Application of MASSC to a Survey of Health Related Behaviors among Military Personnel

G. H. Dunteman, F. Yu, S. Chen, and D. Wang
RTI International, Research Triangle Park, NC 27709

ABSTRACT

The Department of Defense (DoD) Survey of Health Related Behaviors among Military Personnel is a worldwide survey that collects data on a wide range of health related behaviors and conditions, including alcohol use, illicit drug use, tobacco use, mental health status, physical health status, and sexual activity. These variables are used as outcome variables in various statistical models. Because the sensitive information collected in this survey may have the potential to affect or even end a participant's military career, it must be protected from disclosure. The survey also collects personal background data on variables including pay grade, age, gender, education level, and race, among others. These are defined as *identifying variables*, since an intruder could use them to identify a participant and disclose sensitive information. Identifying variables are also typically used as predictor variables in modeling outcome variables. A wide range of identifying variables and sensitive outcome variables need to be included in a public use file (PUF) for it to have analytic utility. Applying the statistical disclosure limitation method MASSC to the creation of a PUF for this survey minimizes disclosure risk while preserving analytic utility. Objective measures of both disclosure risk and analytic utility are calculated for the treated data.

Key words: Disclosure risk, Information loss, MASSC, DoD, PUFs.

1. INTRODUCTION

The 2002 Department of Defense (DoD) Survey of Health Related Behaviors among Military Personnel database, like any other database containing personal, confidential information, requires protection from disclosure before being released as a public use file (PUF) for researchers. This is, in fact, obligatory in view of the confidentiality pledge given to the participant before conducting the interview. Also, in view of regulations such as HIPAA (for protected health information) and CIPSEA (for various kinds of government data), data producers are increasingly more concerned about protecting confidentiality.

It is possible that for some participants, confidential information collected in the 2002 DoD survey is at risk for disclosure. The direct identifiers do not pose any problems because the details can be suppressed. However, an important disclosure scenario known as "disclosure by response knowledge" (see Bethlehem, Keller, & Pannekoek, 1990) is of great concern to survey participants. A participant may be concerned about protecting his/her confidential information from an acquaintance, such as a family member, friend, coworker, supervisor, or installation commander, who knows about the presence of the participant's information in the database. Such an individual could possibly find values of sensitive variables (SVs), such as past-month cigarette, alcohol, and illicit drug use, in the participant's record by using a combination of values of several indirect identifying/intrusion variables (IVs), such as age, gender, race, and marital status. Examples of individuals interested in this kind of information could be spouses concerned about sexual activities, supervisors interested in stress level or suicide ideation, and installation commanders concerned about drug and alcohol abuse. Sample members respond voluntarily. If they cannot be convinced that their answers are confidential, they will be less likely to participate and less likely to give truthful answers if they do participate. We refer to an intruder with response knowledge as an *inside intruder*, as opposed to an *outside intruder*, who does not have knowledge of the presence of the target's information in the database.

There are at least two ways the inside intruder might find values of SVs: first, the combination of IVs makes the target unique in the database; second, the combination of IVs may make the target nonunique, but all records in the nonunique group may have common values for at least one SV. The above scenario of inside intrusion provides a practical way of finding out what records could possibly be at risk and then finding ways of introducing uncertainty about their presence and identity in the database. The fact that some records could be at risk is of great concern to the data producer, who needs participants' trust. It is not possible for the producer to release the original untreated survey data to users for analysis purposes unless some form of a licensing agreement is in place for the users to access the original database. This option of a signed licensing agreement may be viable for some users, but for general users, there is a need for PUFs.

For disclosure treatment of any microdata such as those from the 2002 DoD survey, most of the available methods involve some form of perturbation of IV values (such as swapping, recoding, or adding noise), suppression of IV or SV

values (such as deleting a field or the whole record), or both. Every form of disclosure treatment introduces some loss of information. Under the inside intrusion scenario mentioned above, a large number of records could be at risk if the intruder has knowledge of a participant's IVs. Treatment for records at risk via perturbation or suppression may introduce high information loss so that an unacceptably large bias could be present in the analysis of resulting data. Therefore, a method that simultaneously protects both confidentiality and analytical quality of data is desirable so that a suitable balance between disclosure risk and information loss can be achieved.

RTI's MASSC method offers such an option, and a version of it was used for National Survey on Drug Use and Health (NSDUH) data for the years 1999 through 2002. Alternative methods in the class of nonsynthetic disclosure treatment methods use a deterministic selection of records for treatment (i.e., all the records identified to be at risk with respect to a set of IVs are treated). However, MASSC uses a stochastic selection of records for treatment in which all records are subject to treatment but only a small *random* subset is actually treated; this leads to low information loss and to protection against new IVs that an intruder might know. Thus, under a probabilistic framework, MASSC introduces sufficient uncertainty about the presence and identity of a record, and it provides measures of disclosure risk and information loss without any modeling assumptions.

The MASSC acronym signifies the four steps of **M**icro **A**gglomeration, **S**ubstitution, **S**ubsampling, and **C**alibration. MASSC is grounded in the theory of survey sampling; it uses a subtle analogy between releasing an untreated database and conducting a census. It is designed to minimize disclosure cost while controlling loss of information. Micro Agglomeration creates risk strata and checks for records at risk. This step controls the number of records initially at risk by determining the level of IV details to be released. Substitution uses optimal sampling rates for selecting records at random for perturbation, subject to substitution bias constraints. This step introduces uncertainty about the identity of a target. Subsampling uses optimal sampling rates to select records from the substituted database at random for nonsuppression, subject to precision constraints. This step introduces uncertainty about the presence of a target. Calibration uses optimal weight calibration to adjust subsampling weights, subject to preserving key estimates from the original database. This step reduces bias due to substitution and variance due to subsampling. The MASSC treatment adds a second phase to the 2002 DoD survey data. By making the selection of records independent from one PSU to another for substitution and subsampling, under general conditions the commonly used single phase survey data analysis methods such as the ones in the SUDAAN software can be used to analyze the MASSC treated data. For more details on MASSC, see Singh, Yu, and Dunteman (2003).

In the following sections, we discuss in detail MASSC data disclosure treatment for the 2002 DoD survey.

2. INITIAL RISK ASSESSMENT AND VARIABLE REVIEW

Variables related to disclosure risk are termed *IVs* and *SVs* in the database. After we evaluate an intruder's ease of accessing all possible IVs in the data set and these variables' analytical utilities., eight variables are selected as core IVs, which are expected to be known to the intruder, such as age, race, and services, and two variables are defined as noncore IVs, which are expected to be less easily obtained by the intruder. Fourteen SVs are identified for confidentiality protection, such as alcohol use, illicit drug use, tobacco use, and stress.

MASSC requires that all the variables used in the process have complete information. Among these ten IVs, five have missing values and need imputation. All SVs have missing values and require imputation. Weighted sequential hot-deck imputation technique is applied to impute missing values.

Using a set of IVs, we define a profile (which is a combination of the values of a set of IVs) to be unique or nonunique (double, triple, or other [4+]) as follows:

Uniques: All records in the data set whose profiles, for a given set of identifying variables, are unique

Doubles: All records in the data set whose profiles, for a given set of identifying variables, are shared by only one other record in the data set

Triples: All records in the data set whose profiles, for a given set of identifying variables, are shared by only two other records in the data set

Others (4+): All records in the data set whose profiles, for a given set of identifying variables, are shared by at least three other records in the data set

A record is called *at risk* if it is unique in the database with respect to a set of identifying variables and at least one SV is sensitive, or if the participant falls into a nonunique group but all records in the nonunique group have common values for at least one sensitive variable.

To evaluate and identify the records at risk, we first group IVs into the nested sequence

$$IVgp1 \subset IVgp2 \subset IVgp3,$$

where IVgp1 is the set of core IVs, IVgp2 is the second set of IVs (which consists of IVgp1 and the first noncore IV), and IVgp3 consists of the IVgp2 and the second IV. After defining these IV groups, we are able to determine the records at risk at different levels (i.e., uniques with respect to IVgp1 have higher potential disclosure risk than extra uniques with respect to IVgp2 and so on). For nonuniques, we can assess the disclosure risk by checking the distributions of all the SVs. For the 2002 DoD survey, the percentage of records at risk is rather high.

3. MASSC TREATMENT

3.1 Micro Agglomeration with Initial Categorization

In this step, we recode some IVs to reduce the number of uniques. These variables are recoded into broader categories, reserving the analytical utility. After recording, we evaluate the number of uniques and the records at risk. Six risk strata are defined based on the ten core and noncore IVs for risk analysis. Risk stratum 1 is for uniques with respect to IVgp1, risk stratum 2 is for new uniques with respect to IVgp2, and risk stratum 3 is for new uniques with respect to IVgp3, which consists of all core and noncore IVs. The remaining records are nonuniques with respect to all core and noncore IVs and are divided further into nonunique doubles for risk stratum 4, nonunique triples for risk stratum 5, and nonunique others for risk stratum 6. If the number of records is still too high, we make the categories broader. After recoding, the amount of uniques at risk is reduced to 22.20 percent of all the records.

Outcome variables are transformed into 0/1 variables for subsequent processes. Variables with more than two categories are collapsed and then recoded into 0/1 variables.

3.2 Substitution

Prior to substitution, mean squared error (MSE) constraints have to be defined to control bias in the substitution process. Four domains and fourteen outcome variables are used to construct 240 bias constraints.

Nineteen pre-defined Quadratic Entropy (QE) distance functions were used to select substitution partners. For every distance function, a record with the smallest nonzero distance within that particular group is chosen as the donor for the corresponding record in the data set. Among nineteen potential donor sets, biases are measured if all records get substituted based on the 240 bias constraints. Bias constraints used in the screen process are the same as the above substitution constraints. The bias diagnostic test shows that the thirteenth donor set has the smallest bias for all the bias constraints, so we choose it as the final donor set for substitution.

For each risk stratum, records are further partitioned into subgroups such that biases cancel one another at the maximum level within the subgroup. Same study variables are used for clustering the subgroups among each risk stratum except for four outcome variables. We use four substrata for every risk stratum, for a total of 24 (=6*4) substrata. Optimal selection probabilities for each substratum are found such that disclosure loss is minimized subject to 240 MSE constraints. According the optimal substitution probabilities, records are selected randomly for substitution with variables provided by the corresponding donor. The actual overall substitution rate is 8.78 percent.

3.3 Subsampling

Subsampling constraints are defined for the subsampling in order to control for variance inflation. We considered 23 domains and 14 outcome variables to construct 368 variance constraints.

For each risk stratum, records are further partitioned into subgroups such that variances are minimal within the subgroup. Same study variables are used for clustering the subgroups among each risk stratum except for the four outcome variables. We use four substrata for every risk stratum, for a total of 24 (=6*4) substrata. Optimal subsampling probabilities are found such that the disclosure loss is minimized, subject to 368 variance constraints. The actual subsampling overall rate is 95.12 percent.

3.4 Calibration

The weight after subsampling is calibrated using RTI's Generalized Exponential Model (GEM) (see Folsom & Singh, 2000) technique such that the key estimates on social and demographic variables are preserved. The weight calibration has been made in 210 calibration constraints.

3.5 Confidentiality Diagnostics (δ's)

δ measures the probability that a record survived MASSC treatment and that its values of SVs are sensitive. δ's for uniques, doubles, triples, and others with respect to all core and noncore IVs are computed to check for the disclosure risk of the final treated database assuming different intruders' knowledge about the IVs. They are summarized in Table 1. The results with respect to only eight core IVs are also calculated.

Table 1. Measure of Disclosure Risk on the Treated Database

Intruder's Knowledge	δ			
	Uniques	Doubles	Triples	Others
All 10 IVs	0.12	0.10	0.07	0.19
8 IVs	0.06	0.04	0.03	0.15

4. ANALYTICAL QUALITY

We assess the analytical quality of the data through comparing the analytical results using treated data and original data. We define

$$\text{Ratio Est} = \frac{\text{Estimate After Treatment}}{\text{Estimate Before Treatment}}$$

and

$$\text{Ratio SE} = \frac{\text{Standard Error After Treatment}}{\text{Standard Error Before Treatment}}.$$

Tables 2 and 3 display the quartiles of Ratio Est and Ratio SE for all 47 reporting domains. The results demonstrate that the analytical quality has been preserved very well.

Table 2. Before/After Bias Assessment

Variables	Q1	Median	Q3
BINGE DRINKING	0.990	0.998	1.022
SMOKING STATUS	0.993	1.004	1.028
STRESS – MILITARY	0.997	0.999	1.003
STRESS – FAMILY	0.992	0.997	1.002
SUICIDE IDEATION	0.998	1.027	1.047
ALCOHOL TREATMENT	0.995	1.025	1.056
DRUG USE – 12 MONTHS	0.992	1.036	1.058
OVERWEIGHT	0.982	1.000	1.034
CONDOM USE	0.988	1.005	1.031
ANXIETY	0.964	1.004	1.022
DEPRESSION	0.973	0.997	1.028
MENTAL HEALTH COUNSELING	0.975	1.003	1.037
DRUG USE – 30 DAYS	0.997	1.030	1.051

Table 3. Before/After Inflation of Standard Error

Variables	Q1	Median	Q3
BINGE DRINKING	0.927	1.007	1.074
SMOKING STATUS	0.994	1.046	1.151
STRESS – MILITARY	0.978	1.007	1.103
STRESS – FAMILY	0.985	1.015	1.058
SUICIDE IDEATION	1.019	1.059	1.107
ALCOHOL TREATMENT	0.976	1.034	1.077
DRUG USE – 12 MONTHS	0.994	1.032	1.074
OVERWEIGHT	1.019	1.062	1.167
CONDOM USE	0.936	1.023	1.089
ANXIETY	0.880	0.977	1.082
DEPRESSION	0.955	1.024	1.098
MENTAL HEALTH COUNSELING	0.978	1.042	1.090
DRUG USE – 30 DAYS	1.021	1.037	1.070

5. CONCLUSIONS

This practice on the 2002 DoD survey data has demonstrated the usefulness of MASSC data disclosure treatment. The results in Table 1 and Table 2 display the great reduction of the data disclosure risk. Also, by working under a probabilistic framework, MASSC treatment introduces uncertainty to all released data so that even those records still at risk for disclosure after treatment may not be the real recodes that intruders are seeking. This benefit of MASSC will definitely deter potential intruders, successfully protecting confidentiality. At the same time, the MASSC procedure preserves analysis utility very well. The comparison of analysis results for before and after treatment data (Tables 2 and 3) provides evidence that the effect of MASSC treatment is very minor in this practice. Therefore, the MASSC procedure simultaneously protects both confidentiality and analytical quality of data.

ACKNOWLEDGEMENTS

The authors would like to thank Vincent G. Iannacchione for presenting this paper at the JSM '04 meeting and for his many useful suggestions. The authors are also grateful to Dr. Laurel Hourani and Dr. Robert M. Bray for their valuable suggestions. We specially thank Dr. Avinash C. Singh. This paper would not be possible without his brilliant ideas on MASSC and his enthusiasm on supporting the data disclosure session at the JSM '04 meeting.

References

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

Folsom Jr., R.E., and Singh, A.C. (2000). "A generalized exponential model for sampling weight calibration for a unified approach to nonresponse, poststratification, and extreme weight adjustments," *Proceedings of the American Statistical Association*, the Session on Survey Research Methods, 598-603.

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