Journal of Privacy and Confidentiality (2009) **1**, Number 2, pp. 223–233

Multiple Imputation for Disclosure Limitation: Future Research Challenges

Jerome P. Reiter*

1 Introduction

Statistical agencies that disseminate data to the public are ethically and often legally required to protect the confidentiality of respondents' identities and sensitive attributes. To satisfy these requirements, Rubin (1993), Little (1993), and Fienberg (1994) proposed that agencies utilize multiple imputation. 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. Multiple imputation for protecting confidentiality is often called the synthetic data approach.

In recent years, agencies have begun to use synthetic data approaches to create public use data for major surveys. In 2007, the U.S. Census Bureau released a synthetic, public use file for the Survey of Income and Program Participation that includes imputed values of social security benefits information and dozens of other highly sensitive variables (Abowd et al., 2006). 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 Community Survey by replacing demographic data for people at high disclosure risk with imputations (Hawala, 2008). Synthetic, public use datasets are in the development stage in the U.S. for the Longitudinal Business Database (Kinney and Reiter, 2007), the Longitudinal Employer-Household Dynamics survey, and the American Community Survey veterans and full sample data. Statistical agencies in Australia, Canada, Germany (Drechsler et al., 2008), and New Zealand (Graham and Penny, 2005) are also investigating the approach. Other examples of synthetic data are described by Kennickell (1997), Fienberg et al. (1998), Abowd and Woodcock (2001, 2004), Reiter (2002, 2005a), Little et al. (2004), Mitra and Reiter (2006), and An and Little (2007).

In this article, I describe some open research questions in synthetic data. Arguably, these questions must be resolved if the promises of synthetic data are to be realized on a wide scale. I do not review the methods for obtaining inferences from the multiple synthetic datasets, which differ from the usual combining rules for multiple imputation for missing data. For a summary of inferential methods for various adaptations of multiple imputation, see Reiter and Raghunathan (2007).

*Department of Statistical Science, Duke University, Durham, NC, mailto:jerry@stat.duke.edu

2 Review of synthetic data methods

Synthetic data approaches come in two main flavors: fully and partially synthetic. These are described briefly here.

2.1 Fully synthetic data

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

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

To illustrate how a secondary data analyst utilizes these ten datasets, suppose that the analyst seeks to fit a logistic regression of disease status on income, race, and sex. The analyst estimates the regression coefficients and their variances in each simulated dataset using standard likelihood-based estimates and software. The analyst averages the estimated coefficients and variances across the simulated datasets. These averages are used to form 95% confidence intervals based on the formulas developed by Raghunathan et al. (2003).

Releasing fully synthetic data makes it difficult for data snoopers to identify originally sampled units and learn their sensitive values. Almost all of the released units are not in the original sample, having been randomly selected from the sampling frame, and their values of survey data are simulated. The synthetic records cannot be matched meaningfully to records in other datasets, such as administrative records, because the values of released survey variables are simulated rather than actual. Releasing fully synthetic data is subject to attribute disclosure risk—the risk that the released data can be used to estimate unknown sensitive values very closely—when the models used to simulate data are "too accurate." For example, when data are simulated from a regression model with a very small mean square error, analysts can estimate outcomes precisely using the model, if they know predictors in that model. Or, if all people in a certain demographic group have the same, or even nearly the same, value of an outcome variable, the imputation models likely will generate that value for imputations.

Agencies can reduce these types of risks by using less precise models when necessary.

Fully synthetic datasets can have positive analytic features. When data are simulated from distributions that reflect the distributions of the collected data, frequency-valid inferences can be obtained from the multiple synthetic datasets for a wide range of estimands. These inferences can be determined by combining standard likelihood-based or survey-weighted estimates (Raghunathan et al., 2003; Reiter, 2005b); the analyst need not learn new statistical methods or software programs to adjust for the effects of the disclosure limitation. Synthetic datasets can be sampled by schemes other than the typically complex design used to collect the original data, so that analysts can ignore the design for inferences and instead perform analyses based on simple random samples. Additionally, the data generation models can incorporate adjustments for nonsampling errors and can borrow strength from other data sources, thereby resulting in inferences that can be even more accurate than those based on the original data. Finally, because all units' data are simulated, geographic identifiers can be included in the synthetic datasets, facilitating estimation for small areas.

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

2.2 Partially synthetic data

Partially synthetic data comprise the units originally surveyed with some collected values replaced with multiple imputations. To illustrate a partially synthetic strategy, we can adapt the setting used in Section 2.1. Suppose the agency wants to replace income when it exceeds $100,000 and is willing to release all other values. The agency generates replacement values for the incomes over $100,000 by randomly simulating from the distribution of income conditional on race, sex, and disease status. To avoid bias, this distribution also must be conditional on income exceeding $100,000. The distribution is estimated using the collected data and possibly other relevant information. The result is one synthetic data set. The agency repeats this process multiple times and releases the multiple datasets to the public.

As with fully synthetic data, when the replacement imputations are generated ef-

fectively, analysts can obtain valid inferences for a wide class of estimands with simple combining rules (Reiter, 2003). An advantage of partially synthetic data relative to fully synthetic data is that only a fraction of the data are imputed, so that analysts' inferences are generally less sensitive to the agency's model specification. Unlike fully synthetic data, partially synthetic data must be analyzed in accordance with the original sampling design.

The protection afforded by partially synthetic data depends on the nature of the synthesis. Replacing key identifiers with imputations makes it difficult for users to know the original values of those identifiers, which reduces the chance of identifications. Replacing values of sensitive variables makes it difficult for users to learn the exact values of those variables, which can prevent attribute disclosures. Nonetheless, there remain disclosure risks in partially synthetic data no matter which values are replaced. Analysts can utilize the released, unaltered values to facilitate disclosure attacks, for example via matching to external databases, or they may be able to estimate genuine values from the synthetic data with reasonable accuracy.

When some data are missing, multiple imputation can be used to fill in missing data and replace confidential values simultaneously with a two stage imputation approach. See Reiter (2004b) for details.

3 Research challenges

The key research challenges specific to synthetic data can be classified in four broad areas: (1) developing accurate synthesis models, (2) developing methods for selecting which values to synthesize, (3) developing ways to provide feedback on the quality of synthetic data inferences, and (4) developing methods that enable users to do multivariate and other analyses. Each of these is discussed below.

3.1 Flexible synthesis models

The key to the success of synthetic data approaches, especially when replacing many values, is the data generation model. Current practice for generating synthetic data uses sequential modeling strategies based on parametric or semi-parametric models, similar to those for imputation of missing data in Raghunathan et al. (2001). The basic idea is to impute X_1 from a regression of X_1 on $(X_2, X_3, etc.)$, impute X_2 from a regression of X_2 on $(X_1, X_3, etc.)$, impute X_3 from a regression of X_3 on $(X_1, X_2, etc.)$, and so on. An advantage of this strategy is that it is generally easier to specify plausible conditional models than plausible joint distributions. A disadvantage is that the collection of conditional distributions is not guaranteed to correspond to a proper joint distribution, particularly when the models use different conditioning sets.

Even when replacing only a fraction of values, specifying imputation models can be daunting in surveys with hundreds of variables available for conditioning. The data frequently include numerical, categorical, and mixed variables, some of which may not

be easy to model with standard parametric tools. Therefore, it may be advantageous to use non-parametric methods to generate imputations. Progress on this front has been made already. Reiter (2005c) modified classification and regression trees (CART) into methods for synthesizing numerical and categorical variables. He applied sequential versions of these methods to create partially synthetic data with high utility and low risk for a subset of the Current Population Survey.

In other contexts, recently developed methods from machine learning, namely support vector regression (Drucker et al., 1997) and random forests regression (Breiman, 2001), have been shown to outperform CART on out-of-sample predictive accuracy, especially in high dimensional data. These methods can capture complex relationships that might not be easily found or estimated with standard parametric techniques, and they avoid stringent parametric assumptions. They are computationally fast and easy to implement, with little tuning required by the user. Because of these features, they are widely and successfully used for prediction in high dimensional problems in data mining and bioinformatics (Hastie et al., 2001).

This suggests that, suitably adapted, these techniques from machine learning have great potential as methods for generating model-free synthetic data with high analytic validity. To illustrate how this might work in practice, suppose that the agency seeks to synthesize some categorical variable Y given other variables X. The agency estimates a random forest for Y as a function of X with the entire dataset. For any record with values x^*, where x^* may differ from x because of previous synthetic data replacements, the agency traces down each tree in the forest to find the terminal leaf corresponding to x^*. Each value of Y in the terminal leaves is a plausible synthetic data replacement. The agency simply draws one of these leaves at random to obtain a synthetic value y^*, or equivalently it draws from a multinomial distribution where the terminal values of Y represent the "data." This type of approach can be reasonably automated, an advantage for agencies under time and resource pressure. Developing and evaluating sequential versions of these techniques for generating partially synthetic data is a key research area.

3.2 Synthesis design strategies

Given their reduced reliance on imputation models, partially synthetic data may be more appealing to agencies than fully synthetic data. With partial synthesis, agencies must decide which values to replace with imputations. It may be sufficient from the perspective of disclosure risk to replace only some (not all) quasi-identifiers, as is proposed for the American Community Survey full sample synthesis. For example, a person might possess a unique combination of age, race, sex, marital status, and county. This person might no longer be at risk if either (i) county is not released exactly, (ii) age is not released exactly, or (iii) sex and race both are not released exactly. These represent three different synthesis choices for the agency. The task of selecting among choices is complicated by the fact that replacing values for some records could impact risks for other records. For example, suppose the intruder knows that record 1 has a larger income than record 2, and the intruder could identify these records from released

exact incomes. Synthesizing income (with sufficient variability) for either record might reduce risk for both records. As another example, it may prudent to synthesize some low risk records in addition to the high risk ones, so that the presence of synthesized variables does not automatically imply that a record was at high risk (Liu and Little, 2002).

The choice of synthesis should not depend on risk alone; data utility should affect decisions. Some quasi-identifying variables may have greater impact on data utility than others. For example, county may be deemed less critical to analyses than age, sex, and race, so that synthesizing county may be preferred to the other choices. Utility might depend on individual values. For example, suppose that a record has high leverage for a regression of interest to analysts, and that leverage is attributable to one variable X. Altering X has greater impact on the coefficients of that regression than altering some other values.

Research is needed to develop and evaluate strategies for selecting values to synthesize, based on risk and utility trade-offs. As an example of the general types of strategies that could be investigated, the agency synthesizes enough quasi-identifiers for each record until the probability of identification dips below some threshold. The order of synthesis is selected to minimize the loss in data utility. As another example, the agency improves protection by synthesizing data for low risk records, selecting those that reduce risk for high risk records without severely impacting data utility. For these strategies, the research involves (i) finding suitable measures of risk and utility, and (ii) finding computationally feasible heuristics for determining the ordering of variables or selection of records.

This research requires record-level risk measures—such as those of Reiter and Mitra (2009) and Drechsler and Reiter (2008)—and utility measures for synthetic data. In the context of traditional disclosure limitation methods, Woo et al. (2009) suggest trying to discriminate between the original and altered (synthetic) data. When it is possible to discriminate easily, the altered and original data do not have similar distributions, so that the altered data presumably have lower quality. This metric, while useful, is not sensitive enough to gauge the impact of altering individual values. Hence, to develop methods for selecting the values to synthesize, new utility measures are needed.

3.3 Confidence in synthetic data

Many potential data analysts are reluctant to trust synthetic data. This is understandable, particularly when large amounts of data are being replaced. In such cases, the validity of the results are almost entirely dependent on the validity of the synthetic data models. It is therefore necessary for proponents of synthetic data to demonstrate the validity of synthetic data to the public with real-world examples.

The most extensive testing of the analytical validity of synthetic data has been done for the Survey of Income and Program Participation (SIPP). In 2001, the Census Bureau, the Internal Revenue Service, and the Social Security Administration decided to supplement the information on SIPP panels from 1990 – 1996 with detailed earnings

and Social Security benefits histories. Because of the highly sensitive nature of these supplemental data, the three agencies agreed to release a version of the linked data only if sensitive and identifying information was synthesized. In the end, the agencies determined that it was necessary to synthesize all but four out of over six hundred variables in the linked data. Abowd et al. (2006) compare the observed and synthetic confidence intervals for a large number of estimands, including linear regressions, logistic regressions, means, sub-domain means, and time series analysis. These estimands were taken from the analyses in the published literature and were not explicitly derived from the synthesis models. For most but not all estimands, they find a high degree of overlap in the confidence intervals computed from the synthetic and observed data.

Such empirical evidence aside, some inferences will deteriorate significantly because of imperfect imputation models. When simulating high fractions of data, even small biases can cause substantial reductions in frequentist validity. These biases may be hard to detect from any meta-data released by the agency describing the synthesis process. For this reason, it is arguably essential that agencies develop ways to provide feedback to users about the quality of the synthetic data inferences for specific estimands. One possibility is to build a verification server, as suggested by Reiter et al. (2009). The basic idea is as follows: The data user performs an analysis of the synthetic data, using whatever software she wishes. She then submits a description of the analysis to the verification server; for example, regress attribute 5 on attributes 1, 2, 4 and the logarithm of attribute 6. The verification server performs the analysis on both the confidential data and the synthetic data, and from the results calculates analysis-specific measures of the fidelity of the one to the other. For example, for any regression coefficient, measure the overlap in its confidence intervals (Karr et al., 2006) computed from the real and synthetic data. If the intervals largely overlap, the synthetic data have high utility for that analyses. The verification server returns the value of the fidelity measure to the user. With such feedback, analysts can avoid publishing—in the broad sense—results with poor quality and be confident about results with good quality.

Verification servers are not a panacea. As illustrated by Reiter et al. (2009), fidelity measures provide intruders with information about the real data, albeit in a convoluted form, that could be used for disclosure attacks. As a simple example, suppose the intruder requests and receives the fidelity measure for an analysis that uses one record with a synthetic value of income and ten records with original (not replaced) values of income. To learn the true value of income for the confidential record, the intruder can compute the fidelity measure for each of many guesses at that true income value. The confidential record's true income is the guess that results in the fidelity measure reported by the server.

It may be possible to blunt these attacks by providing coarse fidelity measures or by limiting the types of queries that the server answers. Assessing and reducing the risks of providing fidelity measures are topics of ongoing research. Such research would benefit public use data dissemination in general, regardless of the disclosure limitation methods.

3.4 Expansion of analysis methods

The analysis of multiply-imputed, synthetic datasets involves combining point and variance estimates from the multiple datasets. Currently, methods exist for obtaining interval estimates for scalar quantities and for performing large sample tests of multi-component hypotheses. Recent work by Kinney (2007) suggests methods for performing automated model selection for linear models. Methods do not exist, however, for a wide range of complex analyses often done with rich datasets. These include, for example, multivariate analyses such as cluster and factor analysis, hierarchical (multi-level) models, or Bayesian models. We need to expand the types of analysis that can be done with synthetic data.

3.5 Opportunities for interdisciplinary research

The synthetic data agenda has many opportunities for interdisciplinary research. The development of flexible modeling strategies lies at the boundary of computer science and statistical science. A key issue is to develop methods that reproduce appropriate variability in the synthetic data: predictions of conditional means do not suffice. Developing heuristics and efficient algorithms for selecting values to synthesize is as much a computer science problem as a statistical one. The verification server requires the expertise of systems and data security researchers, statistical researchers who can propose metrics of risk and utility, and subject-matter researchers who can evaluate the impacts on usefulness of alteration of fidelity measures.

4 Concluding remarks

As resources available to malicious data users continue to expand, the alterations needed to protect public use data with traditional disclosure limitation techniques—such as swapping, adding noise, or microaggregation—may become so extreme that, for many analyses, the released data are no longer useful. Synthetic data, on the other hand, has the potential to enable public use data dissemination while preserving data utility. Ultimately, a statistical disclosure limitation strategy that combines restricted data access for sophisticated analyses and synthetic data for a wide range of simple analyses, such as regressions and comparisons of means, should meet the needs of most secondary data users.

References

Abowd, J., Stinson, M., and Benedetto, G. (2006). Final report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project. Technical report, U.S. Census Bureau Longitudinal Employer-Household Dynamics Program. Available at `http://www.bls.census.gov/sipp/synth\_data.html`.

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.

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Drechsler, J., Dundler, A., Bender, S., Rässler, S., and Zwick, T. (2008). 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. 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.

Drucker, H., Burges, J. C., Kaufman, L., Smola, A., and Vapnik, V. (1997). Support vector regression machines. In *Advances in Neural Information Processing Systems 9, NIPS 1996*, 155–161. Cambridge: MIT Press.

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

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

Graham, P. and Penny, R. (2005). Multiply imputed synthetic data files. Technical report, University of Otago. Republished in the *Official Statistics Research Series*, Volume 1, 2007. Available at `http://www.statisphere.govt.nz/official-statistics-research/series/volume-1-2007.aspx`.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2001). *The Elements of Statistical Learning*. Springer.

Hawala, S. (2008). Producing partially synthetic data to avoid disclosure. In *Proceedings of the Joint Statistical Meetings*. Alexandria, VA: American Statistical Association.

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.

Kinney, S. (2007). *Model Selection and Multivariate Inference Using Data Multiply Imputed for Disclosure Limitation and Nonresponse*. PhD thesis, Duke University, Dept. of Statistical Science.

Kinney, S. K. and Reiter, J. P. (2007). Making public use, synthetic files of the Longitudinal Business Database. In *Proceedings of the Joint Statistical Meetings*. Alexandria, VA: American Statistical Association.

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 Perspectives*, 141–152. New York: John Wiley & Sons.

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

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.

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

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

Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29:181–189.

Reiter, J. P. (2004a). New approaches to data dissemination: A glimpse into the future (?). *Chance*, 17(3):12–16.

Reiter, J. P. (2004b). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology*, 30:235–242.

Reiter, J. P. (2005a). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society, Series A*, 168:185–205.

Reiter, J. P. (2005b). Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. *Journal of Statistical Planning and Inference*, 131:365–377.

Reiter, J. P. (2005c). Using CART to generate partially synthetic, public use microdata. *Journal of Official Statistics*, 21:441–462.

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., 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. (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.