Using Multiple Imputation to Integrate and Disseminate Confidential Microdata

Jerome P. Reiter

Department of Statistical Science, Duke University

Box 90251, Durham, NC, 27708, USA.

Summary

In data integration contexts, two statistical agencies seek to merge their separate databases in one file. The agencies also may seek to disseminate data to the public based on the integrated file. These goals may be complicated by the agencies' need to protect the confidentiality of database subjects, which could be at risk during the integration or dissemination stage. This article proposes several approaches based on multiple imputation for disclosure limitation, usually called synthetic data, that could be used to facilitate data integration and dissemination while protecting data confidentiality. It reviews existing methods for obtaining inferences from synthetic data and points out where new methods are needed to implement the data integration proposals.

Key words: Confidentiality; disclosure; fusion; matching; sharing; synthetic

1 Introduction

In many contexts, statistical agencies, survey organizations, businesses, and other data owners (henceforth all called agencies) with related databases can enhance analyses by combining their data. For example, one agency might have demographic and health data, and a second agency might have income data. An integrated database enables predictions of health outcomes from demographic and income variables, which is more informative than predictions from the health and demographic data alone.

Data integration may be complicated by ethical or legal obligations to protect confidentiality of database subjects. These obligations may prevent agencies from sharing the records in their databases with each other or, for agencies charged with disseminating data, with the broader public. For some analyses, the agencies can work around the first constraint using techniques from secure computation, which allows agencies to compute specific quantities from the integrated data without actually sharing individual records with each other. Secure computation algorithms have been developed for linear regression (Du *et al.*, 2004; Karr *et al.*, 2005, 2007, 2009), data mining with association rules (Kantarcioglu and Clifton, 2002; Vaidya and Clifton, 2002; Evfimievski *et al.*, 2004), model based clustering (Vaidya and Clifton, 2003; Lin *et al.*, 2004), and adaptive regression splines (Ghosh *et al.*, 2007). The literature on privacy-preserving data mining (Agrawal and Srikant, 2000; Lindell and Pinkas, 2000) contains related results. Secure computation techniques do not provide methods for sharing an integrated database with the public.

This article discusses how multiple imputation for disclosure limitation, usually called synthetic data, can be adapted to facilitate inter-agency data sharing and public data dissemination. The basic idea is for agencies to construct an integrated database that satisfies inter-agency confidentiality concerns by sharing simulated datasets with each other. To protect confidentiality in further dissemination to the public, the agencies simulate sensitive values in the integrated database to create synthetic data. At each step, the simulations are done multiple times to enable the ultimate users of the integrated data—analysts of public use files—to obtain valid inferences, at least for analyses congenial (Meng, 1994) to the models used in the simulation steps.

Multiple imputation approaches are proposed for two flavors of data integration. To describe these flavors, we suppose that there are two datasets, $D_1 = (Z_1, X_1)$ owned by Agency 1 and $D_2 = (Z_2, Y_2)$ owned by Agency 2, to be integrated by the two agencies. Let D_{com} be the integrated dataset. The first integration setting is to create $D_{com} = (Z_1, X_1, Y_1)$, i.e. to append the values of Y_2 to D_1 for those records common to both datasets (with missing values of Y_1 for records in D_1 but not in D_2). We call this the "add-on" setting. The add-on setting is a standard record linkage problem, where the linking variables might be common identifiers or values in Z. In this article, we ignore the effects of potential matching errors on inferences. The second setting is to create $D_{com} = (Z, X^*, Y^*)$, where $Z = (Z_1, Z_2)$, $X^* = (X_1, X_2^*)$, $Y^* = (Y_1^*, Y_2)$, and X_2^* and Y_1^* are imputed values. We call this the "complete-it" setting. This can be viewed as a missing data problem, where the complete dataset has

(X, Y, Z) for all records in D_1 and D_2. To enable correct estimation of uncertainty, the agencies can create several completed datasets, $(D_{com}^{(1)}, D_{com}^{(2)}, \ldots, D_{com}^{(m)})$, each containing independent draws of (X_2^*, Y_1^*), which can be analyzed using the methods of Rubin (1987). Under strong assumptions, it is possible to create these datasets even when the records in D_1 and D_2 do not overlap. This is called statistical matching (D'Orazio *et al.*, 2006) or data fusion (Rässler, 2003).

The remainder of the article is organized as follows. Section 2 reviews the use and benefits of multiple imputation approaches for disclosure limitation. These approaches are the building blocks of the proposals for integrating and sharing data via multiple imputation, which are described in Section 3. Section 4 illustrates one of the data sharing proposals using genuine data. Section 5 has some concluding remarks about implementation of these proposals.

2 Description of synthetic data methods

Before discussing fully and partially synthetic data approaches, we begin with a general overview of data confidentiality in the context of public use data dissemination.

2.1 Data confidentiality and public use dissemination

Wide dissemination of data greatly benefits society, enabling broad subsets of the research community to access and analyze the collected data. Often, however, data disseminators cannot release data as collected, because doing so could reveal survey respondents' identities or sensitive attributes. Failure to protect confidentiality can have serious consequences for data disseminators, since they may be violating laws passed to protect confidentiality. 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, releasing safe public use data seems straightforward: simply strip unique identifiers like names, tax identification numbers, and exact addresses before releasing data. However, these actions alone may not suffice when quasi-identifiers, such as demographic variables, employment/education histories, or establishment sizes, 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.

Agencies therefore further limit what they release, typically by altering the collected data (Willenborg and de Waal, 2001). 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 quasi-identifiers for at-risk records with those for other records 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.

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 obtained from the released data, since these methods distort relationships among the variables (Duncan *et al.*, 2001; Gomatam *et al.*, 2005; Shlomo, 2007). For example, intensive data swapping severely attenuates correlations between the swapped and unswapped 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, in part because agencies rarely release detailed information about the disclosure limitation strategy. Even when such information is available, adjusting for the data alteration to obtain valid inferences 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), which are difficult to use for non-standard estimands. Moreover, as resources for ill-intentioned data users continue to expand, the alterations needed to protect data with traditional disclosure limitation techniques may become so extreme that, for many analyses, the released data are no longer useful.

2.2 Fully synthetic data

Motivated by these problems, Rubin (1993) proposed an alternative approach to protecting confidentiality in public use data files: release multiply-imputed, synthetic datasets. In this approach, the agency (i) randomly and independently samples units from the sampling frame to comprise each synthetic dataset, (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 datasets to the public. A similar approach was suggested by Fienberg (1994). This can preserve confidentiality, since identification of units and their sensitive data is nearly impossible when the released data are not actual, collected values. Fully synthetic, public use data products are being developed by statistical agencies in Germany (Drechsler *et al.*, 2007) and New Zealand (Graham and Penny, 2005).

To fix notation for describing the generation and analysis of synthetic data, consider the integrated dataset, D_{com}, to be a random sample from a finite population D of size N. We write $D_{com} = (D_{obs}, D_{mis})$. The D_{obs} is the portion of D_{com} that is observed, e.g., (Z, X_1, Y_2). The D_{mis} is the portion of D_{com} that is missing either because of the design, e.g. (X_2, Y_1), or due to nonresponse. Let D_{exc} be all values not observed in D_{com}, including those values for records not in the integrated database. Finally, the entire population of values is $D = (D_{obs}, D_{exc})$.

The agency constructs fully synthetic datasets based on D_{obs} in a two-part process. First, the agency imputes values of D_{exc} to obtain a completed-data population, $D^{(i)}$. For reasons discussed in Rubin (1987) and Raghunathan *et al.* (2003), imputations should be generated from the Bayesian posterior predictive distribution $f(D_{exc}|D_{obs})$, or some approximation of it. One convenient approach is to generate imputations with a sequence of conditional models, also called chained equations, as is frequently done in multiple imputation for missing data (Van Buuren and Oudshoorn, 1999; Raghunathan *et al.*, 2001). The agency may choose to impute values of D for all N units so that the completed-data contain no real values, thereby avoiding the release of any respondent's actual values in D_{obs}. Second, to reduce the size of the file released to the public, the agency samples n_{syn} units from $D^{(i)}$ using a simple random sample. These sampled units are released as public use data, so that the released dataset, $d^{(i)}$, contains the values of $D^{(i)}$ only for units in the synthetic sample. This entire process is repeated independently $i = 1, \ldots, m$ times to get m different synthetic datasets, which are released to the public. In practice, it is not necessary to generate completed-data populations for constructing the $d^{(i)}$. The agency need only generate values of D for units in the synthetic samples.

From these synthetic datasets, the analyst seeks inferences about some estimand Q, for example the population mean of Y or the population regression coefficients of Y on X. In each synthetic dataset, 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 $i = 1, \ldots, m$, let $q^{(i)}$ and $v^{(i)}$ be respectively the values of q and v computed with $d^{(i)}$. Under assumptions described by 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:

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

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 $\bar{q}_m$. For large m, inferences can be based on a normal distribution, $(\bar{q}_m - Q) \sim N(0, T_f)$. This variance estimator differs from the one in Rubin (1987) for standard missing data, because full synthesis involves the additional step of sampling new records off the frame; see Reiter and Raghunathan (2007) for a detailed explanation.

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 are determined by combining standard likelihood-based or survey-weighted estimates; the analyst need not learn new statistical methods or software programs or worry about adjusting for the disclosure limitation method in inferences. Synthetic data sets are analyzed as simple random samples, even when the observed data are collected with a complex sampling design. 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. Because all units are simulated, geographic identifiers can be included in the synthetic data sets, facilitating estimation for small areas. Other benefits are discussed in Raghunathan *et al.* (2003) and Reiter (2002, 2005a,b).

There is a cost to these benefits: the validity of synthetic data inferences depends critically on 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 certain relationships accurately, analysts' inferences also will not reflect those relationships. Similarly, incorrect distributional assumptions built into the models are passed on to the users' analyses. Practically, this dependence means that some analyses cannot be performed accurately, and that agencies need to release information that helps analysts decide whether or not the synthetic data are reliable for their analyses. For example, agencies can include the models as attachments to public releases of data. Or, they can include generic descriptions of the imputation models, such as "Main effects and first order interactions for all other variables are included in the imputation models for income." Another approach is for the agency to build a verification server (Reiter *et al.*, 2009) that users can query for feedback on the differences between the results of analyses done on the synthetic data and the genuine data. Analysts who desire finer detail than afforded by the imputations may have to apply for special access to the observed data.

2.3 Partially synthetic data

To reduce the sensitivity of inferences to the specifications of the imputation models, some statistical agencies have opted for a variant of Rubin's approach called partially synthetic data (Little, 1993; Reiter, 2003). These comprise the units originally surveyed with only some collected values replaced with multiple imputations. For example, the agency might simulate sensitive variables or quasi-identifiers for units in the sample with rare combinations of quasi-identifiers; or, the agency might replace all data for selected sensitive variables or quasi-identifiers. The former strategy has been employed by the U.S. Federal Reserve Board in the Survey of Consumer Finances. They replace monetary values at high disclosure risk with multiple imputations, releasing a mixture of these imputed values and the unreplaced, collected values (Kennickell, 1997). It also is used by the U.S. Bureau of the Census to protect the identities of people in group quarters (e.g., prisons, shelters) in the American Communities Survey. They replace quasi-identifiers for records at high disclosure risk with imputations. The latter strategy has been employed by the U.S. Bureau of the Census to protect data in longitudinal, linked business datasets. They replace all values of some sensitive variables with multiple imputations and leave other variables at their actual values (Abowd and Woodcock, 2001, 2004). It also has been used to create synthesized origin-destination matrices, i.e. where people live and work, available to the public as maps via the web (On The Map, http://lehdmap.did.census.gov/). In the U.S., partially synthetic, public use datasets are in the development stage for the

Survey of Income and Program Participation, the Longitudinal Business Database, the Longitudinal Employer-Household Dynamics survey, and the American Communities Survey veterans and full sample data. Other examples of partially synthetic data are in Abowd and Lane (2004), Little *et al.* (2004), Reiter (2004, 2005c), and Mitra and Reiter (2006).

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 with reasonable accuracy from the synthetic values and any information released about the data generation model.

The methods for generating partially synthetic data depend on whether there is any missing data, i.e. whether or not $D_{com} = D_{obs}$.

2.3.1 Partially synthetic data when $D_{com} = D_{obs}$

Assuming no missing data, the agency constructs partially synthetic datasets by replacing selected values from the observed data with imputations. Let $R_j = 1$ if unit

j is selected to have any of its observed data replaced with synthetic values, and let $R_j = 0$ for those units with all data left unchanged. Let $R = (R_1, \ldots, R_n)$. Let $D_{rep}^{(i)}$ be all the imputed (replaced) values in the ith synthetic dataset, and let D_{nrep} be all unchanged (unreplaced) values of D_{com}. The $D_{rep}^{(i)}$ are generated from the posterior predictive distribution of $(D_{rep} \mid D_{com}, R)$, or a close approximation of it. The values in Y_{nrep} are the same in all synthetic datasets. Each synthetic dataset, $D^{(i)}$, then comprises $(D_{rep}^{(i)}, D_{nrep}, R)$. Imputations are made independently $i = 1, \ldots, r$ times to yield r different partially synthetic data sets, which are released to the public.

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

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

to estimate the variance of $\bar{q}_r$. Inferences for scalar Q are based on t-distributions with degrees of freedom $\nu_p = (r-1)(1 + \bar{v}_r/(b_r/r))^2$. The variance estimator in (5) differs from the one in Rubin (1987) for standard missing data, because the imputations are done on a complete dataset rather than a dataset with missing values; see Reiter and Raghunathan (2007) for further details.

2.3.2 Partially synthetic data when $D_{com} \neq D_{obs}$

When some data are missing, it seems logical to impute the missing and partially synthetic data simultaneously. In general, D_{mis} and D_{rep} are imputed from different distributions. For example, suppose univariate data from a normal distribution have some values missing completely at random. Further, suppose the agency seeks to replace all values larger than some threshold with imputations. The imputations for missing data are based on a normal distribution fit using all of D_{obs}. However, the imputations for replacements must be based on a posterior distribution that conditions on values being larger than the threshold.

Imputing D_{mis} and D_{rep} separately generates two sources of variability, in addition to the sampling variability in D_{com}, that the user must account for to obtain valid inferences. To allow analysts to estimate the total variability correctly, agencies can employ a two stage procedure for generating imputations. First, the agency fills in D_{mis} with draws from $f(D_{mis} \mid D_{obs})$, resulting in m completed datasets, $D^{(1)}, \ldots, D^{(m)}$. Second, the agency selects the units whose values are to be replaced, i.e. whose $R_j = 1$. In each $D^{(l)}$, the agency imputes values $Y_{rep}^{(l,i)}$ for those units with $R_j = 1$, using $f(D_{rep} \mid D^{(l)}, R)$. 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, $D^{(l,i)} = (D_{nrep}, D_{mis}^{(l)}, D_{rep}^{(l,i)}, R)$, includes a label indicating the l of the $D^{(l)}$ from which it was drawn. These M datasets are released to the public.

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 $D^{(l,i)}$ was in fact collected data from a random sample of D based on the original sampling design. For $l = 1, \ldots, m$ and $i = 1, \ldots, r$, let $q^{(l,i)}$ and $u^{(l,i)}$ be respectively the values of q and v in dataset $D^{(l,i)}$. The following quantities are needed for inferences about scalar Q:

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

$$\bar{w}_M = \sum_{l=1}^{m} \sum_{i=1}^{r} (q^{(l,i)} - \bar{q}^{(l)})^2/m(r-1) = \sum_{l=1}^{m} w^{(l)}/m \qquad (7)$$

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

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

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

$$T_M = (1 + 1/m)b_M - \bar{w}_M/r + \bar{v}_M. \qquad (10)$$

Inferences are based on the t-distribution, $(Q - \bar{q}_M) \sim t_{\nu_M}(0, T_M)$, with degrees of freedom

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

Reiter and Raghunathan (2007) explain why this two stage approach requires a different variance estimator than (5). Reiter (2008) discusses the effects on inferences of different selections of m and r.

3 Synthetic data for integration and dissemination

In the data integration and dissemination context, confidentiality can be at risk because of inter-agency data sharing or dissemination of the integrated file. For simplicity, we assume that Agency 1 shares the identifiers of the records in D_1 with Agency 2 to facilitate record linkage. Hence, the primary risk for both agencies at the integration stage is that one agency could learn sensitive attribute values of the records owned by the other agency. Methods of record linkage that do not involve direct sharing of identifiers are described by Churches and Christen (2004) and O'Keefe *et al.* (2004). When D_1 and D_2 have no common records, as in statistical matching, Agency 1 need not share identifiers with Agency 2.

We categorize integration settings in two cases. Case 1 occurs when the two agencies are willing to share all their values with each other, so that confidentiality risks arise only when disseminating the integrated data to the public. Case 2 occurs when at least one agency is unwilling to share its data with the other agency. We describe methods from the perspective of Agency 1, i.e. it initiates all integration protocols and is responsible for releasing a public use version of the integrated data. Table 1 summarizes the scenarios that we consider.

3.1 Case 1: Full access

When both agencies are willing to share all data with each other, the integration is straightforward. The only task is for Agency 1 to disseminate a safe dataset to the

Table 1: Summary of methods with relevant section numbers in parentheses. In the table, F means full synthesis; P means partial synthesis; and M means missing data imputation. Subscripts indicate which dataset is synthesized; no subscript means both datasets are synthesized/completed. For example, MP_2 means use multiple imputation for any missing data in (D_1, D_2), followed by partial synthesis of sensitive values in D_2. The * means that Agency 1 protects D_1 before inter-agency sharing.

Scenario	Add-on setting	Complete-it setting
Case 1: Inter-agency sharing of D_2 and D_1 (3.1)		
No real data for public	F (2.2)	F
Some real data for public	P (2.3.1)	MP (2.3.2)
Case 2, Scenario A: Limited sharing of D_2, unrestricted sharing of D_1 (3.2.2)		
All D_1 for agency and public	P_2	$D_1 \cap D_2 \neq \emptyset$: MP_2
		$D_1 \cap D_2 = \emptyset$: $\mathrm{P}_2\mathrm{M}$
Case 2, Scenario B: Limited sharing of D_2, limited sharing of D_1 (3.2.3)		
All D_1 for agency, some for public	$\mathrm{P}_2\mathrm{P}_1$	$D_1 \cap D_2 \neq \emptyset$: $\mathrm{MP}_2\mathrm{P}_1$
		$D_1 \cap D_2 = \emptyset$: $\mathrm{P}_2\mathrm{MP}_1$
Some D_1 for agency, some for public	$\mathrm{P}_2\mathrm{P}_1^*$	$D_1 \cap D_2 \neq \emptyset$: $\mathrm{MP}_2\mathrm{P}_1^*$
		$D_1 \cap D_2 = \emptyset$: $\mathrm{P}_2\mathrm{MP}_1^*$
Case 2, Scenario C: No sharing of D_2, no sharing of D_1 (3.2.4)		
No real data for public	$\mathrm{P}_2\mathrm{F}^*$	$D_1 \cap D_2 \neq \emptyset$: $\mathrm{MP}_2\mathrm{F}^*$
		$D_1 \cap D_2 = \emptyset$: $\mathrm{P}_2\mathrm{MF}^*$

public using the methods described in Section 2.

If some but not all values are sensitive to disclose, Agency 1 can release a partially synthetic dataset. For the add-on setting, this involves the process and inferential methods described in Section 2.3.1, or the methods in Section 2.3.2 if there are missing values in the integrated D_{com}. For the complete-it setting, this involves the process and inferential methods described in Section 2.3.2. Multiple imputation is used to create D_{com} first, and sensitive values are replaced second.

If all values are sensitive to disclose, Agency 1 can release a fully synthetic dataset. This is the same process regardless of the integration setting. Agency 1 specifies a distribution for all variables based on (D_1, D_2), and whatever other constraints exist, and simulates new data for all variables. Analysts of these data use the inferential methods described in Section 2.2.

3.2 Case 2: Limited Access

The first case presumes no inter-agency concerns about their database subjects' confidentiality. We now turn to the more complex case where these concerns exist.

3.2.1 Synthetic data for vertically partitioned data

We first describe an approach for the add-on setting where D_1 and D_2 have exactly the same records but different variables, known as vertical partitioning (Karr *et al.*, 2007). This approach can be adapted for more general add-on scenarios. The approach was developed by Kohnen (2005) and illustrated by Kohnen and Reiter (2009). We assume no missing data in D_1 or D_2; methods for handling missing data and integration simultaneously are a subject for future research.

To start the protocol, Agency 1 sends a masked version of D_1 to Agency 2 that protects the confidentiality of any sensitive values in D_1. For example, Agency 1 might apply standard disclosure limitation techniques to sensitive values in D_1 like adding noise. Or, Agency 1 might transform the variables with sensitive data to standard normal distributions, for example via Box-Cox transformations; see Kohnen and Reiter (2009) for an application of this approach. A third masking approach was described by Kohnen (2005). Agency 1 creates $k-1$ "disguiser" copies of D_1, for example by adding different amounts of noise to the sensitive values of D_1. Agency 1 then includes D_1 with the disguisers and passes the collection of all k datasets to Agency 2, which is not told which dataset is the genuine D_1. With good disguisers, Agency 2 has only a $1/k$ chance of guessing the label of the true D_1 and hence learning the exact values of these records' sensitive attributes.

In the next step of the protocol, Agency 2 determines which values of D_2 are too sensitive to reveal to Agency 1. Let $D_{2,rep}$ indicate those values, which could be some or all of D_2. Let k be the number of datasets that Agency 2 receives from Agency 1. If it receives a single dataset with perturbed/transformed variables, then $k = 1$. For each dataset $D_1^{(l)}$ that it receives, where $l = 1, \ldots, k$, Agency 2 estimates the distribution $f(D_{2,rep}|D_1^{(l)}, D_2)$. Agency 2 then simulates new values of $D_{2,rep}$ from these distributions as in Section 2.3.1, generating m datasets with each $D_1^{(l)}$. Agency 2 passes the km datasets back to Agency 1. When $k > 1$, it also includes labels indicating which $D_1^{(l)}$ each was created from. Agency 2 could pass the k imputation models to Agency 1 rather than generate data, although with complicated models it might be easier and pose fewer confidentiality risks to pass simulations from those models. If $k > 1$, Agency 1 discards the $k - 1$ disguiser datasets it receives from Agency 2. In the end, Agency 1 is left with partially synthetic data, $\{D^{(1)}, \ldots, D^{(m)}\}$, including actual values of D_1 and simulated values of $D_{2,rep}$.

The agencies may not be willing to share these partially synthetic datasets with the public without further protections. Furthermore, releasing a public use version of the partially synthetic datasets, or even publishing results of analyses based on them, could disclose sensitive values in D_1 to Agency 2. To limit these risks, Agency 1 can release synthetic data, treating each $D^{(l)}$ as the "observed data" from which to synthesize. These datasets can be fully or partially synthetic.

For the fully synthetic case, in each $D^{(l)}$ the agency completes the population by filling in D_{exc} with r independent draws from $f(D_{exc}|D^{(l)})$. For each $D^{(l)}$, the agency then takes a simple random sample from each completed population. The agency releases the mr datasets to the public. Under conditions described in Kohnen (2005), the analyst can use $\bar{q}_M$ from (6) to estimate Q. An estimate of the variance of $\bar{q}_M$ is

$$T_{sf} = b_M/m + \bar{w}_M - \bar{v}_M, \tag{12}$$

where b_M, $\bar{w}_M$, and $\bar{v}_M$ are defined in (7)–(10). This variance estimator differs from

(4) because Agency 1 generates synthetic data from already synthesized data; see Kohnen and Reiter (2009) for derivations. When n, m, and r are large, inferences can be based on the normal distribution, $(Q - \bar{q}_M) \sim N(0, T_{sf})$.

For the partially synthetic case, in each $D^{(l)}$ the agency replaces any sensitive values in D_1 with r independent draws from $f(D_{rep}|D^{(l)}, R)$. Kohnen (2005) shows that the analyst can estimate Q with $\bar{q}_M$, and estimate of the variance of $\bar{q}_M$ with

$$T_{sp} = \bar{v}_M + b_M/m \tag{13}$$

where b_M and $\bar{v}_M$ are defined in (7)–(10). When n, m, and r are large, inferences can be based on the normal distribution, $(Q - \bar{q}_M) \sim N(0, T_{sp})$.

3.2.2 Scenario A: Some values in D_2 are sensitive, but D_1 is not

Suppose that Agency 2 is not willing to share some values in D_2 with Agency 1 nor with the public. Agency 1 is willing to share D_1 with Agency 2 and with the public. As an example, Agency 1 might have non-sensitive demographic values, and Agency 2 might have sensitive health or economic data. We note that the opposite scenario, i.e. Agency 2 is willing to share all but Agency 1 is not, is included in Case 1.

For the add-on setting, Agency 1 sends D_1 to Agency 2. Agency 2 then generates r replacements of its sensitive values, $Y_{2,rep}$, by drawing from $f(Y_{2,rep}|D_1, D_2)$. This results in r partially synthetic versions of D_{com}. Agency 2 sends the r versions to Agency 1, which releases them to the public. When D_{com} has missing values, Agency 2 uses the two-stage process of Section 2.3.2 before passing data to Agency 1. Analysts of the released data use the corresponding inferential methods of Section 2.3.2.

For the complete-it setting, the protocol depends on whether or not some of the same records are in D_1 and D_2. If so, Agency 1 sends D_1 to Agency 2, who then generates imputations as in Section 2.3.2. First, Agency 2 creates m completed versions of D_{com} using standard missing data techniques. Second, in each completed dataset $D_{com}^{(i)}$, Agency 2 replaces sensitive values, $D_{2,rep}$, with r draws from $f(D_{2,rep}|D_{com}^{(i)})$. Agency 2 sends the mr datasets to Agency 1, who releases them to the public. Analysts of these mr datasets base inferences on the methods in Section 2.3.2.

When there are no overlapping records, Agency 1 does not send anything to Agency 2. Instead, Agency 2 simulates new values for its sensitive elements, $D_{2,rep}$, by drawing from $f(D_{2,rep}|D_2)$. Agency 2 does this m times, creating $D_2^{(l)} = (D_{2,nrep}, D_{2,rep}^{(l)})$, for $l = 1, \ldots, m$. It sends these m copies of $D_2^{(l)}$ to Agency 1. Agency 1 appends each copy to D_1, creating $D^{(l)} = (D_1, D_2^{(l)})$ for $l = 1, \ldots, m$. Agency 1 then fills in the values of D_{mis} in each $D^{(l)}$ using statistical matching techniques (Rässler, 2003; D'Orazio *et al.*, 2006), creating r multiple imputations for each $D^{(l)}$. These mr imputations are released to the public. This nesting structure resembles the one in Section 2.3.2; however, the replacement data is simulated before the missing data. This ordering differs from that used in Section 2.3.2, which implies that new methods of combining the point and variance estimates are needed for valid inference with this approach.

3.2.3 Scenario B: Some values in D_1 and D_2 are sensitive

This scenario assumes that Agency 2 follows the same behavior as in Scenario A: it won't share all of its data with anyone. Agency 1 now views some of its data as sensitive. We consider two possibilities for Agency 1, specifically (i) it is willing to share all values of D_1 with Agency 2 but only some values with the public, and (ii) it is willing to share only some values of D_1 with Agency 2 and with the public.

For the add-on setting where Agency 1 is willing to share all of D_1 with Agency 2, the agencies proceed as in the add-on setting for Scenario A. Once Agency 1 gets the partially synthetic datasets from Agency 2, it simulates new values of its sensitive data, $D_{1,rep}$, from $f(D_{1,rep}|D^{(l)})$. When there are no missing values, this is identical to the final stage of the protocols in Section 3.2.1. If the agencies are not willing to release any values to the public, Agency 1 generates fully synthetic data, and secondary data analysts base inferences on (12). If the agencies are willing to release some of D_1 or D_2 to the public, Agency 1 generates partially synthetic data, and analysts base inferences on (13). When there are missing data, Agency 2 could use the two stage approach of Section 2.3.2 to generate datasets. Agency 1 would then generate fully or partially synthetic data based on each dataset it gets from Agency 2. Effectively, this creates a three stage imputation approach. Currently, there are no methods of inference for three stage imputation approaches. We note that the variance formulas in (12) and (13) are not correct, since they presume Agency 2 imputes replacements in one stage.

For the add-on setting where Agency 1 shares only some of D_1 with Agency 2, Agency 1 can adopt the protocols of Section 3.2.1. That is, it applies disclosure protection to D_1 before passing it to Agency 2. Agency 2 then applies the strategies described in the previous paragraph using each of the k datasets sent by Agency 1. Agency 1 discards any disguisers (if $k > 1$) and proceeds as in the previous paragraph.

For the complete-it setting where Agency 1 is willing to share with Agency 2 but not the public, the agencies proceed as in the approaches for Scenario A. When the records in D_1 and D_2 overlap, after receiving the mr datasets from Agency 2, Agency 1 makes the public use file by generating replacement values for the sensitive $D_{1,rep}$ in all mr datasets, $D^{(l,i)}$ for $l = 1, \ldots, m$ and $i = 1, \ldots, r$, with draws from $f(D_{1,rep}|D^{(l,i)})$. This could be done multiple times in each $D^{(l,i)}$, effectively creating a three stage imputation procedure.

When records overlap and Agency 1 is not willing to share with Agency 2, Agency 1 can apply disclosure protection to D_1 before passing it to Agency 2. Agency 2 then follows the two-stage process of Section 2.3.2 on each of the k datasets passed by Agency 1, passing the entire collection back to Agency 1. When $k > 1$, Agency 1 discards the disguiser datasets. Agency 1 then simulates new values of sensitive D_1 for the public use file by simulating from $f(D_{1,rep}|D^{(l,i)})$, as in the previous paragraph.

For the complete-it setting without overlapping records, Agency 1 and Agency 2 proceed as in the corresponding approaches for Scenario A with one modification. Before release to the public, Agency 1 simulates the sensitive values of D_1. This approach requires new inferential procedures.

3.2.4 Scenario C: All values in D_1 and D_2 are sensitive

This scenario assumes that both agencies are unwilling to share any values with each other or with the public. For the add-on setting, Agency 1 can initiate the full synthesis protocol in Section 3.2.1. For the complete-it setting, Agency 1 and Agency 2 proceed as in Scenario A. However, Agency 1 now simulates all data values before release. Once again, this additional simulation requires new inferential procedures.

4 Illustration of data sharing protocol

We now illustrate one of the synthetic data approaches to data integration and dissemination. In particular, we presume two agencies own the same records but different variables, which is the setting of Section 3.2.1. For the illustration we use a subset of public release data from the March 2000 U.S. Current Population Survey. The data comprise nine variables measured on 51,016 heads of households; see Table 2. Similar data are used by Reiter (2005a) to illustrate and evaluate releasing fully synthetic data without data integration.

We partition the data so that Agency 1 owns all variables except income and Agency 2 owns only income. We presume that Agency 1 is willing to share the values of all variables with Agency 2, but Agency 2 is not willing to share any genuine values of income with Agency 1. Thus, Agency 1 does not disguise the data that it sends to

Table 2: Description of variables used in the empirical studies

Variable	Label	Range
Sex	X	male, female
Race	R	white, black, American Indian, Asian
Marital status	M	7 categories, coded 1–7
Highest attained education level	E	16 categories, coded 31–46
Age (years)	G	0 – 90
Number of people in household	H	1 – 16
Number of people in household under age 18	Y	0, 1 – 11
Household property taxes ($)	P	0, 1 – 99,997
Household income ($)	I	-21,011 – 768,742

Agency 2. Agency 2 disguises what it sends to Agency 1 by simulating all records' incomes. See Kohnen and Reiter (2009) for an illustration of techniques that enable Agency 1 to disguise its data before passing to Agency 2.

We also presume that Agency 1 releases a public use file of the integrated data that protects the confidentiality of respondents' identities. We consider age, race, marital status, and sex to be quasi-identifiers that intruders can know precisely. To make the public use file, Agency 1 simulates all records' values of age, race, and marital status in the integrated data. This is arguably more data synthesis than necessary. There are only 521 records with unique combinations of age, race, marital status, and sex; and, there are only 284 combinations of the four variables with two cases. Thus, to protect confidentiality it may be sufficient for Agency 1 to simulate the quasi-identifiers for only a subset of the full sample. Nonetheless, we simulate all values to illustrate heavy synthesis. Intruders might have access to property taxes, in which case Agency 1 may want to simulate those variables as well.

We generate synthetic datasets in two stages. First, Agency 2 replaces all values of income and passes $m = 5$ partially synthetic datasets to Agency 1. Second, in each of these five datasets, Agency 1 generates $r = 5$ datasets in which age, marital status, and race are synthesized. Agency 1 does not change the values of sex. These 25 synthetic datasets are what would be released to the public. The synthetic data are generated using regression trees (CART models), as we now describe.

4.1 CART Synthesis Models

CART models are a flexible tool for estimating the conditional distribution of a univariate outcome given multivariate predictors. Essentially, the CART model partitions the predictor space so that subsets of units formed by the partitions have relatively homogeneous outcomes. The partitions are found by recursive binary splits of the predictors. The series of splits can be effectively represented by a tree structure, with leaves corresponding to the subsets of units. Reiter (2005c) describes how CART models can be used to generate partially synthetic data.

To synthesize all values of income, we first separate the data into four groups defined by the four distinct values of race. In each group, using D_{com} we fit a regression tree of income on all other variables (except race). Label the tree in any group as $\mathcal{Y}_{(I)}$. We require a minimum of five records in each leaf of the trees and use a minimum deviance splitting criterion of 0.001; see Reiter (2005c) for discussion of specifying tree parameters. Let L_{Iw} be the wth leaf in some $\mathcal{Y}_{(I)}$, and let $Y_{(I)}^{L_{Iw}}$ be the $n_{L_{Iw}}$ values of $Y_{(I)}$ in leaf L_{Iw}. In each L_{Iw} in the tree, we generate a new set of values by drawing from $Y_{(I)}^{L_{Iw}}$ using the Bayesian bootstrap (Rubin, 1981). These sampled values are the replacement imputations for the $n_{L_{Iw}}$ units that belong to L_{Iw}. Repeating the Bayesian bootstrap in each leaf of the income trees results in the ith set of synthetic ages, $Y_{(I)\text{rep},i}$. We repeat this process $m = 5$ times to generate the five partially synthetic datasets, $\{D^{(1)}, \ldots, D^{(5)}\}$, to be shared with Agency 1.

To avoid releasing values of the observed incomes in each leaf, we could take an additional step suggested in Reiter (2005c). In each leaf, we would estimate the density of the bootstrapped values using a Gaussian kernel density estimator with support over the smallest to the largest value of $Y_{(I)}$. Then, for each unit, we would sample randomly from the estimated density in that unit's leaf using an inverse-cdf method. We do not take this extra step here.

For the synthesis of age, race, and marital status, we proceed sequentially. First, in each $D^{(i)}$, we fit the age tree, $\mathcal{Y}_{(Gi)}$, with all variables except race and marital status as predictors. In each L_{Giw} in the age tree, we generate a new set of values by drawing from $Y_{(Gi)}^{L_{Giw}}$ using the Bayesian bootstrap. We next simulate values of marital status. In each $D^{(i)}$, we fit the marital status tree with all variables except race as predictors. To maintain consistency with $Y_{(Gi)\text{rep},i}$, units' leaves in $\mathcal{Y}_{(Mi)}$ are located using $Y_{(G)\text{rep},i}$. Occasionally, some units may have combinations of values that do not belong to one of the leaves of $\mathcal{Y}_{(Mi)}$. For these units, we search up the tree until we find a node that contains the combination, then treat that node as if it were the unit's leaf. Once each unit's leaf is located, values of $Y_{(M)\text{rep},i}$ are generated using the Bayesian bootstrap. Imputing races follows the same process: we fit the tree $\mathcal{Y}_{(Ri)}$ using all variables as predictors, place each unit in the leaves of $\mathcal{Y}_{(Ri)}$ based on their synthesized values of age and marital status, and sample new races using the Bayesian bootstrap. The entire process is repeated independently $r = 5$ times for each $D^{(i)}$, resulting in $mr = 25$ datasets that would be released to the public.

All CART models are fit in S-Plus using the "tree" function. It takes about two hours of computer time to generate all 25 synthetic datasets. The sequence of imputations is $G-M-R$; see Reiter (2005c) for a discussion of imputation sequencing.

4.2 Analytic usefulness

We now illustrate the analytic usefulness of the resulting synthetic datasets. We estimate the coefficients in a regression of the logarithm of income on a function of all the predictors, including non-linear effects in age and interactions among marital status and sex. Table 3 summarizes the inferences for the coefficients when fitting the model on the observed data, on the first stage of synthetic data with only income replaced, and on the final stage of synthetic data with income, age, marital status, and race replaced. In general, the estimated coefficients and 95% confidence intervals are similar across all three sources of data, even for the interactions and non-linear effects. The least accurate synthetic coefficients involve people who are widowed males and who are married in the armed forces. There are relatively small numbers of people in these categories: 1012 are widowed males and 773 people are married in the armed forces. Agency 2 could improve inferences by fitting separate trees for these groups. In general, Agency 2 should compare synthetic and observed data inferences for many representative analyses, and adjust the synthesizer when large differences exist. Agency 1 can improve the synthesis by comparing inferences from the mr datasets to those from the m datasets sent by Agency 2.

Of course, the results in Table 3 are specious if the synthesis does not substantially alter the original data. To investigate this, for each person j in the data we estimate the actual income, I_j, as the average of the five synthetic incomes, $\hat{I}_j$. We then compute the absolute relative prediction error for each person, $|(\hat{I}_j - I_j|)/(I_j + 0.5)|$, where the 0.5 is added to avoid division by zero. The median and first quartile of these relative prediction errors are 0.72 and 0.18 respectively, indicating the synthetic data averages tend to differ substantially from the actual incomes.

We next examine the quasi-identifiers. We estimate each person's actual age as the most frequently occurring value among that unit's 25 imputations. We similarly estimate each person's actual marital status and race. We then compare each person's

Table 3: Point estimates and 95% confidence intervals for coefficients in regression of $\log(I)$ using observed data, synthetic data with only income replaced, and synthetic data with income, age, marital status, and race replaced.

		Synthetic Data	
Estimand	Observed Data	I Only	I, A, M, R
Intercept	4.9 (4.8, 5.0)	5.1 (4.9, 5.3)	5.1 (4.8, 5.3)
Black	-.17 (-.19, -.15)	-.18 (-.21, -.15)	-.18 (-.21, -.15)
American Indian	-.25 (-.31, -.18)	-.27 (-.34, -.19)	-.23 (-.35, -.12)
Asian	-.01 (-.05, .04)	.01 (-.03, .06)	.01 (-.05, .06)
Female	.00 (-.02, .03)	.00 (-.02, .03)	-.00 (-.04, .03)
Married in armed forces	-.03 (-.11, .06)	-.10 (-.22, .01)	-.13 (-.20, -.07)
Widowed	-.02 (-.07, .04)	-.10 (-.17, -.03)	-.12 (-.20, -.05)
Divorced	-.16 (-.20, -.13)	-.19 (-.24, -.15)	-.20 (-.25, .15)
Separated	-.24 (-.31, -.17)	-.23 (-.32, -.15)	-.27 (-.36, -.19)
Single	-.17 (-.20, -.14)	-.16 (-.21, -.12)	-.16 (-.20, -.11)
Education	.11 (.108, .113)	.11 (.103, .110)	.11 (.101, .114)
Household size > 1	.50 (.48, .52)	.49 (.45, .52)	.48 (.43, .54)
Females married in armed forces	-.52 (-.64, -.41)	-.35 (-.48, -.22)	-.32 (-.48, -.16)
Widowed females	-.31 (-.37, -.25)	-.26 (-.32, -.19)	-.24 (-.34, .-14)
Divorced females	-.31 (-.35, -.26)	-.30 (-.35, -.25)	-.26 (-.35, -.17)
Separated females	-.52 (-.61, -.43)	-.43 (-.53, -.33)	-.38 (-.53, -.23)
Single females	-.32 (-.36, -.28)	-.29 (-.34, -.24)	-.29 (-.36, -.21)
Age $\times 10$	.43 (.41, .46)	.41 (.38, .44)	.41 (.38, .43)
Age^2 $\times 1000$	-.44 (-.47, -.42)	-.41 (-.44, -.38)	-.41 (-.43, -.38)
Property tax $\times 10000$	.37 (.34, .40)	.38 (.35, .41)	.38 (.35, .41)

Income regression fit using records with $I > 0$.

estimated quasi-identifiers to their counterparts in the original data. The values in the synthetic and original data exactly match on all three variables for only 7% of the records, indicating that age, martial status, and race are substantially altered. For a formal approach to measuring identification disclosure risks in synthetic data, see Drechsler and Reiter (2008).

5 Concluding remarks

This article proposes approaches for integrating and disseminating data using multiple imputation. The data sharing proposals have not been implemented in practice, and there are many challenges to doing them well. The confidentiality protection for methods based on disguisers is sensitive to the properties of the disguisers. Unrealistic disguisers may not afford any protection, and detailed domain knowledge on the part of the receiving agency (Agency 2) can defeat them. For example, when D_1 contains incomes, Agency 2 can identify D_1 if it knows the exact income of at least one record on the file, and there are no duplicates of that income in the disguisers datasets. These risks lead Kohnen and Reiter (2009) to conclude that perturbing or transforming data generally provides greater protection than sharing multiple disguisers.

Creating synthetic data, especially full synthesis, is challenging for large datasets with many variables. With large amounts of simulation, results are sensitive to assumptions used in the synthesis models. Nonetheless, as the illustration and published applications show, it is possible to generate synthetic data that are analytically useful for a wide class of (but not all) estimands. We also note that in complete-it data integration settings with a relatively small amount of overlap in records, agencies have to make strong assumptions about the distributions of the data regardless of confidentiality concerns, as there is little information about the relationships among variables. Such assumptions can be used to generate synthetic data in those settings.

Legal and ethical concerns over data sharing and dissemination seem to be only growing. In the future, it is conceivable that agencies may not be allowed to share or release any genuine data. Yet, there are potentially enormous benefits to agencies of integrating data from different sources, and to the broader public if agencies disseminate the integrated data. The techniques proposed in this article have the potential to handle data integration and dissemination simultaneously while respecting confidentiality constraints. The next steps in developing these methods are clear: derive methods for inferences where needed, and investigate disclosure risks and analytical validity at both the inter-agency and dissemination stages. Empirical investigations on genuine data will help agencies and analysts to understand the benefits and limitations of multiple imputation approaches to data integration and dissemination.

References

Abowd, J. M. and Lane, J. I. (2004). New approaches to confidentiality protection: Synthetic data, remote access and research data centers. In J. Domingo-Ferrer and V. Torra, eds., *Privacy in Statistical Databases*, 282–289. New York: Springer-

Verlag.

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.

Abowd, J. M. and Woodcock, S. D. (2004). Multiply-imputing confidential characteristics and file links in longitudinal linked data. In J. Domingo-Ferrer and V. Torra, eds., *Privacy in Statistical Databases*, 290–297. New York: Springer-Verlag.

Agrawal, R. and Srikant, R. (2000). Privacy-preserving data mining. In *Proceedings of the 2000 ACM SIGMOD on Management of Data*, 439–450. Dallas: ACM Press.

Churches, T. and Christen, P. (2004). Some methods for blindfolded record linkage. *BMC Medical Informatics and Decision Making* **4**, 9, Available at http://www.pubmedcentral.nih.gov/tocrender.fcgi?iid=10563.

D'Orazio, M., Di Zio, M., and Scanu, M. (2006). *Statistical Matching: Theory and Practice*. New York: Wiley.

Drechsler, J., Dundler, A., Bender, S., Rässler, S., and Zwick, T. (2007). A new approach for disclosure control in the IAB establishment panel–Multiple imputation for a better data access. Tech. rep., IAB Discussion Paper, No.11/2007.

Drechsler, J. and Reiter, J. P. (2008). Accounting for intruder uncertainty due to sampling when estimating identification disclosure risks in partially synthetic data. In J. Domingo-Ferrer and Y. Saygin, eds., *Privacy in Statistical Databases (LNCS 5262)*, 227–238. New York: Springer-Verlag.

Du, W., Han, Y., and Chen, S. (2004). Privacy-preserving multivariate statistical analysis: Linear regression and classification. In *Proceedings of the Fourth SIAM Conference on Data Mining*, 222–233.

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.

Evfimievski, A., Srikant, R., Agrawal, R., and Gehrke, J. (2004). Privacy-preserving mining of association rules. *Information Systems* **29**, 343–364.

Fienberg, S. E. (1994). A radical proposal for the provision of micro-data samples and the preservation of confidentiality. Tech. rep., Department of Statistics, Carnegie-Mellon University.

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

Ghosh, J., Reiter, J. P., and Karr, A. F. (2007). Secure computation with horizontally partitioned data using adaptive regression splines. *Computational Statistics and Data Analysis* **51**, 5813–5820.

Gomatam, S., Karr, A. F., Reiter, J. P., and Sanil, A. P. (2005). Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access servers. *Statistical Science* **20**, 163–177.

Graham, P. and Penny, R. (2005). Multiply imputed synthetic data files. Tech. rep., University of Otago, http://www.uoc.otago.ac.nz/departments/pubhealth/pgrahpub.htm.

Kantarcioglu, M. and Clifton, C. (2002). Privacy-preserving distributed mining of association rules on horizontally-partitioned data. In *Proceedings of Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 24–31. Edmonton, Canada: ACM Press.

Karr, A. F., Fulp, W. J., Lin, X., Reiter, J. P., Vera, F., and Young, S. S. (2007). Secure, privacy-preserving analysis of distributed databases. *Technometrics* **49**, 335–345.

Karr, A. F., Lin, X., Reiter, J. P., and Sanil, A. P. (2009). Privacy preserving analysis of vertically partitioned data using secure matrix protocols. *Journal of Official Statistics* **25**, 125–138.

Karr, A. F., Lin, X., Sanil, A. P., and Reiter, J. P. (2005). Secure regressions on distributed databases. *Journal of Computational and Graphical Statistics* **14**, 263–279.

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.

Kohnen, C. N. (2005). *Using Multiply-Imputed, Synthetic Data to Facilitate Data Sharing*. Ph.D. thesis, Duke University, Institute of Statistics and Decision Sciences.

Kohnen, C. N. and Reiter, J. P. (2009). Multiple imputation for combining confidential data owned by two agencies. *Journal of the Royal Statistical Society, Series A* **172**, 511–528.

Lin, X., Clifton, C., and Zhu, Y. (2004). Privacy-preserving clustering with distributed EM models. *Knowledge and Information Systems* **8**, 68–81.

Lindell, Y. and Pinkas, B. (2000). Privacy-preserving data mining. In *Advances in Cryptology: CRYPTO2000*, 36–54. New York: Springer-Verlag.

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

Little, R. J. A., Liu, F., and Raghunathan, T. E. (2004). Statistical disclosure techniques based on multiple imputation. In A. Gelman and X. L. Meng, eds., *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Persepectives*, 141–152. New York: John Wiley & Sons.

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

Mitra, R. and Reiter, J. P. (2006). Adjusting survey weights when altering identifying design variables via synthetic data. In J. Domingo-Ferrer and L. Franconi, eds., *Privacy in Statistical Databases*, 177–188. New York: Springer-Verlag.

O'Keefe, C. M., Yung, M., Gu, L., and Baxter, R. (2004). Privacy preserving data linkage protocols. In V. Atluri, P. Syverson, and S. De Capitani di Vimercati, eds., *Proceedings of the 2004 ACM Workshop on Privacy in the Electronic Society*, 94–102. Washington, DC: ACM Press.

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.

Rässler, S. (2003). A non-iterative Bayesian approach to statistical matching. *Statistica Neerlandica* **57**, 58–74.

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. (2004). 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* **21**, 441–462.

Reiter, J. P. (2008). Selecting the number of imputed datasets when using multiple imputation for missing data and disclosure limitation. *Statistics and Probability Letters* **78**, 15–20.

Reiter, J. P., Oganian, A., and Karr, A. F. (2009). Verification servers: Enabling analysts to assess the quality of inferences from public use data. *Computational Statistics and Data Analysis* **53**, 1475–1482.

Reiter, J. P. and Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association* **102**, 1462–1471.

Rubin, D. B. (1981). The Bayesian bootstrap. *The Annals of Statistics* **9**, 130–134.

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.

Shlomo, N. (2007). Statistical disclosure control for census frequency tables. *International Statistical Review* **75**, 199–217.

Sweeney, L. (1997). Computational disclosure control for medical microdata: the Datafly system. In *Proceedings of an International Workshop and Exposition*, 442–453.

Vaidya, J. and Clifton, C. (2002). Privacy-preserving association mining over vertically-partitioned data. In *Proceedings of Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 639–644. Edmonton, Canada: ACM Press.

Vaidya, J. and Clifton, C. (2003). Privacy-preserving k-means clustering over vertically-partitioned data. In *Proceedings of Ninth ACM SIGKDD International*

Conference on Knowledge Discovery and Data Mining, 206–215. Washington DC: ACM Press.

Van Buuren, S. and Oudshoorn, C. (1999). Flexible multivariate imputation by MICE. Tech. rep., Leiden: TNO Preventie en Gezondheid, TNO/VGZ/PG 99.054.

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