Satisfying Disclosure Restrictions With Synthetic Data Sets

Jerome P. Reiter*

Key Words: Confidentiality, Disclosure, Multiple Imputation, Simulation, Synthetic Data

Abstract

To avoid disclosures, Rubin proposed creating multiple, synthetic data sets for public release so that (i) no unit in the released data has sensitive data from an actual unit in the population, and (ii) statistical procedures that are valid for the original data are valid for the released data. In this article, I show through simulation studies that valid inferences can be obtained from synthetic data in a variety of settings, including simple random sampling, probability proportional to size sampling, two-stage cluster sampling, and stratified sampling. I also provide guidance on specifying the number and size of synthetic data sets and demonstrate the benefit of including design variables in the released data sets.

1 Introduction

When considering the release of data sets to the public, statistical agencies face competing objectives. They seek to provide users with sufficiently detailed data and also to guard the confidentiality of survey respondents. Commonly used methods for meeting these objectives include cell suppression, data masking, and data swapping (e.g., see Willenborg and de Waal, 2001). However, these methods can compromise estimation by distorting relationships among variables in the data set.

*Institute of Statistics and Decision Sciences, Duke University, Durham, NC 27708-0251. E-mail: jerry@stat.duke.edu.

Another approach is to create multiple, synthetic data sets for public release, as proposed by Rubin (1993). In this approach, the agency selects units from the sampling frame and imputes their data using models fit with the original survey data. The approach has three potential benefits. First, it can preserve confidentiality, since identification of units and their sensitive data can be difficult when the data for some or all of the variables in the data set are not actual, collected values. Second, with appropriate estimation methods based on the concepts of multiple imputation (Rubin, 1987), the approach can allow data users to make valid inferences for a variety of estimands without placing undue burdens on these users. Third, synthetic data sets can be sampled by schemes other than the typically complex design used to collect the original data, so that users of synthetic data can ignore the design for inferences.

Variations of the synthetic approach have been suggested or implemented by several authors. Rubin (1993) proposes full simulation, in which (i) units are randomly sampled from the sampling frame for each synthetic data set, and (ii) unknown data values for units in the synthetic samples are imputed. Inferential methods for analyzing such data sets have been developed by Raghunathan *et al.* (2003). Fienberg *et al.* (1998) use the sample cumulative distribution functions and bootstrapping to construct synthetic, categorical data. Little (1993), in a general discussion of the analysis of masked data, presents the possibility of simulating only variables that are potential identifiers. Kennickell (1997) protects several monetary variables in the Survey of Consumer Finances by releasing a mixture of the original survey data and multiple imputations of values that are high disclosure risks. He constrains the imputation models so that the imputed values are reasonably close to the actual values. Abowd and Woodcock (2001) generate synthetic data to avoid disclosures in longitudinal, linked data sets. Each replication in their synthetic data consists of the units originally surveyed, but all units' values are imputed.

A distinguishing feature of Rubin's (1993) full simulation approach is that the released units differ from the units originally surveyed. No actual values or near actual values of sensitive variables are purposefully released. Furthermore, since the released units differ across synthetic data sets, intruders should have more difficulty using the multiple imputations to assist their identification efforts. For these reasons, the full simulation approach promises to guard confidentiality more closely than releasing actual or imputed data for the units originally surveyed. However, this extra protection

comes at a cost: the validity of inferences rely critically on the accuracy of the imputation model. As noted by a referee of this article, this cost may partially explain the dearth of research on the full simulation approach.

In this article, I present results of some applied research on the full simulation approach. Using simulation studies, I show that valid inferences can be obtained in a variety of sampling designs, including simple random sampling, stratified sampling, probability proportional to size sampling, and two-stage cluster sampling. This is encouraging news, since the validity of multiple imputation procedures in design settings other than simple random samples has been questioned by some researchers (see Fay, 1996; Rubin, 1996). I also provide guidance on specifying the number and size of synthetic data sets and demonstrate that including design variables, such as stratum indicators, in the released data sets facilitates inferences.

This article is organized as follows. Section 2 describes the inferential methods proposed by Raghunathan *et al.* (2003). Section 3 presents the simulation studies I use to investigate the performance of these methods. Section 4 examines the sensitivity of inferences obtained from these methods to changes in the number of synthetic data sets, the number of synthetic units, and the inclusion of design variables in the synthetic data. Section 5 concludes with some remarks on this approach.

2 Inferences from Multiple Synthetic Data Sets

To describe construction of and inferences from multiple synthetic data sets, we use notation similar to that of Raghunathan *et al.* (2003). Let $I_j = 1$ if unit j is selected in the original survey, and $I_j = 0$ otherwise. Let $I = (I_1, \ldots, I_N)$. Let Y_{obs} be the $n \times p$ matrix of collected (real) survey data for the units with $I_j = 1$; let Y_{nobs} be the $(N - n) \times p$ matrix of unobserved survey data for the units with $I_j = 0$; and, let $Y = (Y_{obs}, Y_{nobs})$. For simplicity, we assume that all sampled units fully respond to the survey. Let X be the $N \times d$ matrix of design variables for all N units in the population (e.g, stratum or cluster indicators or size measures). We assume that such design information is known at least approximately, for example from census records or the sampling frames.

The agency releasing synthetic data, henceforth abbreviated as the *im-*

puter, constructs synthetic data sets based on the observed data (X, Y_{obs}, I) in a two-part process. First, the imputer imputes values of Y for the $N-n$ unobserved units to obtain a completed-data set. The imputer also may choose to impute values of Y for all N units so that the completed-data contains no real values of Y, thereby avoiding the release of any respondent's value of Y. We assume that imputations are generated from the Bayesian posterior predictive distribution of $(Y|X, Y_{obs}, I)$. Second, the imputer samples units randomly from the completed-data population. These sampled units are released as public use data, so that the released data set contains the values of Y only for units in the synthetic sample. This process is repeated independently m times to get m different synthetic data sets.

This process of data creation differs from the inverse sampling methods of Hinkins *et al.* (1999). In inverse sampling, simple random samples are generated from the units collected in the survey. In this method, simple random samples of new units are taken from the sampling frame itself.

We now specify a formal notation for the process of synthetic data construction. Let $(X, Y_{com,i})$ be the completed-data population from which n_{syn} units are sampled to obtain synthetic data set i. Let $Z_{ij} = 1$ if unit j is selected in synthetic data set i, and $Z_{ij} = 0$ otherwise. Let $Z_i = (Z_{i1}, \ldots, Z_{iN})$. Let $Y_{syn,i}$ be the $n_{syn} \times p$ vector of released, synthetic data for units with $Z_{ij} = 1$. The released synthetic data set i is expressed as $(X, Y_{syn,i}, Z_i)$, where all of X is included since design information is assumed known for all units. In practice, it is not necessary to generate completed-data populations for constructing $Y_{syn,i}$. Instead, the imputer need only generate values of Y for units with $Z_{ij} = 1$.

From these synthetic data sets, some user of the publicly released data, henceforth abbreviated as the *analyst*, seeks inferences about some estimand $Q = Q(X, Y)$, where the notation $Q(X, Y)$ means that the estimand Q is a function of (X, Y). For example, Q could be the population mean of Y or the population regression coefficients of Y on X. In each synthetic data set i, the analyst estimates Q with some estimator $q_i = Q(X, Y_{syn,i}, Z_i)$ and estimates the variance of q_i with some estimator $v_i = V(X, Y_{syn,i}, Z_i)$. We assume that the analyst determines the q_i and v_i as if the synthetic data were in fact collected data from a simple random sample of (X, Y).

In this article, we assume that the imputer and analyst both use the actual posterior predictive distribution of Y. Under this assumption, the analyst can obtain valid inferences for Q by combining the q_i and v_i. Specifically, the

following quantities are needed for inferences:

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

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

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

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

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

to estimate the variance of $\bar{q}_m$. The $b_m - \bar{v}_m$ is an unbiased estimator of the variance of $q_{obs} = Q(X, Y_{obs}, I)$, and the $\frac{1}{m}b_m$ adjusts for using only a finite number of synthetic data sets. When $T_s > 0$, and n, n_{syn}, and m are large, inferences for scalar Q can be based on normal distributions. For moderate m, inferences can be based on t-distributions with degrees of freedom

$$\nu_s = (m-1)(1 - r_m^{-1})^2 \tag{5}$$

where $r_m = (1 + m^{-1})b_m/\bar{v}_m$, so that a $(1-\alpha)\%$ interval for Q is

$$\bar{q}_m \pm t_{\nu_s}(\alpha/2)\sqrt{T_s}. \tag{6}$$

Although not in Raghunathan *et al.* (2003), this reference t-distribution was presented by Raghunathan and Rubin at the International Society for Bayesian Analysis conference in June 2000. Extensions for multivariate Q are not presented here.

Because there may be some estimators for which T_s is negative, particularly when m is modest, it is necessary to have some condition that forces the estimator of $Var(\bar{q}_m)$ to be positive. Thus, I replace (4) with the modified variance estimator,

$$T_s^* = max(0, T_s) \; + \; \delta \; * \left(\frac{n_{syn}}{n}\bar{v}_m\right) \tag{7}$$

where $\delta = 1$ if $T_s < 0$, and $\delta = 0$ otherwise. Negative values of T_s generally can be avoided by increasing m or n_{syn}.

The variance of $\bar{q}_m$ in the synthetic data setting differs from the variance of the analogous $\bar{q}_m$ in the setting of multiple imputation for nonresponse. In the synthetic data setting, the variance calculation involves the distribution used to generate the $(X, Y_{com,i})$ and the additional step of randomly sampling units from this completed-data population. In the usual multiple imputation setting, the variance calculation involves only the distribution used to create imputations for the units with missing data. In fact, as shown in the simulations, the usual variance formula for multiple imputations, $T_m = (1 + \frac{1}{m})b_m + \bar{v}_m$, tends to overestimate significantly the variance of the synthetic $\bar{q}_m$.

3 Simulation Studies

We investigate the performance of these methods in simulation studies of four settings:

- estimate a population mean from a simple random sample,
- estimate a population mean from a stratified simple random sample,
- estimate a regression coefficient from a probability proportional to size sample,
- estimate a regression coefficient from a two-stage cluster sample.

The investigations focus on the coverage of asymptotic 95% confidence intervals; they do not examine the potential of the synthetic data approach to preserve confidentiality.

In all simulations, we use the correct posterior predictive distribution to draw synthetic data sets. Of course, in actual implementations, the correct posterior predictive distribution is not known, and an imputer-constructed approximation is used. Nonetheless, these idealized simulations help us gauge the promise of releasing synthetic data sets.

3.1 Simple Random Sampling

Assume that we want to estimate the mean of some variable, Y, in a population of size N from a simple random sample of size $n = 100$. Let

$Y \sim N(0, 100)$. Further, we assume that $N >> n$, so that the finite population correction factor can be ignored when estimating variances.

For each of 500 replications, we construct a collected data set, $Y_{obs} = (Y_1, \ldots, Y_{100})$, by drawing randomly from $Y_j \sim N(0, 100)$ for $j = 1, \ldots, 100$. The Bayesian posterior predictive distribution of Y is,

$$f(Y|Y_{obs}) = \int f(Y|\theta) f(\theta|Y_{obs}) d\theta, \tag{8}$$

where $\theta = (\mu, \sigma^2)$ are the parameters of the normal distribution. To construct each synthetic data set i, we use standard noninformative priors on all parameters and draw $n_{syn} = 100$ values from (8). This process is repeated independently in $m = 100$ data sets for each replication.

Following the prescription for analyzing multiple synthetic data sets, in synthetic data set i we let

$$q_i = \bar{y}_{syn,i} \tag{9}$$

$$v_i = \frac{\sum (y_{ij} - \bar{y}_{syn,i})^2}{(n_{syn} - 1) n_{syn}}. \tag{10}$$

A summary of the actual coverages of 95% confidence intervals for the mean of Y is shown in Table 1. In that table and other tables that follow, the "Observed Data method" constructs 95% confidence intervals with $q_{obs} \pm 1.96\sqrt{u_{obs}}$, where u_{obs} is the estimate of $Var(q_{obs})$ obtained from the observed data; the "T_s method" uses $\bar{q}_m \pm t_{v_s}\sqrt{T_s}$; and, "Method T_m" uses $\bar{q}_m \pm 1.96\sqrt{T_m}$. The column labeled "Avg. $\hat{q}$" contains the averages across all replications of the point estimates of Q. The column labeled "Avg. Est. Var." contains the averages across all replications of the estimated variances. The column labeled "95% CI cov." contains the percentages of confidence intervals that cover Q.

Table 1: Results for SRS simulation (m=100)

Method	Avg. $\hat{q}$	Avg. Est. Var.	95% CI cov.
Observed Data	.04	1.00	94.2
T_s	.04	1.08	94.0
T_m	.04	3.10	100.0

The average point estimate of the population mean is close to the population value of zero whether we use the actual data or the synthetic data. This is a benefit of using the correct posterior distribution when drawing synthetic data. The actual variance of $\bar{q}_{100}$ across the 500 replications is 1.09, so that the simulations verify that T_s is unbiased. $T_s > 0$ in all 500 replications. Confidence coverage of intervals constructed with T_s mirror those of the observed data, and both coverage percentages are within simulation error of nominal 95% coverage.

The 95% confidence intervals constructed by using T_m are too wide. As discussed previously, the distributions used in the development of the variance formulae for multiple imputation differ from the distributions used to create synthetic data sets, so that T_m overestimates $Var(\bar{q}_m)$.

3.2 Stratified Simple Random Sampling

Assume again that we wish to estimate a population mean of some variable Y. Let each unit j be a member of only one stratum h, where $h = 1, \ldots, 10$ and for all h the size of the stratum, N_h, equals 1,000. We construct the population by drawing values from $Y_{hj} \sim N(10 * h, h^2)$. The actual mean of the 10,000 observations in the generated data is 54.94.

Because of the substantial differences in the means and variances across strata, a stratified simple random sample should yield more accurate estimates of the population mean than a simple random sample of the same number of units. That is, the usual unbiased estimator with a stratified random sample, $\bar{Y}_{strat} = \sum_h \frac{N_h}{N} \bar{Y}_h$, has smaller variance than the usual unbiased estimator with a simple random sample, $\bar{Y}$.

In each of 500 replications, we sample a collected data set from this population by taking a simple random sample of 20 units from each stratum. To construct synthetic each data set i, we draw a simple random sample of $n_{syn} = 200$ stratum indicators from the population of 10,000 units. The value of Y_{hj} for sampled synthetic unit j in stratum h is drawn from the full Bayesian posterior predictive distribution,

$$f(Y_{hj}|Y_{obs}, X) = \int f(Y|\theta_h, Y_{obs}, X) f(\theta_h|Y_{obs}, X) d\theta_h, \tag{11}$$

where $\theta_h = (\mu_h, \sigma_h^2)$ are the parameters of the normal distribution in stratum h, and X is a vector of stratum indicators for all N units. Standard

noninformative priors are used for all parameters. This process is repeated in $m = 100$ data sets for each replication.

We assume that the values of the N_h are available to the analyst, for example from census tabulations. Following the prescription for analyzing multiple synthetic data sets, in synthetic data set i we let

$$q_i = \sum_{h=1}^{10} \frac{N_h}{N} \bar{y}_{ih} \tag{12}$$

$$v_i = \sum_{h=1}^{10} \left(1 - \frac{n_{ih}}{1000}\right) \left(\frac{N_h}{N}\right)^2 \frac{\sum_j (y_{ihj} - \bar{y}_{ih})^2}{(n_{ih} - 1)n_{ih}} \tag{13}$$

where n_{ih} is the number of units in stratum h in synthetic data set i. A summary of the actual coverages of 95% confidence intervals for the population mean of Y is shown in Table 2. The observed data inferences are based on the usual unbiased variance estimator for stratified simple random sampling.

Table 2: Results for STRS simulation (m=100)

Method	Avg. $\hat{q}$	Avg. Est. Var.	95% CI cov.
Observed Data	54.95	.20	94.6
T_s	54.96	.23	96.0
T_m	54.96	.69	100.0

The average point estimates from the observed and synthetic data sets are close to the actual population mean. All $T_s > 0$. The actual variance across 500 replications of $\bar{q}_{100}$ is .21, so that T_s is slightly biased. This bias results from the variation in the n_{ih} across i, which we have not accounted for in the v_i. Coverage rates for the T_s method are slightly larger than those of the observed data because of the inflated variances. With large n_{syn}, the effect of variation in the n_{ih} on inferences is minimized, and the synthetic data inferences can be expected to match the observed data inferences. As in the SRS simulation, the multiple imputation variance estimator leads to substantial over-coverage.

3.3 Probability Proportional to Size Sampling

We now estimate a regression coefficient in a probability proportional to size sample. The hypothetical population is constructed of N=1,000 units with

4 survey variables, $(X1, X2, X3, X4)$. We draw $X1$ from an exponential distribution, draw $X2 \sim N(0, 3.5)$, draw $X3 \sim N(X1, 3.5)$, and draw $X4 \sim N(X1 + X2 + X3, 100)$. The estimand of interest is the regression coefficient of $X3$ in the regression of $X4$ on $(X1, X2, X3)$, which in the generated population equals 1.07. We assume that $X1$ is known for all units and is available for sampling the collected data and for creating synthetic data sets.

In each of 500 replications, we draw collected data by sampling 100 units with probability proportional to X1, without replacement, using the scheme of Sunter (1977) as described in Sarndal *et al.* (1992, pp. 93–96). The ratio of the largest to smallest value of X1 is 42/2, so that the design differs noticeably from simple random sampling.

To create synthetic data, we take $m = 100$ simple random samples of $n_{syn} = 100$ units from the created population. Since $X1$ is assumed known for all units, we use the actual values of $X1$ for the units in the synthetic data set. To create values of $X2, X3$, and $X4$, we draw from a series of conditional regressions derived from full Bayesian posterior predictive distributions. That is, $X2$ is drawn from its regression on $X1$; $X3$ is drawn from its regression on the synthetically drawn values of $(X1, X2)$; and, $X4$ is drawn from its regression on the synthetically drawn values of $(X1, X2, X3)$. Standard noninformative priors are assumed for all regression parameters.

Following the prescription for analyzing multiple synthetic data sets, in synthetic data set i we let q_i equal the estimated regression coefficient of $X3$ in the ordinary least squares regression of $X4$ on $(X1, X2, X3)$, and we let v_i equal the usual estimated variance of this estimated regression coefficient. A summary of the actual coverages of 95% confidence intervals for the regression coefficient is shown in Table 3.

Table 3: Results for PPS simulation (m=100)

Method	Avg. $\hat{q}$	Avg. Est. Var.	95% CI cov.
Observed Data	1.15	.29	96.6
T_s	1.15	.30	96.8
T_m	1.15	.90	100

The average point estimates from the synthetic data match those from the observed data, and both are slightly biased for the population regression coefficient. The actual variances across the 500 replications of q_{obs} and of

$\bar{q}_{100}$ are both .25. The average of the estimated variances for the observed and synthetic methods are larger than .25 because both variance estimators do not account for sampling from a finite population without replacement. The average value of T_s is close to the average value of u_{obs}, suggesting that a version of T_s that corrects for finite-population sampling would be unbiased. Confidence coverage for the synthetic data inferences is similar to the observed data coverage, and T_s is never negative. Once again, the multiple imputation variance estimator is very inefficient.

3.4 Two-stage Cluster Sampling

We now estimate a regression coefficient in a two-stage cluster sample. To construct the population, we use the values of $(X1, X2, X3, X4)$ for the 1,000 units in the PPS simulation and randomly form 20 clusters of size 50. For each unit j in cluster r, we add a cluster effect ω_r to $X4_{rj}$, where each ω_r is drawn independently from $\omega_r \sim N(0, 25)$. The estimand of interest is the regression coefficient of $X3$ in the regression of the new $X4$ on $(X1, X2, X3)$, which in the generated data remains 1.07 after accounting for the clustering.

In 500 replications, we create collected data by sampling in two stages: (i) a simple random sample of 10 clusters; and, (ii) within selected clusters, a simple random sample of 10 units. We assume that cluster indicators and $X1$ are known for all units and are released in the synthetic data sets.

To create synthetic data, we take a simple random sample of $n_{syn} = 100$ units from the population. Since $X1$ is assumed known for all units, we can use the values of $X1$ for the units in the synthetic data set. To create $X2$ and $X3$, we draw values from sequential regressions as is done in the PPS simulation. To draw $X4$, we use a three part process. First, we fit a random effects model to the collected data,

$$X4_{rj} = \beta_0 + \beta_1 X1_{rj} + \beta_2 X2_{rj} + \beta_3 X3_{rj} + \omega_r + \epsilon_{rj}, \tag{14}$$

where $\epsilon_{rj} \sim N(0, \sigma^2)$ and $\omega_r \sim N(0, \tau^2)$. We use this model to determine the posterior distribution of $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)$ and the posterior modes of τ and the ω_r for observed clusters. Second, to estimate ω_r for unobserved clusters, we randomly draw a cluster effect from a normal distribution with mean zero and variance equal to the posterior mode of τ. Finally, we draw β from its posterior distribution, and draw new $X4$ from its regression on $(X1, X2, X3)$, conditional on the estimated values of the cluster effects and the drawn values of β.

Following the prescription for analyzing multiple synthetic data sets, in synthetic data set i we let q_i equal the estimated coefficient of $X3$ in the random effects regression of the new $X4$ on $(X1, X2, X3)$. We let v_i equal the estimated variance of this estimated regression coefficient. A summary of the actual coverages of 95% confidence intervals for the regression coefficient across the 500 replications is shown in Table 4. In that table, both the observed data and synthetic data inferences are from random effects models of $X4$ on $(X1, X2, X3)$.

Table 4: Results for CLUS simulation (m=100)

Method	Avg. $\hat{q}$	Avg. Est. Var.	95% CI cov.
Observed Data	1.07	.32	96.4
T_s	1.07	.32	96.4
T_m	1.07	1.01	100

The average point estimates of the regression coefficient are close to the population value in both the observed and synthetic data. The actual variances of q_{obs} and $\bar{q}_{100}$ across the 500 replications are .30, and the average of the estimated variances for both observed and synthetic data are similar and slightly larger than .30. This increase is again a product of not accounting for sampling from a finite population without replacement. $T_s > 0$ in all replications, and confidence coverage for the synthetic data mirrors the coverage for the actual data. Again, the actual coverage and variance estimates are slightly inflated from ignoring finite population corrections. Finally, as in all three previous studies, using T_m leads to overcoverage.

4 Implementation Guidelines

In this section, I offer guidance on three implemention issues for agencies considering the synthetic data approach. First, I discuss selecting the number of released data sets and show that the usual $m = 5$ rule-of-thumb for multiple imputation may not be valid for synthetic data. Second, I discuss selecting the size of released data sets. Third, I discuss the benefits of releasing design information in the synthetic data sets.

4.1 The number of synthetic data sets

To reduce demands on users' storage and processing needs, imputers may want to release as few data sets as possible. When using multiple imputation for missing data, often the release of $m = 5$ imputed data sets provides sufficient inferential accuracy (Rubin, 1987). Unfortunately, this rule-of-thumb may not apply when releasing synthetic data and using the variance estimator in (4).

This can be demonstrated by repeating the simulation studies of Section 3 using $m = 5$ instead of $m = 100$ data sets. As shown in Table 5, point estimates continue to track the observed data point estimates, but synthetic data variance estimates are problematic. T_s is negative in 15% to 20% of the replications. Excluding the negative variance estimates does not help: the averages of the T_s for replications with positive variance estimates are about 20% larger than the actual $Var(\bar{q}_5)$. Using T^* from (7), which estimates $Var(\bar{q}_5)$ with $\bar{v}_m$ when $T_s < 0$, compounds this over-estimation.

Table 5: Point and variance estimates when m=5

Simulation	Avg. q_{obs}	Avg. $\bar{q}_5$	$Var(\bar{q}_5)$.	% of $T_s < 0$	Avg. T_s^*
SRS	-.010	-.018	1.43	20	1.72
STRS	54.92	54.94	.28	15	.46
PPS	1.16	1.16	.38	21	.54
CLUS	1.01	1.00	.40	16	.70

The problems with T_s in these settings are not due to bias; averaging over all replications in each simulation confirms that each T_s remains unbiased for its corresponding $Var(\bar{q}_5)$. Rather, the problems with T_s stem from the relatively large variances of b_5, which lead to substantial probabilities that $(1 + \frac{1}{m})b_m < \bar{v}_m$. For example, the variances of $\bar{v}_5$ and b_5 across all replications are 0.03 and 2.31, respectively, and their means are 1.02 and 2.05, respectively.

These probabilities can be approximated from the distribution of $(b_m|X, Y, I)$ over repeated draws of synthetic data. In the SRS simulation, the sampling distribution of $(b_m|X, Y, I)$ is a scaled chi-squared distribution with degrees of freedom $m - 1$,

$$\left(\frac{(m-1)b_m}{B}|X, Y, I\right) \sim \chi^2_{m-1} \tag{15}$$

where $B = u_{obs} + E(\bar{v}_m|X, Y, I)$. Using $B = 2.05$ and $\bar{v}_5 = 1.02$, the probability that $(1 + \frac{1}{5})b_5 < \bar{v}_5$ is approximately

$$Pr(1.2b_5 < 1.02) = Pr\left(\chi^2_4 < \frac{(1.02)(4)}{(1.2)(2.05)}\right) = .20, \tag{16}$$

which matches the rate of negative variance estimates in the simulation.

These results naturally lead to the question: for a given n_{syn} and estimator q_i, what is a reasonably small value of m that still allows users to obtain accurate inferences? Intuitively, imputers should select m so that T_s is reasonably close to $Var(\bar{q}_m)$. This can be achieved by requiring the variance of T_s to be substantially less than $Var(\bar{q}_m)$. This requirement is similar to some of the conditions for randomization validity of multiple imputation inferences developed by Rubin (1987, Ch. 4).

For a given estimator, the variance of T_s can be approximated as follows. Under the assumptions that (i) the sampling distribution in (15) remains valid for the estimator of interest, and (ii) the $Var(\bar{v}_m|X, Y, I)$ is small relative to $Var(b_m|X, Y, I)$,

$$\begin{aligned} Var(T_s|X, Y, I) &\approx \left(1 + \frac{1}{m}\right)^2 Var(b_m|X, Y, I) && (17) \\ &= \left(1 + \frac{1}{m}\right)^2 \frac{2B^2}{m-1}. && (18) \end{aligned}$$

These assumptions are satisfied in all four simulation studies of Section 3. The B can be estimated by simulating a very large number of synthetic data sets based on the observed data.

The value of m then can be selected so that (18) is some small fraction, say one-tenth, of u_{obs}. We compare to u_{obs} because it is easier to compute than $Var(\bar{q}_m)$. Since $u_{obs} \leq Var(\bar{q}_m)$, values of m chosen this way should be larger than values of m chosen with comparsions to $Var(\bar{q}_m)$. In the simulations of Section 3, requiring the approximate $Var(T_s|X, Y, I)$ to be one-tenth of u_{obs} suggests that $m \approx 40$ in the SRS example, $m \approx 20$ in the STRS example, and $m \approx 30$ in the PPS and CLUS examples. Simulation studies with these values of m produce $T_s < 0$ in less than 1% of the replications.

Since $Var(T_s)$ and $Var(\bar{q}_m)$ depend on the properties of the analyst's estimator and the imputer's method of data generation, reasonable values of m vary from setting to setting. For any collected data set, imputers can

use the method of this section to derive values of m for a variety of likely analysts' estimators. Then, the imputers can select an m that satisfactorily balances costs and inferential accuracy.

4.2 The number of units in synthetic data sets

The number of released units, n_{syn}, affects the values of b_m and $\bar{v}_m$. This is easily realized by noting that $\bar{v}_m = 0$ when $n_{syn} = N$. When $\bar{v}_m = 0$, it is always true that $T_s > 0$. Of course, releasing N units is not practical for many surveys. Thus, imputers need to consider how many units to release in the synthetic data sets.

To assess this question, we can examine how $Var(\bar{q}_m)$ and $Var(T_s)$ change as n_{syn} increases. Both of these quantities depend on $E(\bar{v}_m|X, Y, I)$ through $B = u_{obs} + E(\bar{v}_m|X, Y, I)$. Since $E(\bar{v}_m|X, Y, I)$ decreases with $\frac{1}{n_{syn}}$, increasing n_{syn} reduces $Var(\bar{q}_m)$ and $Var(T_s)$. However, when n_{syn} is relatively large, increases in n_{syn} do not decrease $E(\bar{v}_m|X, Y, I)$ substantially, so that reductions in $Var(\bar{q}_m)$ and $Var(T_s)$ are likely to be small. This is demonstrated in Table 6, which displays results of 1,000 replications of the SRS simulation using $m = 100$ and $n_{syn} = 100$, $n_{syn} = 1,000$, and $n_{syn} = 10,000$.

Table 6: Variances in SRS simulation for different n_{syn}

n_{syn}	Avg. $Var(\bar{q}_{100})$	Avg. T_s	Var. T_s
100	1.051	1.042	0.115
1,000	1.028	1.028	0.052
10,000	1.023	1.030	0.045

As this table shows, $Var(\bar{q}_{100})$ decreases slightly when going from $n_{syn} = 100$ to $n_{syn} = 1,000$, and there is hardly any change when going from $n_{syn} = 1,000$ to $n_{syn} = 10,000$. For all values of n_{syn}, T_s remains unbiased. Its variance decreases by about 55% when going from $n_{syn} = 100$ to $n_{syn} = 1,000$, but it decreases only by about 15% when going from $n_{syn} = 1,000$ to $n_{syn} = 10,000$.

Reductions in variance of T_s reduce the risk that $T_s < 0$. They also increase the degrees of freedom, ν_s, in the reference t-distribution. Thus, it is advantageous inferentially to release as many units as is feasible in the data set. However, releasing more units increases storage costs and increases

the likelihood that the same unit appears multiple times, which could have ramifications for disclosure protection. For large n_{syn}, these results suggest that the gains in estimation accuracy from increasing n_{syn} may not be worth these costs.

4.3 The inclusion of design information in synthetic data sets

By generating synthetic data from simple random samples of completed data populations, the imputer can release data that can be analyzed with standard likelihood or Bayesian approaches that ignore the design. However, there are still advantages to releasing design information. When design information is related to the survey variables, as in the PPS and CLUS simulations, analysts with access to the design information can properly include this information in their models. Additionally, releasing stratum or cluster indicators makes it easier for analysts to perform within-stratum or within-cluster analyses.

Releasing design information also can help analysts make synthetic data inferences about estimands that do not explicitly depend on such information. For example, consider estimating the population mean in the STRS simulation with a simple expansion estimator, $\bar{Y}_i$, as analysts might do if stratum indicators are not released. Because the synthetic data are generated from the right models, $\bar{Y}_{srs} = \sum_i \bar{Y}_i/m$ is unbiased. However, the value of m needed to estimate $Var(\bar{Y}_{srs})$ accurately increases dramatically. This can be seen in Table 7, which displays the results of simulations of the STRS setting with $m = 100$. The estimated variances for $\bar{Y}_{srs}$ are often negative, whereas the estimated variances for $\bar{Y}_{strat}$ from (12) are always positive. Based on the method outlined in Section 4.1, to obtain adequate variance estimates using $\bar{Y}_{srs}$ we require $m \approx 1,000$. Imputers may find it more manageable to release the stratum indicators instead of releasing this many data sets.

Table 7: Point and variance estimates when m=100 in STRS simulation

Est.	Avg. $\bar{q}_{100}$	$Var(\bar{q}_{100})$	% of $T_s < 0$	Avg. of pos. T_s	Avg. T_s^*
$\bar{Y}_{strat}$	54.91	.20	0	.23	.23
$\bar{Y}_{srs}$	54.91	.25	36	.62	1.92

Inferences using $\bar{Y}_{srs}$ are poor in this example because its associated $Var(T_s) = .38$, which is larger than $Var(\bar{Y}_{srs}) = .25$. Besides increasing

m, we can reduce $Var(T_s)$ by increasing n_{syn}. In fact, in 100 replications of this simulation with $n_{syn} = 1,000$, the $Var(T_s) = .07$ and all of the $T_s > 0$. Still, releasing design information may be preferable to releasing larger data sets.

5 Concluding Remarks

The results of this article contribute to a growing interest in the release of synthetic data for disclosure avoidance. The simulations studies suggest that valid inferences can be obtained from synthetic data sets for many designs and estimands. The studies also suggest that agencies considering this approach should not select the number of released data sets blindly, for the usual $m = 5$ advice does not always hold. On the other hand, it appears that inferences are relatively insensitive to the choice of n_{syn}, so long as it is reasonably large. Agencies also should consider releasing design information to help analysts obtain valid inferences. Of course, agencies considering the release of synthetic data should assess how following these implementation guidelines affects costs and confidentiality.

The flexibility of the synthetic data approach provides further advantages. Imputations can be corrected for measurement or nonsampling error, and released data can include geographic information to facilitate small area estimation (Raghunathan *et al.*, 2003). Synthetic data sets can serve as training data sets for researchers who require special access to highly confidential data (General Accounting Office, 2001). Agencies can provide synthetic administrative records linked to released data in place of actual records.

There are formidable challenges to implementing the full simulation approach in practice. Imputation models must reflect the structure of the data with reasonable accuracy. The public must be convinced to use multiply-imputed, synthetic data. Ongoing research on nonparametric imputation models may help overcome these challenges.

References

Abowd, J. M. and Woodcock, S. D. (2001). Disclosure limitation in longitudinal linked data. In P. Doyle, J. Lane, L. Zayatz, and J. Theeuwes, eds.,

Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, 215–277. Amsterdam: North-Holland.

Fay, R. E. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association* **91**, 490–498.

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.

General Accounting Office (2001). *Record Linkage and Privacy: Issues in Creating New Federal Research and Statistical Information*. United States General Accounting Office.

Hinkins, S., Parsons, V., and Scheuren, F. (1999). Inverse sampling algorithm for NHIS confidentiality protection. In *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 485–502.

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.

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

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics* **9**, 462–468.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association* **91**, 473–489.

Sarndal, C., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Sunter, A. B. (1977). List sequential sampling with equal or unequal probabilities without replacement. *Applied Statistics* **26**, 261–268.

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