exc}^{(l)}; l = 1, 2, \ldots, M)$ as independent draws from the posterior predictive distribution, $Pr(Y_{exc}|X, Y_{inc})$, i.e., conditional on the observed data $\mathcal{D}$ and the model assumptions. If neither X nor Y can be released, the whole population can be generated based on the posterior predictive distribution of "super" or "future" populations, $Pr(X_f, Y_f|\mathcal{D})$, again conditional on the observed data (which includes design and administrative variables) and model assumptions. Thus, which variables are to be synthesized depends upon the specific confidentiality constraints.

Usually, the population size N is too large to make it feasible to release the M synthetic populations. The second step ensures the practicality of this approach. In this step, a simple random sample of size k is taken from each synthetic population, $\mathcal{D}^{(l)} = (x_{i_l}^{(l)}, y_{i_l}^{(l)}, i_l = 1, 2, \ldots, k)$, for $l = 1, 2, \ldots, M$. Then, the corresponding M synthetic samples $\mathcal{D}_{Syn} = \{\mathcal{D}^{(l)}, l = 1, 2, \ldots, M\}$ are released. There may be other versions of what is ultimately released. For instance, a portion of X may not have any confidentiality constraints and so can be released for the entire population. Those variables can be appended to $\mathcal{D}^{(l)}$ for the units not in $\mathcal{D}^{(l)}$. If X is completely confidential and cannot be released at all, one may use X to create the synthetic data sets but release only $(y_{i_l}^{(l)}, i_l = 1, 2, \ldots, k)$. In some circumstances, it may be desirable

to create the synthetic samples using a design other than a simple random sample, but this increases the analysis burden for typical users.

2.2 Analysis of Synthetic Samples

Suppose that an analyst seeks inferences about a scalar population quantity $Q = Q(X, Y)$ that may depend upon both X and Y. Suppose that with a simple random sample the analyst would use a point estimate q and an associated measure of uncertainty v. For example, q could be the maximum likelihood estimate of the model parameter Q, and v could be the inverse of the observed information. Alternatively, q and v could be the posterior mean and variance, respectively, of Q based on the actual sample D. A frequentist could construct an unbiased estimate, q, of Q with v as its sampling variance.

Let $(q^{(l)}, v^{(l)}), l = 1, 2, \ldots, M$ be the values of q and v computed on the M synthetic data sets.Our approach is to consider $(q^{(l)}, v^{(l)}, l = 1, 2, \ldots, M)$ as sufficient summaries of the synthetic databases $\mathcal{D}_{Syn}$, and construct approximations to the posterior density $Pr(Q|\mathcal{D}_{Syn})$. The suggested simplest approximation is the normal distribution with the average of the estimates,

$$\bar{q}_M = \sum_l q^{(l)}/M$$

as the posterior mean of Q, and

$$T_M = (1 + M^{-1})b_M - \bar{v}_M,$$

where $\bar{v}_M = \sum_l v^{(l)}/M$ and $b_M = \sum_l (q^{(l)} - \bar{q}_M)^2/(M-1)$ as the approximate posterior variance.

The minus sign for the average within variance is not a typographical error. It arises formally as shown in Section 3. Intuitively, it occurs because the situation with sythetic samples includes another level of sampling not present in the usual multiple imputation setting: the random sampling of the units that compose the synthetic samples from each multiply-imputed synthetic population. Because of this sampling, the between imputation variance already reflects the usual within imputation variability.

A disadvantage of T_M as the variance estimate is that it can be negative. Though T_M is useful and seems to work well in basic simulations, numerical routines can be used to calculate the integrals involved in the construction of T_M very precisely, as outlined in Section 4.

3 Evaluation of Inferences from Synthetic Data Sets

In this section, we describe three sets of simulations used to compare the properties of inferences from multiply-imputed synthetic data and actual data. The first two simulations involve multivariate normal populations. Two approaches are used to construct synthetic data sets: a parametric approach using a multivariate normal imputation model and a nonparametric approach using the approximate Bayesian bootstrap. In the third simulation, the 1994 Consumer Expenditure Survey data is used as the target population, and synthetic data sets are created using the approximate Bayesian bootstrap.

3.1 Simulation Study 1

We create a population of size $N = 1000$ by drawing 1000 values from a 5-variate, normal distribution with means equal to 0, variances equal to 1, and a common correlation equal to 0.5. Next, 500 independent random samples of size $n = 100$ are drawn from this population. Each such sample is considered to be the observed data $\mathcal{D}$. For each sample, $M = 5$ synthetic populations of size 1000 are created using the procedure described below, and then a simple random sample of size $k = 250$ is drawn from each synthetic population. Thus, for each actual sample, $\mathcal{D}$, of size 100, five synthetic samples of size 250 each are created; these are the released data, $\mathcal{D}_{Syn} = \{\mathcal{D}^{(1)}, \mathcal{D}^{(2)}, \ldots, \mathcal{D}^{(5)}\}$.

The synthetic data sets are created by drawing values from the posterior predictive distribution under the following model assumptions. The data are assumed to follow a multivariate normal distribution with unknown mean vector μ and the unstructured and unknown covariance matrix Σ. A noninformative prior, $\pi(\mu, \Sigma) \propto |\Sigma|^{-1/2}$, is used for the unknown parameters. Here the assumed model matches the true model in terms of the form of the distribution, but the parametric structure of the assumed model is more general. Suppose that $\overline{y}$ is the sample mean and S is the sample covariance matrix for a particular sample. Standard Bayesian calculations lead to the following procedure for creating synthetic data sets:

1. Simulate a Synthetic Population

    - Generate a random variate, W, from a Wishart distribution with $n-1 = 99$ degrees of freedom and the associated matrix $S^{-1}/(n-1)$. Define $\Sigma^* = W^{-1}$.
    - Generate μ^* from a multivariate normal distribution with mean $\overline{y}$ and covariance matrix Σ^*/n.
    - Generate $N = 1000$ independent multivariate normal random vectors with mean μ^* and covariance matrix Σ^*.

2. Repeat this process $M = 5$ times to create five synthetic populations of size 1000 each.

3. Obtain a simple random sample of size $k = 250$ from each of the five synthetic populations.

We assume that the estimand of primary interest is the regression coefficient of the first variable on the other four. We perform ordinary linear regression analyses to obtain q and v based on the actual data, D and each of the corresponding synthetic samples, $D^{(l)}, l = 1, 2, \ldots, 5$. We obtain 95% confidence intervals for the regression coefficient from the synthetic data sets using the normal distribution approximation discussed in Section 2, and from the actual data using the standard t-distribution.

Figure 1 displays the scatter plot of the 500 pairs of estimated regression coefficients from the actual samples and the corresponding synthetic data samples along with a 45 degree line. The sampling distributions of the actual sample and synthetic sample estimates of the regression coefficients are practically the same. Also provided in Figure 1 are the proportions of 95% confidence intervals that contain the true value of the coefficient and the average length of these intervals. For confidence coverage, we use the average of the 500 actual sample estimates of the regression coefficients as the true value. There are no meaningful differences in the coverage properties of the synthetic sample and actual sample intervals. Hence, in this simulation, the repeated sampling properties of the inferences from the actual and synthetic samples are very similar, except that the intervals based on synthetic samples are slightly wider than the intervals based on the actual data.

3.2 Simulation Study 2

The simulation study in Section 3.1 assumes multivariate normality when creating synthetic data sets, which matches perfectly with the true model used to generate the target population. Thus, this comparison of inferences from the actual and synthetic data sets is under the "best" scenario where the imputer's assumed model is also the correct model.

In this study, we create synthetic data sets for the target population described in Section 3.1 without relying on the true model. We use an approximate Bayesian bootstrap where Step 1 in Section 3.1 is replaced with Step 1‘ as follows. Steps 2 and 3 are unchanged.

1‘ Simulate a Synthetic Population

- Draw $n - 1$ uniform random numbers, and sort them in increasing order. Label this ordered sequence as $a_0 = 0, a_1, a_2, \ldots, a_{n-1}, a_n = 1$.
- Draw N uniform random numbers $u_1, u_2, \ldots, u_N$. Select unit j (row j) if $a_{j-1} < u_r \leq a_j$ where $r = 1, 2, \ldots, N$. The resulting $N \times p$ matrix is a synthetic population.

Figure 2 compares the sampling distributions of the 500 estimated regression coefficients from the actual and synthetic data sets. The confidence intervals from the synthetic samples are slightly wider than those constructed with the imputation method in Section 3.1; however, there is no apparent bias as the points are clustered around the 45 degree line.

3.3 Simulation Study 3

The simulation studies in Sections 3.1 and 3.2 use multivariate normality to generate the data and a linear regression model for the complete-data analysis to be applied to the actual data. In realistic situations, the exact model that generated the population is not known, and the model of interest may not be a linear regression.

To evaluate the synthetic data approach in a more realistic setting, we use the 1994 Consumer Expenditure Survey. We consider the complete data on $N = 7,630$ units in the 1994 CES data set to comprise the target population, $\mathcal{P}$. The $p = 28$ variables of interest include expenditure, income and demographic variables. The actual sample,$\mathcal{D}$, is a simple random sample of size $n = 500$ from this population. We create $M = 5$ synthetic populations, $(\mathcal{P}^{(1)}, \mathcal{P}^{(2)}, \ldots, \mathcal{P}^{(5)})$, using the approximate Bayesian bootstrap procedure. Multiply-imputed synthetic samples, $(\mathcal{D}^{(1)}, \mathcal{D}^{(2)}, \ldots, \mathcal{D}^{(5)})$, of $k = 250$ records are randomly selected from each of the five synthetic populations. This entire process is replicated 500 times.

The inferential model of interest is a Tobit model– a censored regression model for a semi-continuous outcome variable (Tobin (1958), Amemiya (1973)). The outcome variable is annual food expenditures away from home, and there are 26 predictors including demographics (e.g., age race, sex, education, region, etc.), family characteristics (e.g., number of earners, family structure, etc.), employment status and log-income before taxes (primary variable of interest). This model was suggested by a staff member at the Bureau of Labor Statistics as a model of substantive interest.

Figure 3 displays the scatter plot of the estimates from the 500 actual samples and the corresponding synthetic data estimates. The sampling distributions of the estimates are very similar to each other. The poor coverage of the actual sample intervals reflects misspecification in the Tobit model. The synthetic sample confidence intervals have slightly better coverage because of inflated variance estimates. This is a reflection of the limited efficiency in the bootstrap approach for creating synthetic data sets.

4 Theoretical Results

This section develops a Bayesian approach for combining inferences about $Q = Q(X, Y)$ from a set of synthetic samples $\mathcal{D}_{Syn}$. This is achieved by constructing an approximate posterior distribution of Q given $\mathcal{D}_{Syn}$ in analogy with the theory

of multiple imputation for missing data. However, the standard multiple imputation combining rules are not appropriate because of the subsampling of $\mathcal{P}^{(l)}$ to obtain $\mathcal{D}^{(l)}$.

The conceptual framework for creating the synthetic population outlined in Section 2.1 suggests the following natural decomposition,

$$Pr(Q|\mathcal{D}_{Syn}) = \int \left[\int Pr(Q|\mathcal{D}, \mathcal{P}_{Syn}, \mathcal{D}_{Syn}) Pr(\mathcal{D}|\mathcal{P}_{Syn}, \mathcal{D}_{Syn}) d\mathcal{D} \right] Pr(\mathcal{P}_{Syn}|\mathcal{D}_{Syn}) d\mathcal{P}_{Syn},$$

where $\mathcal{P}_{Syn} = (\mathcal{P}^{(l)}, l = 1, 2, \ldots, M)$ is the collection of synthetic populations. Clearly, $\mathcal{P}_{Syn}$ and $\mathcal{D}_{Syn}$ are irrelevant after conditioning on $\mathcal{D}$ because both are random functions of $\mathcal{D}$. Similarly, $\mathcal{D}_{Syn}$ is irrelevant after conditioning on $\mathcal{P}_{Syn}$. Thus, $Pr(Q|\mathcal{D}, \mathcal{P}_{Syn}, \mathcal{D}_{Syn}) = Pr(Q|\mathcal{D})$ and $Pr(\mathcal{D}|\mathcal{P}_{Syn}, \mathcal{D}_{Syn}) = Pr(\mathcal{D}|\mathcal{P}_{Syn})$. Thus, the expression for $Pr(Q|\mathcal{D}_{Syn})$ simplifies to

$$\begin{aligned} Pr(Q|\mathcal{D}_{Syn}) &= \int \left[\int Pr(Q|\mathcal{D}) Pr(\mathcal{D}|\mathcal{P}_{Syn}) d\mathcal{D} \right] Pr(\mathcal{P}_{Syn}|\mathcal{D}_{Syn}) d\mathcal{P}_{Syn}, \\ &= \int Pr(Q|\mathcal{P}_{Syn}) Pr(\mathcal{P}_{Syn}|\mathcal{D}_{Syn}) d\mathcal{P}_{Syn} \end{aligned} \tag{1}$$

Throughout this article, we assume that the sample sizes are large enough to permit normal approximations for these posterior distributions. Thus, we require only the first two moments for each distribution. To derive these conditional moments, we use standard large sample Bayesian arguments. For example, to derive $Pr(Q|\mathcal{P}_{Syn})$, we treat the first two moments of Q given $\mathcal{P}_{Syn}$ as unknown and use $\mathcal{P}_{Syn}$ as the data. Similarly, for the first two moments of $Pr(\mathcal{P}_{Syn}|\mathcal{D}_{Syn})$, we treat the first two moments based on $\mathcal{P}_{Syn}$ as unknown and use $\mathcal{D}_{Syn}$ as the data. Diffuse priors are assumed for all parameters.

4.1 $Pr(Q|\mathcal{P}_{Syn})$

For $l = 1, 2, \ldots, M$, let $Q^{(l)} = Q(X^{(l)}, Y^{(l)})$ denote the computed value of the population quantity $Q(X, Y)$ based on $\mathcal{P}^{(l)}$. The nonsampled units can be treated as missing data and $\mathcal{D}$ as the observed data; hence, the standard multiple imputation framework (Rubin, 1987) can be applied. Specifically, in equation (3.1.3) of Rubin (1987, p. 76) set the average within-imputation variance $\bar{U}_M = 0$. This is true since each $U_{*l} = 0$ because each completed data set is an entire population. Then, based on equations (3.1.5) and (3.1.6) from Rubin (1987, pp. 76-77),

$$Q|\mathcal{P}_{Syn} \sim t_{M-1}(\bar{Q}_M, (1 + M^{-1})B_M), \tag{2}$$

where $\bar{Q}_M = \sum_l Q^{(l)}/M$, $B_M = \sum_l (Q^{(l)} - \bar{Q}_M)^2/(M-1)$ and $t_\nu(\mu, \sigma^2)$ denotes a t distribution with ν degrees of freedom, location μ and scale σ^2.

4.2 $Pr(\bar{Q}_M, B_M|\mathcal{D}_{Syn})$

In practice, only the set of M synthetic samples $\mathcal{D}_{Syn}$, not the synthetic populations $\mathcal{P}_{Syn}$, will be made available. The next step, therefore, is to derive the conditional distribution, $Pr(\bar{Q}_M, B_M|\mathcal{D}_{Syn})$, and then construct

$$Pr(Q|\mathcal{D}_{Syn}) = \int Pr(Q|\bar{Q}_M, B_M) Pr(\bar{Q}_M, B_M|\mathcal{D}_{Syn}) d\bar{Q}_M dB_M.$$

We derive $Pr(\bar{Q}_M, B_M|\mathcal{D}_{Syn})$ by treating the $\{q^{(l)}, v^{(l)}, l = 1, 2, \ldots, M\}$ as sufficient summaries of the synthetic data bases $\mathcal{D}_{Syn}$ and $\bar{Q}_M$ and B_M as parameters. To obtain the sampling distribution for $\{q^{(l)}, v^{(l)}, l = 1, 2, \ldots, M\}$, we assume that estimates from each synthetic sample are valid in the following sense:

(i) For each l, the estimate $q^{(l)}$ is unbiased for Q_l and asymptotically normal, with respect to repeated sampling from the synthetic population $\mathcal{P}^{(l)}$, with sampling variance $V^{(l)} = V(X^{(l)}, Y^{(l)})$.

(ii) The sampling variance estimate $v^{(l)}$ is unbiased for $V^{(l)}$, and the sampling variability in $v^{(l)}$ is negligible. That is, $v^{(l)}|\mathcal{P}^{(l)} \approx V^{(l)}$. Thus, $v^{(l)}$ and $V^{(l)}$ are interchangeable.

We also make the simplifying assumption that the variation in $V^{(l)}$ across the M synthetic populations is negligible; that is, $V^{(l)} \approx V$. Then, by (ii), $\bar{v}_M \approx V$.

Since the samples from the synthetic populations are independently drawn, the $q^{(l)}|\mathcal{P}^{(l)}, \bar{v}_M \sim ind\ N(Q^{(l)}, \bar{v}_M)$.

Using the standard Bayesian arguments based on these sampling distributions, it follows that

$$Q^{(l)}|q^{(l)}, \bar{v}_M \sim ind\, N(q^{(l)}, \bar{v}_M),$$

and the posterior distribution of $\bar{Q}_M$ is,

$$\bar{Q}_M|\mathcal{D}_{Syn} \sim N(\bar{q}_M, \bar{v}_M/M), \tag{3}$$

where $\bar{q}_M = \sum_l q^{(l)}/M$. Using the standard one-way analysis of variance setup (see for example, Box and Tiao (1973)), the posterior distribution of B_M is,

$$\frac{\sum_l (q^{(l)} - \bar{q}_M)^2}{B_M + \bar{v}_M}|\mathcal{D}_{Syn} \sim \chi^2_{M-1}. \tag{4}$$

4.3 Approximation of $Pr(Q|\mathcal{D}_{Syn})$

To obtain the posterior distribution of Q given $\mathcal{D}_{Syn}$, we should integrate the t-posterior distribution in equation (2) with respect to the posterior distributions of $\bar{Q}_M$ and B_M in equations (3) and (4). Although this integration can be carried out numerically or using analytical approximations related to those used in Barnard

and Rubin (1999), a basic approximation suitable for large M is useful in practice. Specifically, we approximate the posterior distribution of Q given $\mathcal{D}_{Syn}$ by a normal distribution with mean $E(Q|\mathcal{D}_{Syn})$ and variance $Var(Q|\mathcal{D}_{Syn})$.

Using the results in Sections 4.1 and 4.2,

$$E(Q|\mathcal{D}_{Syn}) = E[E(Q|\bar{Q}_M)|\mathcal{D}_{Syn}] = E(\bar{Q}_M|\mathcal{D}_{Syn}) = \bar{q}_M. \tag{5}$$

Similarly,

$$\begin{aligned} Var(Q|\mathcal{D}_{Syn}) &= E[Var(Q|\mathcal{P}_{Syn})|\mathcal{D}_{Syn}] + Var[E(Q|\mathcal{P}_{Syn})|\mathcal{D}_{Syn}] \\ &\approx (1 + M^{-1})E(B_M|\mathcal{D}_{Syn}) + \bar{v}_M/M \end{aligned} \tag{6}$$

Based on the posterior distribution of B_M given in Equation (4),

$$E(B_M|\mathcal{D}_{Syn}) = \frac{\int_0^\infty B_M(B_M + \bar{v}_M)^{-\frac{M-1}{2}-1} \exp\left(-\frac{(M-1)b_M}{2(B_M+\bar{v}_M)}\right) dB_M}{\int_0^\infty (B_M + \bar{v}_M)^{-\frac{M-1}{2}-1} \exp\left(-\frac{(M-1)b_M}{2(B_M+\bar{v}_M)}\right) dB_M}.$$

After substituting, $u = (M-1)b_M/(B_M + \bar{v}_M)$ in the above integral and simplifying, we obtain

$$\begin{aligned} E(B_M|\mathcal{D}_{Syn}) &= \frac{(M-1)\Gamma_{M-3}(r_M)}{2\Gamma_{M-1}(r_M)} b_M - \bar{v}_M \\ &\approx b_M - \bar{v}_M, \end{aligned} \tag{7}$$

at least for large M, where $r_M = (M-1)b_M/\bar{v}_M$ and $\Gamma_\nu(x) = \int_0^x \exp(-x)x^{\nu/2-1}dx$. The substitution of (7) into equation (6) yields

$$T_M = (1 + M^{-1})b_M - \bar{v}_M$$

for the approximate posterior variance of Q given $\mathcal{D}_{Syn}$.

5 Randomization Validity

The inferential procedures in Section 4 are developed from a Bayesian perspective. This section provides theoretical results concerning the conditions for randomization validity of the procedures developed in Section 4.

We assume the conditions (i) and (ii) used in the previous section. We also assume two conditions similar to those imposed by Rubin (1987) for randomization validity of multiple imputation inferences. Condition (iii) requires the procedures that the analyst would have used if the actual data $\mathcal{D}$ were available to be valid from the randomization perspective. Specifically, $\widehat{Q}_D$ is an unbiased estimate of $Q(X, Y)$ with respect to repeated sampling from the fixed population $\mathcal{P} = (X, Y)$. And, the variance estimate $\widehat{U}_D$ is an unbiased estimate of the sampling variance of $\widehat{Q}_D$, which we label as U, with negligible sampling variability relative to the variability of $\widehat{Q}_D$. In the notation of Rubin (1987), we require

(*iii*) $\widehat{Q}_D|X,Y \sim N(Q,U)$, and $\widehat{U}_D|X,Y \sim (U, << U)$.
Equivalently, $\widehat{Q}_D|X,Y \sim N(Q,\widehat{U}_D)$.

Condition (*iv*) involves the randomization validity of the inferences based on the synthetic populations $\mathcal{P}_{Syn}$. Since this is a particular case of the standard multiple imputation framework where the nonsampled units are treated as missing data, the conditions for proper multiple imputations (Rubin 1987, pp. 118-119) are required. Effectively, the synthetic data imputation procedures are proper when:

(*iv*) $Q^{(l)}|\mathcal{D} \sim N(\widehat{Q}_D, \widehat{U}_D)$. That is, the computed value of Q based on the synthetic population $\mathcal{P}^{(l)}$ is an unbiased estimate of $\widehat{Q}_D$.

Under conditions (*i*)-(*iv*), it can be shown that (1) $\bar{q}_M$ is an unbiased estimate of $Q(X,Y)$, and (2) T_M is its asymptotically unbiased variance estimate.

5.1 Unbiasedness of $\bar{q}_M$

For the first assertion, note that

$$E(\bar{q}_M|\mathcal{P}) = E[E\{E(\bar{q}_M|\mathcal{P}_{Syn})|\mathcal{D}\}|\mathcal{P}].$$

This expectation is determined by the assumed conditions. (*i*) implies $E(\bar{q}_M|\mathcal{P}_{Syn}) = \bar{Q}_M$. Then, (*iv*) implies $E(\bar{Q}_M|\mathcal{D}) = \widehat{Q}_D$. Finally, (*iii*) implies $E(\widehat{Q}_D|\mathcal{P}) = Q$. Hence, $\bar{q}_M$ is unbiased for Q.

5.2 Unbiasedness of T_M

The second assertion involves determining the sampling variance of $\bar{q}_M$ with respect to repeated sampling from $\mathcal{P}$ and showing T_M is an unbiased estimator of this variance. The derivation of the sampling variance involves repeated use of the standard decomposition of the marginal variance as the sum of the variance of the conditional mean and mean of the conditional variances. First, note that

$$Var(\bar{q}_M|\mathcal{P}) = E[Var(\bar{q}_M|\mathcal{P}_{Syn})|\mathcal{P}] + Var[E(\bar{q}_M|\mathcal{P}_S)|\mathcal{P}]. \tag{8}$$

For all l, define $E(V^{(l)}|\mathcal{P}) = V_o$, where $V^{(l)}$ is as in (*i*). Since the samples are drawn independently from each synthetic population, it follows from (*i*) that $Var(\bar{q}_M|\mathcal{P}_{Syn}) = \sum_l v^{(l)}/M$. Taking the expectation with respect to $Pr(\mathcal{P}_{Syn}|\mathcal{P})$, the first term in equation (8) equals V_o/M.

Since $E(\bar{q}_M|\mathcal{P}_{Syn}) = \bar{Q}_M$, the second term in equation (8) equals $Var(\bar{Q}_M|\mathcal{P})$. Using the usual variance decomposition,

$$Var(\bar{Q}_M|\mathcal{P}) = Var[E(\bar{Q}_M|\mathcal{D})|\mathcal{P}] + E[Var(\bar{Q}_M|\mathcal{D})|\mathcal{P}]. \tag{9}$$

The pieces in this variance are determined from the assumed conditions. (*iv*) implies that $E(\bar{Q}_M|\mathcal{D}) = \widehat{Q}_D$. Hence, from (*iii*), the first term in (9) equals U. (*iv*) implies

that $Var(\bar{Q}_M|\mathcal{D}) = \hat{U}_D/M$. Hence, from (iii), the second term in (9) equals U/M. Putting all the pieces of (8) together, the sampling variance of $\bar{q}_M$ equals

$$Var(\bar{q}_M|\mathcal{P}) = (1 + M^{-1})U + V_o/M. \tag{10}$$

The variance estimate T_M is valid, if $E(T_M|\mathcal{P}) = Var(\bar{q}_M|\mathcal{P})$. Using standard one-way analysis of variance calculations,

$$E(b_M|\mathcal{P}) = E[E(b_M|\mathcal{P}_{Syn})|\mathcal{P}] = E(\bar{V}_M + B_M|\mathcal{P}) = V_o + E(B_M|\mathcal{P})$$

where $\bar{V}_M = \sum_l V_l/M$. Since $E(B_M|\mathcal{P}) = E[E(B_M|\mathcal{D})|\mathcal{P}]$ and $E(B_M|\mathcal{D}) = \hat{U}_D$ from (iv), it follows from (iii) that $E(B_M|\mathcal{P}) = U$. Thus, b_M is an unbiased estimate of $V_o + U$. Under (ii), $\bar{v}_M$ and $\bar{V}_M$ are essentially equivalent, so that $E(\bar{v}_m|\mathcal{P}) \approx E(\bar{V}_m|\mathcal{P}) = V_o$. That is, $\bar{v}_M \sim (V_o, << V_o)$. Thus,

$$\begin{aligned} E(T_M|\mathcal{P}) &= (1 + M^{-1})E(b_M|\mathcal{P}) - M^{-1}E(\bar{v}_M|\mathcal{P}) \\ &= (1 + M^{-1})U + V_o/M \\ &= Var(\bar{q}_M|\mathcal{P}). \end{aligned}$$

All the distributions in conditions (i)-(iv) are normal for location quantities, and the distributions for scale quantities have lower order variability than the variability for corresponding location quantities. Therefore, implied convolutions in all the expectation and variance calculations involve normal distributions. Thus, asymptotically, the sampling distribution of $\bar{q}_M$ is normal with mean Q and variance given in equation (10). Since T_M is an unbaised estimate of the actual sampling variance,

$$T_M^{-1/2}(\bar{q}_M - Q) \sim N(0, 1).$$

Thus, the large sample frequentist and Bayesian confidence intervals are identical. Intervals constructed in accordance with conditions (i)-(iv) have valid confidence coverage in large samples.

6 Discussion

In this article, we have evaluated multiple imputation as a framework for creating synthetic databases that can be shared without compromising the confidentiality of responses. Each synthetic data set is a plausible reflection of the target population based on the collected data. We have evaluated a fully parametric approach and a nonparametric approach using the approximate Bayesian bootstrap for creating plausible populations. In these simulation studies, using artificial and genuine data sets, the sampling properties of inferences from synthetic databases and the actual sample data sets are very similar. Further evidence of the effectiveness of the approach is demonstrated by Reiter (to appear), who conducts simulation studies with complex survey designs involving stratification, clustering and unequal probabilities of

selection. For all these designs, the approach discussed in this article obtains approximately valid point and interval estimates of population means, as well as regression coefficients.

The quality of inferences from the synthetic data clearly depend upon the imputation models. As shown in these simulations, it is possible to obtain valid inferences from synthetic data from relationships accurately modeled in the imputation models. On the other hand, inferences derived from inaccurately modeled relationships may not be valid. This is illustrated in Reiter (2002), who shows how small biases in imputation models can lead to synthetic point estimates with small mean squared error but less than nominal confidence coverage.

A related issue is that some synthetic data sets may produce extreme estimates. This possibility can be mitigated by constraining the imputations. For example, agencies could constrain the imputations to match certain one-way and two-way margins in the observed data. Abowd and Woodcock (2001) try to prevent extreme data sets by constraining draws of parameters from posteriors to lie within three standard deviations of the observed data posterior means. Another approach is for agencies to examine the inferences for several key estimands before releasing the M synthetic data sets. If the $\bar{q}_M$ for several widely-used means and regression coefficients are far away from their corresponding q_{obs}, the agency can redraw synthetic data. Agencies also could check the synthetic data sets to verify protection of confidentiality, just as they do for all other disclosure methods (e.g., swapping, adding noise, cell aggregation).

The synthetic data framework has many advantages. Agencies can borrow strength from other data sources when generating imputations and even release synthetic copies of such combined databases. Design information and nonsampling and measurement error models can be utilized when creating synthetic databases. Regional, county and community level information can be released, facilitating small area or community level analysis. Importantly, since each synthetic database can be a simple random sample from the target population, users need only apply simple, unweighted complete-data analysis techniques to each synthetic data set.

A disadvantage of this approach is the need to store and process each of the multiple synthetic data sets. There now exist several macros in SAS and STATA that allow simple processing of multiply-imputed data sets. The new versions of SUDAAN (8.2 avilable now only as beta release), SAS (version 8.1) and WESVAR (to be released soon) also incorporate multiple imputation analysis. However, these packages as yet do not incorporate the modified combining rules presented here. In general, the number of completed data sets should be larger in the synthetic data context than in the standard missing data context because the fractions of missing information can be larger.

Of the two data generation approaches discussed in this article, the parametric model-based approach should protect confidentiality more effectively than the approximate Bayesian bootstrap. In the latter approach, each synthetic database contains several repeats of the observed records, whereas in the former approach the imputa-

tions are all from a smooth distribution and do not contain any fully observed records. The parametric model, however, is far more susceptible to model misspecification. A compromise is to use a semiparametric approach, and work along these lines is in progress.

References

Abowd, J. M. and Woodcock, S. D. (2001), 'Disclosure limitation in longitudinal linked data'. In P. Doyle, J. Lane, L. Zayatz, and J. Theeuwes, eds., *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, North-Holland, Amsterdam, 215–277.

Amemiya, T. (1973), 'Regression analysis when the dependent variable is truncated normal', *Journal of Econometrics*, 24, 3-61.

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

Box, G. E. P., and Tiao, G. C. (1973), **Bayesian Inference in Statistical Analysis**, Addison-Wesley, Reading, Massachusetts.

Clogg, C. C., Rubin, D. B., Schenker, N., Schultz, B., and Weidman, L. (1991), 'Multiple Imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression', *Journal of the American Statistical Association*, 86, 68–78.

Cox, L. (1980), 'Suppression methodology and statistical disclosure control', *Journal of the American Statistical Association*, 75, 377–385.

Cox, L. (1994), 'Matrix masking methods for disclosure limitation in microdata', *Survey Methodology*, 20, 165–169.

Dalenius, T, and Reiss, S. P. (1982), 'Data swapping: A technique for disclosure control', *Journal of Statistical Planning and Inference*, 6, 73–85.

Duncan, G. T. and Lambert, D. (1989), 'Disclosure-limited data dissemination', *Journal of the American Statistical Association*, 81, 10–28.

Ezzati-Rice, T. M., Cohen, S. B. , Khare, M., Moriarity, C. L. (1999), 'Using the National Health Interview Survey as a Sampling Frame for Other Health-Related Surveys', *Proccedings of the Section on Survey Research Methods.*, Alexandria, VA.: American Statistical Association, 728-737.

Fienberg, S. E., Makov, U. E., and Sanil, A. P. (1997), 'A Bayesian approach to data disclosure: Optimal intruder behavior for continuous data', *Journal of Official Statistics*, 13, 75–89.

Fienberg, S. E., Steele, R. J., and Makov, U. E. (1996), 'Statistical notions of data disclosure avoidance and their relationship to traditional statistical methodology: Data swapping and log-linear models', *Proceedings of Bureau of Census 1996 Annual Research Conference.*. US Bureau of the Census, Washington, DC, 87–105.

Fienberg, S.E., Makov, U.E., and Steele, R.J. (1998), 'Disclosure limitation using perturbation and related methods for categorical data (with discussions),' *Journal of Official Statistics*, 14, 485-511.

Fuller, W. (1993), 'Masking procedures for microdata disclosure limitation', *Journal of Official Statistics*, 9, 383-406.

Keller, W. J. and Bethlehem, J. G. (1992), 'Disclosure protection of microdata: Problems and solutions', *Statistical Neerlandica*, 46, 5–19.

Kennickell, A. B. (1999), 'Multiple imputation and disclosure protection: The case of the 1995 Survey of Consumer Finances'. In W. Alvey and B. Jamerson, eds., *Record Linkage Techniques, 1997*, 248–267. National Academy Press, Washington, D.C.

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

Raghunathan, T. E. and Rubin, D. B. (2000) 'Bayesian multiple imputation to preserve confidentiality in public-use data sets.' *ISBA 2000 – The Sixth World Meeting of the International Society for Bayesian Analysis.* Presentation.

Reiter, J. P. (to appear), 'Satisfying Disclosure Restrictions With Synthetic Data Sets', *Journal of Official Statistics.*

Reiter, J. P. (2002) 'Protecting Confidentiality by Releasing Synthetic Microdata' Technical Report, Institute of Statistics and Decision Sciences, Duke University.

Reiter, J. P. and Raghunathan, T. E. (2002) 'Multiple Imputation for Missing Data in Surveys with Complex Designs' Technical Report, Institute of Statistics and Decision Sciences, Duke University.

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

Rubin, D. B. (1993), 'Satisfying confidentiality constraints through use of synthetic multiply-imputed microdata', *Journal of Official Statistics*, 9, 461–468.

Schafer, J. L., Ezaati-Rice, Y. M., Johnson, W., Khare, M., Little, R. J. A., and Rubin, D. B. (1996), 'The NHANES III multiple imputation project', Technical report, The National Center for Health Statistics.

Schenker, N., Raghunathan, T. E., Pei-Lu Chiu, Makuc, D.M. and Guangyu Zhang (2002), 'Multiple imputation of missing income and earnings items in the National Health Interview Survey', Technical Report, Institute for Social Research, University of Michigan.

Tobin, J. (1958), 'Estimation of relationships for limited dependent variables', Econometrika, 26, 24-36.