Producing a Public Use File from the 2015 National Survey on Drug Use and Health

Neeraja Sathe*
RTI International, Research Triangle Park, North Carolina, USA – nss@rti.org

Feng Yu
RTI International, Research Triangle Park, North Carolina, USA – fyu@rti.org

Lanting Dai
RTI International, Research Triangle Park, North Carolina, USA – dai@rti.org

Jonaki Bose
Substance Abuse and Mental Health Services Administration, Rockville, Maryland, USA – Jonaki.Bose@samhsa.hhs.gov

Arthur Hughes
Formerly with Substance Abuse and Mental Health Services Administration, Rockville, Maryland, USA – arthugh80@gmail.com

Abstract

A 2015 questionnaire redesign of the National Survey on Drug Use and Health (NSDUH) resulted in the collection of additional private and sensitive information, including rare health outcomes, besides substance use and mental health data. Other redesign changes occurred in 2015, including a complete revision of the prescription drug modules. This redesign made the disclosure limitation process and production of a public use file (PUF) a continued challenge because, similar to the existing measures, these new measures contain not only private, but also identifying information that might be known to both internal and external intruders. Besides using a probabilistic-based disclosure avoidance technique called MASSC (which stands for **M**icro **A**gglomeration, optimal **S**ubstitution, optimal **S**ubsampling, and optimal sampling weight **C**alibration) to "mask" the data, customized procedures have been developed to treat identifying outcomes for additional confidentiality protection. These procedures include variable recoding and local suppression (i.e., set individual values to missing if deemed a risk) to minimize identification of a respondent's sensitive information. In this paper, rare cancer outcomes and prescription drugs are used as examples to illustrate MASSC and post-MASSC treatment procedures for the 2015 NSDUH data. Several quality assessment measures have been used to assess the impact of treatment on such data in addition to the substance use and mental health outcomes, such as comparing estimates and standard errors of the outcomes and comparing multivariate relationships via regression models before and after treatment. Results show that data confidentiality is adequately protected and information loss is relatively low on the PUF, even for rare cancer-related outcomes. This paper is aimed at describing the disclosure avoidance techniques implemented on the 2015 NSDUH and empirically demonstrating that substance use, mental health, and new prescription drug misuse and cancer-related estimates from the PUF are similar to the estimates from the restricted-used file, so that analysts can be confident in using these data from NSDUH's PUFs.

Keywords: NSDUH disclosure limitation; MASSC; cancer data; quality assessment.

1. Introduction

The National Survey on Drug Use and Health (NSDUH) collects data annually on substance use, mental health, and other health outcomes among the U.S. civilian, noninstitutionalized population aged 12 years old or older.[1] These data include personal information of a sensitive nature that respondents may want to keep private, such as substance use behavior, substance use disorders (i.e., abuse and dependence), mental illness, and other health-related outcomes. Disclosure occurs when such information is revealed to the public. More specifically, disclosure occurs when an unauthorized individual (an "intruder") tries to link a record in the microdata file to an identifiable respondent. All NSDUH data are protected under the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA), which ensures that all NSDUH data are used for statistical purposes only and cannot be used for any other purpose (see http://www.eia.gov/cipsea/cipsea.pdf).[2] To protect the respondents' confidentiality, comply with federal regulations, and honor the confidentiality pledge, statistical disclosure treatment has been imposed on all NSDUH public use files (PUFs) to minimize disclosure risk. Also, CBHSQ within SAMHSA is a statistical unit approved by the Office of Management and Budget and has the responsibility to protect confidential data that are freely available to the public from disclosure (i.e., the identification of respondents and their responses).

The 2015 questionnaire underwent a partial redesign aimed at improving data quality and addressing the changing needs of policymakers and researchers regarding substance use and mental health issues. Major changes to the instrument were in the prescription drug questions where the use and misuse of specific prescription drugs (e.g., pain relievers such as oxycodone, fentanyl, and buprenorphine) were added, and the focus shifted from the lifetime period to the past 12 months. Also, the health module was revamped to feature new rare health conditions, such as cancer outcomes (e.g., esophagus cancer, kidney cancer), and an age at first diagnosis. Demographic question changes focused on educational attainment and new questions on sexual identity and attraction. An entire module on marijuana purchases was dropped to keep the interview length to about 1 hour.[3] The new data resulting from the 2015 questionnaire redesign means that CBHSQ must continue to ensure that data included in the PUFs are acceptably secure from disclosure. Collecting new information on rare cancers and the age of diagnosis necessitated additional data treatment. In this paper, we use the 2015 experience to discuss the disclosure risk via intrusion scenarios, the statistical disclosure limitation technique known as MASSC (for details, see Section 3) that is used to treat NSDUH data, the customized treatment for the 2015 NSDUH's new health outcomes and prescription medicines, and the pretreatment and posttreatment data quality assessments that were implemented when the 2015 PUF was produced. In a previous study, we investigated the impact of MASSC and related disclosure limitation procedures on the data quality of the 2002 to 2013 PUFs.[4] In this study, we further empirically test and verify that the 2015 PUF continues to provide high data quality.

2. Disclosure Scenarios: Inside Intrusion versus Outside Intrusion

The variables most likely to be used to identify a given respondent's record are called identifying variables (IVs) and are usually known to others (e.g., age, gender, and race). Typically, intrusion can be either inside intrusion or outside intrusion. Inside intrusion occurs when the intruder knows that a specific respondent participated in the survey and tries to discover sensitive information by using identifying information that is known to him or her and is included in the data file. This is of major

[1] For detailed information on NSDUH and its history, see https://www.samhsa.gov/data/. The survey is sponsored by the Substance Abuse and Mental Health Services Administration (SAMHSA), U.S. Department of Health and Human Services, and is planned and managed by SAMHSA's Center for Behavioral Health Statistics and Quality (CBHSQ).

[2] NSDUH's data collection and analysis are conducted under contract with RTI International (a registered trademark and a trade name of Research Triangle Institute). During data collection, CIPSEA language is included in the lead letter and the informed consent materials that are sent to respondents to assure them that the confidentiality of the answers they provide to the questions will be fully protected under federal law by CIPSEA.

[3] For details on NSDUH's redesign, see the following reports at https://www.samhsa.gov/data/: ***(1)*** Section C of the *2015 National Survey on Drug Use and Health: Methodological Summary and Definitions*; ***(2)*** *National Survey on Drug Use and Health: 2014 and 2015 Redesign Changes*; and ***(3)*** *2015 National Survey on Drug Use and Health: Summary of the Effects of the 2015 NSDUH Questionnaire Redesign: Implications for Data Users*.

[4] See the report on the *National Survey on Drug Use and Health: Quality Assessment of the 2002 to 2013 NSDUH Public Use Files* at https://www.samhsa.gov/data/.

concern in NSDUH because a family member may know the presence of the other member in the sample. The inside intruder's chance of success is increased if the target record is a sample "unique" (i.e., a single case in a cell defined by a set of IVs). For example, a father knows that his son is in the survey and, by using a set of IVs, finds that only one record is in the file whose profile (combination of IVs) matches his son's demographic characteristics; he becomes certain that this record is his son's. Once the father has found his son's record in the dataset, he then knows all the answers provided by his son to these sensitive questions (e.g., illicit drug use).[5]

Outside intrusion occurs when an intruder does not know the presence of his or her target respondent in the sample and tries to identify a record by matching it to an external data source. Identification may be done by matching the IVs of respondents that are also in another external dataset. An outside intruder typically targets a respondent whose profile in the population is unique (or rare). Although an outside intruder does not have the information about a target's record being in the data, he or she is usually more sophisticated than an inside intruder and may be equipped with massive external data and matching software that could lead to a name or address. An outside intruder is usually not targeting a specific individual, but rather is targeting any subject he or she can identify, thus potentially putting all survey respondents at risk of being disclosed. Both inside and outside intrusion can lead to casting doubt on the confidentiality protection offered to survey respondents, thus potentially affecting response rates for future data collection.

3. NSDUH Disclosure Control Procedure: MASSC

MASSC is the abbreviated name of a statistical disclosure limitation method developed at RTI that can be used to treat microdata files for NSDUH confidentiality protection and data dissemination to the public. MASSC uses four steps: (1) **M**icro **A**gglomeration, (2) optimal probabilistic **S**ubstitution, (3) optimal probabilistic **S**ubsampling, and (4) optimal sampling weight **C**alibration (Singh, 2002; Singh, Yu, & Dunteman, 2003; Singh, Yu, & Wilson, 2004).

Similar to other years, the 2015 NSDUH PUF was created from the 2015 NSDUH restricted-use file (RUF). Directly identifying information, such as name, phone number, and address, was not included either in the RUF or PUF, so these identifiers can never be linked with the responses. As part of the PUF creation, almost all geographic identifiers (including census region, state, and county) were removed.[6] Moreover, the household link between respondents from the same household was deliberately excluded from the PUF to reduce the inside intrusion risk because more than one person in the household could have been selected to participate in the survey and been administered the questionnaire via the computer-assisted interviewing method. All of the variables on the file were reviewed for the possibility of identifying the respondent by combining a number of IVs at one time. Variables considered to have a high potential of personal identification, as well as a high value for analysis (i.e., could not be dropped from the PUF), were treated by standard procedures of categorization and top-and-bottom coding.

To apply the MASSC technique using selected key IVs, the data were first partitioned into risk strata in the *micro agglomeration* step, where records were grouped according to their disclosure risk status.[7] This step's purpose is to control for the level of treatment in subsequent MASSC steps so that more rigorous treatment can be applied to higher risk strata and less treatment to the lower risk strata. A sample of records was then randomly drawn from each stratum, and variables were substituted from a similar donor record. This *substitution* step introduces uncertainty about a record's identity and makes it difficult for an intruder to be certain that any record corresponds to a specific individual because some of the variables used to identify the record may have come from other individuals. Next, some records were randomly removed from the file to reduce the probability of determining that any known

[5] NSDUH can collect information from more than one respondent from a household (up to two per household), thereby increasing the risk of identification of a pair of related respondents from the same household.

[6] The only geographic variables included on the PUF are ***(1)*** two variables that identify the population density and the type of metropolitan area of residence (large metro, small metro, and nonmetro), and ***(2)*** an indicator variable that identifies whether the respondent's residence is located on an American Indian tribal land or not.

[7] The disclosure risk status of a respondent was defined by a set of IVs such that uniques (i.e., respondents who can be uniquely identified by the set of IVs) were assigned to higher risk strata and nonuniques (i.e., two or more respondents in a cell defined by the set of IVs, such as doubles and triples) were assigned to lesser risk strata.

respondent was in the PUF. Approximately 20% of the respondents are sampled out each year to maintain acceptable variance inflation while reducing disclosure risk. This *subsampling* step introduces further uncertainty about the presence of a target record in the database. These two steps in combination substantially reduce the risk of someone being identified or targeted. Substitution and subsampling were done while simultaneously constraining adjustments were made to the resulting file to a minimal increase in bias and a minimal decrease in precision for numerous estimates of substance use prevalence across a number of domains. The variables used to form risk strata or substituted in NSDUH are confidential, so they cannot be discussed here. Finally, the weights on the treated data file were recalibrated to known totals from the full RUF so that the decrease in precision due to subsampling was minimized. Note that MASSC procedures were applied to optimize the national estimates. Even though state-level estimation is possible on the RUF, it is not possible on NSDUH's PUFs because state identifiers were removed.

4. Customized Treatment for 2015's New Variables

Disclosure control procedures for NSDUH data further include treating the design variables in all years. The stratum and replicate identifiers used for variance estimation in data analyses were treated by coarsening or collapsing,[8] substitution, and scrambling or random reordering. Also, certain variables were recoded (e.g., by collapsing rare levels of the cigarette brand used most often in the past month) or locally suppressed (e.g., by setting certain cases of the body mass index variable to missing values) for confidentiality reasons. For the cancer variables, all age at first diagnosis variables were dropped from the 2015 PUF. Specific cancers that had very low frequency counts were collapsed with other cancer variables in such a way that retained to the extent possible their logical and analytic value (e.g., "rectum cancer" was collapsed with "colon cancer" to make it "ever had rectum or colon cancer"; also, "prostate cancer" was collapsed with "testis cancer" to make it "ever had prostate or testis cancer" for males). More commonly occurring health conditions (e.g., high blood pressure, asthma) have relatively higher prevalence rates, so no global collapsing between variables was needed. All of the individual prescription drugs that respondents indicated using or misusing (e.g., specific pain reliever brands such as Vicodin® or Norco®, stimulant brands such as Adderall® or Ritalin®) were dropped from the PUF, with the exception of the OxyContin® variables, which were retained due to their high analytic value. Most subtypes[9] of related prescription drugs that respondents used or misused in the past 12 months were retained on the PUF. All collapsing and local suppression were done in addition to the MASSC treatment to ensure that no respondents could be identified with certainty.

5. Quality Control and Assessment for the 2015 NSDUH's PUF

A trade-off is always present when weighing the advantages and disadvantages of disclosure risk and information loss: When disclosure risk decreases, information loss increases. MASSC simultaneously controls and optimizes disclosure risk and information loss in NSDUH. To assess the quality of NSDUH's PUFs, we compared a set of 2015 PUF and RUF estimates on a broad range of outcome and domain combinations. We computed ratios of both point estimates and their standard errors (SEs) before and after disclosure treatment. Similarly, ratios of contrast estimates (e.g., linear combinations of variables whose coefficients add up to zero, which allow for comparisons between two or more domains for a given outcome) were calculated before and after treatment. For multivariate relationship comparisons between a response (i.e., dependent) variable and two or more predictor (i.e., independent) variables (e.g., the relationship between past month tobacco use and age, gender, and race), we developed regression models to compare model fitting via examining the ratios of the regression coefficients before and after treatment. For the contrasts and regression coefficients, the changes in significance (i.e., whether the significance for a particular contrast or covariate changed from being significant to nonsignificant or vice versa) were also compared.

[8] As in 2014, this treatment in 2015 reduced the number of degrees of freedom (*df*) available in hypothesis testing from 750 to 50. However, the reduction was considered to have a minimal effect as the detection of statistical significance at the 5% level (two-sided) based on a critical value of the *t*-distribution changed from 1.96 based on 750 *df* to just 2.01 based on 50 *df*.

[9] Four types of prescription drugs (each with subtypes defined according to common active ingredients) are on the 2015 PUF: ***(1)*** pain relievers (e.g., hydrocodone and oxycodone products), ***(2)*** tranquilizers (e.g., clonazepam and diazepam products), ***(3)*** sedatives (e.g., zolpidem products and barbiturates), and ***(4)*** stimulants (e.g., amphetamine and methylphenidate products).

6. Summary of the 2015 NSDUH's PUF and Data Quality

The 2015 PUF has a total of 2,666 variables and 57,146 respondents after MASSC and subsequent treatment as compared with the RUF file, which has 3,926 variables and 68,073 respondents. To study the impact of disclosure avoidance treatment on the bias and precision of the PUF estimates compared with those from the full file (i.e., the RUF), we reviewed different sets of variables in the evaluation. Substance use variables, substance abuse and dependence variables, adult mental health variables, health condition variables, and prescription drug use and misuse variables for a wide range of domains (such as age, gender, race, marital status, education, and income) were assessed for posttreatment quality. In the multivariate relationship comparisons, different models were fitted for the health outcomes, especially for the rare cancer variables corresponding to low prevalence estimates.

The PUF quality evaluation was focused on the ratios of point estimates[10] or regression coefficients from the full RUF and the PUF. Ratios closer to 1.00 meant that the PUF estimates were closer to the RUF estimates, which indicated less bias. Average ratios or median ratios (across *N* numbers) were calculated for point estimates, contrasts, and regression coefficients from different sets of outcome and domain combinations. These results are summarized in Table 1. Data show that the average ratios over different sets of outcomes were within the 0.98 to 1.01 range, with the exception of the average ratio (0.91) for the special cancer variables, which tended to have low prevalences. For the contrasts and regression coefficients, the medians of the before and after treatment estimate ratios were within the 0.95 to 1.04 range. For all of the SE comparisons, the decreases in precision were no more than 10% on average. Because the 2015 PUF sample size was about 16% smaller than the 2015 RUF sample size, we expected to see an average 9% increase in the SEs of the point estimates.

Table 1
Average Ratios of Before and After Treatment for Estimates, Contrasts, and Regression Coefficients

	Estimates			Contrasts			Regression Coefficients		
Outcome	***N***	**Ratio of Estimates (Mean)**	**Ratio of SE (Mean)**	***N***	**Ratio of Estimates (Median)**	**Ratio of SE (Mean)**	***N***	**Ratio of Estimates (Median)**	**Ratio of SE (Mean)**
Substance Use	340	1.00	1.09	190	1.00	1.06	170	0.99	1.09
Substance Abuse/ Dependence	408	1.00	1.10	228	1.02	1.09	204	0.98	1.10
Adult Mental Health	363	1.00	1.09	176	0.95	1.07	176	0.98	1.09
Chronic Health Condition	625	0.98	1.06	345	1.01	1.05	285	1.01	1.10
Cancer	698	0.91	1.00	404	0.98	1.00	135	1.01	1.09
Prescription Drugs									
Past Year Use	1,519	1.01	1.09	855	1.01	1.08	270	1.01	1.08
Past Year Misuse	1,253	1.01	1.09	722	1.04	1.07	228	1.02	1.08

N = number of estimates.

Note that, in certain cases, extreme ratios were observed from outcomes with low prevalence rates or small domains (data not shown), especially from special cancer variables. For example, the ratio of the before and after treatment of estimates for esophagus or stomach cancer among respondents whose family income was $10,000 or less was 2.87. The small denominator caused the ratio to be large, where the point estimates (prevalence rate of esophagus or stomach cancer in that domain) were 0.1% and 0.2% from the RUF data and the reduced PUF data, respectively. This estimate would be suppressed if the NSDUH precision-based suppression rules[11] were implemented on the RUF. Therefore, users of PUF estimates are cautioned when analyzing and interpreting near zero or low prevalence rates and estimates based on small sample sizes and are encouraged to use suppression rules to eliminate estimates determined to have low precision. In those situations, combined year data

[10] Ratios are calculated as estimates after disclosure avoidance treatment (namely, substitution, subsampling, or suppression) to estimates from the RUF (estimate can refer to a prevalence rate, SE, or estimate of a contrast). Means or medians of these ratios are shown in Table 1.

[11] For a discussion of NSDUH's criteria for suppressing (i.e., not publishing) unreliable estimates, see Section B.2.2 in Section B of the *2015 National Survey on Drug Use and Health: Methodological Summary and Definitions* at https://www.samhsa.gov/data/.

analyses should always be considered to increase precision and reduce suppression (as is done in many NSDUH publications).

To test the impact of disclosure treatment on the statistical inference of the treated database, we also performed *t* tests to discover the change of significance at the 5% level for both the contrasts and regression coefficients. Table 2 shows that changes in significance for the contrasts and regression coefficients were observed either from being significant in the full RUF data to nonsignificant in the PUF data or vice versa, but all of these changes occurred in less than 10% of the cases.

Table 2
Change of Significance at the 5% Level for Contrasts and Regression Coefficients Analysis

		Change of Significance (Contrasts)					**Change of Significance (Regression Coefficients)**			
Outcome	***N***	**Sig to NS**	**NS to Sig**	**Total Changed**	**% Changed**	***N***	**Sig to NS**	**NS to Sig**	**Total Changed**	**% Changed**
Substance Use	190	2	3	5	2.63	170	9	0	9	5.29
Substance Abuse/ Dependence	228	6	4	10	4.39	204	11	2	13	6.37
Adult Mental Health	176	9	4	13	7.39	176	6	1	7	3.98
Chronic Health Condition	345	16	8	24	6.96	285	8	14	22	7.72
Cancer	404	19	16	35	8.66	135	8	4	12	8.89
Prescription Drugs										
Past Year Use	855	32	19	51	5.96	270	8	7	15	5.56
Past Year Misuse	722	21	20	41	5.68	228	10	5	15	6.58

Sig to NS = significant to nonsignificant changes; NS to Sig = nonsignificant to significant changes.

7. Conclusions

Statistical disclosure limitation methods were implemented in the 2015 NSDUH in such a way that the PUF continues to be a representative sample of the civilian, noninstitutionalized population in the United States and reflects the actual data collected from the survey, in spite of being treated for disclosure avoidance. Results show that data confidentiality was adequately protected and information loss was relatively low on the 2015 PUF, even for rare cancer-related outcomes. This paper demonstrates that the disclosure avoidance techniques implemented on the 2015 NSDUH do not hamper data quality significantly and empirically proves that substance use, mental health, prescription drug, and cancer-related estimates from the PUF are similar to the estimates from the RUF, so that analysts can be confident in using data from NSDUH's PUFs.

References

Singh, A. C. (2002). *Method for statistical disclosure limitation* (U.S. Patent Application Pub. No. US 2004/0049517A1). Research Triangle Park, NC: RTI International. The patent was granted in June 2006 (Patent No. US7058638B2).

Singh, A. C., Yu, F., & Dunteman, G. H. (2003). MASSC: A new data mask for limiting statistical information loss and disclosure. *Proc. Joint UNECE/EUROSTAT Work Session on Statistical Data Confidentiality, Luxembourg,* Working Paper No. 23, pp. 373–394. Geneva, Switzerland: United Nations Statistical Commission and Economic Commission for Europe Conference of European Statisticians, European Commission Statistical Office of the European Communities (EUROSTAT).

Singh, A., Yu, F., & Wilson, D. H. (2004). Measures of information loss and disclosure risk under MASSC treatment of micro-data for statistical disclosure limitation. *Proc. 2004 Joint Statistical Meetings, American Statistical Association, Section on Survey Research Methods* (pp. 4374–4381). Alexandria, VA: American Statistical Association.