A Comparative Assessment of Disclosure Risk and Data Quality Between MASSC and Other Statistical Disclosure Limitation Methods

Feng Yu and Neeraja Sathe
RTI International, 3040 Cornwallis Road, RTP, NC 27709

Abstract

MASSC (an acronym for Micro-Agglomeration, optimal probabilistic Substitution, optimal probabilistic Subsampling, and optimal weight Calibration) is a statistical disclosure limitation (SDL) methodology developed at RTI International for simultaneous confidentiality and analytic utility protection. In this paper, MASSC was compared with two other SDL methods by examining the degree to which these methods impact data quality and lower disclosure risk. The other SDL methods are Post Randomization Method (PRAM) and Random Swapping. The sample was taken from the 2006 and 2007 National Survey on Drug Use and Health (NSDUH) public use files (PUFs) as an initial data set for treatment, where the original PUFs were viewed as the "population" and the three methods were compared via simulations. For risk assessment, the matching probability was calculated to discover if a record in a treated sample can be correctly linked to the corresponding record in the "population." For utility assessment, each treatment's impact on direct estimates and its impact on inference and on estimated regression-model parameters were compared.

Key Words: Statistical Disclosure Limitation, MASSC, PRAM, Random Swapping, Disclosure Risk, Information Loss

1. Introduction

1.1 What is Disclosure?

A disclosure problem arises if an individual in the population can be associated with a record in a database containing sensitive values (SVs) (Singh, Yu, & Dunteman, 2003).

There are three types of disclosure scenarios that are of concern to the data user:

a) **Identity Disclosure.** This is a disclosure in which a direct identifying variable (IV) can reveal the identity of a particular individual. Direct IVs that can reveal the identity of an individual (e.g., Social Security Numbers, names, etc.) are typically not included in public use files (PUFs); therefore, this scenario of risk is unlikely to cause a problem. Indirect IVs, which are a combination of several attributes, could identify a person (e.g., combination of Race, Gender and occupation resulting in a rare response like Asian, Female, astronaut) and are a more likely disclosure scenario than direct IVs.

b) **Attribute Disclosure.** This is a disclosure in which confidential information (typically a SV) about a data subject is revealed and can be attributed to the subject on the PUF. Attribute disclosure may occur when confidential information is revealed exactly or when it can be closely estimated.

c) **Inferential Disclosure.** This is a disclosure in which it becomes possible to determine the value of some characteristic related to the data subject more accurately than would have been otherwise possible. Inferential disclosure occurs when information can be inferred with high confidence from properties of the released data (e.g., the type of car or house you own may reflect your income).

Two types of intrusion scenarios are associated with these three types of disclosure scenarios: inside intrusion and outside intrusion. Inside intruders know that their targets are in the database, while outside intruders do not. Inside intrusion occurs when an unauthorized person tries to link a record in a microdata file to an identifiable respondent who the intruder knows is in the file. Outside intrusion occurs when an intruder tries to identify a sample record by matching it to an external database without prior knowledge of who is in the sample.

1.2 Statistical Disclosure Limitation (SDL)

SDL is a set of statistical techniques applied to publicly released data that minimizes the risk of individuals being identified. Before releasing statistical tables or microdata files, federal agencies use a variety of statistical methods to protect their data and to ensure that the risk of disclosure is very small. In addition to SDL techniques minimizing disclosure risk, assuring confidentiality, being ethical, and ensuring adequate survey response rates, SDL may also be required under CIPSEA (Confidential Information Protection and Statistical Efficiency Act) [1] or other such federal regulations.

SDL techniques are broadly classified into two types:

- Those applied to tabular data: like cell rounding, complementary cell suppression, synthetic substitution, etc. In the cell suppression technique the value of the cell that is sensitive (at disclosure risk) is suppressed, along with a few non sensitive cells (called as complementary cells) (Cox, 1995). Synthetic Substitution (also called Controlled Tabular Adjustment, CTA) was developed by Dandekar and Cox (2002) as an alternative to complementary cell suppression. This procedure uses a threshold rule(s) to determine how cells can be modified. The CTA function replaces the value of each risky cell by a close, safe value (i.e., the cell value plus or minus a protection limit).

- SDL techniques applied to microdata files or PUFs include
    - restricting data dissemination (i.e., making the PUF available to only licensed researchers with confidentiality agreements);
    - stripping off direct identifiers (i.e., removing direct identifiers such as name, address, phone number, etc.);
    - Random Swapping (discussed in more detail in Section 2);
    - substituting (i.e., replacing values of certain SVs with values of a donor);
    - data coarsening including re-categorizing IVs into broader categories and applying top and bottom coding of SVs for categorical variables that have sparse frequency counts (if the highest categories are collapsed, it's

[1] This statute applies to data collected by the BLS and prohibits disclosure or release, for nonstatistical purposes, of information collected under a pledge of confidentiality. Under CIPSEA, data may not be released to unauthorized persons.

  called top coding, and if lowest categories are collapsed, it's called bottom coding);
- generating synthetic data;
- PRAM; and
- MASSC.

PRAM and MASSC are discussed in detail in Section 2.

This paper compares MASSC with two commonly used SDL techniques: Random Swapping and PRAM. The degree to which MASSC and the other two methods impact data quality and lower disclosure risk was examined. For implementing Random Swapping, SAS was used, and PRAM was implemented using the sdcMicro package in R.

Section 2 briefly introduces the three SDL techniques investigated in this paper. Section 3 describes the methods of risk and data quality assessment utilized to compare these three techniques. Section 4 discusses the findings, and Section 5 presents the conclusions.

2. Random Swapping, PRAM, and MASSC

2.1 Random Swapping

Random Swapping (Dalenius & Reiss, 1978) is an SDL technique used to treat categorical variables. Data containing SVs or IVs are swapped so that it is difficult for an intruder to definitively identify an individual. Confidentiality is protected and disclosure risk is minimized by introducing some uncertainty in the released microdata file. Random Swapping ensures that all univariate frequency distributions are maintained and all marginal totals remain unchanged.

Random Swapping can be implemented as follows:

- Defining a distance function, such as an L2-norm, to find swapping partners, using certain IVs and SVs. The goal is to find partners that have similar (but not the same) IV and SV characteristics
- Randomly selecting pairs of records for swapping with a specified target swap rate
- Swapping values of specified IVs and SVs between swap partners
- Defining consistency checks and ensuring that the selected swaps were executed logically.

2.2 PRAM

The Post-RAndomization Method (PRAM) is a perturbative method for disclosure protection of categorical variables (Gouweleeuw et al, 1998). Applying PRAM means that for each record in a microdata file the responses for selected variables are changed (perturbed) according to a specified probability mechanism. Since the original data file is perturbed, it is difficult for an intruder to definitely identify an individual.

On the other hand, since the probability mechanism that is used when applying PRAM is known, characteristics of the (latent) true data can be estimated from the perturbed data file. Hence, it is still possible to perform unbiased statistical analyses after PRAM has been applied.

In the simplest form of this method, let X be a categorical variable with categories $c_1, \cdots, c_k$ on the original file. Then, applying PRAM to X is accomplished by replacing the value c_l ($l = 1, \cdots, k$) of X by value c_i with probability p_{il}, where p_{il} (i, $l = 1, \cdots, k$) are prespecified probabilities that satisfy the condition: sum $\sum p_{il} = 1$ when summed over $l = 1, \cdots, k$.

The transformed variable, X* on the perturbed file satisfies this condition (in expectation): $P(X^* = c_i|X = c_l) = p_{il}$. The choice of P, where $P = [p_{il}]$, affects the degree of disclosure control and information masking.

2.3 MASSC

MASSC is an SDL method developed at RTI that can be used to treated microdata files for data dissemination. MASSC (Singh, 2002; 2006) consists of the following four major steps:

Micro-Agglomeration. In this step, variables that pose high disclosure risk but also have high analytic utilities are identified and re-categorized into broader categories to reduce sample "unique." The data set is then partitioned into risk strata based on the selected IVs so that treatment in later steps can be differentiated based upon the risk status of the individuals in the file. Levels of risk are classified as follows: uniques with respect to core IVs (IVs that are commonly known), uniques with respect to all IVs (core IVs plus other IVs that may not be commonly known), non-unique doubles, non-unique triples, and non-unique others. If a record is at higher risk, the probability of this record getting treated is higher.

Substitution. This step involves replacing values of IVs for a group of randomly selected records with those of substitution donors subject to a set of predefined bias constraints. The substitution donors are identified using a distance function so that the donor and the recipient are similar in terms of IVs, but differ in at least one IV. In addition, bias constraints are defined so that the expected squared bias for an outcome/domain combination relative to the squared population total for the same outcome/domain combination is no larger than a parameter, α, the bias upper bound.

Subsampling. This step involves deleting some records from the file subject to a set of predefined variance constraints. Similar to the bias constraints, they are defined so that the variances introduced by subsampling relative to the squared population total are no larger than a parameter, β, the variance inflation upper bound.

Calibration. In this step, the weights in the subsample are calibrated for some of the social and demographic domains so that certain weighted totals in the treated file are equal to the corresponding weighted totals in the full untreated file.

The purposes of substitution and subsampling are to introduce sufficient uncertainty into the data for public release so that either the identity of an individual has been disturbed or a particular survey respondent being in the file is no longer certain. Both substitution and subsampling steps have built-in optimization algorithms (i.e., minimizing disclosure cost subject to utility constraints); therefore, MASSC has simultaneous control on information loss and disclosure risk.

3. Comparing MASSC with Random Swapping and PRAM

To compare performance of MASSC with Random Swapping and PRAM, simulation studies were conducted. The data set used is a random sample of combined 2006 and 2007 National Survey on Drug Use and Health (NSDUH) public use files (PUFs). Consistent sets of IVs and treatment parameters were used across the three methods investigated in the simulations. The treated data were then evaluated in terms of disclosure risk and information loss. For risk assessment, the rates for a record in the treated sample that could be correctly matched to the corresponding record in the "population" were calculated. For utility assessment, the effects on estimated means and regression-model parameters were compared.

3.1 Data Preparation

The NSDUH is an annual cross-sectional national survey on drug use and related health issues. The survey is conducted by RTI International for the Substance Abuse and Mental Health Services Administration (SAMHSA). NSDUH provides information about the use of illicit drugs, alcohol, and tobacco among members of the noninstitutionalized U.S. civilian population aged 12 or older. The survey also presents measures associated with mental health problems, including data on depression and on the co-occurrence of substance use and mental health problems, as well as health conditions and health care-related issues.

For this simulation study, the 2006 and 2007 NSDUH PUFs were downloaded from the University of Michigan Web site: http://www.icpsr.umich.edu/SAMHDA/archive.html. Variables selected include some demographic variables, certain substance use variables, and survey design variables. The 2006 and 2007 NSDUH data were combined and treated as the "population." A simple random sample of 30,000 records was drawn from this population. This sample, that was selected and subsequently treated as a hypothetical sample, was treated for confidentiality protection. The weights in the sampled data were calibrated so that key estimates from the sample were preserved to those computed from the combined 2006 and 2007 NSDUH PUF data (the hypothetical population).

3.2 Treatment Parameters for the Simulation Runs

The same set of IVs, which include age, gender, race, marital status, etc., was selected, and the same level of categorization was applied to these variables before further treatment. Overall treatment rates were controlled to be identical across treatment methods (i.e., total treatment rate of substitution and subsampling in MASSC, swapping rate in Random Swapping, and total perturbation rate in PRAM). Two levels of total treatment rates were investigated: 10% and 20%. Because of the randomization feature of each method, simulations were run for each treatment rate to obtain summary statistics of the risk measure and quality measure in the treated data sets. For each treatment rate and SDL method, 100 simulations were run.

The perturbation parameter in Random Swapping is the swap rate. It is straightforward and can be simply defined as an input parameter in the SAS swapping program. So, for an overall treatment rate of 10%, the swapping rate was set to 5% so that 5% of the records would be selected as target records and another 5% of the records would be identified as "swapping donors."

Because treatment rate cannot be defined explicitly in PRAM and MASSC, several tests were conducted to determine parameters before performing simulation runs. In the sdcMicro package from R, PRAM has two parameters: *pd*, which is the minimum diagonal entry for the generated transition matrix P, and *alpha*, which is the amount of perturbation for the invariant PRAM method (Templ, 2008). The combination of these two parameters determines the overall treatment rate. Similarly, in MASSC, the control parameters are the bias upper bound, α, for substitution and the variance upper bound, β, for subsampling. Parameter values determined from test runs to result in desired treatment rates for PRAM and MASSC are displayed in Table 1.

Table 1: Parameters Used in the Simulations

Overall Treatment	*Random Swapping*	*PRAM*		*MASSC*	
Rate	*Swap rate*	*pd*	*alpha*	α	β
Approximately 10%	5%	0.8	0.5	0.015	0.015
Approximately 20%	10%	0.5	0.5	0.04	0.04

Swapping donors for the swapping treatment and the substitution donors for MASSC also were controlled. To do this, the initial data set was divided into subgroups based on certain characteristics so that records with the same attributes (e.g., same age) were classified into the same subgroup. Random Swapping and substitution donors were chosen within a subgroup using a distance function. Features of the distance functions used in swapping and MASSC were similar (i.e., the same set of variables was used; variables were ordered with importance of preserving the analytic utility values; high weights were assigned to the most important variables; low weights were assigned to the least important variables, etc.). Pairs that were closest in the distance measure but not identical to each other (distance greater than 0) were chosen as the final swapping or substitution partners. To be consistent, PRAM was implemented within the same subgroups used for swapping and MASSC.

When a record was selected for swapping or substitution, the entire set of predefined variables was swapped in Random Swapping or substituted in MASSC. The sdcMicro package does not offer multivariate PRAM. As a result, to maintain multivariate relationships between variables, a compound variable was first constructed by concatenating the values from all the variables to be treated, and PRAM was then applied to this compound variable (de Wolf et al., n.d.; Shlomo, 2010). Once the treatment was completed, the compound variable was decomposed into the original individual variables.

3.3 Risk Assessment

The disclosure risk in the treated data was quantified by matching the treated data with the population data (the combined 2006 and 2007 NSDUH PUFs). Matching was considered successful if a record in the treated data could be correctly linked to the corresponding record in the population data. Average matching rates were calculated from the 100 simulation runs. Three types of matching rates were calculated: exact matching rate, probability matching rate, and distance matching rate.

An **exact match** refers to a match of a sample record to a population record for a given set of identifying variables. If a sample unique or double matches to the population unique or double, it is considered to be an exact match. After perturbation, such match may not be a true match because profiles were altered in the treated database. Therefore, an exact match was also required to match the respondent ID in addition to the values of IVs. An exact matching rate was calculated as follows:

$$\text{Exact Matching rate} = \frac{\text{sum of matched records from the sample}}{\text{sample size}} \qquad (1)$$

A **probability match** is based on the frequency count in the sample as well as the frequency count in the population for a given cell from a given set of IVs. The matching probability was calculated as follows:

$$\text{Matching Probability} = \frac{1}{n}\sum_{i=1}^{K}\frac{f_i}{F_i} \qquad (2)$$

Where f_i is the cell count in the sample from cell i; F_i is the cell count in the population from cell i; K is the total number of cells (or combinations) for a given set of IVs; and n is the sample size.

This method is analogous to mu-Argus's risk function (Hundepool et al., 2008), where individual risk is assessed based on the posterior mean of $1/F_i$ under distribution of $F_i|f_i$, and where F_i is usually unknown but can be estimated by the sampling weights assuming that $F_i|f_i$ follows a negative binomial distribution. Since the population is known in this case, F_i was obtained directly from the population data.

If the sample is perturbed either by swapping, substitution, or PRAM and if the treated record is not in the same cell as the untreated record (i.e., if a record has been altered such that it does not belong to the original cell in the *i*th combination of IVs) or a record has been sampled out from MASSC, then f_i is modified by removing the treated record in the calculation.

A **distance match** determines a match based on the distance between a sample member and a population member. First a distance function was defined, in terms of the selected IVs, as the weighted sum of the absolute differences of those identifying variables. Similar to the approach used in calculating exact match rates, only uniques or doubles in the sample (*w.r.t.* the IVs) were considered when calculating the distance match rate. Then the distance between each risky record in the sample was calculated relative to each individual in the population. The sample record was deemed a match if the rank of the distance to itself in the population from low to high was two or less. In other words, a sample member was deemed disclosure safe if at least two individuals in the population were as close as or closer to the potential risky record in the sample than the record itself; otherwise, the record was deemed at risk of disclosure. The matching rate was calculated as described in Equation (1).

3.4 Utility Assessment

For utility assessment, relatively simple statistics, such as means and standard errors of the means, of the key outcomes were compared. Impact of treatment parameters such as perturbation rates on inference under different methods were assessed via simulation using regression models. Distribution of the model parameters (i.e., means and standard errors of the model coefficients) were calculated and compared. In addition, tests of significance were performed for regression coefficients and changes of significance were evaluated.

4. Findings

As noted earlier, to capture the randomization feature of each method, 100 simulations were run for each method and each target treatment rate. Risk assessment and utility assessment were based on the average behaviors from the simulation runs.

4.1 Risk Assessment

Using the risk assessment approaches described in Section 3.3, matching experiments were conducted based on IVs, such as age, gender, and race, to link the treated data file with the population data (the original 2006 and 2007 NSDUH PUFs). Two sets of matching rates were computed: one with respect to core IVs, and the other with respect to all IVs. Means and ranges of the match rates under the three different treatment methods were calculated from the simulations, and the results are shown in Figure 1.

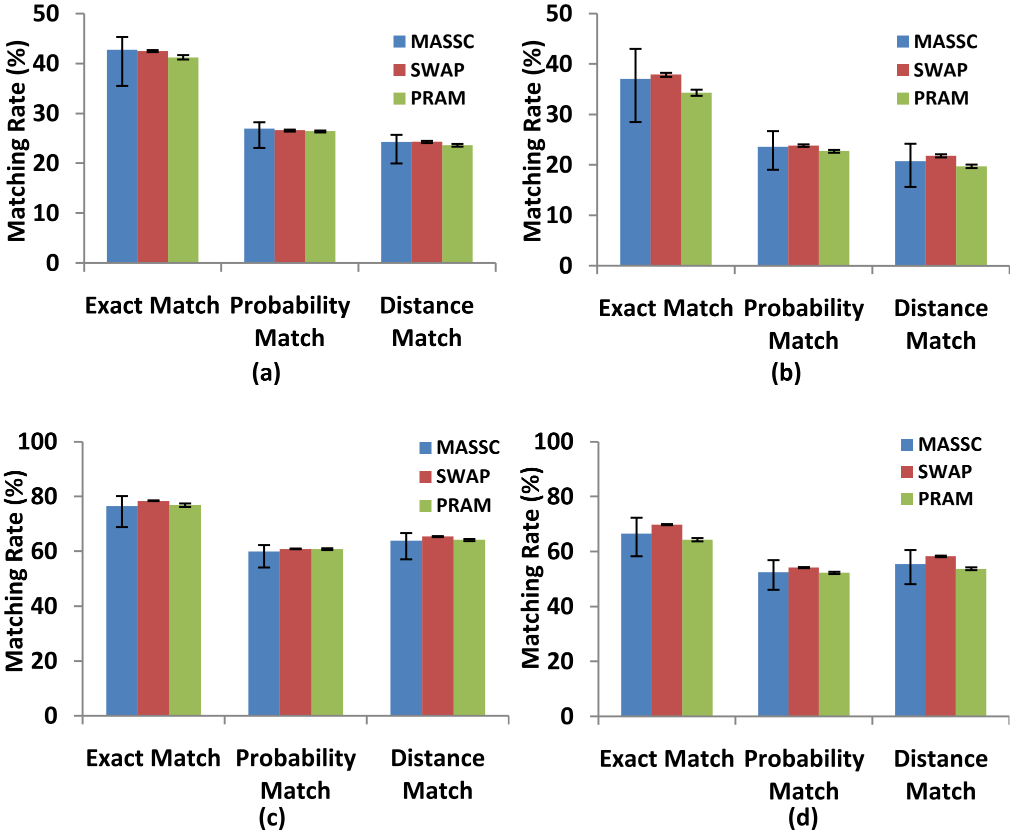

Figure 1. Average Matching Rate after Simulation Runs
(a). 10% Treatment Rate, Matching *w.r.t.* Core IVs
(b). 20% Treatment Rate, Matching *w.r.t.* Core IVs
(c). 10% Treatment Rate, Matching *w.r.t.* All IVs
(d). 20% Treatment Rate, Matching *w.r.t.* All IVs

Figure 1 shows that the average matching rates from the simulations are comparable across all three SDL methods. In most of the cases, PRAM demonstrates the lowest matching rates among all matching methods (Figures 1a, 1b, and 1d). The differences between Random Swapping and PRAM are very small, but MASSC displays larger variations.

Selection of IVs has significant impact on matching rates. A larger set of IVs (Figures 1c and 1d) results in higher matching rates than smaller set of IVs (Figures 1a and 1b). This is because more risky records (i.e., uniques or doubles) can be identified by a combination of a larger set of IVs. Therefore, a larger set of IVs increases the matching rates under the same level of treatment.

Matching rates decrease as the overall treatment rates increase for all three SDL methods, especially when using all IVs, rather than that using just the core IVs. For the three matching methods, exact match presents higher matching rates than probability match and distance match. The exact match method is conservative because the matching rate is based on the individual matches of IVs. Both probability match and distance match are based on average measures and the matching rates are fairly close. These two matching methods may provide more realistic measures for disclosure protection.

Most of the matching rates are high in this matching experiment (i.e., greater than 25%; some are even larger than 60%) because of the conservative assumption of inside intrusion; that is, the intruder knows his target record is in the sample, and in addition, each matched case is a correct match. However, in reality, a match could be a false match. Therefore, a true match probability could be much lower. To find the true match rate, one would need to premultiply by a probability that the target is in the population and postmultiply by a probability that the match is a correct match. This would reduce the true matching probability substantially. For example, if both probability of a record being in the sample and probability of a matching being correct is 50%, then after pre- and postmultiplication, the matching probability is reduced by a multiplicative factor of 25%. Finding true match rates was not of interest here; this matching exercise was done only to compare the three methods under identical conservative assumptions of inside intrusion.

4.2 Data Quality Assessment

The primary purpose of SDL methods is to minimize disclosure risk while maintaining data utility. Here the term maintaining data utility is used in the sense of preserving estimates and statistical inferences and conclusions. In this study, before/after treatment estimates, regression coefficients, and change of significance of regression coefficients under the three methods were examined.

4.2.1 Estimate comparisons

Average before/after ratios of treatment estimates on each substance use variables were calculated (e.g., for past month alcohol use) in combination with domains such as age, gender, and marital status (340 total estimates) from the 100 simulations. The following quantities were defined:

Ratio of the Estimates: $$\text{Ratio\_EST} = \frac{p_i}{p_{io}} \quad (3)$$

Ratio of the Standard Errors: $$\text{Ratio\_SE} = \frac{SE_i}{SE_{i0}} \quad (4)$$

Relative Root Mean Square Errors: $$\text{RRMSE} = \frac{\sqrt{\text{Var}_\text{i} + (p_i - p_{io})^2}}{p_{i0}} \qquad (5)$$

Where p_i and SE_i are the treated sample estimate and standard error, respectively; $p_{i\,0}$ and $SE_{i\,0}$ are the untreated sample estimate and standard error; and Var_i is the treated sample variance.

Summary statistics were computed for the average ratios for all 340 substance use/domain combinations. Results are presented in Table 2. Increase in bias can be evaluated via ratio of the estimates, and decrease in precision can be evaluated via the SE ratios. The Relative Root Mean Square Error (RRMSE) provides a comprehensive measure of bias and SE.

Table 2: Before/After Treatment Estimates Comparisons

Trt Rate	*Stats. (340x100)*	*MASSC*			*Random Swapping*			*PRAM*		
		Ratio _EST	*Ratio _SE*	*RRMSE*	*Ratio _EST*	*Ratio _SE*	*RRMSE*	*Ratio _EST*	*Ratio _SE*	*RRMSE*
10%	Max	1.05	1.10	1.03	3.67	3.67	4.70	13.70	13.11	18.44
	Median	1.00	1.02	0.07	1.00	1.00	0.07	1.00	1.00	0.08
	Min	0.93	0.93	0.01	0.94	0.95	0.01	0.90	0.88	0.01
	Mean	1.00	1.02	0.14	1.01	1.01	0.15	1.07	1.06	0.23
20%	Max	1.10	1.21	1.13	1.40	1.56	1.42	43.40	41.70	59.24
	Median	1.00	1.04	0.08	1.00	1.00	0.08	1.00	1.00	0.09
	Min	0.83	0.83	0.01	0.89	0.91	0.01	0.81	0.76	0.01
	Mean	1.00	1.04	0.15	1.00	1.01	0.15	1.16	1.16	0.37

Table 2 shows that, on average, MASSC preserves the estimates (the mean of the ratio of estimates - Ratio_EST is 1.00, which is the desired ratio), in comparison to Random Swapping, which has a 1% increase in estimates at 10% treatment rate (mean=1.01) and PRAM which has a substantial average increase in bias, 7% (mean=1.07) and 16% (mean=1.16) under 10% and 20% treatment rates, respectively. Because of subsampling in MASSC, treated data from MASSC result in a 2% (from a 10% treatment rate) and 4% (from 20% treatment rate) decrease in precision on average (mean of Ratio_SE), while only a 1% decrease in precision (or increase in SE) after Random Swapping at both 10% and 20% treatment rates was observed. PRAM, however, produced data with the highest ratio of SEs on average, 1.06 and 1.16, indicating a 16% decrease in precision at 20% treatment rate for PRAM. When variance and bias are combined, MASSC and Random Swapping demonstrate the lowest RRMSE on average (0.14 and 0.15 for MASSC and 0.15 and 0.15 for Random Swapping), whereas PRAM has much higher RRMSE values (0.23 and 0.37). Although all three methods achieve the same minimal value of RRMSE (0.01), the maximum RRMSE values of MASSC (1.03 and 1.13) are much lower than those of Random Swapping (4.70 and 1.42), and maximum RRMSE values of PRAM (18.44 and 59.24) are much higher than those of MASSC and Random Swapping.

4.2.2 Regression comparisons

Logistic regression models were fit for each of X substance use outcomes (e.g., past month alcohol use) using demographic domains, including age, gender, race, marital status, etc. as predictors. The models were fit on data sets before treatment and after treatment. Average ratios of before/after treatment regression coefficients for each model were calculated based on the 100 simulation runs for MASSC, Random Swapping, and PRAM. Similar to estimates comparisons, the following quantities were calculated:

Ratio of Betas:
$$\text{Ratio\_Beta} = \frac{\beta_i}{\beta_{io}} \qquad (6)$$

Ratio of Standard Error of Betas:
$$\text{Ratio\_SEBeta} = \frac{\text{SE of } \beta_{\text{i}}}{\text{SE of } \beta_{\text{i0}}} \qquad (7)$$

Relative Bias:
$$\text{Bias/SE} = \frac{\beta_i - \beta_{i0}}{\text{SE of } \beta_{\text{i0}}} \qquad (8)$$

Where β_{i0} and β_i are before and after treatment regression coefficients, respectively.

Summary statistics were computed for these average ratios of all 220 regression coefficients. The results are presented in Table 3.

Table 3: Before/After Treatment Regression Coefficients Comparisons

Trt	*Stats.*	*MASSC*			*Random Swapping*			*PRAM*		
Rate	*(220 x 100)*	*Ratio _Beta*	*Ratio _SEBeta*	*Bias /SE*	*Ratio _Beta*	*Ratio _SEBeta*	*Bias /SE*	*Ratio _Beta*	*Ratio _SEBeta*	*Bias /SE*
10%	Max	2.04	1.11	0.29	2.15	1.44	0.72	2.42	1.50	1.35
	Median	1.00	1.02	0.00	1.00	1.00	0.00	0.95	0.99	0.00
	Min	-1.68	0.96	-0.22	0.69	0.96	-1.03	-2.11	0.90	-1.85
	Mean	0.98	1.02	0.01	0.98	1.00	-0.02	0.94	0.99	-0.01
20%	Max	2.81	1.53	0.63	3.42	1.80	1.45	3.63	1.98	2.91
	Median	0.99	1.03	0.02	0.99	1.00	-0.01	0.85	0.98	0.05
	Min	-4.20	0.95	-0.48	0.05	0.92	-1.99	-2.61	0.79	-3.95
	Mean	0.97	1.04	0.02	0.96	1.00	-0.04	0.85	0.98	-0.01

Table 3 shows that the median ratio of betas (Ratio_Beta) for MASSC and Random Swapping are both close to the ideal value of 1.00 under both treatment rates, but PRAM shows values of 0.95 and 0.85 for 10% and 20% treatment rates, respectively. The maximum ratios of SEs of beta (Ratio_SEBeta) for MASSC are much lower (1.11, and 1.53), as compared to those for Random Swapping (1.44 and 1.80) and PRAM (1.50 and 1.98), indicating that Random Swapping and PRAM are more likely to result in false indicators of significant differences. Overall, the SE ratios of the betas (before and after treatment) for MASSC have a lower spread (narrower range) as compared to the other two methods.

All three methods show no noticeable median relative bias (Bias/SE) under 10% treatment rate, but some bias under 20% treatment for the median values (i.e., 0.02 for MASSC, -0.01 for Random Swapping, and 0.05 for PRAM). For all the absolute values of minimum and maximum relative bias under both treatment rates, MASSC has the smallest bias, Random Swapping is in the middle, and PRAM has the largest bias.

Treatment impact on inference is more important for regression comparisons. To account for this, tests of significance at the 5% level were performed for regression coefficients from all models. Changes of significance from significant to nonsignificant or vice versa were counted before/after treatment. Range and average changes for both directions based on the simulations were calculated. Results are displayed in Table 4.

Table 4: Change of Significance Comparison for Regression Coefficients (Mean and Ranges)

Treatment rate (n=220x100)	*MASSC*		*Random Swapping*		*PRAM*	
	Sig. to Non-Sig.	*Non-Sig. to Sig.*	*Sig. to Non-Sig.*	*Non-Sig. to Sig.*	*Sig. to Non-Sig.*	*Non-Sig. to Sig.*
10%	3.29 (0 - 8)	1.99 (0 - 6)	3.02 (0 – 8)	2.13 (0 – 8)	6.26 (0 – 15)	3.26 (0 – 12)
20%	5.62 (1 – 11)	3.16 (0 – 10)	4.84 (1 – 14)	2.91 (0 – 10)	11.38 (5 – 22)	3.99 (1 – 11)

The values in Table 4 are the average number of changes of significance tests (from significant to nonsignificant or vice versa) for all 220 betas from the 100 simulations. Included in parenthesis is the range of these numbers of changes. Table 4 shows that the average number of changes in significance using MASSC is slightly larger than for Random Swapping, except for the number of changes from nonsignificant to significant under the 10% treatment rate, and PRAM shows consistently larger numbers of change than both MASSC and Random Swapping.

5. Conclusions

The simulation studies show that the following can be deduced from the results in comparing MASSC, Random Swapping, and PRAM:

- All three methods provide a certain degree of confidentiality protection to the data; as the overall treatment rate increases, the matching rate decreases.
- With all three methods, the data quality decreases as the overall perturbation rate increases.
- When Random Swapping is properly designed, it performs similar to MASSC in terms of data quality protection.
- PRAM results indicate that it produces more information loss and reduces precision.
- MASSC has a strong theoretical grounding, and it provides simultaneous protection of data confidentiality and data quality. MASSC tends to provide more opportunities for better disclosure treatment (minimum matching rates in MASSC were much lower than other methods), and the quality of the treated data can be better preserved.
- Since MASSC involves a subsampling step, the suppressed records are guaranteed to have no disclosure risk. Thus, this method is better than the others at protecting against inside intrusion.

- Because of the interactive features of MASSC, it requires more labor and computation time than the other two methods.

Acknowledgements

This research was funded by RTI International through a professional development award. The authors would like to thank RTI International as well as the following people for their valuable comments and suggestions: Dr. Roy Whitmore, Dr. David Wilson, Dr. Philip Kott, Dr. W. Jeremy Aldworth, Dr James Chromy, Dr. Moshe Feder, and Dr. Avinash Singh.

References

Cox, L. H. (1995). Network models for complementary cell suppression. *Journal of the American Statistical Association, 90,* 1453-1462.

Dalenius, T., & Reiss, S. P. (1978). Data-swapping: A technique for disclosure control (extended abstract). *American Statistical Association, Proceedings of the Section on Survey Research Methods*, Washington, DC, 191–194. Retrieved from http://www.amstat.org/sections/srms/proceedings/papers/1978_038.pdf

Dandekar, R. A., & Cox, L. H. (2002). Synthetic tabular data: An alternative to complementary cell suppression. Washington, DC: Energy Information Administration, Department of Energy. Unpublished manuscript.

de Wolf, P. P., Gouweleeuw, J. M., Kooiman P., & Willenborg L. (n.d.) Reflections on PRAM. The Netherlands: Statistics Netherlands. Retrieved from http://neon.vb.cbs.nl/casc/related/Sdp_98_2.pdf

Gouweleeuw, J. M., Kooiman, P., Willenborg, L. C. R. J., & de Wolf, P. P. (1998). Post randomization for statistical disclosure control: Theory and implementation. *Journal of Official Statistics, 14,* 463–478.

Hundepool, A., van de Wetering, A., Ramaswamy, R., Franconi, L., Polettini, S., Capobianchi, A., et al. (2008), µ-Argus User's Manual, version 4.2. Netherlands, Voorburg, The Netherlands: Methodology Department, Statistics. Retrieved from http://neon.vb.cbs.nl/casc/Software/MuManual4.2.pdf

Shlomo, N. (2010). Releasing microdata: Disclosure risk estimation, data masking and assessing utility. *Journal of Privacy and Confidentiality*, *2*(1), 73-91.

Singh, A. C. (2002, 2006). *US Patent No. US7058638B2.* Method for statistical disclosure limitation. Patent granted June 2006. Washington, DC: U.S. Patent and Trademark Office.

Singh, A. C., Yu, F., & Dunteman, G. H. (2003). MASSC: A new data mask for limiting statistical information loss and disclosure. *Proceedings of the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*, Luxembourg. Working Paper No. 23.

Templ, M. (2008). Statistical disclosure control for microdata using the R-Package sdcMicro. *Transactions on Data Privacy, 1,* 67-85.