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

Jörg Drechsler*and J. P. Reiter†

Abstract

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

*Institute for Employment Research, Regensburger Str. 104, 90478 Nuremberg, Germany. E-mail: joerg.drechsler@iab.de.

†Department of Statistical Science, Box 90251, Duke University, Durham, NC 27708-0251, E-mail: jerry@stat.duke.edu, phone: 919 668 5227, Fax: 919 684 8594.

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 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 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.*, 2007), and New Zealand (Graham and Penny, 2005) also are 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 generally are 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 paper 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 improvement in risk-utility profile. 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 $D^{(i)} = (Y_{nrep}, Y_{rep}^{(i)})$ for $i = 1, \ldots, m$. and releases $D = \{D^{(1)}, \ldots, D^{(m)}\}$ 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 $i = 1, ..., m$. 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 $T_m = b/m + \bar{v}_m$, 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 $\nu_m = (m-1)(1 + r_m^{-1})^2$, 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 (forthcoming). 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, $\mathbf{t}$, on a particular target unit in the population which may or may not correspond to a unit in the m released simulated datasets, $\mathbf{D} = \{D^{(i)}, \ldots, D^{(m)}\}$. 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 $Pr(J = j|\mathbf{t}, \mathbf{D}, M)$ for $j = 1, \ldots, s + 1$. He or she then would 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

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

This construction suggests a Monte Carlo approach to estimating each $Pr(J = j|\mathbf{t}, \mathbf{D}, M)$. First, sample a value of Y_{rep} from $Pr(Y_{rep}|\mathbf{t}, \mathbf{D}, M)$. Let Y^{new} represent one set of simulated values. Second, compute $Pr(J = j|\mathbf{t}, \mathbf{D}, Y_{rep} = Y^{new}, M)$ 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 $Pr(J = j|\mathbf{t}, \mathbf{D}, Y_{rep} = Y^{new}, M)$. 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 $D^{(i)}$, where $i = 1, \ldots, m$ 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 non-simulated quasi-identifiers. For any j, $Pr(J = j|\mathbf{t}, \mathbf{D}, M) = (1/m) \sum_i (1/N^{(i)})(Y_j^{new,i} = \mathbf{t})$, where $(Y_j^{new,i} = \mathbf{t}) = 1$ 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 $\mathbf{t}$ the record j with the highest value of $Pr(J = j|\mathbf{t}, \mathbf{D}, M)$, if a unique maximum exists. We consider two disclosure risk measures: the *expected match risk* and the *true match risk*. To calculate these measures, 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 $j = 1, ..., s$; 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 data on the original and released files, 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 paper, 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 $\frac{U_i - L_i}{(U_o - L_o)} = 1$ 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 the approach of Woo *et al.* (2007), 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 attempts 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 data set, 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 the (unweighted) average number of employees by industry, since it is based solely on the two variables we synthesized. Tables 1 and 2 display the $\bar{q}_m$ and the interval overlap measures for different values of m. For most estimates, increasing m moves point estimates closer to their original values and increases the overlaps in the confidence intervals. Increasing $m = 3$ to $m = 10$ results in the largest increase in data utility, as the relative confidence interval overlap averaged over all sixteen estimates increases from 0.754 to 0.815. Increasing $m = 50$ to $m = 100$ does not have much impact on data utility.

Each entry in Table 1 and 2 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 3 presents the average confidence interval overlap over all sixteen

industry categories 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.

We also estimated the coefficients in a probit regression that appeared in a paper by Zwick (2005). The response is a binary variable indicating if firms offer vocational training. There are twelve explanatory variables including number of employees and industry. The results show similar trends: increasing m results in point estimates that are closer to the observed data estimates, higher CI overlap, and lower variability between the replications. Tables for this regression are omitted for brevity.

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 their true values for the number of employees and industry. 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 $\mathbf{t}$, 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 the target's true number of employees. Acting as the agency, we define the interval as follows. We divide the transformed (true) 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 at least some variation within each group. Using a larger number of quantiles results in groups with only one value of 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 doesn't change dramatically. For more than twenty quantiles, the number of identifications starts to decline again.

Table 4 displays the average true matching risk and expected matching risk over the ten simulation runs used in Table 3. There is clear evidence that a higher number of imputations leads to a higher risk of disclosure. 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.

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 $m = 50$. None of the industries had a percentage of identified establishments exceeding 4%. 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. 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 disclosure risk never went beyond 2.3%. If these risks are too high, the agency could collapse some of the industry categories.

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 (2008).

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 1 and Y_b are the values to be replaced in stage 2. 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 $f(Y_a \mid Y, Z)$. 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 $f(Y_b \mid Y, Z, Y_a^{(i)})$. Each synthetic dataset, $D^{(i,j)}$, comprises $(Y_a^{(i)}, Y_b^{(i,j)}, Y_{nrep})$. The entire collection of $M = mr$ data sets, $D_{syn} = \{D^{(i,j)}, i = 1, \ldots, m; j = 1, \ldots, r\}$, 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 $i = 1, ..., m$ and $j = 1, ..., r$. The following quantities are necessary for

inferences

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

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

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

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

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, 2008).

4.2 Application for 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 5 displays the average confidence interval overlap over all industries and the two disclosure risk measures for the different settings averaged over all ten replications. 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 $M = 10$ to $M \approx 50$. The two stage results with $M = 9$ (average overlap of .819) are slightly better than the one stage results with $m = 10$ (average overlap of .806). The two stage results with $M \approx 50$ are always slightly better then the one stage results for $m = 50$ (average overlap of .817).

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 approaches with the same number of synthetic datasets. For example, releasing two stage synthetic data with $M = 9$ carries an average true match risk of 67, whereas releasing one stage synthetic data with $m = 10$ has a true match risk of 82. Risks also are lower for $M \approx 50$ as compared to one stage 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 high data quality without breaking the pledge of confidentiality under which many of the data are collected. As with most disclosure control methods the risk of disclosure is not zero however, since true values remain in the released datasets and intruders can try to guess true values from the synthetic values.

In this paper we demonstrated that both data utility and disclosure risk increases 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 few number of

synthetic datasets. This can be problematic for inferences, particularly when synthesizing many values. As we have shown, it is possible to simultaneously reduce disclosure risks and improve data utility by using a two stage imputation approach.

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 could be to develop methods to identify the "best" number of imputations without the need of time consuming simulation studies. Another important issue is to develop measures that help to decide which variables should be imputed on stage one and which can be imputed on the second stage.

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.

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., 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.

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. Tech. rep., 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. In W. Alvey and B. Jamerson, eds., *Record Linkage Techniques, 1997*, 248–267. Washington, D.C.: National Academy Press.

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

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.

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. (2008). Releasing multiply-imputed, synthetic data generated in two stages to protect confidentiality. Tech. rep., IAB Discussion Paper, No.20/2007.

Reiter, J. P. and Mitra, R. (forthcoming). Estimating risks of identification disclosure in partially synthetic data. *Journal of Privacy and Confidentiality* .

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. (2007). Global measures of data utility for microdata masked for disclosure limitation. *Journal of Privacy and Confidentiality* forthcoming.

Zwick, T. (2005). Continuing vocational training forms and establishment productivity in Germany. *German Economic Review* **6**, 2, 155–184.

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

Table 2: 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

Table 3: Average confidence interval overlap for average number of employees for 10 independent simulations of one stage synthesis.

	m=3	m=10	m=50	m=100
Simulation 1	0.754	0.815	0.816	0.812
Simulation 2	0.818	0.818	0.808	0.820
Simulation 3	0.812	0.813	0.820	0.816
Simulation 4	0.854	0.796	0.819	0.817
Simulation 5	0.823	0.808	0.808	0.824
Simulation 6	0.796	0.801	0.823	0.807
Simulation 7	0.787	0.778	0.819	0.819
Simulation 8	0.785	0.799	0.815	0.821
Simulation 9	0.770	0.829	0.823	0.821
Simulation 10	0.808	0.804	0.821	0.809
Average	0.801	0.806	0.817	0.817

Table 4: Averages of disclosure risk measures over ten repetitions of the simulation.

	m=3	m=10	m=50	m=100
Expected match risk	67.8	94.8	126.9	142.5
True match risk	35.2	82.5	126.1	142.4

Table 5: Average CI overlap and match risk for different two stage imputations (10 simulation runs).

m,r	Avg. overlap	Expected match risk	True match risk
m=3,r=3	0.819	83.1	67.6
m=3,r=16	0.819	98.0	91.8
m=3,r=33	0.822	99.8	96.3
m=5,r=10	0.823	106.1	101.2
m=10,r=5	0.824	113.8	109.4
m=16,r=3	0.824	119.9	116.4