

# Protecting Data Confidentiality in Public Release Datasets: Approaches Based on Multiple Imputation

Jerome P. Reiter\*

## Abstract

When releasing data to the public, data disseminators typically are required to protect the confidentiality of survey respondents' identities and attribute values. Removing direct identifiers such as names and addresses generally is not sufficient to eliminate disclosure risks, so that statistical disclosure limitation strategies must be applied to the data before release. This chapter provides an overview of how multiple imputation, originally devised to handle missing data, can be adapted for disclosure limitation. It reviews recent literature on inferential methods for analyzing such datasets.

## 1 Introduction

Many national statistical agencies, survey organizations, and researchers disseminate microdata, i.e. data on individual units, to the public. Wide dissemination greatly benefits society, enabling broad subsets of the research community to access and analyze the collected data. Often, however, data disseminators cannot release microdata in their collected form, because doing so would reveal some survey respondents' identities or values of sensitive attributes. Failure to protect confidentiality can have serious consequences for the data disseminators, since they may be violating laws passed specifically to protect confidentiality, such as the recently enacted HIPPA and CIPSEA

---

\*Institute of Statistics and Decision Sciences, Duke University, Durham, NC, USA.

(Wallman and Harris-Kojetin, 2004) in the U.S. Additionally, when confidentiality is compromised, the data collectors may lose the trust of the public, so that potential respondents are less willing to give accurate answers, or even to participate, in future surveys.

At first glance, protecting confidentiality seems a straightforward task: simply strip unique identifiers like names, social security numbers, and exact addresses before releasing data. However, these actions alone may not suffice when key identifying variables, such as age, sex, race, and marital status, remain on the file. These keys can be used to match units in the released data to other databases. For example, Sweeney (1997) showed that 97% of the records in a medical database for Cambridge, MA, could be identified using only birth date and 9-digit ZIP code by linking them to a publicly available voter registration list.

Data disseminators therefore further limit what they release, typically by altering the collected data. Common strategies include recoding variables, such as releasing ages or geographical variables in aggregated categories; reporting exact values only above or below certain thresholds, for example reporting all incomes above 100,000 as “100,000 or more”; swapping data values for selected records, e.g. switch the sexes of some men and women, to discourage users from matching, since matches may be based on incorrect data; and, adding noise to numerical data values to reduce the possibilities of exact matching on key variables or to distort the values of sensitive variables. Most public use datasets analyzed by researchers have undergone at least one of these methods of statistical disclosure limitation. See Willenborg and de Waal (2001) for a general overview of common methods.

These methods can be applied with varying intensities. Generally, increasing the amount of alteration decreases the risks of disclosures, but it also decreases the accuracy of inferences obtainable from the released data (since these methods distort relationships among the variables). Unfortunately, it is difficult—and for some analyses impossible—for data users to determine how much their particular estimation has been compromised by the data alteration, because disseminators rarely release detailed information about the disclosure limitation strategy. Even when such information is available, adjusting for the data alteration may be beyond some users’ statistical knowledge. For example, to analyze properly data that include additive random noise, users should apply measurement error models (Fuller, 1993) or the likelihood based approach of (Little, 1993), which are difficult to use for non-standard estimands.

Motivated by these problems, Rubin (1993) proposed an alternative approach to protecting confidentiality in public use data files: release multiply-imputed, synthetic data sets. In this approach, the data disseminator (i) randomly and independently samples units from the sampling frame to comprise each synthetic data set, (ii) imputes the unknown data values for units in the synthetic samples using models fit with the original survey data, and (iii) releases multiple versions of these data sets to the public. These are called fully synthetic datasets. Releasing fully synthetic data can preserve confidentiality, since identification of units and their sensitive data is nearly impossible when the released data are not actual, collected values. Furthermore, with appropriate data generation and estimation methods based on the concepts of multiple imputation (Rubin, 1987), the approach can allow data users to make valid inferences for a variety of estimands using standard, complete-data statistical methods and software.

Recently, statisticians in both academia and governmental agencies have begun to develop and extend Rubin's proposal. Raghunathan *et al.* (2003) and Reiter (2005b) derive methods for obtaining valid inferences from multiple fully synthetic datasets. Reiter (2002) illustrates the impact of the sampling design and the number and size of synthetic datasets on inferences. Raghunathan (2003) describes a semi-parametric approach to simulating data. Reiter (2005a) generates fully synthetic data for a subset of the U.S. Current Population Survey. Several researchers (Little, 1993; Kennickell, 1997; Abowd and Woodcock, 2001; Liu and Little, 2002; Reiter, 2003, 2004b, 2005c) investigate a variant of Rubin's approach: release the units originally surveyed but replace only some of these units' data with multiple imputations. These are called partially synthetic datasets. Other discussions and variants of synthetic data approaches include those in Little (1993), Fienberg *et al.* (1998), Dandekar *et al.* (2002a,b), Franconi and Stander (2002, 2003), Polettini *et al.* (2002), and Polettini (2003).

In this chapter, I review the multiple imputation framework for statistical disclosure limitation. The remainder of the chapter is organized as follows. Section 2 describes the general framework of the synthetic data approach. Section 3 summarizes recent research on methods for obtaining inferences from multiple synthetic datasets. Finally, Section 4 lays out some of the challenges to implementing the fully synthetic approach in practice.

## 2 Description of synthetic data methods

In what follows, the organization releasing synthetic public use data is abbreviated as the imputer, and the user of the publicly released data is abbreviated as the analyst. We outline the ideas underpinning fully synthetic data in Section 2.1 and partially synthetic data in Section 2.2.

### 2.1 Fully synthetic data

To illustrate how fully synthetic data might work in practice, we modify the setting described by Reiter (2004a). Suppose the imputer has collected data on a random sample of 10,000 people. The data comprise each person’s race, sex, income, and indicator for the presence of a disease. We assume the imputer has a list containing all people in the population, including their race and sex. This list could be the one used when selecting the random sample of 10,000, or it could be manufactured from census tabulations of the race-sex joint distribution. We assume the imputer knows the income and disease status only for the people who respond to the survey.

To generate synthetic data, first the imputer randomly samples some number of people, say 20,000, from the population list. The imputer then generates values of income and disease status for these 20,000 synthetic people by randomly simulating values from the joint distributions of income and disease status, conditional on their race and sex values. These distributions are estimated using the collected data and possibly other relevant information. The result is one synthetic data set. The imputer repeats this process say ten times, each time using different random samples of 20,000 people, to generate ten synthetic data sets. These ten data sets are then released to the public.

When the incomes and disease statuses for the synthetic people are simulated from the true joint probability distributions, the synthetic data should have similar characteristics on average as the collected data. There is an analogy here to random sampling. Some true distribution of income and disease status exists in the population. The observed data are just a random sample from that population distribution. If we generate synthetic data from that same distribution, we essentially create different random samples from the population. Hence, the analyst using these synthetic samples essentially analyzes alternative samples from the population.

The on average caveat is important: parameter estimates from any one

simulated data set are unlikely to equal exactly those from the observed data. The synthetic parameter estimates are subject to three sources of variation, namely (i) sampling the collected data; (ii) sampling the synthetic units from the population; and, (iii) generating values for those synthetic units. It is not possible to estimate the three sources of variation from only one released synthetic data set. However, it is possible to do so from multiple synthetic data sets, which explains why the multiple imputation framework applies. To account for the three sources of variability, the analyst estimates parameters and their variances in each of the synthetic data sets, and combines these results using the methods of Raghunathan *et al.* (2003), described in Section 3.1.

Releasing fully synthetic data makes identification of units and their sensitive data from synthetic samples very difficult. Almost all of the released, synthetic units are not in the original sample, having been randomly selected from the sampling frame, and their values of survey data are simulated. The synthetic records cannot be matched meaningfully to records in other data sets, such as administrative records, because the values of released survey variables are simulated rather than actual. Releasing fully synthetic data is subject to attribute disclosure risk—the risk that the released data can be used to estimate unknown sensitive values very closely—when the models used to simulate data are “too accurate.” For example, when data are simulated from a regression model with a very small mean square error, analysts can estimate outcomes precisely using the model, if they know predictors in that model. Or, 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 can reduce these types of risks by using less precise models when necessary.

Fully synthetic data sets can have positive data utility features. When data are simulated from distributions that reflect the distributions of the observed data, frequency-valid inferences can be obtained from the multiple synthetic data sets for a wide range of estimands. These inferences can be determined by combining standard likelihood-based or survey-weighted estimates; the analyst need not learn new statistical methods or software programs. Synthetic data sets can be sampled by schemes other than the typically complex design used to collect the original data, so that analysts can ignore the design for inferences and instead perform analyses based on simple random samples. Additionally, the data generation models can incorporate adjustments for nonsampling errors and can borrow strength from other data

sources, thereby resulting in inferences that can be even more accurate than those based on the original data. Finally, because all units are simulated, geographic identifiers can be included in the synthetic data sets, facilitating estimation for small areas.

There is a cost to these benefits: the validity of synthetic data inferences depends critically on the validity of the models used to generate the synthetic data. This is because the synthetic data reflect only those relationships included in the data generation models. When the models fail to reflect accurately certain relationships, analysts' inferences also will not reflect those relationships. Similarly, incorrect distributional assumptions built into the models will be passed on to the users' analyses. This dependence is a potentially serious limitation to releasing fully synthetic data. Practically, it means that some analyses cannot be performed accurately, and that imputers need to release information that helps analysts decide whether or not the synthetic data are reliable for their analyses. For example, imputers can include the models as attachments to public releases of data. Or, they can include generic statements that describe the imputation models, such as "Main effects for age, sex, and race are included in the imputation models for education." Analysts who desire finer detail than afforded by the imputations may have to apply for special access to the observed data.

Releasing or describing the imputation models is necessary, but it is not sufficient: imputers also should release synthetic data generated from the models. Some analysts are not able to generate synthetic data given the models; they need imputers to do it for them. Even when analysts can do so, it is a cumbersome burden to place on them. Additionally, analysts may desire some function of the synthetic data that is hard to estimate from the model parameters, but easy to determine from the synthetic data.

## **2.2 Partially synthetic data**

As of this writing, no agencies have adopted the fully synthetic approach, although the U.S. Bureau of the Census is currently developing synthetic datasets for the Survey of Income and Program Participation and the Longitudinal Business Database. However, some agencies have adopted partially synthetic data approaches.

Partially synthetic data comprise the units originally surveyed with some collected values replaced with multiple imputations. For example, the imputer may simulate sensitive variables or identifiers only for units in the

sample with rare combinations of identifiers; or, the imputer may replace all data for selected sensitive variables or identifiers. The former strategy has been employed by the U.S. Federal Reserve Board. They protect 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 un-replaced, collected values (Kennickell, 1997). It also underlies SMiKE, an algorithm for simulating multiple values of key identifiers for selected units developed by Liu and Little (2002). The latter strategy has been employed by U.S. Bureau of the Census. They protect data in a longitudinal, linked business dataset by replacing all values of some sensitive variables with multiple imputations and leaving other variables at their actual values (Abowd and Woodcock, 2001).

To illustrate a partially synthetic strategy, we can adapt the setting used in Section 2.1. Suppose the imputer wants to replace income when it exceeds \$100,000 and is willing to release all other values. The imputer generates replacement values for the incomes over \$100,000 by randomly simulating from the distribution of income conditional on race, sex, and disease status. To avoid bias, this distribution also must be conditional on income exceeding \$100,000. The distribution is estimated using the collected data and possibly other relevant information. The result is one synthetic data set. The imputer repeats this process say ten times to generate ten synthetic data sets. These ten data sets are then released to the public.

As with fully synthetic data, when the replacement imputations are generated from the true posterior distribution, each synthetic dataset is essentially a random sample from the population. Inferences are straightforward: the analyst estimates parameters and their variances in each of the synthetic data sets, and combines these results using the methods of Reiter (2003), described in Section 3.2. An advantage of partially synthetic data relative to fully synthetic data is that only a fraction of the data are imputed, so that analysts' inferences are generally less sensitive to the imputer's model specification. Unlike fully synthetic data, partially synthetic data must be analyzed in accordance with the original sampling design.

The protection afforded by partially synthetic data depends on the nature of the synthesis. Replacing key identifiers with imputations makes it difficult for users to know the original values of those identifiers, which reduces the chance of identifications. Replacing values of sensitive variables makes it difficult for users to learn the exact values of those variables, which can prevent attribute disclosures. Nonetheless, there remain disclosure risks

in partially synthetic data no matter which values are replaced. Analysts can utilize the released, unaltered values to facilitate disclosure attacks, for example via matching to external databases, or they may be able to estimate genuine values from the synthetic data with reasonable accuracy.

### 3 Inferential methods

This section summarizes recent research on methods of obtaining inferences from synthetic, multiply-imputed datasets. It focuses on inferences for scalar estimands. See Reiter (2005b) for inferences for multi-component estimands.

We use the following notation for all inferential methods. 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  $j = 1, 2, \dots, N$ . Let  $I = (I_1, \dots, I_N)$ . 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  $R = (R_1, \dots, R_N)$ . 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  $Y = (Y_{obs}, Y_{exc})$  be the  $N \times p$  matrix of survey data for all units in the population. 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 for all population units, for example from census records or the sampling frame(s). When it is not known for some units,  $X$  can be treated as part of  $Y$  for those units. Finally, we write the observed data as  $D = (X, Y_{obs}, I, R)$ .

#### 3.1 Fully Synthetic Data

The imputer constructs synthetic data sets based on the observed data,  $D$ , in a two-part process. First, the imputer imputes values of  $Y_{exc}$  to obtain a completed-data population,  $(X, Y_{com,i})$ . For reasons discussed in Rubin (1987) and Raghunathan *et al.* (2003), imputations should be generated from the Bayesian posterior predictive distribution of  $(Y|D)$ . The imputer also may choose to impute values of  $Y$  for all  $N$  units so that the completed-data contain no real values of  $Y$ , thereby avoiding the release of any respondent's actual value of  $Y$ . Second, the imputer samples  $n_{syn}$  units randomly from the completed-data population  $(X, Y_{com,i})$ , using a simple random sample.

These sampled units are released as public use data, so that the released data set,  $d_i = (X, Y_{syn,i})$ , contains the values of  $Y$  only for units in the synthetic sample. This entire process is repeated independently  $i = 1, \dots, m$  times to get  $m$  different synthetic data sets. In practice, it is not necessary to generate completed-data populations for constructing the  $Y_{syn,i}$ . The imputer need only generate values of  $Y$  for units in the synthetic samples.

From these synthetic data sets the analyst seeks inferences about some estimand  $Q = Q(X, Y)$ , where the notation  $Q(X, Y)$  means that the estimand  $Q$  is a function of  $(X, Y)$ . For example,  $Q$  could be the population mean of  $Y$  or the population regression coefficients of  $Y$  on  $X$ . In each synthetic data set, the analyst estimates  $Q$  with some estimator  $q$  and the variance of  $q$  with some estimator  $v$ . It is assumed that the analyst specifies  $q$  and  $v$  by acting as if the synthetic data were in fact collected data from a simple random sample of  $(X, Y)$ .

For  $i = 1, \dots, m$ , let  $q_i$  and  $v_i$  be respectively the values of  $q$  and  $v$  in synthetic data set  $i$ . Under assumptions described in Raghunathan *et al.* (2003), the analyst can obtain valid inferences for scalar  $Q$  by combining the  $q_i$  and  $v_i$ . Specifically, the following quantities are needed for inferences:

$$\bar{q}_m = \sum_{i=1}^m q_i/m \tag{1}$$

$$b_m = \sum_{i=1}^m (q_i - \bar{q}_m)^2/(m-1) \tag{2}$$

$$\bar{v}_m = \sum_{i=1}^m v_i/m. \tag{3}$$

The  $\bar{q}_m$  is the average of the point estimates; the  $b_m$  is the variance of these point estimates; and, the  $\bar{v}_m$  is the average of the variance estimates. These quantities are identical to those defined by Rubin (1987) for multiple imputation for missing data.

The analyst can use  $\bar{q}_m$  to estimate  $Q$  and

$$T_f = (1 + 1/m)b_m - \bar{v}_m \tag{4}$$

to estimate the variance of  $\bar{q}_m$ . The  $b_m - \bar{v}_m$  is an approximately unbiased estimator of the variance of  $q_{obs} = Q(D)$ , and the  $b_m/m$  adjusts for using only a finite number of synthetic data sets. Although it is possible for  $T_f < 0$ ,

negative values generally can be avoided by making  $m$  and  $n_{syn}$  large. A more complicated variance estimator that is always positive is described in Raghunathan *et al.* (2003). When  $T_f > 0$ , and  $n$ ,  $n_{syn}$ , and  $m$  are large, inferences for scalar  $Q$  can be based on a normal distribution, so that a synthetic 95% confidence interval for  $Q$  is  $\bar{q}_m \pm 1.96\sqrt{T_f}$ . An approximate t-distribution for small  $m$  is described by Reiter (2005b).

The variance for fully synthetic data differs from the standard variance formula from multiple imputation for missing data, which is  $T_m = (1 + 1/m)b_m + \bar{u}_m$ . In fully synthetic data, the  $b_m$  reflects two sources of uncertainty: sampling the collected units and sampling the synthetic units. Hence, we subtract  $\bar{u}_m$ , which reflects sampling of the synthetic units, from  $b_m$  to obtain an appropriate estimate of  $\text{Var}(q_{obs})$ . In multiple imputation for missing data, the  $b_m + \bar{u}_m$  is an appropriate estimate of  $\text{Var}(q_{obs})$ .

### 3.2 Partially synthetic data when $Y_{inc} = Y_{obs}$

Assuming no missing data, i.e.  $Y_{inc} = Y_{obs}$ , the imputer 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  $Z = (Z_1, \dots, Z_n)$ . Let  $Y_{rep,i}$  be all the imputed (replaced) values in the  $i$ th 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  $(Y_{rep,i} | D, Z)$ , 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  $i = 1, \dots, r$  times to yield  $r$  different partially synthetic data sets, which are released to the public.

Inferences from partially synthetic datasets are based on the quantities defined in Equations (1)-(3). We assume the analyst specifies the point and variance estimators,  $q$  and  $u$ , by acting as if each  $d_i$  was in fact collected data from a random sample of  $(X, Y)$  based on the original sampling design  $I$ . As shown by Reiter (2003), under certain conditions the analyst can use  $\bar{q}_r$  to estimate  $Q$  and

$$T_p = b_r/r + \bar{u}_r \tag{5}$$

to estimate the variance of  $\bar{q}_r$ . Inferences for scalar  $Q$  can be based on t-distributions with degrees of freedom  $\nu_p = (r - 1)(1 + \bar{u}_r/(b_r/r))^2$ . In many

cases,  $\nu_p$  will be large enough that a normal distribution provides an adequate approximation to the t-distribution.

$T_p$  differs from the variance estimator for multiple imputation of missing data,  $T_m = (1 + 1/m)b_m + \bar{u}_m$ . In the partially synthetic data context, the  $\bar{u}_r$  estimates  $\text{Var}(q_{obs})$  and the  $b_r/r$  estimates the additional variance due to using a finite number of imputations. In the missing data context, the  $b_m/m$  has the same interpretation, but now  $b_m + \bar{u}_m$  estimates  $\text{Var}(q_{obs})$ . The additional  $b_m$  is needed to average over the nonresponse mechanism (Rubin, 1987, Ch. 4). This additional averaging is unnecessary in partially synthetic data settings with no missing data, so that using  $T_m$  can severely overestimate variances.

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

When some data are missing, i.e.  $Y_{inc} \neq Y_{obs}$ , 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 imputer seeks to replace all values larger than some threshold 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 threshold. 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 analysts to estimate the total variability correctly, imputers can employ a three-step procedure for generating imputations, as described by

Reiter (2004b). First, the imputer fills in  $Y_{mis}$  with draws from the posterior distribution for  $(Y_{mis} | D)$ , resulting in  $m$  completed datasets,  $D^{(1)}, \dots, D^{(m)}$ . 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  $(Y_{rep} | D^{(l)}, Z)$ . This is repeated independently  $i = 1, \dots, r$  times for  $l = 1, \dots, m$ , 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, 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  $l = 1, \dots, m$  and  $i = 1, \dots, r$ , 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  $Q$ :

$$\bar{q}_M = \sum_{l=1}^m \sum_{i=1}^r q_i^{(l)} / (mr) = \sum_{l=1}^m \bar{q}^{(l)} / m \quad (6)$$

$$\bar{b}_M = \sum_{l=1}^m \sum_{i=1}^r (q_i^{(l)} - \bar{q}^{(l)})^2 / m(r-1) = \sum_{l=1}^m b^{(l)} / m \quad (7)$$

$$B_M = \sum_{l=1}^m (\bar{q}^{(l)} - \bar{q}_M)^2 / (m-1) \quad (8)$$

$$\bar{u}_M = \sum_{l=1}^m \sum_{i=1}^r u_i^{(l)} / (mr). \quad (9)$$

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 Reiter (2004b), the analyst can use  $\bar{q}_M$  to estimate  $Q$ . An estimate of the variance of  $\bar{q}_M$  is:

$$T_M = (1 + 1/m)B_M - \bar{b}_M/r + \bar{u}_M. \quad (10)$$

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

$$\nu_M = \left( \frac{((1 + 1/m)B_M)^2}{(m - 1)T_M^2} + \frac{(\bar{b}_M/r)^2}{m(r - 1)T_M^2} \right)^{-1}. \quad (11)$$

The behavior of  $T_M$  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  $T_M \approx T_m$ . 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 Concluding remarks

There are many challenges to implementing fully or partially synthetic data approaches for disclosure limitation. These challenges represent opportunities for statistical researchers. In this concluding section, I lay out some of these challenges.

For both fully and partially synthetic data, the main challenge is specifying imputation models that give valid results for a wide range of analyses. 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 fully and partially synthetic data. When large fractions of data are replaced, for example entire variables, analyses involving the imputed 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 some applications of the partially synthetic approach.

Certain data characteristics can be especially challenging to handle with 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 issue for research: developing and evaluating methods for generating synthetic data, including semi-parametric and non-parametric approaches.

For partially synthetic data, 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, these decisions should consider tradeoffs between disclosure risk and data utility (Duncan *et al.*, 2001). Guidance on selecting values for replacement is a high priority for research in this area.

Given the ever-increasing resources available to those seeking to achieve disclosures—the proliferation of readily available databases, and advances in computing and record linkage technologies—the risks of unintended and/or illegal disclosures in many datasets are high and rising. In the future, it is conceivable that agencies may not be allowed or willing to release any genuine data in public use files. If so, the synthetic data approach may be one of the only ways to provide society with public use data. However, synthetic datasets should not replace all access to the original data. Such access, which can continue through licensing and other special arrangements, is needed if researchers are to make “surprise” discoveries from data. Ultimately, a statistical disclosure limitation strategy that combines restricted data access for sophisticated analyses and synthetic data for a wide range of simple analyses, such as regressions and comparisons of means, should meet the needs of a large fraction of the users of public-use data.

## 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.
- Dandekar, R. A., Cohen, M., and Kirkendall, N. (2002a). Sensitive micro data protection using Latin hypercube sampling technique. In J. Domingo-Ferrer, ed., *Inference Control in Statistical Databases*, 117–125. Berlin: Springer-Verlag.
- Dandekar, R. A., Domingo-Ferrer, J., and Sebe, F. (2002b). LHS-based hybrid microdata versus rank swapping and microaggregation for numeric microdata protection. In J. Domingo-Ferrer, ed., *Inference Control in Statistical Databases*, 153–162. Berlin: Springer-Verlag.
- 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.
- Franconi, L. and Stander, J. (2002). A model based method for disclosure limitation of business microdata. *The Statistician* **51**, 1–11.
- Franconi, L. and Stander, J. (2003). Spatial and non-spatial model-based protection procedures for the release of business microdata. *Statistics and Computing* **13**, 295–306.
- 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.
- Polettini, S. (2003). Maximum entropy simulation for microdata protection. *Statistics and Computing* **13**, 307–320.
- Polettini, S., Franconi, L., and Stander, J. (2002). Model-based disclosure protection. In J. Domingo-Ferrer, ed., *Inference Control in Statistical Databases*, 83–96. Berlin: Springer-Verlag.
- Raghunathan, T. E. (2003). Evaluation of Inferences from Multiple Synthetic Data Sets Created Using Semiparametric Approach. Tech. rep. National Academy of Sciences Panel on Access to Confidential Research Data.
- 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* **29**, 181–189.
- Reiter, J. P. (2004a). New approaches to data dissemination: A glimpse into the future (?). *Chance* **17**, 3, 12–16.
- Reiter, J. P. (2004b). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology* **30**, 235–242.
- 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.

- Reiter, J. P. (2005b). Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. *Journal of Statistical Planning and Inference* **131**, 365–377.
- Reiter, J. P. (2005c). Using CART to generate partially synthetic, public use microdata. *Journal of Official Statistics* 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.
- Sweeney, L. (1997). Computational disclosure control for medical microdata: the Datafly system. In *Proceedings of an International Workshop and Exposition*, 442–453.
- Wallman, K. K. and Harris-Kojetin, B. A. (2004). Implementing the Confidentiality Information Protection and Statistical Efficiency Act of 2002. *Chance* **17**, 3, 21–25.
- Willenborg, L. and de Waal, T. (2001). *Elements of Statistical Disclosure Control*. New York: Springer-Verlag.