Simultaneous Use of Multiple Imputation for Missing Data and Disclosure Limitation

Jerome P. Reiter*

Abstract

Several statistical agencies use, or are considering the use of, multiple imputation to limit the risk of disclosing respondents' identities or sensitive attributes in public use data files. For example, agencies can release partially synthetic datasets, comprising the units originally surveyed with some collected values, such as sensitive values at high risk of disclosure or values of key identifiers, replaced with multiple imputations. This article presents an approach for generating multiply-imputed, partially synthetic datasets that simultaneously handles disclosure limitation and missing data. The basic idea is to fill in the missing data first to generate m completed datasets, then replace sensitive or identifying values in each completed dataset with r imputed values. This article also develops methods for obtaining valid inferences from such multiply-imputed datasets. New rules for combining the multiple point and variance estimates are needed because the double duty of multiple imputation introduces two sources of variability into point estimates, which existing methods for obtaining inferences from multiply-imputed datasets do not measure accurately. A reference t-distribution appropriate for inferences when m and r are moderate is derived using moment matching and Taylor series approximations.

KEY WORDS: Confidentiality, Missing data, Public use data, Survey, Synthetic data

*Jerome P. Reiter, Institute of Statistics and Decision Sciences, Duke University, Box 90251, Durham, NC 27708-0251.

1 Introduction

Many statistical agencies disseminate microdata, i.e. data on individual units, in public use files. These agencies strive to release files that are (i) safe from attacks by ill-intentioned data users seeking to learn respondents' identities or attributes, (ii) informative for a wide range of statistical analyses, and (iii) easy for users to analyze with standard statistical methods. Doing this well is a difficult task. The proliferation of publicly available databases, and improvements in record linkage technologies, have made disclosures a serious threat, to the point where most statistical agencies alter microdata before release. For example, agencies globally recode variables, such as releasing ages in five year intervals or top-coding incomes above $100,000 as "$100,000 or more" (Willenborg and de Waal, 2001); they swap data values for randomly selected units (Dalenius and Reiss, 1982); or, they add random noise to continuous data values (Fuller, 1993). Inevitably, these strategies reduce the utility of the released data, making some analyses impossible and distorting the results of others. They also complicate analyses for users. To analyze properly perturbed data, users should apply the likelihood-based methods described by Little (1993) or the measurement error models described by Fuller (1993). These are difficult to use for non-standard estimands and may require analysts to learn new statistical methods and specialized software programs.

An alternative approach to disseminating public use data was suggested by Rubin (1993): release multiply-imputed, synthetic datasets. Specifically, he proposed that agencies (i) randomly and independently sample units from the sampling frame to comprise each synthetic data set, (ii) impute unknown data values for units in the synthetic samples using models fit with the original survey data, and (iii) release multiple versions of these datasets to the public. These are called *fully synthetic* data sets. Releasing fully synthetic data can protect confidentiality, since identification of units and their sensitive data is nearly impossible when the values in the released data are not actual, collected values. Furthermore, with appropriate synthetic data generation and the inferential methods developed by Raghunathan *et al.* (2003) and Reiter (2004b), it can allow data users to make valid inferences for a variety of estimands using standard, complete-data statistical methods and software. Other attractive features of fully synthetic data are described by Rubin (1993), Little (1993), Fienberg *et al.* (1998), Raghunathan *et al.* (2003), and Reiter (2002, 2004a).

No statistical agencies have released fully synthetic datasets as of this writing, but some have adopted a variant of the multiple imputation approach suggested by Little (1993): release datasets comprising the units originally surveyed with some collected values, such as sensitive values at high risk of disclosure or values of key identifiers, replaced with multiple imputations. These are called *partially synthetic* datasets. For example, the U.S. Federal Reserve Board protects data in the U.S. Survey of Consumer Finances by replacing monetary values at high disclosure risk with multiple imputations, releasing a mixture of these imputed values and the unreplaced, collected values (Kennickell, 1997). The U.S. Bureau of the Census and Abowd and Woodcock (2001) protect data in longitudinal, linked data sets by replacing all values of some sensitive variables with multiple imputations and leaving other variables at their actual values. Liu and Little (2002) present a general algorithm, named SMIKe, for simulating multiple values of key identifiers for selected units.

All these partially synthetic approaches are appealing because they promise to maintain the primary benefits of fully synthetic data–protecting confidentiality while allowing users to make inferences without learning complicated statistical methods or software–with decreased sensitivity to the specification of imputation models (Reiter, 2003). Valid inferences from partially synthetic datasets can be obtained using the methods developed by Reiter (2003, 2004b), whose rules for combining point and variance estimates again differ from those of Rubin (1987) and also from those of Raghunathan *et al.* (2003).

The existing theory and methods for partially synthetic data do not deal explicitly with an important practical complication: in most large surveys, there are units that fail to respond to some or all items of the survey. This article presents a multiple imputation approach that handles simultaneously missing data and disclosure limitation. The approach involves two steps. First, the agency uses multiple imputation to fill in the missing data, generating m multiply-imputed datasets. Second, the agency replaces the values at risk of disclosure in each imputed dataset with r multiple imputations, ultimately releasing $m * r$ multiply-imputed datasets. This double-duty of multiple imputation requires new methods for obtaining valid inferences from the multiply-imputed datasets, which are derived here.

The paper is organized as follows. Section 2 reviews multiple imputation for missing and partially synthetic data. Section 3 presents the new methods for generating partially synthetic data and obtaining valid inferences when some survey data are missing. Section 4 shows a derivation of these methods from a Bayesian perspective, and it discusses conditions under which the resulting inferences should be valid from a frequentist perspective. Section 5 concludes with a discussion of the challenges to implementing this multiple imputation approach on genuine data, with an aim towards stimulating future research.

2 Review of multiple imputation inferences

To describe multiple imputation, 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 survey, and $I_j = 0$ otherwise, where $j = 1, 2, \ldots, N$. Let $I = (I_1, \ldots, I_N)$. Let R_j be a $p \times 1$ vector of response indicators, where $R_{jk} = 1$ if the response for unit j to survey item k is recorded, and $R_{jk} = 0$ otherwise. Let $R = (R_1, \ldots, R_N)$. Let Y be the $N \times p$ matrix of survey data for all units in the population. Let $Y_{inc} = (Y_{obs}, Y_{mis})$ be the $n \times p$ matrix of survey data for the n 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 X 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 approximately for all population units, for example from census records or the sampling frame(s). Finally, we write the observed data as $D = (X, Y_{obs}, I, R)$.

2.1 Multiple imputation for missing data

The agency fills in values for Y_{mis} with draws from the Bayesian posterior predictive distribution of $(Y_{mis} \mid D)$, or approximations of that distribution such as those of Raghunathan *et al.* (2001). These draws are repeated independently $l = 1, \ldots, m$ times to obtain m completed data sets, $D^{(l)} = (D, Y_{mis}^{(l)})$. Multiple rather than single imputations are used so that analysts can estimate the variability due to imputing missing data.

In each imputed data set $D^{(l)}$, the analyst estimates the population quantity of interest, Q, using some estimator q, and estimates 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 (X, 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:

$$
\begin{aligned}
\bar{q}_m &= \sum_{l=1}^{m} q^{(l)}/m & (1)\\
b_m &= \sum_{l=1}^{m} (q^{(l)} - \bar{q}_m)^2/(m-1) & (2)\\
\bar{u}_m &= \sum_{l=1}^{m} u^{(l)}/m. & (3)
\end{aligned}
$$

The analyst then can use $\bar{q}_m$ to estimate Q and $T_m = (1 + 1/m)b_m + \bar{u}_m$ to estimate the variance of $\bar{q}_m$. Inferences can be based on t-distributions with degrees of freedom $\nu_m = (m-1)(1 + \bar{u}_m/((1+1/m)b_m))^2$.

2.2 Multiple imputation for partially synthetic data when $Y_{inc} = Y_{obs}$

Assuming no missing data, i.e., $Y_{inc} = Y_{obs}$, the agency constructs partially synthetic datasets by replacing selected values from the observed data with imputations. Let $Z_j = 1$ if unit j is selected to have any of its observed data replaced with synthetic values, and let $Z_j = 0$ for those units with all data left unchanged. Let $Z = (Z_1, \ldots, Z_n)$. Let $Y_{rep,i}$ be all the imputed (replaced) values in the ith synthetic data set, and let Y_{nrep} be all unchanged (unreplaced) values of Y_{obs}. The $Y_{rep,i}$ are assumed to be generated from the posterior predictive distribution of $(Y_{rep,i} \mid D, Z)$, or a close approximation of it. The values in Y_{nrep} are the same in all synthetic data sets. Each synthetic data set, d_i, then comprises $(X, Y_{rep,i}, Y_{nrep}, I, Z)$. Imputations are made independently $i = 1, \ldots, r$ times to yield r different partially synthetic data sets, which are released to the public. Once again, multiple imputations enable analysts to account for variability due to imputation.

The values in Z can and frequently will depend on the values in D. For example, the agency may simulate sensitive variables or identifiers only for units in the sample with rare combinations of identifiers; or, the imputer may replace only incomes above \$100,000 with imputed values. To avoid bias, the imputations should be drawn from the posterior predictive distribution of Y for those units with $Z_j = 1$. Reiter (2003) illustrates the problems that can arise when imputations are not conditional on Z.

Inferences from partially synthetic datasets are based on quantities defined in Equations (1)-(3). As shown by Reiter (2003), under certain conditions the analyst can use $\bar{q}_r$ to estimate Q and $T_p = b_r/r + \bar{u}_r$ to estimate the variance of $\bar{q}_r$. Inferences for scalar Q can be based on t-distributions with degrees of freedom $\nu_p = (r-1)(1 + \bar{u}_r/(b_r/r))^2$.

3 Partially synthetic data when $Y_{inc} \neq Y_{obs}$

When some data are missing, it seems logical to impute the missing and partially synthetic data simultaneously. However, imputing Y_{mis} and Y_{rep} from the same posterior predictive distribution can result in improper imputations. For an illustrative example, suppose univariate data from a normal distribution have some values missing completely at random (Rubin, 1976). Further, suppose the agency seeks to replace all values larger than some threshhold with imputations. The imputations for missing data can be based on a normal distribution fit using all of Y_{obs}. However, the imputations for replacements must be based on a posterior distribution that conditions on values being larger than the threshhold. Drawing Y_{mis} and Y_{rep} from the same distribution will result in biased inferences.

Imputing the Y_{mis} and Y_{rep} separately generates two sources of variability, in addition to the sampling variability in D, that the user must account for to obtain valid inferences. Neither T_m nor T_p correctly estimate the total variation introduced by the dual use of multiple imputation. The bias of each can be illustrated with two simple examples. Suppose only one value needs replacement, but there are hundreds of missing values to be imputed. Intuitively, the variance of the point estimator of Q should be well approximated by T_m, and T_p should underestimate the variance, as it is missing a b_m. On the other hand, suppose only one value is missing, but there are hundreds of values to be replaced. The variance should be well approximated by T_p, and T_m should overestimate the variance, as it includes an extra b_m.

To allow users to estimate the total variability correctly, agencies can employ a three-step procedure for generating imputations. First, the agency fills in Y_{mis} with draws from the posterior distribution for $(Y_{mis} \mid D)$, resulting in m completed datasets, $D^{(1)}, \ldots, D^{(m)}$. Then, in each $D^{(l)}$, the agency selects the units whose values are to be replaced, i.e. whose $Z_j^{(l)} = 1$. In many cases, the agency will impute values for the same units in all $D^{(l)}$ to avoid releasing any genuine, sensitive values for the selected units. We assume this is the case throughout and therefore drop the superscript l from Z. Third, in each $D^{(l)}$, the agency imputes values $Y_{rep,i}^{(l)}$ for those units with $Z_j = 1$, using the posterior distribution for $(Y_{rep} \mid D^{(l)}, Z)$. This is repeated independently $i = 1, \ldots, r$ times for $l = 1, \ldots, m$, so that a total of $M = mr$ datasets are generated. Each dataset, $d_i^{(l)} = (X, Y_{nrep}, Y_{mis}^{(l)}, Y_{rep,i}^{(l)}, I, R, Z)$, includes a label indicating the l of the $D^{(l)}$ from which it was drawn. These M datasets are released to the public. Releasing such nested, multiply-imputed datasets also has been proposed for handling missing data outside of the disclosure limitation context (Shen, 2000; Rubin, 2003).

Analysts can obtain valid inferences from these released datasets by combining inferences from the individual datasets. As before, let q be the analyst's estimator of Q, and let u be the analyst's estimator of the variance of q. We assume the analyst specifies q and u by acting as if each $d_i^{(l)}$ was in fact collected data from a random sample of (X, Y) based on the original sampling design I. For $l = 1, \ldots, m$ and $i = 1, \ldots, r$, let $q_i^{(l)}$ and $u_i^{(l)}$ be respectively the values of q and u in data set $d_i^{(l)}$. The following quantities are needed for inferences about scalar Q:

$$\begin{aligned}
\bar{q}_M &= \sum_{l=1}^{m}\sum_{i=1}^{r} q_i^{(l)}/(mr) = \sum_{l=1}^{m} \bar{q}^{(l)}/m && (4)\\
\bar{b}_M &= \sum_{l=1}^{m}\sum_{i=1}^{r} (q_i^{(l)} - \bar{q}^{(l)})^2/m(r-1) = \sum_{l=1}^{m} b^{(l)}/m && (5)\\
B_M &= \sum_{l=1}^{m} (\bar{q}^{(l)} - \bar{q}_M)^2/(m-1) && (6)\\
\bar{u}_M &= \sum_{l=1}^{m}\sum_{i=1}^{r} u_i^{(l)}/(mr). && (7)
\end{aligned}$$

The $\bar{q}^{(l)}$ is the average of the point estimates in each group of datasets indexed by l, and the $\bar{q}_M$ is the average of these averages across l. The $b^{(l)}$ is the variance of the point estimates for each group of datasets indexed by l, and the $\bar{b}_M$ is average of these variances. The B_M is the variance of the $\bar{q}^{(l)}$ across synthetic datasets. The $\bar{u}_M$ is the average of the estimated variances of q across all synthetic datasets.

Under conditions described in Section 4, the analyst can use $\bar{q}_M$ to estimate Q. An estimate of the variance of $\bar{q}_M$ is:

$$T_M = (1 + 1/m)B_M - \bar{b}_M/r + \bar{u}_M. \tag{8}$$

When n, m, and r are large, inferences can be based on the normal distribution, $(Q - \bar{q}_M) \sim N(0, T_M)$.

When m and r are moderate, inferences can be based on the t-distribution, $(Q - \bar{q}_M) \sim t_{\nu_M}(0, T_M)$, with degrees of freedom

$$\nu_M = \left(\frac{((1 + 1/m)B_M)^2}{(m-1)T_M^2} + \frac{(\bar{b}_M/r)^2}{m(r-1)T_M^2} \right)^{-1}. \tag{9}$$

The behavior of T_M and ν_M in special cases is instructive. When r is very large, $T_M \approx T_m$. This is because the $\bar{q}^{(l)} \approx q^{(l)}$, so that we obtain the results from analyzing the $D^{(l)}$. When the fraction of replaced values is small relative to the fraction of missing values, the $\bar{b}_M$ is small relative to B_M, so that once again $T_M \approx T_m$. In both these cases, the ν_M approximately equals ν_m, which is Rubin's (1987) degrees of freedom when imputing missing data only. When the fraction of missing values is small relative to the fraction of replaced values, the $B_M \approx \bar{b}_M/r$, so that T_M is approximately equal to T_p with M released datasets.

4 Justification of new combining rules

This section presents a Bayesian derivation of the inferences described in Section 3 and describes conditions under which these inferences are valid from a frequentist perspective. These results make use of the theory developed in Rubin (1987) and Reiter (2003). For the Bayesian derivation, we assume that the analyst and imputer use the same models.

Let $D^m = \{D^{(l)} : l = 1, \ldots, m\}$ be the collection of all multiply-imputed datasets before any observed values are replaced. For each $D^{(l)}$, let $q^{(l)}$ and $u^{(l)}$ be the posterior mean and variance of Q. As in Rubin (1987, Chapter 3), let B_∞ be the variance of the $q^{(l)}$ obtained when $m = \infty$.

Let $d^M = \{d_i^{(l)} : i = 1, \ldots, r \text{ and } l = 1, \ldots, m\}$ be the collection of all released synthetic datasets. For each $d_i^{(l)}$, let $q_i^{(l)}$ be the posterior mean of $q^{(l)}$. For each l, let $B^{(l)}$ be the variance of the $q_i^{(l)}$ obtained when $r = \infty$. Lastly, let B be the average of the $B^{(l)}$ obtained when $m = \infty$.

Using these quantities, the posterior distribution for $(Q \mid d^M)$ can be decomposed as

$$f(Q \mid d^M) \quad = \quad \int f(Q \mid d^M, D^m, B_\infty, B) f(D^m, B_\infty \mid d^M, B) f(B \mid d^M) dD^m \, dB_\infty \, dB \tag{10}$$

The integration is over the distributions of the values in D that are missing and the values in each $D^{(l)}$ that are replaced with imputations; the observed, unaltered values remain fixed. We assume standard Bayesian asymptotics hold, so that complete-data inferences for Q can be based on normal distributions.

4.1 Evaluating $f(Q \mid d^M, D^m, B_\infty, B)$

Given D^m, the synthetic data are irrelevant, so that $f(Q \mid d^M, D^m, B_\infty, B) = f(Q \mid D^m, B_\infty)$. This is the posterior distribution of Q for multiple imputation for missing data, conditional on B_∞. As shown by Rubin (1987), this posterior distribution is approximately

$$(Q \mid D^m, B_\infty) \quad \sim \quad N(\bar{q}_m, (1 + 1/m) B_\infty + \bar{u}_m) \tag{11}$$

where $\bar{q}_m$ and $\bar{u}_m$ are defined as in (1) and (3). In multiple imputation for missing data, we integrate (11) over the posterior distribution of $(B_\infty \mid D^m)$. This is not done here, since we integrate over $(B_\infty \mid d^M)$.

4.2 Evaluating $f(D^m, B_\infty \mid d^M, B) f(B \mid d^M)$

Since the distribution for Q in (11) relies only on $\bar{q}_m$, $\bar{u}_m$, and B_∞, it is sufficient for $f(D^m, B_\infty \mid d^M, B)$ to determine $f(\bar{q}_m, \bar{u}_m, B_\infty \mid d^M, B) = f(\bar{q}_m, \bar{u}_m \mid d^M, B_\infty, B) f(B_\infty \mid d^M, B)$.

Following Reiter (2003), we first assume replacement imputations are made so that, for all i, the sampling distributions of each $q_i^{(l)}$ and $u_i^{(l)}$ are,

$$(q_i^{(l)} \mid D^{(l)}, B^{(l)}) \quad \sim \quad N(q^{(l)}, B^{(l)}) \tag{12}$$

$$(u_i^{(l)} \mid D^{(l)}, B^{(l)}) \quad \sim \quad (u^{(l)}, << B^{(l)}). \tag{13}$$

Here, the notation $F \sim (G, << H)$ means that the random variable F has a distribution with expectation of G and variability much less than H. In actuality, $u_i^{(l)}$ is typically centered at a value larger than $u^{(l)}$, since synthetic data incorporate uncertainty due to drawing values of the parameters. For large sample sizes n, this bias should be minimal. The assumption that $E(q_i^{(l)} \mid D^{(l)}, B^{(l)}) = q^{(l)}$ and the normality assumption should be reasonable when the imputations are drawn from correct posterior predictive distributions, $f(Y_{rep} \mid D^{(l)}, Z)$, and the usual asymptotics hold.

Assuming flat priors for all $q^{(l)}$ and $v^{(l)}$, standard Bayesian theory implies that

$$(q^{(l)} \mid d^M, B^{(l)}) \quad \sim \quad N(\bar{q}^{(l)}, B^{(l)}/r) \tag{14}$$

$$(u^{(l)} \mid d^M, B^{(l)}) \quad \sim \quad (\bar{u}^{(l)}, << B^{(l)}/r) \tag{15}$$

$$\left(\frac{(r-1)b^{(l)}}{B^{(l)}} \mid d^M, B^{(l)}\right) \quad \sim \quad \chi^2_{r-1} \tag{16}$$

where $b^{(l)}$ is defined in (5). We next assume that $B^{(l)} = B$ for all l. This should be reasonable, since the variability in posterior variances tends to be of smaller order than the variability of posterior means. Averaging across l, we obtain

$$(\bar{q}_m \mid d^M, B) \quad \sim \quad N(\bar{q}_M, B/rm) \tag{17}$$

$$(\bar{u}_m \mid d^M, B) \quad \sim \quad (\bar{u}_M, << B/rm) \tag{18}$$

where $\bar{q}_M$ is defined in (4) and $\bar{u}_M$ is defined in (7). The posterior distribution of $(B_\infty \mid d^M, B)$ is

$$\left(\frac{(m-1)B_M}{B_\infty + B/r} \mid d^M, B\right) \sim \chi^2_{m-1} \tag{19}$$

where B_M is defined in (6).

Finally, the posterior distribution of $(B \mid d^M)$ is

$$\left(\frac{m(r-1)\bar{b}_M}{B} \mid d^M\right) \quad \sim \quad \chi^2_{m(r-1)} \tag{20}$$

where $\bar{b}_M$ is defined in (5).

4.3 Evaluating $f(Q \mid d^M)$

We need to integrate the product of (11) and (17) with respect to the distributions in (19) and (20). This can be done by numerical integration, but it is desirable to have simpler approximations for users.

For large m and r, we can replace the terms in the variance with their approximate expectations: the $B_\infty \approx B_M - B/r$, and the $B \approx \bar{b}_M$. Hence, for large m and r, the posterior distribution of Q is approximately:

$$\begin{aligned}(Q \mid d^M) \quad &\sim \quad N(\bar{q}_M, (1+1/m)(B_M - \bar{b}_M/r) + \bar{b}_M/mr + \bar{u}_M) \\ &= \quad N(\bar{q}_M, (1+1/m)B_M - \bar{b}_M/r + \bar{u}_M) = N(\bar{q}_M, T_M). \end{aligned} \tag{21}$$

When m and r are moderately sized, the normal distribution may not be a good approximation. To derive an approximate reference t-distribution, we use the strategies of Rubin (1987) and Barnard and Rubin (1999). That is, we assume that for some degrees of freedom ν_M to be estimated,

$$\left(\frac{\nu_M T_M}{\bar{u}_M + (1+1/m)B_\infty + B/mr} \mid d^M\right) \sim \chi^2_{\nu_M} \tag{22}$$

so that we can use a t-distribution with ν_M degrees of freedom for inferences about Q. We approximate ν_M by matching the first two moments of (22) to those of a chi-squared distribution. The details showing that ν_M is approximated by the expression in (9) are provided in the appendix.

The inferences based on (4) - (9) have valid frequentist properties under certain conditions. First, the analyst must use randomization-valid estimators, q and u. That is, when q and u are applied on D to get q_{obs} and u_{obs}, the $(q_{obs} \mid X, Y) \sim N(Q, U)$ and $(u_{obs} \mid X, Y) \sim (U, << U)$, where the relevant distribution is that of I. Second, the imputations for missing data must be proper in the sense of Rubin (1987, Chapter 4). Essentially, this requires that inferences from the imputations for missing data be randomization-valid for q_{obs} and u_{obs}, under the posited non-response mechanism. Third, the imputations for partially synthetic data must be synthetically proper in the sense of Reiter (2003). This requires that the inferences from the replacement imputations associated with each $D^{(l)}$ be randomization valid for the $q^{(l)}$ and $u^{(l)}$.

In general, it is difficult to verify that imputations for missing data are proper in complex samples (Binder and Sun, 1996). They may be proper for some analyses but not for others. As a result, some confidence intervals centered on unbiased estimators may not have nominal coverage rates; see Meng (1994) for a discussion of this issue. These difficulties exist for the multiple imputation approach used here, and indeed may be compounded because of the additional imputation of synthetic data.

5 Concluding Remarks

There are many challenges to using partially synthetic data approaches for disclosure limitation. Most important, agencies must decide which values to replace with imputations. General candidates for replacement include the values of identifying characteristics for units that are at high risk of identification, such as sample uniques and duplicates, and the values of sensitive variables in the tails of distributions. Confidentiality can be protected further by, in addition, replacing values at low disclosure risk (Liu and Little, 2002). This increases the variation in the replacement imputations, and it obscures any information that can be gained just from knowing which data were replaced. As with any disclosure limitation method (Duncan *et al.*, 2001), these decisions should consider tradeoffs between disclosure risk and data utility. Guidance on selecting values for replacement is a high priority for research in this area.

There remain disclosure risks in partially synthetic data no matter which values are replaced. Users can utilize the released, unaltered values to facilitate disclosure attacks, for example via matching to external databases, or they may be able to estimate actual values of Y_{obs} from the synthetic data with reasonable accuracy. For instance, if all people in a certain demographic group have the same, or even nearly the same, value of an outcome variable, the imputation models likely will generate that value for imputations. Imputers may need to coarsen the imputations for such people. As another example, when users know that a certain record has the largest value of some Y_{obs}, that record can be identified when its value is not replaced.

On the data utility side, the main challenge is specifying imputation models, both for the missing and replaced data, that give valid results. For missing data, it is well known that implausible imputation models can produce invalid inferences, although this is less problematic when imputing relatively small fractions of missing data (Rubin, 1987; Meng, 1994). There is an analogous issue for partially synthetic data. When large fractions of data are replaced, for example entire variables, analyses involving the replaced values reflect primarily the distributional assumptions implicit in the imputation models. When these assumptions are implausible, the resulting analyses can be invalid. Again, this is less problematic when only small fractions of values are replaced, as might be expected in many applications of the partially synthetic approach.

Certain data characteristics can be especially challenging to handle with partially synthetic data. For example, it may be desirable to replace extreme values in skewed distributions, such as very large incomes. Information about the tails of these distributions may be limited, making it difficult to draw reasonable replacements while protecting confidentiality. As another example, randomly drawn imputations for highly structured data may be implausible, for instance unlikely combinations of family members' ages or marital statuses. These difficulties, coupled with the general limitations of inferences based on imputations, point to an important issuse for research: developing and evaluating methods for generating partially synthetic data, including semi-parametric and non-parametric approaches.

We note that building the synthetic data models is generally an easier task than building the missing data models. Agencies can compare the distributions of the synthetic data to those of the observed data being replaced. When the synthetic distributions are too dissimilar from the observed ones, the imputation models can be adjusted. There usually is no such check for the missing data models.

It is, of course, impossible for agencies to anticipate every possible use of the released data, and hence impossible to generate models that provide valid results for every analysis. A more modest and attainable goal is to enable analysts to obtain valid inferences using standard methods and software for a wide range of standard analyses, such as some linear and logistic regressions. Agencies therefore should provide information that helps analysts decide what inferences can be supported by the released data. For example, agencies can include descriptions of the imputation models as attachments to public releases of data. Users whose analyses are not supported by the data may have to apply for special access to the observed data. Agencies also need to provide documentation for how to use the nested data sets. Rules for combining point estimates from the multiple data sets are simple enough to be added to standard statistical software packages, as has been done already for Rubin's (1987) rules in SAS, Stata, and S-Plus.

As constructed, the multiple imputation approach does not calibrate to published totals. This could make some users unhappy with or distrust the released data. It is not clear how to adapt the method—or, for that matter, many other disclosure limitation techniques that alter the original data—for calibration.

Missing data and disclosure risk are major issues confronting organizations releasing data to the public. The multiple imputation approach presented here is suited to handle both simultaneously, providing users with rectangular completed datasets that can be analyzed with standard statistical methods and software. There are challenges to implementing this approach in genuine applications, but, as noted by Rubin (1993) in his initial proposal, the potential payoffs of this use of multiple imputation are high. The next item on the research agenda is to investigate how well the theory works in practice, including comparisons of this approach with other dislosure limitation methods. These comparisons should focus on measures of disclosure risks, obtained by simulating intruder behavior, and on measures of data utility for estimands of interest to users, including properties of point and interval estimates.

Appendix: Derivation of Approximate Degrees of Freedom

Inferences from datasets with multiple imputations for both missing data and partially synthetic replacements are made using a t-distribution. A key step is to approximate the distribution of

$$\left(\frac{\nu_M T_M}{\bar{u}_M + (1+1/m)B_\infty + B/mr} \mid d^M\right) \tag{23}$$

as a chi-squared distribution with ν_M degrees of freedom. The ν_M is determined by matching the mean and variance of the inverted χ^2 distribution to the mean and variance of (23).

Let $\alpha = (B_\infty + B/r)/B_M$, and let $\gamma = B/\bar{b}_M$. Then, $(\alpha^{-1} \mid d^M, B)$ and $(\gamma^{-1} \mid d^M)$ have mean square distributions with degrees of freedom $m-1$ and $m(r-1)$, respectively. Let $f = (1+1/m)B_M/\bar{u}_M$, and let $g = (1/r)\bar{b}_M/\bar{u}_M$. We can write (23) as

$$\frac{T_M}{\bar{u}_M + (1+1/m)B_\infty + B/mr} = \frac{\bar{u}_M(1+f-g)}{\bar{u}_M(1+\alpha f - \gamma g)}. \tag{24}$$

To match moments, we need to approximate the expectation and variance of (24).

For the expectation, we use the fact that

$$E\left(\frac{1+f-g}{1+\alpha f-\gamma g} \mid d^M\right) = E\left(E\left(\frac{1+f-g}{1+\alpha f-\gamma g} \mid d^M, B\right) \mid d^M\right). \tag{25}$$

We approximate these expectations using first order Taylor series expansion in α^{-1} and γ^{-1} around their expectations, which equal one. As a result,

$$E\left(E\left(\frac{1+f-g}{1+\alpha f-\gamma g} \mid d^M, B\right) \mid d^M\right) \approx E\left(\frac{1+f-g}{1+f-\gamma g} \mid d^M\right) \approx 1. \tag{26}$$

For the variance, we use the conditional variance representation

$$E\left(Var\left(\frac{1+f-g}{1+\alpha f-\gamma g} \mid d^M, B\right) \mid d^M\right) + Var\left(E\left(\frac{1+f-g}{1+\alpha f-\gamma g} \mid d^M, B\right) \mid d^M\right). \tag{27}$$

For the interior variance and expectation, we use a first order Taylor series expansion in α^{-1} around its expectation. Since $Var(\alpha^{-1} \mid d^M, B) = 2/(m-1)$, the expression in (27) equals approximately

$$E\left(\frac{2(1+f-g)^2 f^2}{(m-1)(1+f-\gamma g)^4} \mid d^M\right) + Var\left(\frac{1+f-g}{1+f-\gamma g} \mid d^M\right). \tag{28}$$

We now use first order Taylor series expansions in γ^{-1} around its expectation to determine the components of (28). The first term in (28) is,

$$E\left(\frac{2(1+f-g)^2 f^2}{(m-1)(1+f-\gamma g)^4} \mid d^M\right) \approx \frac{2f^2}{(m-1)(1+f-g)^2}. \tag{29}$$

Since $Var(\gamma^{-1} \mid d^M) = 2/(m(r-1))$, the second term in (28) is

$$Var\left(\frac{1+f-g}{1+f-\gamma g} \mid d^M\right) \approx \frac{2g^2}{m(r-1)(1+f-g)^2}. \tag{30}$$

Combining (29) and (30), the variance of (23) equals approximately

$$\frac{2f^2}{(m-1)(1+f-g)^2} + \frac{2g^2}{m(r-1)(1+f-g)^2}. \tag{31}$$

Since a mean square random variable has variance equal to 2 divided by its degrees of freedom, we conclude that

$$\nu_M = \left(\frac{f^2}{(m-1)(1+f-g)^2} + \frac{g^2}{m(r-1)(1+f-g)^2} \right)^{-1}. \tag{32}$$

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*, 215–277. Amsterdam: North-Holland.

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

Binder, D. A. and Sun, W. (1996). Frequency valid multiple imputation for surveys with a complex design. In *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 281–286.

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., Keller-McNulty, S. A., and Stokes, S. L. (2001). Disclosure risk vs. data utility: The R-U confidentiality map. Tech. rep., U.S. National Institute of Statistical Sciences.

Fienberg, S. E., Makov, U. E., and Steele, R. J. (1998). Disclosure limitation using perturbation and related methods for categorical data. *Journal of Official Statistics* **14**, 485–502.

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

Kennickell, A. B. (1997). 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. Washington, D.C.: National Academy Press.

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

Liu, F. and Little, R. J. A. (2002). Selective multiple imputation of keys for statistical disclosure control in microdata. In *ASA Proceedings of the Joint Statistical Meetings*, 2133–2138.

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

Reiter, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics* **18**, 531–544.

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

Reiter, J. P. (2004a). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society, Series A* forthcoming.

Reiter, J. P. (2004b). Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. *Journal of Statistical Planning and Inference* forthcoming.

Rubin, D. B. (1976). Inference and missing data (with discussion). *Biometrika* **63**, 581–592.

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

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

Rubin, D. B. (2003). Nested multiple imputation of NMES via partially incompatible MCMC. *Statistica Neerlandica* **57**, 3–18.

Shen, Z. (2000). *Nested Multiple Imputation*. Ph.D. thesis, Harvard University, Dept. of Statistics.

Willenborg, L. and de Waal, T. (2001). *Elements of Statistical Disclosure Control*. New York: Springer-Verlag.