Journal of Privacy and Confidentiality (2009) **1**, Number 2, pp. 155–182

Maintaining Analytic Utility while Protecting Confidentiality of Survey and Nonsurvey Data

Avinash C. Singh*

Abstract. Consider a complete rectangular database at the micro (or unit) level from a survey (sample or census) or nonsurvey (administrative source) in which potential identifying variables (IVs) are suitably categorized (so that the analytic utility is essentially maintained) for reducing the pretreatment disclosure risk to the extent possible. The pretreatment risk is due to the presence of unique records (with respect to IVs) or nonuniques (i.e., more than one record having a common IV profile) with similar values of at least one sensitive variable (SV). This setup covers macro (or aggregate) level data including tabular data because a common mean value (of 1 in the case of count data) to all units in the aggregation or cell can be assigned. Our goal is to create a public use file with simultaneous control of disclosure risk and information loss after disclosure treatment by perturbation (i.e., substitution of IVs and not SVs) and suppression (i.e., subsampling-out of records). In this paper, an alternative framework of measuring information loss and disclosure risk under a nonsynthetic approach as proposed by Singh (2002, 2006) is considered which, in contrast to the commonly used deterministic treatment, is based on a stochastic selection of records for disclosure treatment in the sense that all records are subject to treatment (with possibly different probabilities), but only a small proportion of them are actually treated. We also propose an extension of the above alternative framework of Singh with the goal of generalizing risk measures to allow partial risk scores for unique and nonunique records. Survey sampling techniques of sample allocation are used to assign substitution and subsampling rates to risk strata defined by unique and nonunique records such that bias due to substitution and variance due to subsampling for main study variables (functions of SVs and IVs) are minimized. This is followed by calibration to controls based on original estimates of main study variables so that these estimates are preserved, and bias and variance for other study variables may also be reduced. The above alternative framework leads to the method of disclosure treatment known as MASSC (signifying micro-agglomeration, substitution, subsampling, and calibration) and to an enhanced method (denoted GenMASSC) which uses generalized risk measures. The GenMASSC method is illustrated through a simple example followed by a discussion of relative merits and demerits of nonsynthetic and synthetic methods of disclosure treatment.

Keywords: Calibration, Disclosure Risk, GenMASSC, Information Loss, Optimal Allocation, Substitution, Subsampling

*Center for Excellence in Survey Research, NORC at the University of Chicago, Chicago, IL, mailto:singh-avi@norc.org

1 Introduction

It is well known that there is a great user demand for administrative, census, and sample survey data at the micro level, although they all are typically collected under a confidentiality pledge to the respondent. In this paper, it is assumed that we have data in the form of a complete rectangular file with records as rows and variables (IVs—Identifying variables and SVs—sensitive variables) as columns. This representation applies to both microdata and macrodata (including tabular data) in that values for each record in a cell could be deemed as common and equal to the cell average value; e.g., 1 in the case of count data. The IVs are the variables that an intruder might know about the target, and provide, in general, information about demographic, geographic, and socio-economic status; while the SVs provide, in general, information about medical, financial, social, and professional status. With a suitable disclosure treatment, any data set can be made available to users. In fact, highly sensitive data not previously available to researchers can also be made available. In data mining applications for detecting rare events or characteristics of small subgroups, surrogate data (created after disclosure treatment) can be used by researchers at large, and then the final analysis on the original data can be performed under tight security. In the context of administrative records and census taking, Scheuren (1995, 1999), in particular, addresses important data privacy issues.

With any public use database, the importance of protecting confidentiality of respondents (with its obvious implications on the credibility of the data producer) has always been of concern, but it has taken on new significance with the introduction of HIPAA (Health Insurance Portability and Accountability Act of 1996) regulations; see Department of Health and Human Services (2000). This problem becomes challenging because of demands by data users regarding the analytical utility of the database which conflicts with the need to protect the confidentiality of respondents. There is a tension between protecting utility and confidentiality in that as the data is treated more and more to protect confidentiality before being released for public use, it loses more and more of its analytical utility. In this paper, we will mainly consider nonsynthetic statistical disclosure limitation (SDL) approaches to creating public use files (PUFs) in the sense that only a small fraction of the original records are actually treated; thus these approaches have the appealing feature of preserving the original data to a considerable extent. This is in contrast to synthetic SDL approaches in which all records are treated; an important example of which is the case when all values of IVs (and possibly SVs) are generated from a modeled distribution. The synthetic approach, based on modeling, is theoretically attractive as it provides good analytical utility (there is in fact no information loss if the model is valid), and possibly no disclosure risk (see Section 6 for further discussion). There are, however, several areas of concern: problems with modeling when there are too many variables to be treated in the database, taking account of the underlying sampling design, especially when it is complex due to multistage cluster unequal selection probabilities (see e.g., Reiter et al., 2006), and the general reluctance among users to rely on treated data where a large portion is not original.

For any database, direct identifiers such as date of birth, address, telephone number,

and social security number do not pose any problems for creating PUFs because these details can be suppressed as only broad age categories, and geographic information are generally sufficient in practice for analysts. (However, under some dissemination protocols such as licensing and research data centers, direct identifiers do not need to be suppressed.) In this paper, we consider a special but rather practical problem of inside intrusion where an intruder knows the presence of the target in the database and might know enough indirect identifiers or IVs (such as age group, race, and gender) to narrow down the profile to a unique individual to disclose values of SVs (such as medical conditions). Even with nonunique records, i.e., records with identical profiles on a given set of IVs assumed to be known to an intruder, disclosure could happen with certainty if all the records assume common values or common categories of at least one of the SVs (in practice, it may be sufficient to view continuous SVs as implicitly being categorized corresponding to anticipated knowledge of the intruder), or with high probability if values of at least one SV for most of the records are close to each other. In other words, after removing direct identifiers, we are still concerned about the problem of identity disclosure of the target resulting essentially from attribute disclosure based on remaining (indirect) identifiers.

Under the inside intrusion approach as first introduced by Singh (2002, 2006), an important scenario from the respondent's perspective known as 'disclosure by response knowledge' (see Bethlehem et al., 1990) arises when the respondent identifies his or her own record and is concerned about its disclosure by someone who might know enough about them to identify the record. Although this kind of threat may be more of a perception, it is nevertheless serious enough to put the reputation and credibility of a data producer at stake if the respondents in the database do not have confidence in the producer. Now, in contrast to the commonly used outside intrusion approach (where the intruder does not know the presence of the target in the database with certainty), the inside intrusion approach (although more conservative) may be appealing to practitioners. Under this approach, clearly there is always a need of disclosure treatment. Commonly used techniques to reduce disclosure risk under nonsynthetic approaches are perturbation (or substitution) of records such as recoding, top-coding, swapping, and noise addition for IVs (SVs not perturbed), and suppression (or sampling-out) of records such as removing part or all fields of IVs/SVs for all records deemed to be at risk. However, when all records deemed to be at risk are treated, although the disclosure risk is controlled, there is no control on the resulting information loss. Here, there could be considerable bias introduced by perturbation or suppression of records. Moreover, there is no protection against new IVs that an intruder might know but were not part of the disclosure treatment. With sample survey data, the disclosure problem could be even more challenging because of the availability of potentially new IVs such as stratum/PSU (primary sampling unit) identifiers and knowledge of unequal sampling weights at the record level due to over/under sampling.

It follows that as an alternative to deterministic treatment, there is need for a stochastic framework of disclosure treatment which allows for treatment of a partial set of records (selected at random) so that information loss could be reduced to an

acceptable level. Moreover, it should allow for measuring disclosure risk to check if it is tolerable, thus providing a simultaneous control on both disclosure risk and information loss, as well as some protection (due to stochastic treatment) against new IVs that an intruder might know. In this paper, as introduced by Singh (2002, 2006), we consider the alternative stochastic framework of measuring disclosure risk and information loss in a nonparametric way in which the disclosure treatment is performed through random substitution and subsampling, and all records (and not just those deemed to be at risk) are subject to treatment (with possibly different probabilities), but only a small proportion of them are actually treated. The random treatment feature of the new framework allows for computing bias due to substitution and variance due to subsampling, and thus provides a practical application of the important conceptual framework of risk-utility trade-off introduced by Duncan et al.(2001).

It turns out that the alternative measures of disclosure risk under a random treatment mechanism do not require any modeling assumptions but do require access to the original database. Therefore, these measures cannot be computed by the data intruder but can be calculated in advance for the satisfaction of the data producer before releasing the PUF. The above measures are different from the innovative population model-based measures proposed by Skinner and Holmes (1998), and the nonmodel-based ones proposed by Skinner and Elliot (2002) and Skinner and Carter (2003), where the database is assumed to be selected at random under Poisson sampling, is not subject to disclosure treatment, and the target's presence in the database is not known. The risk measures considered here also complement the simple but creative measures of Reiter (2005), which have the unique feature of being computable from the treated data (and hence by the data intruder), but do require modeling of assumed types of data perturbation and suppression by the intruder.

Using a subtle analogy between releasing an untreated database and conducting a census, the above alternative framework leads to a method of disclosure treatment, termed MASSC by Singh, Yu, and Dunteman (2003), which is discussed in Section 3. The enhanced MASSC, denoted GenMASSC, generalizes risk measures to include cases with partial risk, discussed in Section 2. MASSC uses survey sampling techniques of sample allocation for subsampling, imputation (applicable to the case of missing data but here viewed as perturbation or substitution) for handling item nonresponse, and finally calibration to reduce bias due to substitution and variance due to subsampling. The term MASSC signifies Micro-Agglomeration for creating risk strata, random Substitution for introducing uncertainty about the identity of a target, random Subsampling for introducing uncertainty about the presence of the target, and Calibration for preserving estimates for main study variables. MASSC is applicable to survey (sample or census) and nonsurvey (such as administrative) data. For sample survey data, it adds another phase of sampling, rendering the resulting treated data into a two-phase sample.

With MASSC, among other features, it is possible to assign two measures $(\varepsilon,\ \delta)$ between 0 and 1 as upper bounds on information loss and disclosure risk, respectively.

This serves as a way of providing assurance both to the data producer about confidentiality protection and to the data user about analytical utility protection in the treated database. Here information loss is measured as the maximum mean squared error (i.e., variance plus squared bias) relative to the squared true parameter value over a set of main study variables, and disclosure risk is measured as the average expected risk score (or expected loss—a conventional term) over records in the original database that might resemble a record in the treated database and could be at risk of disclosure.

The organization of this paper is as follows. Section 2 deals with background, motivation, and measures of disclosure risk (under GenMASSC) and information loss; Section 3 describes in detail the method of MASSC followed by a simple illustrative example of GenMASSC in Section 4. Section 5 considers complexity in the analysis of MASSC-treated data as it adds another phase of sampling when the original database is a sample. In Section 6, we consider alternatives based on synthetic and nonsynthetic SDL approaches and discuss relative merits and demerits; Section 7 presents concluding remarks and directions for future work.

2 Background, Motivation, and Measures of Disclosure Risk and Information Loss

Consider a microdata with IVs and SVs. For example, for the Behavior Risk Factor Surveillance Survey (BRFSS) data, IVs could be age group, race, gender, education, income (in broad categories to account for the fact that the intruder's knowledge would be limited with respect to precision), job status, marital status, height, weight, frequency of eating fruits, and flu shot status; the SVs could be asthma condition, diabetes condition, number of permanent teeth removed, use of a car under alcohol influence, reason for HIV test, and method of birth control. For MASSC treatment, we categorize IVs (or if necessary recode them into coarser categories) to reduce the pre-treatment disclosure risk (see Subsection 2.3 below) as long as the analytic utility is not too compromised. It follows that in the MASSC-treated data, it is not possible to release IVs at finer levels than the categories used for treatment without increasing the risk. This may not be a limitation in general except when dealing with variables with highly skewed distributions (such as income); see remarks at the end of Section 3. To be conservative in computing disclosure risk, SVs are also categorized (or recoded if necessary) to account for the fact that reported values close to the true value may be sufficient for disclosure. However, the original values of SVs are reported in the disclosure treated database.

To begin, we assume that the intruder knows a given set of core IVs (also known as keys) about the target such as race, gender, and marital status. Later in Section 3 we also consider the case that the intruder might know noncore IVs (such as completed education and job status). In the following, we first discuss the more conservative inside intrusion scenario, then present a motivation of the MASSC method based on the alternative stochastic framework for measuring disclosure risk and information loss, details of which are presented later on.

2.1 Inside Intrusion Scenario

As mentioned earlier, here the intruder knows the presence of the target in the database; an important example of which is 'disclosure by response knowledge'. Other examples include a child's response to drug behavior in a drug survey (e.g., National Survey on Drug Use and Health as discussed in Singh et al., 2003) where the results are at risk of disclosure to parents because parents know the presence of their child in the database. In the case of an administrative data (such as the Canadian Cancer Registry), a neighbor or a coworker might know about the cancer episode of an individual. A form of inside intrusion scenario (termed coalition intrusion) has been well-known for tabular data where for a cell with small counts, a person's SV (which typically represents one of the dimensions of the table) may be at risk of disclosure if other persons belonging to the cell form a coalition for intrusion; see also Zayatz (2000) for risk of disclosure of units in small areas. Under inside intrusion, the pre-treatment disclosure risk may be 100% for unique records with at least one value of SVs being sensitive as well as for nonunique records with common sensitive values of at least one SV. Using expected loss as the definition of risk, here in the absence of stochastic treatment, the disclosure risk is simply the disclosure loss (or risk score) for the record believed to correspond to the target where the disclosure loss function for the record is assumed to take only two values: 1 if the record is at risk and 0 otherwise. Therefore, some disclosure treatment via perturbation and suppression becomes naturally necessary if the database has records at risk.

Under outside intrusion (in contrast to the inside intrusion), the target's presence in the database (viewed as a sample) is unknown to the intruder. Here the pre-treatment risk for a sample unique with a sensitive value of one of the SVs is not 100% due to uncertainty in assigning the target with the sample unique. It can be estimated under a model for population uniques as proposed by Skinner and Holmes (1998) or under a nonmodel-based approach as in Skinner and Elliot (2002) and Skinner and Carter (2003). If the risk is not serious, then there is no need for disclosure treatment. In the SDL practice, although the outside intrusion scenario is typically assumed, protecting against inside intrusion might be the safer option because it automatically protects against outside intrusion, and in terms of disclosure risk, provides an upper bound for any database. More specifically, under 0/1 loss function, the disclosure risk under outside intrusion can be defined as the probability Pr(ABC) where the event A corresponds to the target being in the database (this would be the case if the database is a sample), event B corresponds to the target being at risk of disclosure, and event C corresponds to the target being correctly matched to an external file for identification. This probability can also be expressed as the product of three probabilities: Pr(A), Pr(B|A), and Pr(C|AB); the third probability is expected to be just Pr(C|A). Observe that Pr(B|A) is simply the disclosure risk under inside intrusion which is pre- and post-multiplied by probabilities to get the risk under outside intrusion. It follows that the risk under outside intrusion is likely to be much smaller than the risk under inside intrusion.

2.2 Motivation of the Method MASSC

As mentioned in the Introduction, the framework of stochastic treatment of substitution and subsampling allows us to construct measures of disclosure risk and information loss; see Subsections 2.3 and 2.4 for details. It leads quite naturally to the method of disclosure treatment termed MASSC whose theoretical foundation rests on the theory of survey sampling based on a subtle analogy between releasing an untreated database and conducting a census for a finite population. Under the inside intrusion approach, the database can be viewed as the finite population. The monetary cost of conducting a census is known to be very high, but there is no loss of information. Similarly, the disclosure cost of releasing an untreated database could be very high, but there is no loss of information. To minimize cost, an optimal survey design is used to take a sample while controlling loss of information. Likewise, MASSC is designed to minimize disclosure cost while controlling loss of information.

The step of micro agglomeration creates risk strata and checks for records at risk; this step controls the number of records initially at risk by making decisions about the level of details of identifying variables to be released. The creation of risk strata allows for over/under treatment and is similar to stratification for over/under sampling. The step of substitution uses optimal sampling rates for selecting records at random for perturbation subject to substitution bias constraints; it introduces uncertainty about the identity of a target. This step of perturbation at random is somewhat like imputation for item nonresponse except that only IVs are subject to imputation. The step of subsampling uses optimal sampling rates for selecting records from the substituted database at random for non-suppression subject to precision constraints; it introduces uncertainty about the presence of a target. This step of random non-suppression is similar to sample selection. Finally, the step of calibration uses optimal weight calibration for adjusting subsampling weights subject to preserving main estimates from the original database; this step reduces bias due to substitution and variance due to subsampling analogous to what is done in survey sampling.

2.3 Measures of Disclosure Risk

Under a deterministic selection of records for treatment, all records at risk (with respect to a given set of IVs deemed to be known to the intruder) are treated by perturbation or suppression. The post-treatment disclosure risk goes to zero but may lead to high information loss in terms of bias in the resulting estimates. Also, there is no protection against new IVs that the intruder might know. In contrast, under the alternative framework considered here, the term stochastic treatment is used to mean that all records are subject to treatment (by randomly selecting a subset for substitution followed by another random selection of a subset for subsampling), but only a small random subset is actually treated. It leads to low information loss as well as protection against new IVs that the intruder might know. However, unlike deterministic treatment, the post-treatment disclosure risk is not zero with respect to a given set of IVs but may be made reasonably small by suitably choosing substitution and subsampling rates. The above

stochastic framework allows the data producer to measure and control disclosure risk and information loss without recourse to modeling; i.e., in a nonparametric way. These measures are now defined below under a more general framework (leading to the method GenMASSC proposed in this paper) than the framework originally used for MASSC in Singh, Yu, and Dunteman (2003) and Singh (2002, 2006).

Let $B^{(0)}$ denote the micro database of records to be treated for disclosure which is assumed to be drawn either under an unknown superpopulation model $m^{(0)}$ if it is a census or administrative data, or under a known randomization design $p^{(0)}$ if it is a sample. With respect to a given set of IVs (categorical), records are classified as U (unique) and NU (nonuniques). Each NU record is further classified as D (double), T (triple), or O (other) where D signifies that there is only one other record in $B^{(0)}$ with the IV profile common with the NU record, T signifies that there are exactly two other records with a common IV profile, and O signifies four or more records with common IV profiles. The database $B^{(0)}$ is partitioned into four risk strata indexed by h where h varies over four types of records, *U, D, T, O*. Also let $\tilde{B}^{(0)}$ denote the database of substitution partners under a substitution model $\tilde{m}^{(0)}$, such as the nearest neighbor imputation. If the substitution partner is obtained at random from a neighborhood, then it will add another randomization stage. (We will denote by $\tilde{m}^{(0)}$ both sources of random variation, one due to substitution/imputation model error and another due to random substitution if that is the case.) Next, let $B^{(1)}$ denote the treated database obtained from $B^{(0)}$ after substitution under a random mechanism ψ, and $B^{(2)}$ denote the further treated database obtained from $B^{(1)}$ after subsampling under a random mechanism ϕ. Further, let $d_k^{(1)}$ denote the substitution indicator for the record k in $B^{(0)}$ taking the value of 1 if the record got substituted; i.e., did not survive the substitution treatment-ψ and 0 otherwise. Similarly, let $d_k^{(2)}$ denote the subsampling indicator for the record k in B^1, taking the value of 1 if the record got sampled; i.e., survived the sampling-out treatment-ϕ and 0 otherwise.

Now, before we define the disclosure loss function or risk score for a record k after $\psi\phi$-treatment, we will define 'similarity' and 'sensitivity' scores for the record given that it survived the $\psi\phi$-treatment. This will generalize the measures proposed earlier by Singh (2002, 2006) by allowing for partial risk score (i.e., the disclosure loss function can now assume values between 0 and 1, and not just the extremes of 0 and 1 used earlier). For this purpose, we first categorize SVs as mentioned earlier. We use it to define similarity scores between any two records. However, the original values (categorical or otherwise) of SVs are reported in the treated database. For a record k in the database ($B^{(0)}$,$B^{(1)}$, or $B^{(2)}$), the term dissimilarity score $\eta_{k(y)}$ for each SV (denoted by y) will be used for the measure of variability of SV values among records with IV profiles being common with the record k. This is analogous to the concept of variance, except that here the SV values are categorical. Following Rao's (1982) Quadratic Entropy measure of diversity, for a record k having a cluster of m_k (≥ 2) records sharing the same IV profile, the distance measure or the dissimilarity score $\eta_{k(y)}$ is defined as

$$\eta_{k(y)} = \frac{1}{2} \left(\begin{array}{c} m_k \\ 2 \end{array} \right)^{-1} \sum_{j<j'} \lambda_{y_j y_{j'}}, \tag{1}$$

where $\lambda_{y_j y_{j'}}$denotes a subjective measure (between 0 and 1) of distance between categorical values of a given SV for the two records j and j' out of a total of m_k records. In the case of a simple random sample of m_k records, the observed counts in the categories (i =1,.., q) of the variable y follow a multinomial distribution with a corresponding probability vector $\boldsymbol{p}$, and the expected value of the distance measure is given by $\frac{1}{2} \sum_{i<i'} \lambda_{ii'} p_i p_{i'}$ or $\frac{1}{2} \boldsymbol{p}' \boldsymbol{\Lambda} \boldsymbol{p}$ where $\boldsymbol{\Lambda}$ is the $q \times q$ matrix of pairwise distances between categories (i, i'), and the diagonal elements λ_{ii} are set to zeros. The distance measure is nonnegative (in fact, between 0 and 1) by construction and reduces to the well-known Gini's measure of inequality $\frac{1}{2}(1 - \sum_{i=1}^{q} p_i^2)$ if $\lambda_{ii'}$ is set to 1 whenever $i \neq i'$. For unique records, $\eta_{k(y)}$ is not defined, but we can assign the value of 0 for computational purposes.

We also define for each record k, a sensitivity score $\zeta_{k(y)}$ (averaged over the cluster of m_k records if $m_k \geq 2$) between 0 and 1 specific to each SV such that it captures (although somewhat subjectively but based on subject matter considerations) the degree of sensitivity attached to the specific SV value in case of disclosure. The risk score for each record k can now be defined over a set of main SVs y as

$$r_k = max_y (1 - \eta_{k(y)}) \zeta_{k(y)} \tag{2}$$

For records in $B^{(0)}$, the risk score will be denoted by $r_k^{(0)}$; for records in $B^{(1)}$, it is $r_k^{(1)}$given that the record survived the substitution treatment-ψ; and for records in $B^{(2)}$, it is $r_k^{(2)}$ given that the record survived both substitution and subsampling treatments $\psi\phi$. We can now define the (unconditional) post-treatment risk score $\delta_k^{(2)}$for each record k in $B^{(0)}$. If it does not appear to be in $B^{(2)}$ by virtue of having no record with a matching IV-profile, then it is clearly zero. Otherwise, it is given by

$$\delta_k^{(2)} = (1 - d_k^{(1)}) d_k^{(2)} r_k^{(2)} \tag{3}$$

The term unconditional implies that we are not assuming that the record has actually survived the treatment process $\psi\phi$; i.e., it may or may not have survived but still appears to have. So $\delta_k^{(2)}$is zero if the record, although appears to be in $B^{(2)}$, gets substituted or sampled out. Note that the pre-treatment risk score $\delta_k^{(0)}$, and the subsequent score $\delta_k^{(1)}$ after the treatment-ψ, can also be defined in an analogous manner. All the risk scores are random due to the random process $\psi\phi$. The risk score $\delta_k^{(2)}$ estimates the parameter $E_{\psi\phi}(\delta_k^{(2)})$ which is the disclosure risk or the expected disclosure loss. A consistent estimate of this parameter can be obtained by a large number R of independent replications of the $\psi\phi$-treatment and is given by the average $R^{-1} \sum_{r=1}^{R} \delta_{k,r}^{(2)}$ where

$\delta^{(2)}_{k,r}$denotes the risk score for the rth replication of the treatment. In practice, it may be more meaningful to consider average post-treatment disclosure risk measures ($\overline{\delta}_p$) at an aggregate level such as the subgroup of records corresponding to an IV-profile 'p' corresponding to an intruder's knowledge about his or her target. We define

$$\overline{\delta}_p = T^{(0)}_{\delta(p)}/N_{p(0)};\ T^{(0)}_{\delta(p)} = \sum_{k \in B^{(0)}_p} \delta^{(2)}_k, \tag{4}$$

where $N_{p(0)}$ is the size of the subgroup $B^{(0)}_p$of records having the IV-profile 'p'. A single measure δ can also be obtained as $\max_p \overline{\delta}_{\rm p}$ over all profiles 'p'. Similarly, global post-treatment risk measures $\overline{\delta}_u$, $\overline{\delta}_d$, $\overline{\delta}_t$, and $\overline{\delta}_o$ for the whole database $B^{(0)}$can be defined which correspond respectively to averages over all records k in $B^{(0)}$ that appear to be 'unique' in $B^{(2)}$, all records that appear to be 'double', all that appear to be 'triple', and finally all that appear to be 'other' in $B^{(2)}$. Here the lower case subscripts (u, d, t, o) are used to indicate that records' features may not correspond to the true ones (U, D, T, O) from $B^{(0)}$. These global measures based on a single disclosure treatment may be of considerable interest to the data producer, and are in general, for large $N_{(0)}$, consistent estimates of the corresponding finite population parameters such as $E_{\psi\phi}(T^{(0)}_{\delta(u)})/N_{(0)}$ for $\overline{\delta}_u$ where $T^{(0)}_{\delta(u)} = \sum_{k \in B^{(0)}} \delta^{(2)}_k 1_{k(u)}$, $1_{k(u)}$ being the indicator for the event that the record k appears as u, and so on. The size $N_{(0)}$ of $B^{(0)}$ as divisor is used in the above definition of parameters because any one of the records in $B^{(0)}$ can appear as (u, d, t, o). Unlike the usual estimation problem in survey sampling, there is no need for sampling weights in estimating the finite population total $\sum_{k \in B^{(0)}} E_{\psi\phi}(\delta^{(2)}_k 1_{k(u)})$, because $\delta^{(2)}_k 1_{k(u)}$ is observed for all k in $B^{(2)}$.

It is noted that it may not be meaningful to define pre-treatment risk measures by averaging $\delta^{(0)}_k$ over subgroups containing unique records because there is no uncertainty in matching a unique record with the target's profile as long as the intruder has the knowledge of the profile. If there is match, then the record's risk score may vary from 0 to 100%. So the pre-treatment disclosure risk can be summarized by its frequency distribution over the $N_{(0)}$ records in $B^{(0)}$. On the other hand, it is indeed meaningful to average post-treatment risk scores $\delta^{(2)}_k$ over subgroups ('p' for example) because even if there is a unique match of the target's profile with a record in $B^{(2)}$, there is the possibility that the record could be any one of the records in $B^{(0)}_p$.

The calculation of above risk measures ($\overline{\delta}_u$, $\overline{\delta}_d$, $\overline{\delta}_t$, and $\overline{\delta}_o$, for example) requires performing the disclosure treatment first. It would be useful, in practice, to be able to compute initial post-treatment values of these measures using working assumptions before the actual treatment. This would be needed at the SDL design stage for choosing suitable substitution and subsampling rates for a desired goal of disclosure risk so that after treatment the resulting risk is close to what was planned. A good design for disclosure treatment before actual treatment would in practice save the number of iter-

ations needed to ensure that the estimated risk did not exceed the desired level. To this end, we consider, for simplified risk computations, disclosure risk for each risk stratum *h=U, D, T, O,* and assume that the substitution (ψ_k) and subsampling (ϕ_k) rates are approximately constant for all k in each risk stratum. Next partition the risk measure ($\overline{\delta}_u$, for example) into components so that suitable working values can be substituted to get a good initial value for $\overline{\delta}_u$. In particular, consider the partition

$$N_{(0)}\ \overline{\delta}_u \triangleq T^{(0)}_{\delta(u)} = T^{(0)}_{\delta(u|U)} + T^{(0)}_{\delta(u|D)} + T^{(0)}_{\delta(u|T)} + T^{(0)}_{\delta(u|O)}, \tag{5}$$

where $T^{(0)}_{\delta(u|U)}$, for example, is $\sum_{k\in B^{(0)}_U} \delta^{(2)}_k 1_{k(u|U)}$, and others are defined similarly. The term $T^{(0)}_{\delta(u|U)}$is simply the total risk score of all records in the risk stratum U that appear as u in the treated database $B^{(2)}$ as implied by the indicator function $1_{k(u|U)}$. Now, the average or the proportion $N^{-1}_{(0)}\ T^{(0)}_{\delta(u|U)}$, to be denoted as $\overline{\delta}_{u|U}$, can be factored as

$$\overline{\delta}_{u|U} = \pi_U \left(1 - \widehat{\psi}_U\right) \hat{\phi}_U \widehat{\chi}_{u|U} \widehat{\xi}_{u|U}, \tag{6}$$

where π_U is the relative stratum U size $N_{U(0)}/N_{(0)}$, $\widehat{\psi}_U$ is the proportion of records in stratum U that did not survive the substitution treatment and is given by $\sum_{k\in B^{(0)}_U} d^{(1)}_k/N_{U(0)}$, $\hat{\phi}_U$ is the proportion of records in stratum U that are sampled-in, given that they were not substituted, and is given by $\sum_{k\in B^{(0)}_U} (1-d^{(1)}_k)d^{(2)}_k / \sum_{k\in B^{(0)}_U} (1-d^{(1)}_k)$, $\widehat{\chi}_{u|U}$ is the proportion of records in stratum U that were classified as u after treatment given that they survived $\psi\phi$-treatment and is given by $\sum_{k\in B^{(0)}_U} (1-d^{(1)}_k)d^{(2)}_k 1_{k(u|U)} / \sum_{k\in B^{(0)}_U} (1-d^{(1)}_k)d^{(2)}_k$, and the last term $\widehat{\xi}_{u|U}$ is the average risk score for records in stratum U that survived the $\psi\phi$-treatment and were not misclassified as well, which is given by $\sum_{k\in B^{(0)}_U} \delta^{(2)}_k 1_{k(u|U)} / \sum_{k\in B^{(0)}_U} (1-d^{(1)}_k)d^{(2)}_k 1_{k(u|U)}$. Denoting the averages of $\widehat{\psi}_U$, $\widehat{\phi}_U$, $\widehat{\chi}_{u|U}$, and $\widehat{\xi}_{u|U}$ over all replications of the $\psi\phi$-treatment by ψ_U, ϕ_U, $\chi_{u|U}$, and $\xi_{u|U}$ respectively, $\overline{\delta}_{u|U}$ can be approximately expressed as $\pi_U\,(1-\psi_U)\,\phi_U \chi_{u|U} \xi_{u|U}$. Thus, before performing the actual treatment, one can get simple initial measures of the disclosure risk $\overline{\delta}_{u|U}$ by using working values of $\psi_U, \phi_U, \chi_{u|U}$, and $\xi_{u|U}$ based on user requirements and past experiences. Initial values for other components of $\overline{\delta}_u$, namely, $\overline{\delta}_{d|U}$, $\overline{\delta}_{t|U}$, and $\overline{\delta}_{o|U}$, can also be obtained along the same lines. These initial post-treatment measures are particularly useful in practice to get a preliminary idea of the extent of disclosure treatment necessary for a given database. If it is deemed too high, then the IVs may have to be recoded into broader categories to reduce the number of uniques which tends to decrease the disclosure risk, $\max\{\ \overline{\delta}_u\ ,\ \ \overline{\delta}_d,\ \ \overline{\delta}_t,\ \ \overline{\delta}_o\}$, which is desired to be less than a prescribed level to be denoted by $\delta > 0$.

2.4 Measures of Information Loss

By regarding the original database $B^{(0)}$ as the finite population, and the treated one $B^{(2)}$ as a random sample under the random mechanism $\psi\phi$, the data producer can compute the relative root mean square error (RRMSE) for a number of population total estimates $\widehat{\theta}_z^*$ (defined below) corresponding to main study variables (z) and then compute the overall information loss as $max_z\{RRMSE\left(\widehat{\theta}_z^*\right)\}$, which is desired to be less than a prescribed level to be denoted by $\varepsilon > 0$. Note that when dealing with many study variables, taking the maximum over RRMSE might be more meaningful to protect against the worst case than taking the average. The study variable z is a function of SVs (with the original values and not the ones after categorization used for computing the disclosure risk) and IVs (with the values after categorization or recoding used in the disclosure treatment and in computing the disclosure risk). In practice, the variables z's would typically be categorical, such as binary indicators of possible categories.

For the purpose of calculating information loss needed for choosing a suitable design of the disclosure treatment, we assume for simplicity that both random mechanisms (ψ) and (ϕ) are stratified simple random without replacement–STSRS. In applications, the actual design implemented may be somewhat different, but STSRS is expected to provide a reasonable approximation; see Section 5. Now for the study variable z, let $\tilde{z}$ denote the corresponding value after substitution, where only IVs (and not SVs) are substituted, and let z^* denote $\tilde{z}$ or z depending upon whether the record was selected for substitution or not. Also, let θ_z denote the total parameter of interest obtained from $B^{(0)}$ and let α, β denote respectively the desired bounds on constraints on relative bias squared and relative variance (under $\psi\phi-$randomization conditional on the substitution model $\tilde{m}^{(0)}$) of the estimator $\widehat{\theta}_z^*$ such that the RRMSE is bounded by ε; here we condition on $\tilde{m}^{(0)}$ for the sake of convenience. If the database is a sample, then θ_z is defined as $\sum_{k\in B^{(0)}} z_k w_k$ where w_k denotes the sampling weight; if not, then w_k is taken as 1. Finally, let θ_z^* denote a census-type estimator of θ_z based on $B^{(1)}$; i.e., $\sum_{k\in B^{(1)}} z_k^* w_k$. The estimator $\widehat{\theta}_z^*$ is obtained as $\sum_{k\in B^{(2)}} z_k^* w_k^*$ where $w_k^* = \phi_k^{-1} w_k$, ϕ_k being the sampling rate for record k. Note that given $\tilde{m}^{(0)}$, $\widehat{\theta}_z^*$ is a biased estimator of θ_z due to substitution. It follows that given $\tilde{m}^{(0)}$, we want the $\psi\phi-$design such that (for notational simplicity the conditioning on $\tilde{m}^{(0)}$ is omitted),

$$MSE\left(\widehat{\theta}_z^*\right) = E_{\psi\phi}\left(\widehat{\theta}_z^* - \theta_z\right)^2 = E_\psi V_{\phi|\psi}\left(\widehat{\theta}_z^*\right) + E_\psi B_{\phi|\psi}^2\left(\widehat{\theta}_z^*\right)$$
$$\leq \left(\beta + \alpha\right)\theta_z^2 = \varepsilon^2\theta_z^2 \quad , \tag{7}$$

where under STSRS with the selection rate ψ_h for substitution of records k in the risk stratum h, we obtain mean squared conditional bias as

$$
\begin{aligned}
E_\psi B^2_{\phi|\psi}\left(\widehat{\theta}^*_z\right) &= E_\psi(\theta^*_z-\theta_z)^2 = V_\psi\left(\theta^*_z\right)+\left(E_\psi\left(\theta^*_z\right)-\theta_z\right)^2 \\
&= \sum_h\sum_k N^2_{h(0)}\left(m^{-1}_{h(0)}-N^{-1}_{h(0)}\right)S^2_{\nu,h}\psi^2_h + \left(\sum_h\left(\sum_k \nu_{hk}\right)\psi_h\right)^2 \\
&= \sum_h\sum_k N_{h(0)}\left(1-\psi_h\right)\psi_h S^2_{\nu,h} + \left(\sum_h\left(\sum_k \nu_{hk}\right)\psi_h\right)^2, \qquad \text{(8a)}
\end{aligned}
$$

where $N_{h(0)}\psi_h = m_{h(0)}$, $\nu_{hk} = (\tilde{z}_{hk}-z_{hk})w_{hk}$, $\overline{\nu_h} = N^{-1}_{h(0)}\sum_k \nu_{hk}$,

$S^2_{\nu,h} = (N_{h(0)}-1)^{-1}\sum_{k=1}^{N_{h(0)}}(\nu_{hk}-\overline{\nu_h})^2$, and the conditional variance as

$$
V_{\phi|\psi}\left(\widehat{\theta}^*_z\right) = \sum_h N^2_{h(0)}\left(m^{-1}_{h(0)}-N^{-1}_{h(0)}\right)S^2_{z^*,h} = \sum_h N_{h(0)}\left(\phi^{-1}_h-1\right)S^2_{z^*,h}, \qquad \text{(8b)}
$$

where $S^2_{z^*,h}$ is defined similarly to $S^2_{\nu,h}$ except that ν_{hk} is replaced by z^*_{hk}. It is not possible to obtain an analytically closed form expression of the mean conditional variance $E_\psi V_{\phi|\psi}\left(\widehat{\theta}^*_z\right)$, although a numerical approximation can be computed from independent replications of treatment-ψ. The conditional variance, $V_{\phi|\psi}\left(\widehat{\theta}^*_z\right)$, however, provides an unbiased estimate.

3 Description of MASSC

We now describe in detail each step of the MASSC method, introduced briefly in Subsection 2.2.

Step I: Micro Agglomeration. It consists of defining risk strata with respect to IVs. As mentioned earlier, we have four broad strata *U, D, T, O* corresponding to uniques, doubles, triples, and others. In practice, one could also include noncore IVs; noncore in the sense that they may not be easily available to the intruder. Now, further strata of new uniques can be defined as one adds a noncore IV one at a time to the earlier selected IVs in a rank order defined by the anticipated difficulty in obtaining the new IV value about a target. After unique risk strata, the risk stratum of nonuniques with respect to core and noncore IVs is further divided into *D, T, O* strata.

The main purpose of this step, besides forming risk strata (needed in the subsequent steps), is to check if the initial post-treatment disclosure risk for $B^{(0)}$ (as defined in Subsection 2.3) with respect to core (and noncore) IVs needs to be reduced further by recoding IVs without jeopardizing the analytical utility of the database. It may be noted that the term micro-agglomerate for risk strata signifies that the records within

a risk stratum could be quite disparate, but the only reason they belong to the same stratum is that they happen to share features of being unique or nonunique with respect to a given set of IVs.

Step II: Substitution. In this step, the set $\tilde{B}^{(0)}$ of substitution partners for all records in $B^{(0)}$ is constructed by using the nearest neighbor imputation idea of survey sampling such that the donor record is closest to the recipient in terms of a distance function based on IVs and main SVs. Note that unlike imputation in survey sampling for missing values, here there are no missing values, but the known values are used in finding a suitable partner. In practice, it is useful to impose restrictions on the set of substitution partners so that multivariate relationships between study variables (which are functions of SVs and IVs) are not too distorted. Once the donor set for each record is selected, the record closest in terms of IVs and SVs (but with at least one field change in IVs) is selected as a substitution partner. The restriction to at least one field change of IVs avoids the possibility that the recipient and the donor may turn out to have identical IVs. As in the case of dissimilarity measure (1) for computing the risk score, we can use Rao's quadratic entropy for defining a standardized distance between the donor value ($y_{j'}$) and the recipient value (y_j) for each categorical variable y (IV or SV), given by $\lambda_{y_j y_{j'}}/p_y' \Lambda_y p_y$ where Λ_y is the matrix of pairwise distances between categories of y, and p_y is the vector of proportions of records in the donor set for the categories of y. Now, the distance between the donor and the recipient with respect to the multivariate set of IVs and SVs is defined as a linear combination of standardized distances for each variable; the weights used in the combination reflect the rank order of importance in the process of finding a similar record; see Singh (2002, 2006).

The above step is one of the most important steps in an attempt to control bias due to substitution in the analysis of a treated database. This bias is further controlled by choosing optimal selection rates for substitution as follows. For this purpose, each risk stratum can be further partitioned into substrata using clustering algorithms such that contributions to the absolute bias over main study variables are small from each substratum. Now, to find optimum selection rates ψ_h (between 0 and 1 for the hth stratum, which could be a substratum), the total expected disclosure cost from nonsubstitution of records is minimized subject to the mean squared conditional bias (8a) constraints for a number of study variables z. That is, the optimization problem is

$$min_{\psi} \sum_{h} c(\psi_h) N_{h(0)} (1 - \psi_h) \quad \text{subject to } \max_{z} E_{\psi}(\theta_z^* - \theta_z)^2/\theta_z^2 \quad \leq \alpha, \qquad (9)$$

where $c(\psi_h)$ is the nonsubstitution disclosure cost function chosen as a decreasing function of ψ_h (e.g., $a_{h(1)} {\psi_h}^{-1}$ with $a_{h(1)}$ serving as tuning parameters to allow for differential cost from stratum to stratum), and $N_{h(0)} (1 - \psi_h)$ is the expected number of records not substituted in stratum h, and α is an upper bound on the mean squared bias relative to the squared population total. Note that the term $c(\psi_h) (1 - \psi_h)$ can be interpreted as the expected disclosure cost due to possible non-substitution of a single record in stratum h.

The selection probabilities ψ_h can be restricted to lie strictly between 0 and 0.25, for example, in the interest of reducing bias, but there is a positive probability for each record to be substituted, and in any (sub)-stratum the proportion of records substituted is no more than 25%. It can be seen by differentiating the mean squared conditional bias (or $E_\psi(\theta_z^* - \theta_z)^2$) with respect to ψ_h and observing that as ψ_h increases from 0 to around 0.5, the mean squared conditional bias, under mild conditions, increases but the disclosure cost decreases; i.e., the two functions tend to move in opposite directions in the process of finding an optimum solution. Given ψ_h, a sample is selected for substitution using stratified simple random sampling without replacement with selection rates ψ_h from the hth stratum of $B^{(1)}$. Note that when a record is selected for substitution, all variables related to IVs are also substituted from the donor in order to maintain internal consistency and multivariate relationships to a certain degree. We remark that if $B^{(0)}$ itself is a multi-stage cluster sample; as may often be the case in practice, it is preferable to perform a nesting modification for selecting records for substitution within each PSU using a pps (probability proportional to size) sampling in the interest of simplified variance estimation as in single phase sampling; see Section 5 for more details.

Step III: Subsampling. In this step, given substituted database $B^{(1)}$ obtained from the previous step, the risk strata of Step I can be partitioned into a set of substrata (generally different from Step II) using a clustering algorithm such that the variability of observations within each substratum with respect to a set of main study variables is small. Now, to find optimum selection probabilities, ϕ_h, for the hth stratum or substratum, the total expected disclosure cost from sampling-in of records is minimized subject to variance constraints on a set of main study variables. That is, the optimization problem is

$$min_\phi \sum_h c(\phi_h) N_{h(0)} \phi_h \quad \text{subject to } \max_z E_\psi V_{\phi|\psi}\left(\widehat{\theta}_z^*\right)/\theta_z^2 \quad \leq \beta, \tag{10}$$

where $c(\phi_h)$ is the sampling-in disclosure cost function chosen as an increasing function of ϕ_h(e.g., $a_{h(2)}(1-\phi_h)^{-1}$ with $a_{h(2)}$ serving as tuning parameters to allow for differential cost from stratum to stratum), $N_{h(0)}\phi_h$ is the expected number of records sampled-in, and β is an upper bound on mean conditional variance relative to the squared population total. Note that the term $c(\phi_h)\phi_h$ can be interpreted as the expected disclosure cost due to possible non-sampling-out of a single record in stratum h.

The selection probabilities ϕ_h can be restricted to lie strictly between 0.5 and 1, for example. In other words, every record has a positive chance of being sampled-out, but the proportion of records sampled out from a given (sub)-stratum is no more than 50%. This restriction helps to keep the variance inflation effect of unequal weighting under control on resulting estimates. It may be noted that the two functions, variance and disclosure cost, move in opposite directions as ϕ_h increases; a desirable condition for finding a unique solution for optimization. Given ϕ_h, a sample is selected using

stratified simple random sampling without replacement with selection rates ϕ_h from the hth stratum of $B^{(1)}$. Clearly, if the database itself is a sample, then subsampling is like a second phase sample. In this case, if the first phase sample is a (multi-stage) cluster, then it is preferable, as in the substitution step, to select a pps (probability proportional to size) sample within each PSU or first phase cluster; see Section 5.

Step IV: Calibration. In this step, the sampling weights obtained from the previous step as inverse of selection probabilities are adjusted so that estimates based on a set of main study variables as well as auxiliary variables (which are selected IVs such as geo-demographic) match estimates obtained from the original database. This process, known as calibration (also termed poststratification), helps to reduce bias caused by substitution and variance due to subsampling. In this context, the method of Folsom and Singh (2000) based on a generalized logit model with bounds on weight adjustment factors can be used.

We end this section with a few remarks. If the database $B^{(0)}$ is large, like a population census or an administrative data, then it may be sufficient to perform a nominal treatment of substitution and subsampling under a simplified MASSC without any need for optimal ψ_h and ϕ_h and yet have a reasonable control on the disclosure risk. Here one can use proportional allocation of subsampling rates ϕ_h to risk strata in order to reduce variability due to unequal weighting after disclosure treatment. If $B^{(0)}$ is a sample survey data, then typically it may not be very large and would require optimized treatment for a good control on disclosure risk and information loss.

With sample survey data, there may be new concerns for data disclosure if the intruder knows stratum/PSU (primary sampling unit) identifiers and can associate his or her target with the stratum containing high or low sampling weights corresponding to under or over sampling; see De Waal and Willenborg (1997). For variance estimation, stratum and PSU subset identifications are needed, which may act as new IVs. However, it is sufficient to have pseudo-identifiers. Moreover, since under MASSC all records, including those in the stratum/PSU subset, are subject to treatment, there would be very little concern that the intruder would be able to identify a pseudo-PSU for the target except probably in cases where the intruder is indeed a respondent and belongs to the same pseudo-PSU; the reason being that having knowledge of one's IVs and SVs, it is possible to identify the record unless it was actually substituted or sampled-out. In such situations, the disclosure risk needs to be recomputed under the condition that the target belongs to the smaller stratum/PSU subset. This risk would undoubtedly be higher, but it would need to be scaled down by the probability that the intruder is also in the same subset and survived the treatment, which is likely to be very small. With regard to sampling weights which are needed for reducing selection bias in analysis (Pfeffermann, 1993), they may also act as new IVs. However, sampling weights in $B^{(0)}$ would be subject to calibration for nonresponse, coverage bias or post-stratification, and extreme values, and these calibrated weights adjusted by ϕ_h^{-1} would be further calibrated after disclosure treatment as part of MASSC. As a result, they are likely to further lose their identifying values.

We remark that for any SV (such as income) taking extreme values, a coarse categorical version of it might serve as an IV also, and as a result its original values cannot be released after MASSC treatment. In other words, if the SV with extreme values has the potential of being an IV, then MASSC essentially truncates it by top/bottom coding, making it difficult to preserve the true distribution in the treated database; see Section 6 for other limitations of MASSC. Finally, we note that any MASSC disclosure treatment is an iterative process in that the process may have to be repeated a few times with possible recoding of IVs and redefining of risk strata, and revised selection probabilities for substitution and subsampling if measures of disclosure risk and information loss turn out to be not sufficiently small or if analytical utility diagnostics such as comparison of estimates before and after treatment turn out to be unsatisfactory. Also, assuming different scenarios about intruders' possible knowledge of extra IVs, the disclosure risk for the final treated database can be recomputed to see if it remains tolerable. This is how protection against new IVs can be checked under MASSC. Once the steps of MASSC are finalized for a given (ε, δ), it is recommended that the treatment be replicated several times in order to perform analytic utility diagnostics such as summary measures of empirical distributions of ratios of estimates for several study variables from databases $B^{(0)}$ and $B^{(2)}$ as in Singh, Yu, and Wilson (2004).

4 An Illustrative Simple Hypothetical Example

It would be useful to explain in terms of a very simple example how GenMASSC introduces uncertainty about the risk status of a record, the four steps of MASSC or GenMASSC, calculations of pre- and post-disclosure treatment risks, and summary measures of information loss and disclosure risk (ε, δ). However, for real applications to large survey data, see Singh, Yu, and Wilson (2004) for creating cross-sectional PUFs at the national level, and Singh, Wright, and Yu (2004) for combined year PUFs at the state level in the context of the National Survey on Drug Use and Health.

Table 1 (a) shows a hypothetical example of 10 cases in a macro data form while Table 1(b) shows the corresponding micro data form. The IVs are age (four age categories: $1 \triangleq 12-16,\ 2 \triangleq 17-20,\ 3 \triangleq 21-24,\ 4 \triangleq 25-29$) and gender (M, F), and the SV is binge alcohol drinking, or 'Alc' for short. Suppose the sensitivity score for this SV is 1 if the response is yes for binge drinking, and 0 if the response is no. For the dissimilarity score of any record having a cluster of records with common IVs, define pairwise distances to be 1 if the SV categories are different and 0 otherwise. Now observe that the record #4 is unique because the profile based on two IVs (age and gender) is unique, and its risk score is 1 because the value of SV (Alc) is yes. Record #1 and #5 form a nonunique double because there are only two records with the IV profile of age at category 4 and gender female. These records are not at risk because values of SVs are not sensitive, although the similarity score is 1. Similarly, risk scores for nonunique triples and others can be defined. Under the inside intrusion scenario using (1) and (2), the pre-treatment disclosure risk measures can be obtained as follows: the three unique records (#4, #6, #8) and the two nonunique doubles (#2, #3) are at maximum risk each with a score of 100%, the three nonunique triples (#7, #9, #10) at moderate risk

each with a score of 2/3 of 2/3; i.e., 44.44%, and the remaining two nonunique doubles (#1, #5) at minimum risk with a score of 0%.

We now explain how the above pre-treatment risk measures can be reduced using MASSC. Basically, we need to introduce minimal distortion via substitution and subsampling, and then fix it via calibration. Table 2 illustrates the Micro Agglomeration step. There are two micro-agglomerates or risk strata corresponding to U (first three records #4, #6, #8) and NU (last seven records). We did not divide the nonunique stratum into *D, T,* and O and further into substrata because of the small size of the database. The creation of micro agglomerates is needed because the amount of disclosure treatment may vary from one agglomerate to another.

Table 2 also illustrates the step of substitution for the small hypothetical data. Although all records are subject to substitution with a positive chance, only three records were selected under a chosen random selection scheme under the MASSC process; assume ψ_h equals 1/3 for the unique stratum and 2/7 for nonuniques for an overall 30% substitution rate. For each record, we create a set of substitution partners (see Table 3) based on the nearest neighbor imputation idea under suitable restrictions such as maintaining age to the extent possible. Observe that gender is substituted for record #6, age for record #10, and both age and gender for record #5. Under subsampling (see Table 2), as in substitution, no substrata were created, again because of the small size of the database; here assume ϕ_h equals 2/3 for the unique stratum and 6/7 for nonuniques. The subsampling rates for the two risk strata are obtained after rounding the proportional allocations of an overall 80% subsampling. Observe that records #4 and #3 get sampled out, although all records are assigned a positive chance of being sampled out or suppressed. After the treatment of substitution and subsampling, all records except #4, #6, #3, #5, and #10 survive the treatment, but some of them get misclassified from being unique to nonunique or vice versa. For example, record #8 was unique and survived treatment but gets misclassified as a nonunique double because another record #10 after substitution assumed the same profile.

Finally, Table 2 also illustrates a simple calibration where after treatment the data consists of only 8 records with two females and eight males. To preserve the before-treatment gender distribution of 50-50, the subsampling weights of 7/6 for the two female records are increased to 2.5 while the weights of 3/2 for the two male records in U are decreased to 0.98, and 7/6 for four male records in NU are decreased to 0.76 so that the treated data also shows a 50-50 proportion distribution for the two gender classes.

The post-treatment global disclosure risk measures $\overline{\delta}_u$, $\overline{\delta}_d$, $\overline{\delta}_t$, and $\overline{\delta}_o$ can be computed as per the description following (4). In particular, from Table 3, we obtain $\overline{\delta}_u = 0.10$, $\overline{\delta}_d = 0.10$, $\overline{\delta}_t = 0.09$, $\overline{\delta}_o = 0$, and therefore $\delta = 10\%$. Thus with Gen-MASSC, the whole database level risk is reduced to 10% compared to being very high for some records and very low for others before any treatment. To measure the information loss, consider the binary study variable z taking the value of 1 if the individual reports binge drinking and is male; 0 otherwise. We have $N_{(0)}^{-1}\theta_z = 0.40$ and $N_{(0)}^{-1}\widehat{\theta}_z^*$=0.35. Using the formula (8a), $S^2_{\nu,1}$= 4/3 based on the U stratum of Table 3 and $S^2_{\nu,2} = 5/21$

based on the NU stratum, we get $E_\psi B^2_{\phi|\psi}\left(\widehat{\theta}^*_z\right)$ as 1.67. Similarly, using (8b), $S^2_{z^*,1}$= 0, and $S^2_{z^*,2} = 5/21$, we get $V_{\phi|\psi}\left(\widehat{\theta}^*_z\right)$ as 5/18, which imply that RRMSE$\left(N^{-1}_{(0)}\widehat{\theta}^*_z\right)$ is estimated as 0.35 or $\varepsilon = 35\%$—rather high due to substitution bias. In the above variance calculation, the Taylor-based variance adjustment for calibration is typically made in practice, although it was not done here.

5 Analysis of MASSC-treated Data

So far we have considered the computation of disclosure risk and information loss from a data producer's point of view where one has complete access to the raw data and design parameters used in the disclosure treatment. In this section, we consider analytic needs of the data user and how they can be fulfilled by MASSC-treated data. In particular, we consider the usual need of point, variance, and interval estimation of total or mean parameters of the underlying finite population (either $B^{(0)}$ if it is a census or the population from which it is sampled from), or parameters of a superpopulation model. Since the MASSC treatment process itself adds a sampling stage, the resulting data forms a two-phase sample if the original database $B^{(0)}$ is a sample or a single phase sample if it is a census or an administrative data. Conceptually, there are four stages of randomization in the MASSC treatment as explained in Subsection 2.3: first, $m^{(0)}$ or $p^{(0)}$ for the database $B^{(0)}$; second, $\tilde{m}^{(0)}$for the database $\tilde{B}^{(0)}$, third, $\psi-$randomization for $B^{(1)}$, and fourth, $\phi-$randomization for $B^{(2)}$.

For large sample inferential estimation, we generally need to estimate the MSE of a calibrated Horvitz-Thompson-type total estimator $\widehat{\theta}^*_z$ for a study variable z about the population total θ_z if $B^{(0)}$ is a census, or about $\Theta_z(=E_{p^{(0)}}(\theta_z))$ if $B^{(0)}$ is a sample, from which normality-based interval estimates could easily be constructed under a central limit theorem. So in this section, we restrict our attention to only estimation of MSE of $\widehat{\theta}^*_z$. In the following, we first condition on $B^{(0)}$; i.e., assume that θ_z is given. Now typically, the calibrated estimator $\widehat{\theta}^*_z$ will be approximately unbiased for θ^*_z (defined in Subsection 2.4) under ϕ given $\tilde{m}^{(0)}\psi$ and hence for $E_{\psi|\tilde{m}^{(0)}}(\theta^*_z)$). Next, the MSE of $\widehat{\theta}^*_z$ about θ_z under $\tilde{m}^{(0)}\psi\phi$ has three parts given by

$$E_{\tilde{m}^{(0)}\psi\phi}\left(\widehat{\theta}^*_z - \theta_z\right)^2 = E_{\tilde{m}^{(0)}}V_{\psi\phi|\tilde{m}^{(0)}}\left(\widehat{\theta}^*_z\right) + V_{\tilde{m}^{(0)}}E_{\psi|\tilde{m}^{(0)}}\left(\theta^*_z\right) + (E_{\tilde{m}^{(0)}\psi}\left(\theta^*_z\right) - \theta_z)^2, \qquad (11)$$

where the third term on the right hand side is negligible due to approximate unbiasedness of $\widehat{\theta}^*_z$ about $\theta_z$under $\tilde{m}^{(0)}\psi\phi$. The first term on the right hand side can be estimated by survey sampling techniques as discussed below. The second term inflates the variance due to substitution and reflects three sources of variation; estimation of parameters of the model $\tilde{m}^{(0)}$, and the other two due to model error and random substitution if the imputation procedure is not deterministic. It is the second term that is difficult to estimate because of its nonstandard nature. The reason for this is that imputation or substitution flags are not part of PUF for obvious confidentiality reasons. So the user

cannot create replicate copies $B_r^{(2)}$ of the treated database $B^{(2)}$. Moreover, it is also not safe for the data producer to release copies $B_r^{(2)}$ because records whose values of IVs do not change from replicate to replicate may be more at risk as they are likely to be the records that survived the treatment process, and hence the corresponding SVs are at risk of disclosure.

Fortunately, for our application, the second term $V_{\tilde{m}^{(0)}} E_{\psi|\tilde{m}^{(0)}}\left(\theta_z^*\right)$ of (11) is expected to be negligible in comparison to the first term $E_{\tilde{m}^{(0)}} V_{\psi\phi|\tilde{m}^{(0)}}\left(\widehat{\theta}_z^*\right)$ because under MASSC only a small portion of the database needs to be actually substituted when $N_{(0)}$ is large, as is usually the case in practice. The desired MSE about θ_z can be approximated by estimating $V_{\psi\phi|\tilde{m}^{(0)}}\left(\widehat{\theta}_z^*\right)$ using known results in survey sampling. In particular, single phase variance estimation methods (Taylor linearization or replication methods) can be used under certain conditions. More specifically, if the original database $B^{(0)}$ is a first phase sample based on PSUs, then with a nesting modification for the second phase, well-known single phase simplified variance estimation methods based on the with-replacement PSU assumption can be used after suitable adjustments for weight calibration step in MASSC; see Singh et al. (2003) and Singh (2008). This modification entails nesting of second phase sampling within each first phase cluster and is performed pps with size measures determined by sample allocations to domains defined by second phase stratification variables. Thus for MASSC, as mentioned in Section 3, using the substitution rates ψ_k, records are selected pps for substitution within each first phase cluster. Similarly, using the subsampling rates ϕ_k, records are selected pps for subsampling within each first phase cluster. The nesting modification for subsampling makes the estimators $\widehat{\theta_{z,i}^*}$ for each PSU i unbiased about $E_{\phi|\psi}(\widehat{\theta_{z,i}^*})$ conditionally on the substituted first phase sample, and for PSUs i and j, $\widehat{\theta_{z,i}^*}$ and $\widehat{\theta_{z,j}^*}$ also conditionally uncorrelated; here conditioning on $\tilde{m}^{(0)}$ is omitted for notational convenience. This, in turn, implies that the nesting modification for substitution makes for each PSU i the estimator $E_{\phi|\psi}(\widehat{\theta_{z,i}^*})$ unbiased about θ_z^* conditionally on the first phase sample, and for PSUs i and j, $E_{\phi|\psi}(\widehat{\theta_{z,i}^*})$ and $E_{\phi|\psi}(\widehat{\theta_{z,j}^*})$ also conditionally uncorrelated. The above follows from the identity,

$$
\begin{aligned}
C\left(\widehat{\theta_{z,i}^*}-\theta_z^*,\ \widehat{\theta_{z,j}^*}-\theta_z^*\right) = {} & C\left(\widehat{\theta_{z,i}^*}-E_{\phi|\psi}\left(\widehat{\theta_{z,i}^*}\right),\quad \widehat{\theta_{z,j}^*}-E_{\phi|\psi}(\widehat{\theta_{z,j}^*})\right) + \\
C\left(\widehat{\theta_{z,i}^*}-E_{\phi|\psi}\left(\widehat{\theta_{z,i}^*}\right),\quad E_{\phi|\psi}\left(\widehat{\theta_{z,j}^*}\right)-\theta_z^*\right) + {} & C\left(E_{\phi|\psi}\left(\widehat{\theta_{z,i}^*}\right)-\theta_z^*,\quad \widehat{\theta_{z,j}^*}-E_{\phi|\psi}(\widehat{\theta_{z,j}^*})\right) + \\
& C\left(E_{\phi|\psi}\left(\widehat{\theta_{z,i}^*}\right)-\theta_z^*,\quad E_{\phi|\psi}\left(\widehat{\theta_{z,j}^*}\right)-\theta_z^*\right), \qquad (12)
\end{aligned}
$$

where C denotes the covariance operator under appropriate randomization schemes. All the terms on the right hand side are zero under nested modifications whenever i is not equal to j. Applicability of simplified single phase variance estimators allows for use of standard software packages such as SUDAAN for analysis of MASSC-treated data. Note that with the above nesting modification, although the sample sizes in the

original second phase strata become random, computations of optimal substitution and subsampling rates remain approximately valid because in expectation the sample sizes nearly match the target allocations.

In the above discussion, we considered estimating MSE about θ_z; i.e., θ_z is treated as given which will be the case if $B^{(0)}$ is not a sample. Even if $B^{(0)}$ were a sample, using an important well-known result in survey sampling for simplifying variance estimation via the working assumption of with-replacement PSUs, the above variance estimator continues to provide an approximately unbiased estimate of the unconditional variance of $\widehat{\theta}_z^*$ about Θ_z. Now, if the first phase sample $B^{(0)}$ is single stage and not based on PSUs, as in the case of element level first phase sampling, and the second phase sampling is STSRS, then assuming that a replicate variance estimator can be obtained for the first phase, the single phase variance estimator still remains approximately valid; see Kim, Navarro, and Fuller (2006), and Singh (2008). Finally, if $B^{(0)}$ is not a sample, then the $V\left(\widehat{\theta}_z^*\right)$ can, of course, be estimated by using standard results on STSRS for single phase designs.

6 Alternative Methods: Review and Comparison

As mentioned in the introduction, the SDL methods can broadly be classified into two types, synthetic and nonsynthetic; see also Cox (1996) and Skinner (2009). The synthetic type consists of using (explicit or implicit) model-based techniques (such as multiple imputation, noise addition, or data swapping) to perturb IVs of all records by replacing the original untreated database with generated data. Important papers on the use of multiple imputation can be attributed to Raghunathan et al. (2003) for complete synthesis (i.e., both IVs and SVs are perturbed) and Liu and Little (2002), and Reiter (2003) for partial synthesis (i.e., only IVs are perturbed); noise addition by Kim (1986), Kim and Winkler (1995), Fuller (1993), and Cox, Kelly, and Patil (2004); and data swapping by Dalenius and Reiss (1982). The nonsynthetic type also involves perturbing IVs based on imputation methods, but only for a small fraction of records (typically those deemed to be at risk), as well as suppression of the whole record (or just some values of IVs/SVs). The MASSC method, considered in this paper, falls under the nonsynthetic SDL type but is quite different, as its framework is based on sampling from a finite population defined by the original database.

The introduction of HIPAA regulations has generated in recent years considerable interest in simple methods of disclosure treatment such as the Safe Harbor rule—a form of global recoding for perturbation which is relatively easy to implement. Among well-known software for nonsynthetic SDL methods are tau-Argus for macro data and mu-Argus for micro data developed at Statistics Netherlands using mainly the techniques of global recoding of IVs for all records and local suppression of IVs/SVs for selected records based on threshold rules for risk assessment; see Hundepool and Willenborg (1999). Moreover, there is the option of using optimization methods for efficient suppression. The Argus software is an important contribution to the practice of SDL.

The main limitations, however, seem to be restrictions in making data available at lower levels (cells in a high dimensional table or statistics for small areas), and the problem of analyzing data in the presence of selection bias due to deterministic treatment (i.e., suppression) of records.

The synthetic approach is theoretically appealing as it has no information loss if the model is analytically valid. Unlike nonsynthetic methods, the IVs do not need to be categorized or recoded to reduce the number of records at risk. Also, the disclosure risk in principle is zero because all IVs are synthetic. However, it is possible that records whose IV categories do not change even after synthetic treatment might be at risk of disclosure as in risk scenarios considered under MASSC because typically it may be reasonable to assume that the IVs/SVs are implicitly being categorized by an intruder due to limited precision in his or her knowledge. Note that the extent of this problem could be checked by computing disclosure risk measures analogous to those for GenMASSC. With multiple copies of synthetic data, suitable analysis can also be performed similar to the case of multiply imputed data. Here, we do not have the problem (mentioned earlier in Section 5) of releasing replicate copies of MASSC-treated data because all IVs are substituted under a synthetic approach. Despite advantages of synthetic SDL techniques, there are areas of concern as mentioned earlier in the introduction.

In nonsynthetic methods, there is more preservation of original data because only a small fraction of records are perturbed or suppressed. However, unlike MASSC, there is in general no simultaneous control of information loss and disclosure risk under other nonsynthetic methods. Also, for nonsynthetic methods based on deterministic treatment, although the disclosure risk for treated records goes to zero, there is no protection of new records at risk if the intruder gains knowledge of extra IVs. Moreover, assuming that the substitution model is valid, the analysis based on treated data is likely to be biased because it may not be reasonable in general to assume that the selection mechanism for treatment is ignorable for the analysis of interest. Finally, it is difficult to measure information loss due to bias because of nonrandom treatment. Thus there are also several areas of concern with nonsynthetic methods in general.

We remark that the (nonsynthetic) SDL method MASSC can overcome above concerns. However, the main limitation of MASSC seems to be the need to recode IVs for limiting number of records at risk, and as a result the IVs may get too coarsened. Note that we cannot release data at finer categories of IVs than what were used for disclosure treatment or else the risk would go up. Another limitation is the need for imputation-adjusted variance estimates especially when the fraction of substituted records is not too small; this could happen if the original database is not too large.

Finally, we remark that it seems difficult to compare different SDL techniques in terms of disclosure risk and information loss. Just as there are a variety of SDL techniques, there are a variety of definitional assumptions underlying each technique. Definitions of disclosure, disclosure risk, and information loss can vary depending upon the particular context of a treatment scenario. However, using the inside intrusion framework, it is possible for the data producer to compute disclosure risk and information

loss for any database along the lines of GenMASSC after suitably categorizing IVs and SVs, taking into account anticipated knowledge of the intruder.

7 Concluding Remarks and Future Work

We have considered the nonsynthetic SDL method MASSC (and GenMASSC) which, being a survey sampling based method, remains applicable to any data which can be at the macro or micro level arising from sample surveys, censuses, or administrative sources. It can also be applied to longitudinal data as well, but substitution and subsampling rates may have to be revised periodically in view of potentially new IVs arising from knowing SVs from previous time points. If there are different levels in the database as in the case of individual, family, and household level data, MASSC treatment would need to be applied at each level. The MASSC method was developed at RTI International in the early 2000s and has been in use for several surveys, including its major application to the yearly data from the National Survey on Drug Use and Health.

The main contributions of MASSC to the SDL literature comprise the inside intrusion framework (a conservative but practical stance), nonparametric measures (ε, δ) of information loss and disclosure risk for their simultaneous control (working values of (0.15, 0.25) as upper bounds for (ε, δ) have been used in practice although more discussion among researchers is needed to recommend acceptable threshold levels), some protection against new IVs not taken into account during the disclosure treatment, and valid analysis of MASSC-treated data for complex surveys by taking account of first and second phase sampling weights if the original database is a sample. Even in the absence of inside intrusion, the proposed inside intrusion framework provides a simple way to find an upper bound on the disclosure risk under outside intrusion. In this paper, we also proposed an enhanced MASSC (denoted GenMASSC) which generalizes risk measures by allowing for partial risk scores to unique and nonunique records. The main limitations of MASSC in relation to synthetic methods, however, seem to be the lack of ability to publish IVs at a level finer than the categorization used in the disclosure treatment, and the difficulty in obtaining imputation (or substitution) adjusted variance estimates for data analysis because records subject to substitution cannot be flagged for confidentiality reasons. It would be useful to investigate ways to overcome these limitations without losing the simplicity of the MASSC methodology.

Finally, MASSC provides an important nonsynthetic alternative to synthetic methods which seem to have limitations mainly with respect to taking full account of nonignorable sample designs, and developing suitable models for high dimensional data. In this context, the method of inverse sampling for undoing the complex design as proposed by Hinkins, Oh, and Scheuren (1997) and further work by Rao, Scott, and Benhin (2003) might be fruitful for synthetic methods in dealing with survey data. We conclude with the remark that the SDL methods considered in this paper were concerned only with the problem of input data treatment, that is, disclosure treatment of the raw data. For future research, we plan to investigate the complementary problem of treating output from data analysis of the original untreated database which, although challenging,

might be more appealing to practitioners with a variety of analytical needs, and users with remote access.

Acknowledgments

This research was partially supported by an individual operating grant from the Natural Sciences and Research Council of Canada held at Carleton University, Ottawa, under an adjunct research professorship in the School of Mathematics and Statistics. The initial work on MASSC was conducted when the author was at RTI International and thanks are due to Feng Yu for her contributions in carrying out extensive tests for implementing MASSC on the NSDUH Data. The current work on MASSC enhancements (denoted GenMASSC in this paper) was conducted at NORC at the University of Chicago.

References

Bethlehem, J.G., Keller, W.J., and Pannekoek, J. (1990). Disclosure control of Microdata. *Journal of the American Statistical Association,* 85, 38-45.

Cox, L.H. (1996). Protecting confidentiality in small population health and environmental statistics. *Statistics in Medicine*, 15, 1895-1905.

Cox, L.H., Kelly, J.P., and Patil, R. (2004). Balancing quality and confidentiality for multivariate tabular data. In J. Domingo-Ferrer and V. Torra (eds.), *Privacy in Statistical Databases.* Lecture Notes in Computer Science, 3050 Berlin: Springer, 87-98.

Dalenius, T. and Reiss, S.P. (1982). Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference,* 6. 73-85.

Duncan, G.T., Fienberg, S.E., Krishnan, R., Padman, R., and Roehrig, S.F. (2001). Disclosure limitation methods and information loss for tabular data. In P. Doyle, J.I. lane, J.J.M. Theeuwes, and L.V. Zayatz (eds.), *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies.*Amsterdam: Elsevier, 135-166.

De Waal, A.G., and Willenborg, L.C.R.J.(1997). Statistical disclosure control and sampling weights. *Journal of Official Statistics*, 13, 417-434.

Folsom, R.E. Jr., and Singh, A.C. (2000). "A Generalized Exponential Model for Sampling Weight Calibration for a Unified Approach to Nonresponse, Poststratification, and Extreme Weight Adjustments." In *Proceedings of the Section on Survey Research Methods,* American Statistical Association, 598-603.

Fuller, W.A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics*, 9, 383-406.

Hindepool, A. and Willenborg, L.C.R.J. (1999). ARGUS: Software for SDC project. *Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*, Thessaloniki, Greece, Working paper #7.

Hinkins, S., Oh, H.L., and Scheuren, F. (1997). Inverse sampling design algorithms. *Survey Methodology*, 23, 11-21.

Kim, J. J. (1986). A Method For Limiting Disclosure in Microdata Based on Random Noise and Transformation. In *Proceedings of the Section on Survey Research Methods,* American Statistical Association, 370-374.

Kim, J. J, and Winkler, W.E. (1995). Masking microdata files. In *Proceedings of the Section on Survey Research Methods,* American Statistical Association, 114-119.

Kim, J.K., Navarro, A., and Fuller, W.A. (2006). Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Association,* 101, 312-320.

Liu, F., and Little, R.J.A. (2002). Selective multiple imputation of keys for statistical disclosure control in microdata. In *Proceedings of the Section on Survey Research Methods,* American Statistical Association, 2133-2138.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *Int. Statist. Rev.*, 61, 317-337.

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

Rao, C.R. (1982). Diversity: its measurement, decomposition, apportionment, and analysis. *Sankhya: Indian Journal of Statistics*, Series A, 44, 1-22.

Rao, J.N.K., Scott, A.J., and Benhin, E. (2003). Undoing complex survey data structures: Some theory and applications of inverse sampling (with discussion). *Survey Methodology*, 29, 107-128.

Reiter, J.P. (2005). Estimating risks of identification disclosure in microdata. *Journal of the American Statistical Association*, 100, 1103-1112.

Reiter, J.P. (2003). Inference for partially synthetic public use microdata sets. *Survey Methdology*, 29, 181-188.

Reiter, J, P., Raghunathan, T.E., and Kinney, S.K. (2006). The Importance of Modeling the Sampling Design in Multiple Imputation for Missing Data. *Survey Methodology*, 32, 143-149.

Scheuren, F. (1995). Private lives and public policies: confidentiality and accessibility of government services. *Journal of the American Statistical Association*, 90, 386-387.

Scheuren, F. (1999). Administrative records and census taking. *Survey Methodology*, 25, 151-160.

Singh, A.C. (2008). Single phase simplified variance estimation approach to two phase-stage hybrid designs. In *Proceedings of the Section on Survey Research Methods,* American Statistical Association, 2501-2508.

Singh, A.C. (2006). Method for Statistical Disclosure Limitation. *United States Patent No. 7058638 B2,* (www.uspto.gov/patft/index.html)

Singh, A.C. (2002). Method for Statistical Disclosure Limitation. *United States Patent Application Pub. No.* US 2004/0049517A1.

Singh, A.C., Yu, F., and Wilson, D.H. (2004). Measures of Information loss and Disclosure risk under MASSC treatment of microdata for statistical disclosure limitation. In *Proceedings of the Section on Survey Research Methods,* American Statistical Association*,* 4374-4381.

Singh, A.C., Wright, D., and Yu, F. (2004). Combined-year state-level and single-year nation-level public use files from the National Household Survey on Drug Use and Health data. In *Proceedings of the Section on Survey Research Methods,* American Statistical Association, 4366-4373.

Singh, A.C., Yu, F., and Dunteman, G.H. (2003). MASSC: A new data mask for limiting statistical information loss and disclosure. In *Proceedings of the Joint UN-ECE/EUROSTAT Work Session on Statistical Data Confidentiality*, Luxembourg, 373-394.

Skinner, C.J. (2009). Statistical Disclosure Control of Survey Data, In: D. Pfeffermann and C.R. Rao, Eds., *Handbook of Statistics: Sample Surveys: Inference and Analysis, Vol. 29A*, The Netherlands: North-Holland, 381-396.

Skinner, C.J. and Carter, R.G.(2003). Estimation of a measure of disclosure risk for survey microdata under unequal probability sampling. *Survey Methodology*, 29, 177-180.

Skinner, C.J., and Elliott, M.J. (2002). A measure of disclosure risk for microdata. *Journal of Royal Statistical Society*, Ser. B., 64, 855-867.

Skinner, C.J., and Holmes, D.J. (1998). Estimating the re-identification risk per record in microdata. *Journal of Official Statistics*, 14, 361-372.

U.S. Department of Health and Human Services (2000). HIPAA – Health Insurance

Portability and Accountability Act. *Federal Register No. 250*, *Rules and Regulations*, Vol.65, 82798-82829.

Zayatz, L. (2003). Disclosure limitation for Census 2000 tabular data. In *Proceedings of Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*, Luxembourg, 230-238.

Table 1(a): Hypothetical Tabular Data (IV: age, gender, SV: Binge Alcohol Drinking–Alc)

	Age 12-16		**Age 17-20**		**Age 21-24**		**Age 25-29**	
Alc	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**
Y	1	1	1	2	2	0	0	0
N	0	0	0	0	1	0	0	2

Table 1(b): Micro Data Representation of the Tabular Data from Table 1(a)

Raw Data before treatment				
Obs	**Age**	**Gender**	**Alc**	**Status before Treatment**
1	4	F	N	Nonunique double; risk score=0
2	2	F	Y	Nonunique double; risk score =1
3	2	F	Y	Nonunique double; risk score =1
4	1	M	Y	Unique; risk score =1
5	4	F	N	Nonunique double; risk score =0
6	1	F	Y	Unique; risk score =1
7	3	M	N	Nonunique triple; risk score =4/9
8	2	M	Y	Unique; risk score =1
9	3	M	Y	Nonunique triple; risk score =4/9
10	3	M	Y	Nonunique triple; risk score =4/9

Table 2: MASSC Method (Illustration of Four Steps)

Data After Micro Agglomeration					**After Substitution**			**After Subsampling**	**After Calibration**
Obs	**Age**	**Gender**	**Alc**	**Wt**	**Age**	**Gender**	**Alc**	**Status after Treatment**	**Adjusted Wt**
Unique Risk Stratum: **U**									
4	1	M	Y	1	1	M	Y	Sampled out; Wt=0	0.00
6*	1	F	Y	1	1	M	Y	Pseudo-unique; Wt=3/2	0.98
8	2	M	Y	1	2	M	Y	Pseudo-nonunique double; Wt=3/2	0.98
Nonunique Risk Stratum: **NU (D and T)**									
2	2	F	Y	1	2	F	Y	Pseudo-unique; Wt=7/6	2.50
3	2	F	Y	1	2	F	Y	Sampled out; Wt=0	0.00
1	4	F	N	1	4	F	N	Pseudo-unique; Wt=7/6	2.50
5*	4	F	N	1	3	M	N	Pseudo-nonunique triple; Wt=7/6	0.76
9	3	M	Y	1	3	M	Y	Pseudo-nonunique triple; Wt=7/6	0.76
7	3	M	N	1	3	M	N	Pseudo-nonunique triple; Wt=7/6	0.76
10*	3	M	Y	1	2	M	Y	Pseudo-nonunique double; Wt=7/6	0.76

Table 3: Variables for the Calculation of Disclosure Risk and Information Loss under GenMASSC

Obs	**Sub**	$\mathbf{1 - d_{hk}^{(1)}}$	$\mathbf{d_{hk}^{(2)}}$	$\mathbf{r_{hk}^{(2)}}$	$\boldsymbol{z_{hk}}$	$\boldsymbol{\tilde{z}_{hk}}$	$\boldsymbol{z_{hk}^{*}}$	$\boldsymbol{v_{hk}}$
4	6	1	0	1	1	0	1	-1
6^*	4	0	1	1	0	1	1	1
8	1, 2, or 3	1	1	1	1	0	1	-1
2	8	1	1	1	0	1	0	1
3	8	1	0	1	0	1	0	1
1	7, 9, or 10	1	1	0	0	0	0	0
5^*	7, 9, or 10	0	1	0	0	0	0	0
9	8	1	1	4/9	1	1	1	0
7	8	1	1	4/9	0	0	0	0
10^*	8	0	1	4/9	1	1	1	0

Notes: 'Sub' denotes the Substitution partner. In selecting partners, priority was given to maintaining age to the extent possible. Superscript '*' in the first column denotes the unit actually selected for substitution. The hk subscript denotes the kth observation in the hth stratum. ψ_h equals 1/3 for U and 2/7 for NU. ϕ_h equals 2/3 for U and 6/7 for NU. The variable z_{hk} equals 1 if the unit is male and reports binge drinking. $\tilde{z}_{hk}$ denotes the value after substitution where only IVs and not SVs are substituted. z^*_{hk} denotes $\tilde{z}_{hk}$ or z_{hk} depending upon whether the record was selected for substitution or not, and $v_{hk} = (\tilde{z}_{hk} - z_{hk})w_{hk}$ where $w_{hk} = 1$ for this example.