Journal of Official Statistics, Vol. 25, No. 4, 2009, pp. 589–603

Disclosure Risk and Data Utility for Partially Synthetic Data: An Empirical Study Using the German IAB Establishment Survey

Jörg Drechsler[1] *and J. P. Reiter*[2]

Statistical agencies that disseminate data to the public must protect the confidentiality of respondents' identities and sensitive attributes. To satisfy these requirements, agencies can release the units originally surveyed with some values, such as sensitive values at high risk of disclosure or values of key identifiers, replaced with multiple imputations. These are called partially synthetic data. In this article, we empirically examine trade-offs between inferential accuracy and confidentiality risks for partially synthetic data, with emphasis on the role of the number of released datasets. We also present a two-stage imputation scheme that allows agencies to release different numbers of imputations for different variables. This scheme can result in lower disclosure risks and higher data utility than the typical one-stage imputation with the same number of released datasets. The empirical analyses are based on partial synthesis of the German IAB Establishment Survey.

Key words: Confidentiality; disclosure; multiple imputation; synthetic data.

1. Introduction

Statistical agencies and other organizations that disseminate data to the public are ethically, practically, and often legally required to protect the confidentiality of respondents' identities and sensitive attributes. To satisfy these requirements, Rubin (1993) and Little (1993) proposed that agencies utilize multiple imputation approaches. For example, agencies can release the units originally surveyed with some values, such as sensitive values at high risk of disclosure or values of key identifiers, replaced with multiple imputations. These are called partially synthetic datasets (Reiter 2003).

In recent years, statistical agencies have begun to use partially synthetic approaches to create public use data for major surveys. For example, in 2007 the U.S. Census Bureau released a partially synthetic, public use file for the Survey of Income and Program Participation (SIPP) that includes imputed values of social security benefits information and dozens of other highly sensitive variables (www.sipp.census.gov/sipp/synth_data.html). The U.S. Census Bureau also plans to protect the identities of people in group quarters (e.g., prisons, shelters) in the next release of public use files of the American Communities Survey by replacing demographic data for people at high disclosure risk with imputations. Partially synthetic public use datasets are in the development stage in the

[1] Institute for Employment Research, Regensburger Str. 104, 90478 Nuremberg, Germany. Email: joerg.drechsler@iab.de
[2] Department of Statistical Science, Box 90251, Duke University, Durham, NC 27708-0251, U.S.A. Email: jerry@stat.duke.edu

U.S. for the Longitudinal Business Database, the Longitudinal Employer-Household Dynamics survey, and the American Communities Survey veterans and full sample data. Statistical agencies in Australia, Canada, Germany (Drechsler et al. 2008a), and New Zealand (Graham and Penny 2005) are also investigating the approach. Other applications of partially synthetic data are described by Kennickell (1997); Abowd and Woodcock (2001, 2004); Abowd and Lane (2004); Little et al. (2004); Reiter (2004, 2005c); Mitra and Reiter (2006); An and Little (2007), and Reiter and Raghunathan (2007).

Although these methods are being utilized, there has been little discussion of how many multiply-imputed datasets to release. From the perspective of the secondary data analyst, a large number of datasets is desirable. The additional variance introduced by the imputation decreases with the number of released datasets. For example, Reiter (2003) finds nearly a 100% increase in variance of regression coefficients when going from fifty to two partially synthetic datasets. From the perspective of the agency, a small number of datasets is desirable. The information available to ill-intentioned users seeking to identify individuals in the released datasets increases with the number of released datasets. Thus, agencies considering the release of partially synthetic data are generally confronted with a trade-off between disclosure risk and data utility.

In this article, we examine the impact of the number of imputations, m, on data utility and disclosure risk when releasing partially synthetic data. We do so by generating partially synthetic datasets for a German establishment survey, the Establishment Panel of the Institute for Employment Research (IAB). We find that, for the estimands we examine, the disclosure risks increase more rapidly with m than the data utility does. This leads us to examine an alternative approach to generating partially synthetic data based on imputation in two stages. We find that, compared to the equivalent number of datasets from a one-stage approach, this can reduce disclosure risks without sacrificing data utility.

The remainder of the article is organized as follows. In Section 2, we describe the methodological background for one-stage partially synthetic data, including the data utility and disclosure risk measures we employ. In Section 3, we apply the risk and utility measures to partially synthetic data generated from the IAB Establishment Panel. In Section 4, we apply the two-stage imputation approach and illustrate the potential improvements in risk and utility. Finally, in Section 5, we conclude with some remarks on how agencies can go about selecting the number of synthetic datasets to release.

2. Background on Partially Synthetic Data

We first outline the main ideas underpinning partially synthetic data, followed by discussions of disclosure risk and data utility measures for partially synthetic data.

2.1. Inference with Partially Synthetic Data

The partially synthetic data approach is similar to multiple imputation for missing data (Rubin 1987). There is a key difference, however: the imputations replace the originally observed values rather than fill in missing values. This difference leads to different formulas for combining the point and variance estimates from the multiple datasets.

Following Reiter (2003, 2004), let $Z_j = 1$ if unit j is selected to have any of its observed data replaced, and let $Z_j = 0$ otherwise. Let $Z = (Z_1, \ldots, Z_s)$, where s is the number of

records in the observed data. Let $Y = (Y_{rep}, Y_{nrep})$ be the data collected in the original survey, where Y_{rep} includes all values to be replaced with multiple imputations and Y_{nrep} includes all values not replaced with imputations. Let $Y_{rep}^{(i)}$ be the replacement values for Y_{rep} in synthetic dataset i. Each $Y_{rep}^{(i)}$ is generated by simulating values from the posterior predictive distribution $f(Y_{rep}^{(i)}|Y, Z)$, or some close approximation to the distribution such as those of Raghunathan et al. (2001). The agency repeats the process m times, creating $D^{(i)} = \left(Y_{nrep}, Y_{rep}^{(i)}\right)$, for $i = 1, \ldots, m$, and releases $\mathbf{D} = \{D^{(1)}, \ldots, D^{(m)}\}$ to the public.

To get valid inferences, secondary data users can use the combining rules presented by Reiter (2003). Let Q be an estimand, such as a population mean or regression coefficient. Suppose that, given the original data, the analyst would estimate Q with some point estimator q and the variance of q with some estimator v. Let $q^{(i)}$ and $v^{(i)}$ be the values of q and v in synthetic dataset $D^{(i)}$, for $i = 1, \ldots, m$. The analyst computes $q^{(i)}$ and $v^{(i)}$ by acting as if each $D^{(i)}$ is the genuine data.

The point estimate of Q is $\bar{q}_m = \sum_i q^{(i)}/m$. The estimated variance of $\bar{q}_m$ is $T_m = b_m/m + \bar{v}_m$, where $b_m = \sum_i (q^{(i)} - \bar{q}_m)^2/(m-1)$ and $\bar{v}_m = \sum_i v^{(i)}/m$. Inferences for scalar Q can be based on t-distributions with degrees of freedom $\nu_m = (m-1)\left(1 + r_m^{-1}\right)^2$, where $r_m = (m^{-1}b_m/\bar{v}_m)$. Methods for multivariate inferences are developed in Reiter (2005b).

2.2. *Disclosure Risk*

To evaluate disclosure risks, we compute probabilities of identification by following the approach of Reiter and Mitra (2009). Related approaches are described by Duncan and Lambert (1989); Fienberg et al. (1997) and Reiter (2005a). Roughly, in this approach we mimic the behavior of an ill-intentioned user of the released data who possesses the true values of the quasi-identifiers for selected target records (or even the entire database). To illustrate, suppose the malicious user has a vector of information, $\boldsymbol{t}$, on a particular target unit in the population which may or may not correspond to a unit in the m released simulated datasets, $\mathbf{D} = \{D^{(1)}, \ldots, D^{(m)}\}$. Let t_0 be the unique identifier (e.g., establishment name) of the target, and let d_{j0} be the (not released) unique identifier for record j in $\mathbf{D}$, where $j = 1, \ldots, s$. Let M be any information released about the simulation models.

The malicious user's goal is to match unit j in $\mathbf{D}$ to the target when $d_{j0} = t_0$, and not to match when $d_{j0} \neq t_0$ for any $j \in \mathbf{D}$. Let J be a random variable that equals j when $d_{j0} = t_0$ for $j \in \mathbf{D}$ and equals $s+1$ when $d_{j0} = t_0$ for some j $\notin \mathbf{D}$. The malicious user thus seeks to calculate the $Pr(J = j|\mathrm{t}, \mathbf{D}, M)$ for $j = 1, \ldots, s+1$. He or she would then decide whether or not any of the identification probabilities for $j = 1, \ldots, s$ are large enough to declare an identification. Because the malicious user does not know the actual values in Y_{rep}, he or she should integrate over its possible values when computing the match probabilities. Hence, for each record in $\mathbf{D}$ we compute

$$Pr(J = j|\mathrm{t}, \mathbf{D}, M) = \int Pr(J = j|\mathrm{t}, \mathbf{D}, Y_{rep}, M) Pr(Y_{rep}|\mathrm{t}, \mathbf{D}, M) dY_{rep} \quad (1)$$

This construction suggests a Monte Carlo approach to estimating each $Pr(J = j|\mathrm{t}, \mathbf{D}, M)$. First, sample a value of Y_{rep} from $Pr(Y_{rep}|\mathrm{t}, \mathbf{D}, M)$. Let Y^{new} represent one set of simulated

values. Second, compute $Pr(J = j|\mathrm{t}, \mathbf{D}, Y_{rep} = Y^{new}, M)$ using exact or, for continuous synthesized variables, distance-based matching assuming Y^{new} are collected values. This two-step process is iterated R times, where ideally R is large, and (1) is estimated as the average of the resultant R values of $Pr(J = j|\mathrm{t}, \mathbf{D}, Y_{rep} = Y^{new}, M)$. When M has no information, the malicious user can treat the simulated values as plausible draws of Y_{rep}.

To illustrate, suppose that region and employee size are the only quasi-identifiers in a survey of establishments. A malicious user seeks to identify an establishment in a particular region of the country with 125 employees. The malicious user knows that this establishment is in the sample. Suppose that the agency releases m datasets after simulating only employment size, without releasing information about the imputation model. In each $D^{(i)}$, the malicious user would search for all establishments matching the target on region and having synthetic employee size within some interval around 125, say 110 to 140. The agency selects the intervals for employment size based on its best guess of the amount of uncertainty that intruders would be willing to tolerate when estimating true employee sizes. Let $N^{(i)}$ be the number of records in $D^{(i)}$ that meet these criteria. When no establishments with all of those characteristics are in $D^{(i)}$, set $N^{(i)}$ equal to the number of establishments in the region, i.e., match on all nonsimulated quasi-identifiers. For any j, $Pr(J = j|\mathrm{t}, \mathbf{D}, M) = (1/m)\sum_i (1/N^{(i)})\left(Y_j^{new,i} = \mathrm{t}\right)$, where $\left(Y_j^{new,i} = \mathrm{t}\right) = 1$ when record j is among the $N^{(i)}$ matches in $D^{(i)}$ and equals zero otherwise. Similar computations arise when simulating region and employee size: the malicious user exactly matches on the simulated values of region and distance-based matches on employee size to compute the probabilities.

Following Reiter (2005a), we quantify disclosure risk with summaries of these identification probabilities. It is reasonable to assume that the malicious user selects as a match for t the record j with the highest value of $Pr(J = j|\mathrm{t}, \mathbf{D}, M)$, if a unique maximum exists. We consider two risk measures: the *expected match risk* and the *true match risk.* To calculate these, we need some further definitions. Let c_j be the number of records in the dataset with the highest match probability for the target t_j for $j = 1, \ldots, s$; let $I_j = 1$ if the true match is among the c_j units and $I_j = 0$ otherwise. Let $K_j = 1$ when $c_j I_j = 1$ and $K_j = 0$ otherwise. The *expected match risk* can now be defined as $\sum_j (1/c_j) I_j$. When $I_j = 1$ and $c_j > 1$, the contribution of unit j to the expected match risk reflects the intruder randomly guessing at the correct match from the c_j candidates. The *true match risk* equals $\sum_j K_j$.

2.3. Data Utility

It is important to quantify the analytic usefulness of the synthetic datasets. Research on utility measures for synthetic data, and for disclosure limitation in general, is less developed than research on risk assessment. Existing utility measures are of two types: (i) comparisons of broad differences between the original and released data, and (ii) comparisons of differences in specific models between the original and released data. Broad difference measures essentially quantify some statistical distance between the distributions of the original and released data, for example a Kullback-Leibler or Hellinger distance. As the distance between the distributions grows, the overall quality of the released data generally drops.

In this article, we focus on utility measures for specific estimands. We use the interval overlap measure of Karr et al. (2006). For any estimand, we first compute the 95% confidence intervals for the estimand from the synthetic data, (L_s, U_s), and from the collected data, (L_o, U_o). Then, we compute the intersection of these two intervals, (L_i, U_i). The utility measure is:

$$I = \frac{U_i - L_i}{2(U_o - L_o)} + \frac{U_i - L_i}{2(U_s - L_s)} \tag{2}$$

When the intervals are nearly identical, corresponding to high utility, $I \approx 1$. When the intervals do not overlap, corresponding to low utility, $I = 0$. The second term in (2) is included to differentiate between intervals with $(U_i - L_i)/(U_o - L_o) = 1$ but different lengths. For example, for two synthetic data intervals that fully contain the collected data interval, the measure I favors the shorter interval. The synthesis is successful if we obtain large values of I for many estimands. To compute one number summaries of utility, we average the values of I over all estimands.

There do not exist published broad utility measures that account for all m synthetic datasets. The U.S. Census Bureau has adapted an approach described by Woo et al. (2009), which is based on how well one can discriminate between the original and disclosure protected data. In this approach, the agency stacks the original and synthetic datasets in one file and estimates probabilities of being "assigned" to the original data conditional on all variables in the dataset. When the distributions of probabilities are similar in the original and synthetic data, the distributions of the variables are similar – this fact comes from the literature on propensity scores – so that the synthetic data have high utility. This approach is especially useful as a diagnostic for deficiencies in the synthesis methods (variables with significant coefficients in the logistic regression have different distributions in the original and synthetic data). It is not as useful for evaluating the impacts of increasing m, which is the objective of our empirical investigations.

3. Application to the IAB Establishment Panel

To assess the impact of different numbers of imputations, we generate partially synthetic datasets from the German IAB Establishment Panel. We first describe the survey and synthesis plan, then evaluate the trade-off between risk and utility as a function of m.

3.1. The IAB Establishment Panel

The IAB Establishment Panel, conducted since 1993, contains detailed information about German firms' personnel structure, development, and policy. Considered one of the most important business panels in Germany, there is high demand for access to these data from external researchers. Because of the sensitive nature of the data, researchers desiring direct access to the data have to work on site at the IAB. Alternatively, researchers can submit code for statistical analyses to the IAB research data center, whose staff run the code on the data and send the results to the researchers. To help researchers develop code, the IAB provides remote access to a publicly available "dummy dataset" with the same structure as the Establishment Panel. The dummy dataset comprises random numbers generated without any attempt to preserve the distributional properties of the variables in the Establishment

Panel data. For all analyses done with the genuine data, researchers can publicize their analyses only after IAB staff check for potential violations of confidentiality.

Releasing public use files of the Establishment Panel would allow more researchers to access the data with fewer burdens, stimulating research on German business data. It also would free up staff time from running code and conducting confidentiality checks. Because there are so many sensitive variables in the dataset, standard disclosure limitation methods like swapping or microaggregation would have to be applied with high intensity, which would severely compromise the utility of the released data. Therefore, the IAB decided to develop synthetic datasets for public release.

For this simulation study, we synthesize two variables in the Establishment Panel for 1997: the number of employees and the industry coded in 16 categories. For both variables, all 7,332 observations are replaced by imputed values. Employment size and industry code are high-risk variables since (i) they are easily available in other databases and (ii) the distribution for the number of employees is heavily skewed. Imputations are based on linear models with more than 100 explanatory variables for the number of employees and on a multinomial logit model with more than 80 explanatory variables for the industry. We use large numbers of predictors in hopes of reducing problems from uncongeniality (Meng 1994). Some variables for the multinomial logit model are dropped for multicollinearity reasons.

3.2. *Data Utility for the Panel*

We investigate data utility for some descriptive statistics and a probit regression. The descriptive statistics are the (unweighted) average numbers of employees by industry; they are based solely on the two variables we synthesized. The probit regression, which originally appeared in an article by Zwick (2005), uses the 1997 wave of the Establishment Panel to determine why some firms offer vocational training and others do not. This model has been used by others to evaluate the utility of synthetic datasets (Drechsler et al. 2008a, b).

Tables 1–4 display point estimates and the interval overlap measures for different values of m. For most parameters, increasing m moves point estimates closer to their values in the original data and increases the overlaps in the confidence intervals. Increasing $m = 3$ to $m = 10$ results in the largest increase in data utility, as the average confidence interval overlap over all 31 parameters in Tables 3 and 4 increases from 0.828 to 0.867. Increasing $m = 50$ to $m = 100$ does not have much impact on data utility.

Each entry in Tables 1–4 results from one replication of a partially synthetic data release strategy. To evaluate the variability across different replications, we repeated each simulation ten times. Table 5 presents the average confidence interval overlap over all 31 estimands for the ten simulations. The variation in the overlap measures decreases with m. This is because the variability in $\bar{q}_m$ and T_m decreases with m, so that results stabilize as m gets large. We believe most analysts would prefer to have stable results across different realizations of the synthesis and hence favor large values of m.

3.3. *Disclosure Risk for the Panel*

To assess disclosure risk, we assume that the intruder knows which establishments are included in the survey and the true values for the number of employees and industry.

Table 1. Average number of employees by industry for one-stage synthesis

	Original data	$m = 3$	$m = 10$	$m = 50$	$m = 100$
Industry 1	71.5	84.2	84.2	82.6	82.4
Industry 2	839.1	919.4	851.2	870.2	852.9
Industry 3	681.1	557.7	574.5	594.4	593.1
Industry 4	642.9	639.9	644.8	643.5	649.6
Industry 5	174.5	179.8	176.0	183.5	187.4
Industry 6	108.9	132.4	121.8	120.8	120.7
Industry 7	117.1	111.6	112.9	117.1	119.6
Industry 8	548.7	455.3	504.3	514.2	513.0
Industry 9	700.7	676.9	689.4	711.8	713.4
Industry 10	547.0	402.4	490.3	499.3	487.7
Industry 11	118.6	142.7	130.2	132.1	131.0
Industry 12	424.3	405.6	414.9	424.5	425.2
Industry 13	516.7	526.1	549.1	550.2	551.9
Industry 14	128.1	185.8	167.1	160.0	159.0
Industry 15	162.0	292.8	233.4	221.9	238.1
Industry 16	510.8	452.8	449.9	441.5	439.3

This is a conservative scenario but gives, in some sense, an upper bound on the risk for this level of intruder knowledge. Intruders might also know other variables on the file, in which case the agency may need to synthesize them as well.

The intruder computes probabilities using the approach outlined in Section 2.2. We assume that the agency does not reveal the synthesis model to the public, so that the only information in M is that employee size and industry were synthesized. For a given target t, records from each $D^{(i)}$ must meet two criteria to be possible matches. First, the record's synthetic industry code exactly matches the target's true industry code. Second, the record's synthetic number of employees lies within an agency-defined interval around

Table 2. Point estimates for vocational training regression for one-stage synthesis

	Original data	$m = 3$	$m = 10$	$m = 50$	$m = 100$
Intercept	−1.319	−1.323	−1.322	−1.323	−1.324
Redundancies expected	0.253	0.268	0.262	0.264	0.264
Many emp. expected on maternity leave	0.262	0.334	0.316	0.312	0.314
High qualification need exp.	0.646	0.636	0.640	0.640	0.639
Appr. training reaction on skill shortages	0.113	0.098	0.106	0.110	0.112
Training reaction on skill shortages	0.540	0.529	0.538	0.542	0.543
Establishment size 20–199	0.684	0.718	0.709	0.705	0.701
Establishment size 200–499	1.352	1.363	1.333	1.339	1.343
Establishment size 500–999	1.346	1.315	1.386	1.377	1.367
Establishment size 1,000 +	1.955	1.782	1.800	1.778	1.776
Share of qualified employees	0.787	0.787	0.788	0.784	0.785
State-of-the-art technical equipment	0.171	0.183	0.178	0.174	0.174
Collective wage agreement	0.255	0.268	0.264	0.267	0.268
Apprenticeship training	0.490	0.501	0.510	0.507	0.507
East Germany	0.058	0.038	0.033	0.042	0.044

Table 3. Confidence interval overlap for average number of employees for one-stage synthesis

	$m = 3$	$m = 10$	$m = 50$	$m = 100$
Industry 1	0.778	0.770	0.777	0.782
Industry 2	0.844	0.893	0.853	0.874
Industry 3	0.730	0.776	0.797	0.800
Industry 4	0.983	0.992	0.995	0.971
Industry 5	0.920	0.935	0.863	0.817
Industry 6	0.605	0.749	0.764	0.767
Industry 7	0.809	0.820	0.863	0.876
Industry 8	0.692	0.862	0.894	0.890
Industry 9	0.926	0.966	0.968	0.963
Industry 10	0.660	0.876	0.897	0.871
Industry 11	0.609	0.804	0.773	0.792
Industry 12	0.903	0.912	0.916	0.918
Industry 13	0.946	0.814	0.809	0.799
Industry 14	0.408	0.589	0.655	0.664
Industry 15	0.586	0.639	0.654	0.638
Industry 16	0.666	0.645	0.583	0.566
Average	0.754	0.815	0.816	0.812

the target's true number of employees. Acting as the agency, we define the interval as follows. We divide the transformed (cubic root) number of employees into twenty quantiles and calculate the standard deviation of the number of employees within each quantile. The interval is $t_e \pm sd_s$, where t_e is the target's true value and sd_s is the standard deviation of the quantile in which the true value falls. When there are no synthetic records that fulfill both matching criteria, the intruder matches only on the industry code.

We use 20 quantiles because this is the largest number of groups that guarantees some variation within each group. Using more than 20 results in groups with only one value of

Table 4. Confidence interval overlap for vocational training probit regression for one-stage synthesis

	$m = 3$	$m = 10$	$m = 50$	$m = 100$
Intercept	0.987	0.989	0.986	0.984
Redundancies expected	0.931	0.958	0.946	0.948
Many emp. expected on maternity leave	0.808	0.856	0.867	0.861
High qualification need exp.	0.965	0.977	0.978	0.976
Appr. training reaction on skill shortages	0.928	0.964	0.984	0.996
Training reaction on skill shortages	0.946	0.989	0.989	0.982
Establishment size 20–199	0.802	0.856	0.879	0.902
Establishment size 200–499	0.934	0.939	0.935	0.933
Establishment size 500–999	0.926	0.907	0.928	0.953
Establishment size 1,000 +	0.731	0.763	0.727	0.723
Share of qualified employees	0.995	0.997	0.989	0.993
State-of-the-art technical equipment	0.919	0.953	0.976	0.977
Collective wage agreement	0.926	0.952	0.934	0.927
Apprenticeship training	0.937	0.883	0.899	0.899
East Germany	0.872	0.840	0.899	0.912
Average	0.907	0.922	0.928	0.931

Table 5. Average confidence interval overlap for all 31 estimands for ten independent simulations of one-stage synthesis

	$m = 3$	$m = 10$	$m = 50$	$m = 100$
Simulation 1	0.828	0.867	0.870	0.870
Simulation 2	0.864	0.869	0.869	0.874
Simulation 3	0.858	0.866	0.873	0.868
Simulation 4	0.881	0.861	0.874	0.871
Simulation 5	0.872	0.865	0.866	0.875
Simulation 6	0.845	0.862	0.869	0.865
Simulation 7	0.849	0.851	0.871	0.873
Simulation 8	0.841	0.862	0.871	0.873
Simulation 9	0.841	0.877	0.873	0.872
Simulation 10	0.861	0.865	0.874	0.867
Average	0.854	0.865	0.871	0.871

employment, which forces exact matching for targets in those quantiles. On the other hand, using a small number of quantiles does not differentiate adequately between small and large establishments. For small establishments, we want the potential matches to deviate only slightly from the original values. For large establishments, we accept higher deviations.

We studied the impact of using different numbers of groups for $m = 50$. We found a substantial increase in the risks of identifications, especially for the small establishments, when going from exact matching to five quantiles. Between five and twenty quantiles, the disclosure risk does not change dramatically. For more than twenty quantiles, the number of identifications starts to decline again.

Table 6 displays the average true matching risk and expected matching risk over the ten simulation runs used in Table 5. Since the largest establishments are usually considered as the records most at risk of identification, we also include the risk measures for the largest 25 establishments in parentheses. There is clear evidence that a higher number of imputations leads to a higher risk of disclosure, especially for the largest establishments. This is because, as m increases, the intruder has more information to estimate the distribution that generated the synthetic data. It is arguable that the gains in utility, at least for these estimands, are not worth the increases in disclosure risks.

The establishments that are correctly identified vary across the 10 replicates. For example, for $m = 50$, the total number of identified records over all 10 replicates is 614. Of these records, 319 are identified in only one simulation, 45 are identified in more than five simulations, and only 10 records are identified in all 10 replications. For $m = 10$, no records are identified more than seven times.

Table 6. Averages of disclosure risk measures over ten simulations of one-stage synthesis. Measures for the 25 largest establishments are reported in parentheses

	$m = 3$	$m = 10$	$m = 50$	$m = 100$
Expected match risk	67.8 (3.2)	94.8 (5.2)	126.9 (6.9)	142.5 (7.1)
True match risk	35.2 (2.0)	82.5 (4.9)	126.1 (6.8)	142.4 (7.1)

The risks are not large on an absolute scale. For example, with $m = 10$, we anticipate that the intruder could identify only 83 establishments out of 7,332. This assumes that the intruder already knows the establishment size and industry classification code and also has response knowledge, i.e., he knows which establishments participated in the survey. Furthermore, the intruder will not know how many of the unique matches (i.e., $c_j = 1$) actually are true matches.

We also investigated the disclosure risk for different subdomains for $m = 50$. Four of the sixteen industry categories had less than 200 units in the survey. For these categories, the percentage of identified records ranged between 5% and almost 10%. For the remaining categories, the percentage of correct identifications never went beyond 2.3%. If these risks are too high, the agency could collapse some of the industry categories.

The percentage of identified establishments was close to 5% for the largest decile of establishment size and never went beyond 2.5% for all the other deciles. The identification risk is higher for the top 25 establishments, but still moderate. When $m = 3$ only two of these establishments are correctly identified; this increases to seven establishments with $m = 100$. The intruder also makes many errors when declaring matches for these establishments. In fact, of all times when the intruder finds a unique match ($c_j = 1$) for these top establishments, it is the correct match only 13% of the time for $m = 3$, 23% of the time for for $m = 10$, and approximately 30% of the time for $m = 50$ and $m = 100$. None of the top 10 establishments are identified in all ten simulations.

The largest establishment's size is reduced by at least 10% in all synthetic datasets. We note that this can be viewed as reduction in data utility, since the tail is not accurate at extreme values. It may be possible to improve tail behavior with more tailored synthesis models, such as CART approaches (Reiter 2005c).

As noted previously, these risk computations are in some ways conservative. First, they presume that the intruder knows which records are in the survey. This is not likely to be true, since most establishments are sampled with probability less than one. However, large establishments are sampled with certainty, so that the risk calculations presented here apply for those records. Drechsler and Reiter (2008) show how to adjust the identification disclosure probabilities for intruder uncertainty due to sampling. In their application, the true match rate is 6% when the intruder knows which records are in the sample, and only 1% when the intruder does not know which records are in the sample. Second, the risk measurements presume that the intruder has precise information on establishment size and industry code. In Germany, it is not likely that intruders will know the sizes of all establishments in the survey, because there is no public information on small establishments. However, intruders can obtain size and industry type for large companies from public databases. They also can purchase large private databases on establishments, although the quality of these databases for record linkage on employee size is uncertain. Thus, except for possibly the largest establishments, the risk measures here could overstate the probabilities of identification.

4. A Two-Stage Approach for Imputation

The empirical investigations indicate that increasing m results in both higher data utility and higher risk of disclosures. In this section, we present and investigate an alternative

synthesis approach that can maintain high utility while reducing disclosure risks. The basic idea behind this approach is to impute variables that drive the disclosure risk only a few times and other variables many times. This can be accomplished by generating data in two stages, as described by Reiter and Drechsler (forthcoming). In general, two-stage and one-stage approaches require similar amounts of modeling efforts; however, in some settings, the two-stage approach can reduce computational burdens associated with generating synthetic data and thereby speed up the process; see (Reiter and Drechsler forthcoming) for further discussion of this point.

4.1. Inference for Synthetic Datasets Generated in Two Stages

The agency first partitions $Y_{rep} = (Y_a, Y_b)$, where Y_a are the values to be replaced in stage one and Y_b are the values to be replaced in stage two. The agency seeks to release fewer replications of Y_a than of Y_b, yet do so in a way that enables the analyst of the data to obtain valid inferences with standard complete data methods. To do so, the agency first replaces confidential values of Y_a with draws from $f(Y_a|Y, Z)$. Let $Y_a^{(i)}$ be the values imputed in the first stage in nest i, for $i = 1, \ldots, m$. Second, in each nest, the agency generates $Y_b^{(i,j)}$ by drawing from $f\left(Y_b|Y, Z, Y_a^{(i)}\right)$. Each synthetic dataset, $D^{(i,j)}$, comprises $\left(Y_a^{(i)}, Y_b^{(i,j)}, Y_{nrep}\right)$. The entire collection of $M = mr$ datasets, $\mathrm{D}_{syn} = \{D^{(i,j)}, i = 1, \ldots, m; j = 1, \ldots, r\}$, with labels indicating the nests, is released to the public.

To get valid inferences from two-stage synthetic data, new combining rules for the point and variance estimate are necessary. Let $q^{(i,j)}$ and $v^{(i,j)}$ be the values of q and v in synthetic dataset $D^{(i,j)}$, for $i = 1, \ldots, m$ and $j = 1, \ldots, r$. The following quantities are necessary for inferences

$$\bar{q}_r^{(i)} = \sum_j q^{(i,j)}/r \tag{3}$$

$$\bar{q}_m = \sum_i \bar{q}_r^{(i)}/m = \sum_j \sum_i q^{(i,j)}/mr \tag{4}$$

$$b_m = \sum_i \left(\bar{q}_r^{(i)} - \bar{q}_m\right)^2/(m-1) \tag{5}$$

$$\bar{v}_m = \sum_{ij} v^{(i,j)}/mr \tag{6}$$

The analyst can use $\bar{q}_m$ to estimate Q and $T_{2st} = \bar{v}_m + b_m/m$ to estimate the variance of $\bar{q}_m$. Inferences can be based on a t-distribution with $\nu_{2st} = (m-1)(1 + m\bar{v}_m/b_m)^2$ degrees of freedom (Reiter and Drechsler forthcoming).

Because these combining rules are more complicated than standard one-stage combining rules, some analysts may have difficulty applying them correctly. We suggest that agencies work with academic researchers and software developers to write software routines that implement the rules. These routines can be disseminated with the synthetic datasets. Similar routines have been developed for standard multiple imputation and for one-stage partial synthesis and are part of leading statistical software packages, including SAS, Stata, and R.

Table 7. Average CI overlap and match risk for two-stage synthesis based on ten simulations. Match risk for largest 25 establishments is in parentheses

m,*r*	Avg. overlap	Expected match risk	True match risk
$m = 3, r = 3$	0.867	83.1 (4.0)	67.6 (3.4)
$m = 3, r = 16$	0.868	98.0 (4.1)	91.8 (4.0)
$m = 3, r = 33$	0.870	99.8 (3.8)	96.3 (3.8)
$m = 5, r = 10$	0.869	106.1 (4.6)	101.2 (4.4)
$m = 10, r = 5$	0.875	113.8 (5.0)	109.4 (5.0)
$m = 16, r = 3$	0.874	119.9 (5.2)	116.4 (5.2)

4.2. Application to the IAB Establishment Panel

We impute the industry in stage one and the number of employees in stage two. Exchanging the order of the imputation does not materially impact the results. We consider different values of m and r. We run ten simulations for each setting and present the average estimates over these ten simulations.

Table 7 displays the average confidence interval overlap for all 31 parameters and the two disclosure risk measures for the different settings. As with one-stage synthesis, there is not much difference in the data utility measures for different M, although there is a slight increase when going from $M = 9$ to $M \approx 50$. The two-stage results with $M = 9$ (average overlap of .867) are slightly better than the one-stage results with $m = 10$ (average overlap of .865). The two-stage results with $M \approx 50$ are approximately on the same level or slightly above the one-stage results for $m = 50$ (average overlap of .871).

The improvements in data utility when using the two-stage approach are arguably minor, but the reduction in disclosure risks is more noticeable. The measures are always substantially lower for the two-stage approach compared to the one-stage approach with approximately the same number of synthetic datasets. For example, releasing two-stage synthetic data with $M = 9$ carries an average true match risk of 67 (3.4 for the top 25 establishments), whereas releasing one-stage synthetic data with $m = 10$ has a true match risk of 82 (4.9). Risks are lower for $M \approx 50$ as compared to one-stage with $m = 50$ as well. Additionally, for the top 25 establishments, the percentage of unique matches that are true matches is lower for the two-stage approach. When $M = 9$, this percentage is 17% for the two-stage approach compared to around 23% for one-stage synthetic data with $m = 10$. When $M \approx 50$, this percentage varied between 18% and 22%, whereas it is around 30% for one-stage synthetic data with $m = 50$.

The two-stage methods have lower disclosure risks at any given total number of released datasets because they provide fewer pieces of data about industry codes. This effect is evident in the two-stage results with $M \approx 50$. The risks increase monotonically with the number of imputations dedicated to the first stage.

5. Conclusion

Releasing partially synthetic datasets is an innovative method for statistical disclosure control. The released datasets can provide detailed information with good analytic utility without breaking pledges of confidentiality under which many data are collected. As with

most disclosure control methods, the risk of disclosure is not zero. The original records and some original values are released in the synthetic data, and intruders can use information from the synthetic data to approximate the replaced values.

In this article, we have demonstrated that both data utility and disclosure risk increase with the number of synthetic datasets. Thus, agencies have to decide what level of disclosure risk they are willing to accept to provide the highest data utility possible. In general, agencies consider disclosure risks to be primary and so are inclined to release only a small number of synthetic datasets. This can be problematic for inferences, particularly when synthesizing many values. With modest amounts of synthesis, our results suggest that the gains in utility from releasing more than ten or so synthetic datasets may not be worth the increase in disclosure risks.

In our application, we found that a two-stage approach can drive down the disclosure risk while keeping the data utility at the same level. A topic for future research is to develop methods to determine optimal numbers and allocations of two-stage imputations for given data, ideally without the need of extensive simulation studies. Another important issue is to develop measures that help to decide which variables should be imputed in stage one and which should be imputed in stage two.

6. References

Abowd, J.M. and Lane, J.I. (2004). New Approaches to Confidentiality Protection: Synthetic Data, Remote Access and Research Data Centers. Privacy in Statistical Databases, J. Domingo-Ferrer and V. Torra (eds). New York: Springer-Verlag, 282–289.

Abowd, J.M. and Woodcock, S.D. (2001). Disclosure Limitation in Longitudinal Linked Data. Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, P. Doyle, J. Lane, L. Zayatz, and J. Theeuwes (eds). Amsterdam: North-Holland, 215–277.

Abowd, J.M. and Woodcock, S.D. (2004). Multiply-imputing Confidential Characteristics and File Links in Longitudinal Linked Data. Privacy in Statistical Databases, J. Domingo-Ferrer and V. Torra (eds). New York: Springer-Verlag, 290–297.

An, D. and Little, R. (2007). Multiple Imputation: An Alternative to Top Coding for Statistical Disclosure Control. Journal of the Royal Statistical Society, Series A, 170, 923–940.

Drechsler, J. and Reiter, J.P. (2008). Accounting for Intruder Uncertainty Due to Sampling When Estimating Identification Disclosure Risks in Partially Synthetic Data. Privacy in Statistical Databases (LNCS 5262), J. Domingo-Ferrer and Y. Saygin (eds). New York: Springer-Verlag, 227–238.

Drechsler, J., Dundler, A., Bender, S., Rässler, S., and Zwick, T. (2008a). A New Approach for Disclosure Control in the IAB Establishment Panel-Multiple Imputation for a Better Data Access. Advances in Statistical Analysis, 92, 439–458.

Drechsler, J., Bender, S., and Rässler, S. (2008b). Comparing Fully and Partially Synthetic Datasets for Statistical Disclosure Control in the German IAB Establishment Panel. Transactions on Data Privacy, 1, 105–130.

Duncan, G.T. and Lambert, D. (1989). The Risk of Disclosure for Microdata. Journal of Business and Economic Statistics, 7, 207–217.

Fienberg, S.E., Makov, U.E., and Sanil, A.P. (1997). A Bayesian Approach to Data Disclosure: Optimal Intruder Behavior for Continuous Data. Journal of Official Statistics, 13, 75–89.

Graham, P. and Penny, R. (2005). Multiply Imputed Synthetic Data Files. Technical report, University of Otago. http://www.uoc.otago.ac.nz/departments/pubhealth/pgrahpub.htm

Karr, A.F., Kohnen, C.N., Oganian, A., Reiter, J.P., and Sanil, A.P. (2006). A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality. The American Statistician, 60, 224–232.

Kennickell, A.B. (1997). Multiple Imputation and Disclosure Protection: The Case of the 1995 Survey of Consumer Finances. Record Linkage Techniques, W. Alvey and B. Jamerson (eds). Washington, D.C. National Academy Press, 248–267.

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. Applied Bayesian Modeling and Causal Inference From Incomplete-Data Persepectives, A. Gelman and X.L. Meng (eds). New York: John Wiley & Sons, 141–152.

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. Privacy in Statistical Databases, J. Domingo-Ferrer and L. Franconi (eds). New York: Springer-Verlag, 177–188.

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.

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). Estimating Identification Risks in Microdata. Journal of the American Statistical Association, 100, 1103–1113.

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. and Drechsler, J. (forthcoming). Releasing Multiply-imputed, Synthetic Data Generated in Two Stages to Protect Confidentiality. Statistica Sinica 1, Volume 20.

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

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. (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.

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

Zwick, T. (2005). Continuing Vocational Training Forms and Establishment Productivity in Germany. German Economic Review, 6(2), 155–184.

Received May 2008
Revised July 2008