Addressing Disclosure Concerns and Analysis Demands in a Real-Time Online Analytic System

Tom Krenzke¹, Jane F. Gentleman², Jianzhu Li¹, and Chris Moriarity²

This article focuses on methods for enhancing access to survey data produced by government agencies. In particular, the National Center for Health Statistics (NCHS) is developing methods that could be used in an interactive, integrated, real-time online analytic system (OAS) to facilitate analysis by the public of both restricted and public use survey data. Data from NCHS' National Health Interview Survey (NHIS) are being used to investigate, develop, and evaluate such methods. We assume the existence of public use microdata files, as is the case for the NHIS, so disclosure avoidance methods for such an OAS must account for that critical constraint. Of special interest is the analysis of state-level data because health care is largely administered at the state level in the U.S., and state identifiers are not on the NHIS public use files. This article describes our investigations of various possible choices of methods for statistical disclosure control and the challenges of providing such protection in a real-time OAS that uses restricted data. Full details about the specific disclosure control methods used by a working OAS could never be publicly released for confidentiality reasons. NCHS is still evaluating whether to implement an OAS that uses NHIS restricted data, and this article provides a snapshot of a research and developmental project in progress.

Key words: Confidentiality; data utility; disclosure risk; statistical disclosure control.

1. Introduction

To meet data user needs, a battery of systems for facilitating access to data need to be provided. The battery of systems that a government survey organization offers to facilitate data access must be integrated, not just to maximize usefulness and convenience to users,

¹ Westat, 1600 Research Boulevard, Rockville, MD 20850, U.S.A. Emails: tomkrenzke@westat.com; janeli@westat.com

² National Center for Health Statistics, 3311 Toledo Road, Hyattsville, MD 20782, U.S.A. Emails: jgentleman@cdc.gov; cmoriarity@cdc.gov

Acknowledgments: The authors of this article are employees of NCHS and of Westat, NCHS' contractor for investigating, developing, and evaluating disclosure avoidance methods for possible use in a future online analytic system for NCHS survey data. The UCLA Center for Health Policy Research is also serving as a subcontractor. The findings and conclusions in this article are those of the authors and do not necessarily represent the views of the National Center for Health Statistics, Centers for Disease Control and Prevention. The authors gratefully acknowledge the many individuals who are contributing to the project. We thank NCHS staff, including Brenda Baker, our original contract Project Officer, for her attention to this project; Susan Jack, who provided valuable assistance on the extensive preparations of variables in the underlying microdata test files; and Jacqueline Lucas, who assisted in providing the necessary microdata test files. At Westat, we are grateful to Michele Harmon, our Project Director, for her oversight on this project, and to Craig Ray for his leadership and experience relating to system development. At UCLA, special thanks to AJ Scheitler and Alberta Soranzo for sharing their experience and providing their input on the interface design. We are also indebted to the Integrated Health Interview Series staff at the University of Minnesota for providing advice in variable selection and assistance with classifications for variables used in NHIS test files. Lastly, the authors are grateful for the special methods review conducted by Natalie Shlomo.

but also for disclosure avoidance. Maintaining confidentiality (which is mandated by law and which is critical to continuing survey respondent cooperation) is becoming increasingly complex as data user sophistication and access to technology increase. This article describes how the National Center for Health Statistics (NCHS), the official health statistics agency of the United States federal government, is developing methods that would be required to offer to the public an interactive, integrated, real-time online analytic system (OAS) to facilitate analysis of both restricted and public use survey data. Data from NCHS' National Health Interview Survey (NHIS) are being used to investigate, develop, and evaluate such methods. Detailed information about the NHIS is available at: <http://www.cdc.gov/nchs/nhis.htm>.

An OAS must be integrated with existing means of data access and must also within itself deal with the tradeoffs between providing access to data and observing the agency's own rules and laws for protecting confidentiality. The NHIS website releases annual public use NHIS microdata files as well as documentation and numerous analytic reports (NHIS 2010, 2011). Like all of the surveys conducted by the NCHS, the NHIS has a mandate to protect the confidentiality of survey participants. Personally identifiable information (PII), such as persons' names, is not released. Certain "restricted variables," including U.S. state identifiers, lower-level geographical identifiers, and sensitive variables, are not available in public use files (PUFs) because the combination of low-level geographic data and demographic data on an individual microdata record could lead to a breach of confidentiality. In general, this means that detailed versions of indirect identifying variables are not released to the public. By "indirect identifying variables," we refer to variables that are factual in nature, such as age, gender, and geographical area, and that can reveal the identity of an individual when used in combination.

The NCHS alleviates limitations on public access to its restricted data by providing access to NCHS restricted data through its Research Data Center (RDC) (NCHS RDC 2011; Meyer, Robinson and Madans 2012; <http://www.cdc.gov/rdc/>).

The NHIS was designed mainly to produce national estimates. However, the NHIS collects data in all 50 states and the District of Columbia, and the NHIS sample is designed so that NHIS data are representative of the target population (the civilian noninstitutionalized population) at the state level as well as at the national level. As discussed in Gentleman (2012), analysis of state-level data is appropriate and desirable for studying the U.S. health care system because the U.S. health care system is largely administered at the state level. Nevertheless, appreciable numbers of state-level NHIS sample sizes are normally too small for analysts to produce stable state-level estimates from a single year of data. Sometimes, satisfactorily precise state-level estimates can be produced by combining two or more adjacent years of data. However, for confidentiality reasons, NCHS does not include any state identifiers on the NHIS PUFs. Thus, when NHIS state sample sizes are insufficient for a particular state-level analysis, an NHIS data user from outside NCHS who wishes to perform that analysis must currently utilize the NCHS RDC.

In developing a new policy to provide access to restricted microdata, an agency must identify and consider what types of data access already exist (such as through an RDC) so that new uncontrolled or inadvertent risks to confidentiality are not created. In particular, the disclosure avoidance methods needed for an OAS are highly dependent on whether or

not PUFs are also produced, because users of an OAS can compare its results with PUFs and analyses thereof.

The NHIS has been in the field since 1957. Since 1997, the Family Component of the NHIS has collected data about persons of all ages in each family within each interviewed household, and additional data have been collected about one randomly selected child (the “sample child”) and from one randomly selected adult (the “sample adult”) in each family. The existence of a PUF (and any other publicly available linkable file) makes the creation of an OAS for the NHIS particularly challenging, as discussed in this article.

The OAS under consideration for use with the NHIS would use NHIS data starting with the 1997 data year. It would integrate two real-time OAS subsystems in a manner that would be transparent to the user: 1) “Subsystem P,” which uses only public use data; and 2) “Subsystem R,” which uses its own underlying microdata files containing selected restricted data (e.g., state) as well as public use data. All developmental work assumes this system design. In addition, OAS analyses would be performed in real time with no pre-prepared results.

NCHS is still evaluating whether to implement an OAS that uses restricted NHIS data, and this article provides a snapshot of where the OAS research and development project is at the time of writing. Therefore, it is important to note that all aspects of the system discussed in this article are under evaluation, including, as examples, a utility to allow the user to combine categories of variables, the parameters for thresholds on unweighted counts to determine whether or not to deny results from being shown to the user, and the Statistical Disclosure Control (SDC) treatments.

The OAS approaches discussed here are generic for agencies/organizations that conduct household sample surveys, provide PUFs, and have confidentiality constraints. We do not attempt to explore disclosure risks related to administrative records data or to data with highly skewed distributions, such as data from establishment-based surveys.

In Section 2, we discuss OAS features being considered, including general data management for data to be accessed in the integrated system, analysis methods, and transmission and storage of the results of analyses. Section 3 presents a discussion of challenges and trade-offs that impact design decisions. Section 4 describes the main disclosure threats to such a system and investigations of how to protect against such risks. Section 5 presents various statistical disclosure control (SDC) treatments under consideration. An initial evaluation of some SDC approaches is discussed in Section 6 in terms of their impact on disclosure risk and analytic utility. Lastly, concluding remarks are given in Section 7.

2. General OAS Features

The general system architecture for an OAS such as the one under consideration for use with the NHIS would include the distinct system components that are described below and depicted in [Figure 1](#).

User interface. For greatest efficiency and cost effectiveness, it is necessary to include the interface design as an integral part of the system development. For the OAS under development, the user interface is implemented through a web-based, statistical reporting tool, for which the challenge is to arrive at a single interface that provides a

Fig. 1. General System Architecture from Query to Data Access

user-friendly environment that appeals to a wide audience. The interface would accommodate sophisticated queries as well as guide novice users effectively through the system to generate results that meet their needs.

Metadata database. A metadata database is needed to provide the capability to easily define datasets and to group variables into logical “folders,” to enable selected options for statistics and missing values, and to define the contexts in which each variable can be used. A metadata database helps to determine which full-sample weight is needed to request estimates from different sub-questionnaires, to provide information about rounding rules for each variable, and to indicate the data years for which each variable is available. The metadata database is impacted by disclosure risk concerns. For example, it will contain information to help integrate the two subsystems. If a restricted variable (e.g., state) is specified in the query, then Subsystem R, its SDC treatments, and the restricted data would be used. The database would contain information on the subsetting population filter variables for Subsystem R. If a restricted variable is not being used, then Subsystem P, without restriction, would be used. System usage should be tracked within the database, including information about queries to inform system managers about what variables are most frequently requested. If the OAS database were, for example, implemented using a SQL Server, the website would communicate with the database via a metadata and usage class library.

Administrative tools. The system needs to include a suite of parameter-driven administrative tools that allow for rapid configuration of projects using the same interface and analytic engine. A parameter-driven system provides the flexibility to put in values that define certain rules. For example, rounding rules can be turned off, or the amount of rounding can be changed. Under evaluation is an administrative infrastructure to maintain user profiles and project-specific configurations, and to migrate metadata from a staging area to the production environment.

Survey microdata file. The integration of data sources would allow queries to be requested using variables from multiple sources. It may be beneficial to combine data into one large file with multiple sets of weights and sets of survey variables, and include several survey years. The survey microdata files are impacted by disclosure risk since the restricted data files may have been treated by random perturbation, while the corresponding public use files may not have been perturbed for the same variables.

Types of analysis and analytic engine. The types of analyses under consideration for the OAS include descriptive statistics (proportions, means, medians and other quantiles, etc.), multi-way cross-tabulations, ratios, standardized rates, standard errors, confidence intervals, and graphs. The graphing utility will only display estimates that have passed disclosure-related threshold rules set up for the tables (e.g., as discussed in the introduction in Section 5). More advanced statistical options could include choice of confidence interval width and type and provide the capability to do significance tests to compare pairs of cells in a cross-tabulation. Any changes to the types of analyses, such as a new approach to estimating confidence intervals or adding new modeling capabilities, might require enhancements to the analytic engine. With interface panels devoted to each components of the query (e.g., selecting geography, variables for tables, statistical options), a natural progression would be to add a new panel to the interface to devote to regression modeling capabilities. In addition, the methodologies related to reducing disclosure risk are relatively fluid and continue to develop with new approaches. For example, for regression analysis, [Reiter \(2003\)](#) provides an approach to perturbing residuals prior to displaying a diagnostic plot. [Chipperfield and Yu \(2011\)](#) provide a discussion of disclosure rules applied to regression analysis in an OAS. Once a system is launched, if an SDC treatment is enhanced or changed, it could impact on the consistency of results from prior runs. Therefore, a balance between reducing disclosure risk further and the impact on data utility would need to be considered.

A benefit is that since the analytic engine is a portable utility (i.e., un-plug and plug-in capabilities), it would allow for any methodological (statistical or disclosure-related) changes and additions to be developed in-house and uploaded without disruption to users. The analytic engine is optimized for speed of computation, and takes into account the complex sample design when estimating sampling error. A replication method to estimate variances from sample surveys with complex sample designs is being evaluated. Replicates are random subsamples of the full sample, and replicate weights are created for each replicate to account for the sample cases that were dropped from the full sample to form the subsample.

The existence of disclosure risk makes an impact on how the components of the OAS architecture are constructed. We discuss further the disclosure threats with respect to challenges faced during development in the next section, and we discuss the specific disclosure risks, various limitations and SDC treatments later in the article.

3. Beginning the Work: Challenges and Trade-offs Relating to the System Architecture

We describe here some challenges and trade-offs encountered and addressed during the early phases of development of the OAS that may give others a head start in the development of their own systems. The challenges for the architectural structure are classified into three categories: 1) balancing risk, utility, and user friendliness; 2) system issues; and 3) maximizing statistical methodology. We start by describing some decisions that had to be made in balancing risk, utility, and user friendliness.

Decision about providing ways to combine multiple years of data. The utility of Subsystem R can be increased through limited efforts of constructing consistent recodes (new variables created from existing variables), values, and value labels across data years.

Decision about what age groupings to permit. Data analysts would ideally be free to choose whatever age groupings suit their specific subject matter, but the desired age groupings, especially for health data, often are not common (e.g., five-year) age groupings. An intruding user, if allowed to choose any age groupings, could deliberately select age groups that differ by, for example, one year, which would make it possible for a differencing/linking attack to occur. The topic of differencing tables is summarized in [Duncan et al. \(2011\)](#), and an example in the context of official statistics can be found in [ONS \(2006\)](#). Striving for flexibility in defining age ranges to meet the needs of the data user led to the initial development of a dynamic utility for combining (but not disaggregating) variable categories within the OAS.

Decision about the extent to which users should be protected from denials. Various threshold rules were decided upon to deny tables of high risk. Threshold rules are discussed further in Section 5. In an ideal OAS using NHIS data, analysts would be free to request, for example, a state-level cross-tabulation involving any state. But requests for analyses of states with small populations would very likely have to be denied by the system because of unacceptably high risks to confidentiality.

Decision about the number of variables to allow in a given query. Initial investigations show that limiting the number of variables used for subsetting the data so as to define the subpopulation or creating universe of interest helps to reduce disclosure risk. Therefore, decisions were made as to the maximum number of variables to allow (i.e., about a dozen variables) and which ones were analytically important to keep for this purpose.

Decision about how much to tell users about the confidentiality protection methods used. Analysts of perturbed data need to be warned that their analyses will have reduced accuracy and increased inconsistencies due to the application of confidentiality protection methods. A summary of approaches to perturb microdata is given in [Duncan et al. \(2011\)](#).

The following decisions address challenges relating to the subsystems.

Decision about integrating Subsystems R and P. It was decided to integrate Subsystem R and Subsystem P because many users would not be able to or want to specify to which subsystem they want to submit their queries ahead of time. This led to the development of a smart system to determine which underlying microdata file to use (PUF or restricted), depending upon the variables in a specific query. The challenge arises in the handling of changes to the query as the user provides a developing series of specifications for the desired analysis.

Decision about the structure of the underlying files. We were faced with several options relating to file structures underlying the system. The decisions took into account the size of the file, processing of integrated public and restricted data (perturbed), and the various levels of the NHIS data (i.e., file types). Because a query that uses restricted data must use perturbed data, it was decided to make a set of the files containing just the public use data and a set containing perturbed restricted use data. This simplified the system processing at the expense of higher storage costs.

The following decisions address challenges while attempting to maximize the statistical methodologies available.

Decision about dynamic combining of categories. Without the ability to combine categories, the user is constrained by the categorized versions of the variables in the underlying dataset, with no ability to modify the variables. However, the combining of

variable categories increases disclosure risk, which must be attended to in developing the system. The associated increase of disclosure risk due to dynamic combining of categories is touched upon in Section 6.1.1. In addition, the decision to offer dynamic collapsing of categories has several implications for creating estimates for a variety of statistics.

Decision pitting confidentiality against consistency. In an OAS using NHIS data, analysts using Subsystem R to obtain state-level estimates will likely want to compare those state estimates to national estimates. Using Subsystem R's underlying microdata file to compute both state-level and national estimates would provide consistency. For smaller domains, intense exploration by an intruder of the differences between estimates produced from Subsystem R data files and the same estimates produced from PUFs could result in a threat to confidentiality. To avoid such intrusion, the current approach is for Subsystem R to use PUFs to calculate any requested estimate that does not involve restricted variables (e.g., a national estimate). Under that plan, some inconsistencies caused by differences between Subsystem R's underlying data file and the PUFs will need to be tolerated.

Decision about rounding rules. Rounding reduces the amount of detailed and precise information. As an example, the NCES Statistical Standards Section 5.3 ([NCES 2003](#)) addresses rounding rules that should be applied to numbers and percentages in text and summary tables.

Decision about dynamic degrees of freedom. Degrees of freedom are used in calculating confidence intervals and conducting significance tests. They are an indication of the stability of the estimated sampling variance. In the case of an OAS using NHIS data with a focus on providing state-level estimates, and because some states have a small number of primary sampling units, the associated degrees of freedom could vary widely. The dynamic approximation of the degrees of freedom in the OAS depends on the subgroup, number of active replicates, the type of replication approach (which may require the number of unique strata in the subgroup), and the type of statistic.

4. Challenges Relating to Disclosure Risk in an OAS

A major task is controlling risks involved within the data that can be exposed in a real-time OAS. The main disclosure risk elements in an online table generator are 1) cross-tabulations with small unweighted cell sizes, especially in the margins of the tables, 2) table linking, 3) weights that vary and variables with a large number of categories, and 4) table differencing. We expand on each of these risks.

There are different ways to define small cell sizes. By "cell size," we refer to the underlying unweighted number of respondents from which the weighted cell frequency was calculated. Sometimes the "Rule of Three" is used based on the concept of k-anonymity ([Sweeney 2002](#)), which specifies that at least three individuals must be present (a count of at least three). The Rule of Three is based on the premise that a singleton (a table cell that is based on one sample case) can possibly be identified by another individual, while a cell count of two is susceptible to disclosure if a sampled individual finds himself in the cell and deduces the characteristics about the other individual in the cell. Singletons and doubletons (table cells that are based on one and two sample cases, respectively) are typically high risk. A singleton is sometimes referred to as a "sample unique." Sample uniques are not necessarily unique in the population. Ways to

find risky data values by estimating the probability of being a population unique are discussed throughout the literature and summarized in [Hundepool et al. \(2012\)](#).

Table linking is an attacking technique used to string together information, called the “matching key,” from several tables in order to form a microdata record on a singleton. By “matching key,” we refer to a combination of variables from the restricted file that, when used together, can be matched to a publicly available external file in order to merge restricted variables (e.g., detail on geographic location) to the public file. This type of table linking in the context of producing tables (albeit static, as opposed to real-time in this article) when a PUF exists is described in the context of the Census Transportation Planning Products (CTPP) on five-year American Community Survey (ACS) data ([Krenzke et al. 2011b](#)) and the existence of the ACS public-use microdata sample (PUMS). For the OAS, when a singleton exists in the marginal of variable S_0 (e.g., a questionnaire or registry variable such as gender) in cross-tabulations with other variables, and S_0 appears in several explicit table queries, a microdata record can be created for the singleton. Suppose a series of two-way tables (where a “*” means “crossed with”) using other questionnaire or registry variables is requested, such as S_0*S_1 , S_0*S_2 , . . . , S_0*S_3 , and suppose that a category in S_0 is comprised of a singleton. The tables can be linked together to define a microdata record for the single sample unit consisting of values of S_0 , S_1 , S_2 , S_3 , . . . , S_k , for k variables crossed pairwise with variable S_0 . Suppose the series of tables is generated for a subgroup $R=1$ defined by a restricted variable (e.g., State=Maryland). This may allow restricted information, such as detailed geographic information, to be assigned to a public-use record on an external file.

[Fienberg \(2009\)](#) summarizes various disclosure risks attributable to weights. Because sampling weights almost always vary, tables with many cells that are generated from a small universe usually would show unique weighted totals for selected cells that would show up for other tables requested for the same universe, leaving the system susceptible to table linking. By “universe” we refer to a key domain of interest to make inference about, which could be a group or subgroup of persons specified by the user, and based on one or more variables. Suppose S_0 has two levels (such as gender), and there are three records with $S_0=1$. The following series of tables ([Table 1](#)) shows the sum of weights for variable S_0 being crossed pairwise with S_1 (six levels), then with S_2 (eight levels), and then with S_3 (four levels), for the restricted subgroup $R=1$. Other variables can be crossed pairwise in the same manner with S_0 but are not shown here. Moreover, only the row for $S_0=1$ is shown in the illustration.

Suppose the sampling fraction is such that a weight of 1,000 is one of the smallest weights on the file. As one example from the series of tables in [Table 1](#), an intruder can see a very small number 1,000 for $S_0=1$ and $S_1=1$, then 1,000 for $S_0=1$ and $S_2=1$, and lastly 1,000 for $S_0=1$ and $S_3=2$. Therefore, the intruder can link the tables together using the common variable S_0 and determine that $S_1=1$, $S_2=1$, and $S_3=2$. Since 1,000 continues to appear in the tables, the intruder would grow in confidence that the cell is a singleton. Even when two sample records fall into the same cell, as shown in the third part of [Table 1](#) where $S_0=1$ and $S_3=1$, one knows from the prior tables that the two contributing weights are 3,000 and 3,500, since those are the only two that can sum to 6,500. With this intruder approach, the tables can be linked using their sampling weight to arrive at pseudo-microdata records displayed in [Table 2](#), which would serve as a “matching key” to attempt to attach the restricted subgroup to the public file.

Table 1. Series of tables illustrating table linking due to weights for subgroup $R = 1$

S ₁						
S ₀	1	2	3	4	5	6
1	1,000	3,000	0	0	0	3,500
2						

S ₂								
S ₀	1	2	3	4	5	6	7	8
1	1,000	3,000	0	0	0	0	0	3,500
2								

S ₃				
S ₀	1	2	3	4
1	6,500	1,000	0	0
2				

Differencing tables from multiple requests poses another challenging threat. By “differencing,” we refer to an intruder attack where the difference is taken between common cell estimates between two explicit tables displayed in the online tool in order to create an implicit table (perhaps written down on paper). Suppose there are two tables that use the same table variables but differ in the universe specification. Suppose both queries successfully satisfy query threshold rules (such as the Rule of Three, which requires three or more sample cases). These two queries may lead to an implicit table (the difference between the two tables) that would have failed the query threshold rule if explicitly specified. This is also sometimes referred to as “slivering,” since the multiple requests may create “slivers” of universes.

Likewise, differencing tables from multiple requests on the same universe can lead to the derivation of tables on subgroup slivers when “not applicable” populations occur for items in different tables. Given the above risk elements, we formulate the intruder scenario as it would relate to Subsystem R. We define the following groups of variables:

- R=Set of restricted variables that would reside on Subsystem R only.
- D=Set of demographic variables and other indirect identifiers that reside on the public use file (PUF) and within Subsystem R. The source of the variables could come from a registry or the questionnaire. Examples are age, race, sex, and marital status.
- S=Set of questionnaire variables that reside on both the PUF and within Subsystem R, which contain subject matter information, such as body mass index (BMI), duration

Table 2. Illustrated microdata formed by table linking due to weights

Pseudo-microdata record	S ₀	S ₁	S ₂	S ₃
1	1	1	1	2
2	1	2	2	1
3	1	6	8	1

variables, limitation variables, and so on. These are variables that an intruder would likely not know about the individual.

O=Set of other variables that are not available to Subsystem R, but are on the PUF. These variables would include any variables suppressed from Subsystem R. Some variables may be included in the PUF only. For example, if the PUF offers national-level analysis and R offers state-level analysis, the questions that were answered by a very small subset of participants (e.g., due to skip patterns) can only be analyzed at the national level. Since these data are so sparse, collapsing or recoding would not work for these variables. Another example is that Subsystem R may include the collapsed version of a variable, but the more detailed versions are in the PUF only.

L=Set of variables that are not on Subsystem R, nor on the PUF, but are on other files that can be linked to Subsystem R. An example is in reference to in-house files and those files that are accessible through an RDC.

Subsystem R variables can be represented by $R||D||S$, where “||” indicates the concatenation or union of the groups of variables. The PUF variables can be represented by $D||S||O$. All publicly available data that are available through linking IDs on the public files may contain at most the following $D||S||O||L$, where D, S, O, and L are disjoint. Some variables in L are only available in the RDC’s restricted files through Designated Agent agreements. Here we want to be clear that we are not expanding the space of alternative data sources to include nonpublic sources. The situation is illustrated in [Figure 2](#).

A data intruder may try to conduct multiple queries on a real-time OAS using restricted data, using table differencing, in order to link implicit “slivers” (from differencing) to obtain the “matching key” ($D||S$) for a limited number of records contained in the sliver. The matching key can be used to try to attach R to the PUF, which would be a violation of the PUF data use restrictions and may lead to the identity of an individual, or at least increase the chance for inferential disclosure. The set of variables identified by this can be represented by $R||D||S||O$. Certainly, then, it is simple to expand the set by linking other linkable files, which can be represented by $R||D||S||O||L$. In summary, it follows that an objective of the SDC treatment is to address table differencing (TD) to prevent a microdata record from being formed in a real-time OAS using restricted data through linking implicit tables (LIT), and then conducting a record linking (RL) operation to put restricted data on the PUF and other linkable files. Thus, we refer to this as the TD-LIT-RL

Fig. 2. Schematic representation of survey variables and associated files

(or differencing/linking) sequence attack, which comprises the highest chance of success in any intruder action. In such a case, in the absence of SDC procedures, it would be highly likely to have a one-to-one match with the existing NHIS public use files.

5. SDC Treatments

For an OAS, an additional risk is the availability of a PUF, and therefore in general, an objective of the SDC treatments is to protect all variables offered in the OAS in some manner, since any subset of the variables can be used in a record linking operation. The approach for Subsystem R of the NHIS OAS currently under review consists of one-time SDC treatments on the underlying microdata, and modification of analytic results in real time within the OAS' Subsystem R.

The purpose of treating the underlying microdata is to add uncertainty to the identification of an individual in the sample. The following steps would be taken to achieve that goal:

- Initial risk analysis. The goal of the initial risk analysis was to identify high risk data values to inform the data coarsening and random perturbation processes.
- Data coarsening. By “data coarsening” we mean global recoding, including top-coding (trimming high outlier values to a cutoff) and variable suppression. A discussion of data coarsening can be found in the *mu-Argus* 4.2 manual ([Statistics Netherlands 2008](#)), and summarized in [Hundepool et al. \(2012\)](#).
- Random perturbation. By “random perturbation” we imply an SDC treatment used to modify data values using a controlled approach with a random mechanism. Some perturbation approaches are discussed in [FCSM \(2005\)](#). Two examples of random perturbation approaches are data swapping and the constrained hotdeck. Data swapping procedures have been discussed throughout the literature (e.g., [Dalenius and Reiss 1982](#); [Fienberg and McIntyre 2005](#)). Data swapping is used at the U.S. Census Bureau, where in practice pairs of households are swapped across different geographic regions, as discussed in [Ramanayake and Zayatz \(2010\)](#). Controlled random swapping is described in [Kaufman et al. \(2005\)](#) as it relates to National Center for Education Statistics data. The constrained hotdeck is a relatively new approach among data replacement methodologies that was developed through research for the National Academy of Sciences ([Krenzke et al. 2011a](#)).
- Evaluation of the impact on data utility. Reviewing weighted frequencies and reviewing associations between variables before and after random perturbation are typically done. As an example for multivariate associations, [Woo et al. \(2009\)](#) proposed using propensity scores as a global utility measure for microdata.

Within Subsystem R, it would be of extreme importance that proper confidentiality screening be conducted and that the system architecture be constructed to handle it. That is the most critical and the most challenging requirement. In addition to the research being done at NCHS ([Gentleman 2011](#)), other agencies and organizations have attempted or are attempting to address the disclosure risk elements of a dynamic OAS that utilizes restricted data, such as the UCLA Center for Health Policy Research's AskCHIS ([UCLA 2012](#)), with limited suppression and rounding, the U.S. Census

Bureau (Lucero et al. 2011), with subsampling and threshold rules, the Australian Bureau of Statistics (ABS) (Tam 2011), which applies a post-tabular perturbation approach, Statistics Canada (Simard 2011), which applies random rounding, and the U.S. Substance Abuse and Mental Health Services Administration, which also has disclosure limitations embedded in it. By “post-tabular perturbation,” we refer to either a random rounding mechanism or a look-up table that contains the noise to add to the original estimate. In some respects, the SDC approaches under consideration for Subsystem R are influenced by the approach being implemented in the Census Bureau’s developing Microdata Analysis System (MAS), discussed in Lucero et al. (2011) and Freiman et al. (2011). Like the scenarios being considered for Subsystem R, the MAS plans contain rules to restrict universe and table query specifications, apply threshold rules, and include a subsampling approach referred to as the “Drop Q Rule.”

5.1. Goals of SDC Treatments in Subsystem R

A general goal when applying SDC treatments to developing an OAS in the presence of a PUF is to limit distortion in the results while protecting against the TD-LIT-RL attack. With this in mind, we define the following properties when developing an OAS:

- (1) Cell consistency – Across multiple users, if the same set of records contribute to a table cell, the same results are attained.
- (2) Query consistency – Across multiple users using the same query path (e.g., same specification for universe definition and requested table), the same results are attained.
- (3) Additivity – The sum of results from multiple tables is equal to the results directly arrived at for the aggregated table.

It is interesting to note that through our risk evaluations, we have found that attaining cell consistency may not provide adequate protection against an extensive TD-LIT-RL attack under extreme conditions when a PUF exists. For instance, in the extreme scenario of table differencing (TD) for explicit tables that differ by one case, attaining cell consistency results allows for the potential identification of the true attribute since all cells but one have zero sum of weights in the implicit tables from differencing. Therefore, when linking implicit tables (LIT), a set of true attributes are revealed and can be used in record linkage (RL) to a PUF, for example.

Attaining query consistency helps to provide credibility in the results. With such a property however, it must be understood that two users who strive for the same estimate through two different paths may get different results.

Attaining additivity, while achieving consistency, is difficult in a dynamic query system. The Australian Bureau of Statistics (Fraser and Wooten 2005) achieves consistency in the first stage through a post-tabular adjustment, and then applies an additivity algorithm in the second stage using a form of iterative proportional fitting (IPF), which was introduced by Deming and Stephan (1940). The consistency property is lost during the second stage since the IPF is applied to the specific table at hand.

The task is, however, challenging to a lesser degree for surveys considering an OAS without a PUF. When a PUF does not exist, the focus is typically on reducing the risk for a

limited number of indirectly identifying variables (the ones that provide factual information about an individual). As mentioned previously, when a PUF exists, all variables must be protected because it becomes a computer matching operation if a microdata record can be formed through the TD-LIT-RL attack. Table 3 provides a summary of SDC treatments that were considered for a real-time OAS using restricted data.

There are four main SDC treatments being considered:

- (1) Query restrictions.
- (2) Threshold rules.
- (3) Dynamic subsampling.
- (4) Rounding.

The implementation of the approaches in Subsystem R (query restrictions, threshold rules, subsampling, and rounding) took much collaboration among systems staff, computer programmers, statisticians, and program managers. It was necessary that each approach be fully understood, and many times trade-offs were discussed in terms of costs, amount of effort, processing time, and complexity.

5.2. Query Restrictions

Once within the system, the user would specify the items of interest: main topic, geography, universe, and table request. Based on the query, the system would be able to determine if the query should run through Subsystem P or Subsystem R. Within Subsystem R, query restrictions would be employed to reduce disclosure risk. For example, only one restricted variable could be used in a query. As another example, after an evaluation of risk, it was determined that one of the largest risk-reducers was to only make a limited number of “filter” variables available to define a subpopulation. The systems architecture issues relating to query restrictions included how to identify the variables that were allowed to serve as filter variables, and to be able to manage the flow between screen panels as the system tries to determine which subsystem to invoke.

5.3. Threshold Rules

The system would check whether the query would pass threshold rules, such as those given below.

Example Rule 1. Universe Threshold Rule. The purpose of the Universe Threshold Rule is to prevent small cells from occurring in the subsequent table request after the initial request. This rule denies a query result if the specified universe has less than X records, where X is a prespecified and confidential integer.

Example Rule 2. Table Marginal Count Threshold Rule. The purpose of the Table Marginal Count Threshold Rule is to prevent table linking from explicitly specified tables. This rule denies a query request if the specified m -way table has less than Y records in any $m - 1$ marginal, where Y is a prespecified and confidential integer. The cell means or quantiles need to be based on at least Y records in order to help reduce the risk of disclosing specific values of the continuous variable. The systems architecture issues relating to threshold rules were straightforward to address. If the query fails the threshold

Table 3. Examples of possible SDC treatments in a real-time OAS using restricted data

SDC Treatment	Risk-reducing benefit	Utility impact	Performance
Limiting scope of queries (e.g., not allowing more than three variables in a cross-tabulation)	Disallows attempts by intruder to drill down into the details of the microdata file	Reduces the scope of analysis	Negligible impact
Threshold rules	Denies tables with small sample sizes to an intruder attempting a table linking attack	Denies tables with small sample sizes to the user	Negligible impact
Subsampling, Post-tabular perturbation, Post-tabular random rounding	Provides a level of protection against a differencing/linking attack, depending upon the approach	Has potential to retain consistency; however, the additivity property is lost. Impact on variance estimates would be readily retained in subsampling	Initial tests show that the impact of subsampling on performance is minimal. Impact of post-tabular approaches minimal
Post-subsampling calibration to grand total for the universe	Provides additional level of protection against a differencing/linking attack	Reduces the difference for universe totals. May have some negative impact for table cells, depending on how it is applied	Initial tests show that impact on performance due to an extra step depending on the table generator is minimal
Rounding output	Provides some protection against a differencing/linking attack	Reduces the detail in the estimates displayed to the user	Negligible impact

rules, a message would be displayed on the screen to provide some general guidance to the user.

5.4. *Dynamic Subsampling*

Dynamic subsampling developed for Subsystem R is influenced by the U.S. Census Bureau's Drop Q Rule approach and the post-tabular perturbation approach of the Australian Bureau of Statistics (ABS) described in [Fraser and Wooten \(2005\)](#).

The following approach to subsampling within an OAS achieves query consistency and limited additivity, while protecting against a TD-LIT-RL attack. Suppose a two-way table is specified. Suppose U is the sum of random numbers (RND) for records in the defined universe. Suppose $M1$ is the sum of RND for records in the row marginal, $M2$ is the sum of RND for records in the column marginal, C is the sum of RND for records in the table cell of interest. Let SEED be a function of U , $M1$, $M2$, and C . The derived seed is a function of the query request (defined universe and table variables) such that the same seed would be generated for the selection of the same subsample within a table cell.

The dynamic subsampling occurs cell by cell; that is, the resulting subsample for the query is stratified by the table cells. The number of cases to deselect is a function of the cell's sample size. The seeding approach reduces the risk of an intruder making an accurate inference after taking the difference between two tables generated on slightly different universes.

Prior to computing the results, the sampling weights (full sample and replicates) are adjusted within table cells to account for the cases not selected in the subsample. This ensures that the weighted totals are unbiased and that the variances are computed properly. After the within-cell reweighting, another overall reweighting step is used to calibrate to the grand total. Regarding systems architecture, the system needs to first detect that the query should invoke Subsystem R. The analytic engine is programmed to conduct the subsampling and reweighting, and return the results to the interface. Initial test runs indicate that the implementation of dynamic subsampling does not result in a measurable increase in the run time.

5.5. *Rounding Rules*

Rounding rules can be used for the resulting estimates to reduce the chance for linking tables with weighted totals in cells with small sample sizes. The following are examples under consideration:

- Weighted percentages – nearest 10th (0.1).
- Standard errors and confidence intervals of weighted percentages – nearest 100th (0.01).
- Weighted counts – nearest thousand (1,000).
- Standard errors and confidence intervals of weighted counts – nearest hundred (100).

If the statistics of interest are means or percentiles, the rounding rules in [Table 4](#) are being considered.

The systems component most impacted by the rounding is the metadata database. The database provides rounding rule information to the NHIS OAS for each variable in the

Table 4. Rounding rules for means or percentiles

Analysis variable	Rounding rule for means, medians	Rounding rule for standard errors and confidence intervals
If 0 – 1 variables	0.001	0.0001
Else if 25th percentile ≤ 9	0.01	0.001
Else if 25th percentile ≤ 99	0.1	0.01
Else if 25th percentile ≤ 999	1	0.1
Else if 25th percentile $\leq 9,999$	10	1
...	...	...

underlying microdata file. The analytic engine receives the rounding rule information, which depends on the query that is specified, including the variables involved and the statistics requested. The rounding rules (RR) for the corresponding precision measures such as standard errors and confidence intervals can be derived accordingly (dividing RR by 10). If the statistic of interest is the weighted total of an analysis variable, we consider using RR(100) (meaning, round to the nearest multiple of 100) as the rounding rule for the total. If the statistic of interest is the ratio of weighted totals of two analysis variables, we consider using RR_{num}/RR_{denom} as the rounding rule for the ratio, where RR_{num} is the rounding rule associated with the variable in the numerator, and RR_{denom} is the rounding rule associated with the variable in the denominator.

5.6. Limitations of the SDC Approach

We note the following limitations of the SDC approach. First, due to the perturbation approach in data preparation, there is some increase to the mean square error term. Second, the property of additivity will not be preserved due to missing data, the subsampling (or post-tabular perturbation) approach, and rounding. That is, two tables added together may not equal the aggregate if estimated directly. To address this issue, the tables would need to be estimated directly, and not as the sum of individual tables.

6. Analysis of the SDC Impact on Risk and Utility

The impact of the set of SDC treatments that were discussed in the previous section was evaluated in terms of the competing objectives of reducing disclosure risk while retaining data utility. Here we provide a step-by-step approach to the evaluation such that the procedures are emphasized, while the outcomes of the evaluations are still under review and need to be kept confidential.

As discussed previously, the biggest challenge to a real-time OAS using restricted data is to protect against a TD-LIT-RL attack. This occurs when implicit tables with small cells (from table differencing) can be linked together to form a string of identifying characteristics (“matching key”). The matching key can be matched to external databases, such as a corresponding PUF. Recognizing that combinations of just a few variables can lead to a singleton, it was necessary that the initial risk evaluation of the microdata take into account the ability to set system parameters, such as only allowing up to three-way tables, or only allowing a certain level of geography. The main objectives were to

determine the query restrictions, threshold rules, the parameters relating to dynamic subsampling, and the rounding rules.

6.1. Impact on Disclosure Risk

Here we take the risks in turn with a discussion of table differencing (TD) and linking implicit tables (LIT), and then record linkage (RL).

6.1.1. TD and LIT Risk

To reduce the risk due to table differencing and linking implicit tables, the following Subsystem R treatments are being evaluated: query restrictions (e.g., the number of variables allowed to define the universe and tables), threshold rules, dynamic subsampling, and rounding rules.

Query restrictions and threshold rules. To evaluate the impact from query restrictions and threshold rules, restricted NHIS test microdata files were used for six combinations of data years (2009 and 2007–2009) and file types (Person (Family Core), Sample Adult, and Sample Child). Restricted NHIS data were accessed at the NCHS Research Data Center in Hyattsville, Maryland. While the sample size varies across the years, the 2009 Person File had about 88,000 records, which was about four times as much as the Sample Adult File and about eight times as much as the Sample Child File. The six combinations of years and file types were processed first through an initial risk analysis.

Exhaustive tabulations were processed for a given set of variables, which formed m -way tables. Given lower (MINDIM) and upper (MAXDIM) bounds for the table dimension size m , n tables were formed by

$$n = \sum_{m=MINDIM}^{MAXDIM} \binom{p}{m} = \sum_{m=MINDIM}^{MAXDIM} \frac{p(p-1)\cdots(p-m+1)}{m(m-1)\cdots 1} \quad (1)$$

where p is the number of variables used in the table generation. For example, suppose there are $p = 20$ total variables and all two-way, three-way, and four-way tables are created, then 6,175 tables would be generated. If the number of cases in a cell was less than the threshold, the cell was flagged as a violation cell, and the variables/categories used to define the cell were identified as “contributing to cell violations.” Violation counts were the number of times a record was in a violation cell. The percentage of cell violations for a variable/category was computed as follows:

$$V = 100[V_c/k_c] \quad (2)$$

where V_c is the number of violation cells involving this variable/category c , and k_c is the total number of cells formed by this variable/category c . The above formulation can be adapted to define universes in lieu of defining tables. The records were classified into a risk measure with five categories (values 0 through 4), from low to high, based on the frequency that a record was flagged for violations. Category 0 was assigned to the records with the lowest risk, and Category 4 was assigned to records with the highest risk as determined by the violation counts.

The analysis compared different Universe Threshold Rules and Table Marginal Count Threshold Rules. Examples of variables used in the initial risk analysis were state, Core-Based Statistical Area size and status, urban/rural status, income groups, poverty status, employment status, earnings groups, living quarters, family type, education attainment,

house ownership status, gender, race/ethnicity, age, country of birth, citizenship, health insurance information, height and weight of sample children, smoking status, and diabetes status, among others. We note that the number of variables allowed to define the universe, or table, had a large impact on the number of violations. The violation percentage was far greater when three variables were used, and far less with one or two variables. Surprisingly, combining years impacted the number of violations to a lesser extent. In general, threshold rules provided much protection – without rules, tables would show small sample sizes or singletons, and table linking could occur to form microdata records.

It is typically assumed that an intruder does not know if a given person is in the sample, in which case the probability of disclosure for a sample unique is a function of the sample selection rate and the indirectly identifiable data. As an alternative to the above risk measure, many have considered the expected number of correct matches to a population file for a sample unique to be written as:

$$GlobalRisk = \sum_{SU} (1/F_k | f_k = 1) \quad (3)$$

where SU is the set of sample uniques, f_k is the sample frequency in cell k , and F_k is the population frequency in cell k . The aforementioned *mu-Argus 4.2* manual provides a discussion of how disclosure risk is measured as a function of the sampling fraction, when an intruder knows that the unweighted cell count is one (sample unique). Investigation into the use of models to provide more stable estimates of risk has been conducted by researchers (e.g., [Skinner and Shlomo 2008](#)). A more extensive risk measure that takes into account the measurement error (e.g., data swapping) is given in [Shlomo and Skinner \(2010\)](#).

When determining the level of disclosure risk a file may pose, other factors need to be considered in conjunction with any single risk measure, such as the degree of potential harm to individuals if data were disclosed and the presence of risk-reducing factors, such as the level of geographic detail; availability of external data; age of the data; existence of imputation, perturbed data, recodes; and variable suppression.

Dynamic Subsampling. The benefit of dynamic subsampling is displayed in the following example. Suppose an intruder would like to make a table using a sparse category 1 of a three-category variable (A) as shown in [Table 5](#). The intruder would be denied a table due to the threshold rule. However, if the system allows combined categories, then the categories could be combined in different trials: combining categories A=1 with A=2 and then A=1 with A=3 in separate queries. These explicit tables that pass the threshold rules could be differenced to arrive at a table that involves category 1. The first two tables show explicit two-way tables $A' \times X$ and $A'' \times X$, followed by the implicit table from the difference of the two explicit tables. For each cell, the noise added to the cell estimate due to subsampling is provided by a_{ij} and a'_{ij} where i is the row and j is the column. The noise from the subsampling is in unpredictable amounts, and therefore the intruder cannot determine with certainty the value of X in the implicit difference table even though there is a single case with a weight of 4,000 where A=1 and X=1. This is because the a_{ij} and a'_{ij} do not necessarily cancel out since the sum of RND is different in the row marginal, and the amount of the difference between a_{ij} and a'_{ij} does not have a defined limit.

Evaluations were conducted on the data to provide indications of the levels of a_{ij} and a'_{ij} under different subsampling scenarios. To understand the level of risk reduction through

Table 5. Illustration of dynamic combining of categories with subsampling

	X=1	X=2	X=3	X=4
A': A=1,2	$36,000 + a_{11}$	$42,000 + a_{12}$	$16,000 + a_{13}$	$5,000 + a_{14}$
A': A=3	$21,000 + a_{21}$	$36,000 + a_{22}$	$42,000 + a_{23}$	$8,000 + a_{24}$
A'': A=2	$32,000 + a'_{11}$	$42,000 + a'_{12}$	$16,000 + a'_{13}$	$5,000 + a'_{14}$
A'': A=1,3	$25,000 + a'_{21}$	$36,000 + a'_{22}$	$42,000 + a'_{23}$	$8,000 + a'_{24}$
A=1	$4,000 + a_{11} - a'_{11}$	$a_{12} - a'_{12}$	$a_{13} - a'_{13}$	$a_{14} - a'_{14}$
A=1	$-4,000 + a_{21} - a'_{21}$	$a_{22} - a'_{22}$	$a_{23} - a'_{23}$	$a_{24} - a'_{24}$

subsampling, we found it best to incorporate all SDC treatments used by the system (i.e., subsampling, rounding, thresholds), and then evaluate the reduction in disclosure risk through extreme examples where two universes differ by only one case. This allowed us to make clear observations about the impact on risk reduction. Through such effort, we found that when selecting a subsample in an OAS using simple random sampling there is a danger of creating a sliver. Through further evaluation, we also determined that limiting the variables to form universes greatly limits the ability to create implicit slivers with a sample size of 1.

We also note that we evaluated a post-tabular random rounding approach, and if relied upon without subsampling or other post-tabular perturbation, we are skeptical about its protection in the extreme case of a single case with $A = 1$ and $X = 1$. In this extreme case, suppose ε is the resulting amount added or subtracted to result in a rounded value of a weighted cell frequency. Suppose cells are rounded to the nearest 1,000. When table differencing (TD) explicit tables that differ by one case, the results allow a possible identification of the true attribute since all cells but one would have a difference of 1,000, 0, or $-1,000$. If the direction of the rounding is only triggered by the sum of the RND values for the cell, in order to retain cell consistency, then all cells in the implicit difference table, with the exception of one cell, would have a difference equal to 0. Therefore, when linking implicit tables (LIT), a set of true attributes can be revealed, which then can be used in record linkage (RL). The amount of protection is related to, and limited to, the size of ε relative to the magnitude of the estimate.

Table 6. Illustration of dynamic combining of categories with subsampling and rounding applied

	X = 1	X = 2	X = 3	X = 4
A': A=1,2	$36,000 + a_{11} + b_{11}$	$42,000 + a_{12} + b_{12}$	$16,000 + a_{13} + b_{13}$	$5,000 + a_{14} + b_{14}$
A': A=3	$21,000 + a_{21} + b_{21}$	$36,000 + a_{22} + b_{22}$	$42,000 + a_{23} + b_{23}$	$8,000 + a_{24} + b_{24}$
A'': A=2	$32,000 + a'_{11} + b'_{11}$	$42,000 + a'_{12} + b'_{12}$	$16,000 + a'_{13} + b'_{13}$	$5,000 + a'_{14} + b'_{14}$
A'': A=1,3	$25,000 + a'_{21} + b'_{21}$	$36,000 + a'_{22} + b'_{22}$	$42,000 + a'_{23} + b'_{23}$	$8,000 + a'_{24} + b'_{24}$
A=1	$4,000 + a_{11} - a'_{11}$ $+ b_{11} - b'_{11}$	$a_{12} - a'_{12}$ $+ b_{12} - b'_{12}$	$a_{13} - a'_{13}$ $+ b_{13} - b'_{13}$	$a_{14} - a'_{14}$ $+ b_{14} - b'_{14}$
A=1	$-4,000 + a_{21} - a'_{21}$ $+ b_{21} - b'_{21}$	$a_{22} - a'_{22}$ $+ b_{22} - b'_{22}$	$a_{23} - a'_{23}$ $+ b_{23} - b'_{23}$	$a_{24} - a'_{24}$ $+ b_{24} - b'_{24}$

Rounding rules. Deterministic rounding, as illustrated in Table 6, has a risk-reducing effect on the TD-LIT-RL attack. As discussed above, for each cell, the amount added to the cell estimate due to subsampling is provided by a_{ij} and d'_{ij} , and we introduce b_{ij} and b'_{ij} to represent the amount added or subtracted due to the rounding mechanism. The amount of rounding depends on the original cell estimate and the amount of noise added through subsampling. Therefore, as illustrated in Table 6, the b_{ij} and b'_{ij} do not necessarily cancel out.

6.1.2. Record Linkage (RL) Risk

Attempts were made to provide indications of the level of disclosure risk that remain for a possible sample unique record that may leak through the threshold rules, possibly through linking tables due to unusual variation in the weights across queried tables, for example.

Conditional on threshold rules being applied, and under the assumption that a series of table queries were linked to form a microdata record, the risk due to the matchability to the PUF was assessed. For the group of records in the highest risk stratum j determined through state-level tabulations using the Subsystem R data (R||D||S), we use the key D||S to attach R to the PUF. We expect E_{Mj} records to match if there was a one-to-one match, where E_{Mj} is the number of Subsystem R records in risk stratum j . When the match occurs, let A_{Mj} equal the actual number of records on the PUF (national level) that are matched, including all one-to-many matches. Without perturbation or subsampling, then the general risk measure for the highest risk records in risk stratum j was computed as $R_j = E_{Mj}/A_{Mj}$. This assessment for a group of risky records defined by the initial risk analysis provided a general indication of the risk level.

6.2. Evaluating Subsystem R SDC Treatment Impact on Disclosure Utility

To analyze the impact of subsampling on data utility, an extensive analysis was conducted on a wide variety of subsampling approaches. The evaluation included an exhaustive tabulation of state-level two-way tables using selected variables, and the relative difference between the full sample and the subsample estimates was computed. The results of each run used the same rounding rules. Both universe and marginal threshold rules were applied to suppress tables with too few cases in the universe or in any margins. The average subsampling rates within table cells dropped gradually down to a specified rate as the unweighted cell sizes increased. The results can be reviewed in risk-utility charts, such as the one illustrated from one subsampling scheme as presented in Figure 3. The chart summarizes the results from more than 1,500 tables that contained over 20,000 cells. In the graph, the 10th, 50th, and 90th percentiles of the relative differences in weighted counts are displayed at three levels: grand total, one-way tables, and two-way tables. The size strata of the subgroups, one-way tables, and two-way tables are labeled, with 1 indicating the smallest size (higher risk) and 6 indicating larger size (lower risk). Since a calibration adjustment was done to ensure that the subsample sums to the same weighted total as the full sample, the relative differences at the grand total level are zero for universes of all sizes. For two-way table cells, the ratio of full sample to subsample estimates has a median at 1 regardless of the cell sizes. The 10th

Fig. 3. Percentiles of ratios of full sample to subsample estimates by dimension and size

and 90th percentiles of the ratios show that subsample estimates for smaller cells deviate more from the full sample estimates due to higher subsampling rates.

In addition, the impact of subsampling on standard errors can be simply computed as the square root of the reciprocal of the relative difference in sample size. Table 7 provides information on the increase to the standard errors, based on the number of records deselected (q), relative to the number in the original sample (n).

6.3. Other Remarks Related to the Evaluation

If the risk measures indicate that risk levels are not acceptable, as determined by the agency, then the perturbation rates, subsample rates, and other treatments, such as

Table 7. Increase in standard errors when deselecting q records from a sample of size n

Sample size n	Number of deselected records		
	$q = 1$	$q = 10$	$q = 20$
50	1.02%	11.80%	29.10%
100	0.50%	5.41%	11.80%
200	0.25%	2.60%	5.41%
2,000	0.03%	0.25%	0.50%
20,000	0.00%	0.03%	0.05%

rounding rules, will need to be adjusted. Another approach is to be more restrictive; that is, adjust the threshold parameters to deny more tables, reduce the number of filter variables allowed, and reduce the number of variables allowed when defining tables. If the utility measures indicate that deviations are outside of tolerance bounds, then the parameters for the SDC treatments need to be reviewed with the objective of reducing the amount of the deviation. For example, the amount of data perturbation or subsampling could be reduced. If correlations indicate that attenuation is occurring, then the perturbation approach would need to be reviewed and adjusted by involving key variables in the process. As one can see, trade-offs between risk and utility cannot be avoided.

Once a system is up and running, the performance of the system needs to be evaluated. Performance of the system should be measured in terms of the time it takes from the submission of the query to the display of the results, assuming good Internet connectivity. The implementation of SDC treatment methodology, with the addition of dynamic subsampling and/or calibration, may or may not add to the server processing time due to the computations required for any one query. The extent to which the additional processing is actually noticeable to the user should be measured. If the incremental time for each query web request is excessively long, then improvements to performance may be attained by reducing the file size, reworking the SDC code, or upgrading the server platform.

7. Concluding Remarks

Two general objectives of conducting a survey are for the data to be useful and for the data to be used. As technology advances, however, the challenges of preserving confidentiality while at the same time maintaining data access increase. This article has described how we, like other researchers around the world, are striving to adapt systems architecture to facilitate methods for making restricted survey data more accessible for analysis while preserving confidentiality. The development of an OAS is a very large undertaking, but such a system has the very worthwhile potential to simplify and increase accessibility and use of the data.

Several components of systems architecture are impacted by disclosure-limitation methodology. The user interface is impacted because the interface is dynamic to the query being submitted. For example, if a state is selected for a query, then Subsystem R is invoked and fewer and sometimes different variables are available for the analysis. Since

there are differences between what is allowed in a query for restricted data and for PUF data, much thought was needed on how to change the presentation of the screen contents as the user goes back and forth freely in the OAS. The system must be able to adapt to changes that the user makes to the query (e.g., deselecting a state), which can change the operating subsystem (R or P). The metadata database will contain information to help integrate the two subsystems. The administrative tools are impacted by disclosure risk, such that they control the rules relating to the protection necessary for different types of data, such as public and restricted use. The structure of the underlying microdata files is also impacted by disclosure risk if restricted data are randomly perturbed.

This article will aid others developing similar real-time online systems using restricted data, for which demands from data users are increasing. This paper provides future system developers with descriptions of tasks and issues to consider when defining capabilities and estimating costs and level of effort. For those considering an OAS as an option for disseminating survey results, including results based on restricted data, we describe the following experiences and lessons learned.

- If the home organization (e.g., a statistical agency) is not large, it will probably be necessary for it to enlist the services of a contractor to help develop an OAS.
- In implementing a new OAS, there will likely be numerous regulatory obstacles to overcome: It is difficult to implement a new system of a type that has never before been used in one's agency. Paradoxical situations may arise where conflicting demands are being made or where needs never before encountered must be met, and not everyone will fully comprehend the objectives of building an OAS. Issues may not be straightforward to address, such as where and by whom the system will be housed, hosted, and maintained.
- In implementing a new OAS, there will be numerous administrative obstacles to overcome: for example, the system developers are likely more expert in science than in contracting, despite having taken required training in dealing with contracts. Contracts will likely have to be amended and extended because of the many unknowns in the development of an OAS system.
- If more restricted data and/or new capabilities are to be added to an OAS on a periodic basis, this will require expertise that may not be found and retainable within the home organization; the knowledge and experience needed to apply specialized disclosure avoidance methods to updated data files and new system capabilities may not be available in a home organization that is not large.
- An OAS is not likely to be quickly and easily adaptable for use on a distinctly different set of data from the one for which it was originally developed.
- It is helpful for the system developers at the home organization to work very closely with the contractor to ensure that both groups have the same objectives and to obtain input from both groups in the areas of their specific expertise.

NCHS believes that online analytic systems that use restricted data can decrease the need for research data centers by making some analyses of restricted data available in a much more timely and convenient way, but such online analytic systems might also increase the use of research data centers by generating interest of the OAS users for more analyses of the type that must be conducted in a research data center. Because the future

implementation of an OAS that uses NHIS data is still being considered by NCHS, it is not possible at this time to determine the effect of such a system on the overall accessibility and use of NHIS data. However, we are aware that similar systems, such as AskCHIS and the Integrated Health Interview Series' OAS (IHIS 2012), have many users, and we also hear from NHIS data users, in particular from policy makers, that they would appreciate and benefit from quicker and easier access to analyses of NHIS restricted variables as well as to analyses of NHIS public use variables.

8. References

- Chipperfield, J. and Yu, F. (2011). Protecting Confidentiality in a Remote Analysis Server for Tabulation and Analysis of Data. Joint UNECE/Eurostat work session on statistical data confidentiality (Tarragona, Spain, 26–28 October 2011). United Nations Economic Commission for Europe (UNECE). Conference of European Statisticians, European Commission Statistical Office of the European Union (Eurostat).
- Dalenius, T. and Reiss, S. (1982). Data-swapping: A Technique for Disclosure Control. *Journal of Statistical Planning and Inference*, 6, 73–85.
- Deming, W.E. and Stephan, F.F. (1940). On a Least Square Adjustment of a Sampled Frequency Table When the Expected Marginal Totals Are Known. *Annals of Mathematical Statistics*, 11, 427–444.
- Duncan, G.T., Elliot, M., and Salazar-González, J. (2011). *Statistical Confidentiality: Principles and Practice*. New York: Springer.
- FCSM (2005). Report on Statistical Disclosure Methodology. Statistical Policy Working Paper 22 of the Federal Committee on Statistical Methodology, 2nd version. Revised by Confidentiality and Data Access Committee 2005, Statistical and Science Policy, Office of Information and Regulatory Affairs, Office of Management and Budget. Available at: http://www.fcsm.gov/working-papers/SPWP22_rev.pdf (accessed October 21, 2012).
- Fienberg, S.E. (2009). The Relevance or Irrelevance of Weights for Confidentiality and Statistical Analyses. *Journal of Privacy and Confidentiality*: Vol. 1: Iss. 2, Article 4. Available at: <http://repository.cmu.edu/jpc/vol1/iss2/4> (accessed February 1, 2013).
- Fienberg, S.E. and McIntyre, J. (2005). Data Swapping: Variations on a Theme by Dalenius and Reiss. *Journal of Official Statistics*, 21, 309–323.
- Fraser, B. and Wooten, J. (2005). A Proposed Method for Confidentialising Tabular Output to Protect Against Differencing. Paper presented at the November 2005 UNECE Work Session on statistical data confidentiality.
- Freiman, M., Lucero, J., Singh, L., You, J., DePersio, M., and Zayatz, L. (2011). The Microdata Analysis System at the U.S. Census Bureau. Proceedings of the Joint Statistical Meetings, American Statistical Association Section on Survey Research Methods.
- Gentleman, J.F. (2011). A Real-Time Online System for Analyzing Restricted Data from the U.S. National Center for Health Statistics' National Health Interview Survey. Proceedings of the 58th World Statistics Congress of the International Statistical Institute. Available at: <http://isi2011.congressplanner.eu/pdfs/650208.pdf> (accessed February 1, 2013).

- Gentleman, J.F. (2012). Using the National Health Interview Survey to Monitor the Early Effects of the Affordable Care Act. Proc. 2012 Joint Statistical Meetings, San Diego.
- Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E.S., Spicer, K., and de Wolf, P.-P. (2012). Data Access Issues. In *Statistical Disclosure Control*. Chichester, UK: John Wiley & Sons. doi: 10.1002/9781118348239.ch6.
- IHIS (2012). Minnesota Population Center and State Health Access Data Assistance Center, Integrated Health Interview Series: Version 5.0. Minneapolis: University of Minnesota, 2012. Available at: <http://www.ihis.us> (accessed November 23, 2012).
- Kaufman, S., Seastrom, M., and Roey, S. (2005). Do disclosure controls to protect confidentiality degrade the quality of the data? Proceedings of the Joint Statistical Meetings, American Statistical Association Section on Government Statistics. Alexandria, VA: American Statistical Association.
- Krenzke, T., Li, J., Judkins, D., and Larsen, M. (2011a). Evaluating a Constrained Hotdeck to Perturb American Community Survey Data for the Transportation Planning Products. Proceedings of the Joint Statistical Meetings, American Statistical Association Section on Survey Research Methods.
- Krenzke, T., Li, J., Freedman, M., Judkins, D., Hubble, D., Roisman, R., and Larsen, M. (2011b). Producing Transportation Data Products from the American Community Survey that Comply with Disclosure Rules. Washington, DC: National Cooperative Highway Research Program, Transportation Research Board, National Academy of Sciences.
- Lucero, J., Zayatz, L., Singh, L., You, J., DePersio, M., and Freiman, M. (2011). The Current Stage of the Microdata Analysis System at the US Census Bureau. Proceedings of the 58th World Statistics Congress of the International Statistical Institute. Available at: isi2011.congressplanner.eu/pdfs/650103.pdf (accessed February 1, 2013).
- Meyer, P., Robinson, E.S., and Madans, J. (2012). Protecting Confidentiality in a Data Enclave. *Statistical Journal of the IAOS*, 28, 25–30.
- NCHS RDC (2011). National Center for Health Statistics Research Data Center website link <http://www.cdc.gov/rdc/> (accessed February 1, 2013). National Center for Health Statistics, Centers for Disease Control and Prevention. U.S. Department of Health and Human Services.
- NCES (2003). NCES Statistical Standards. NCES 2003-601. Washington, DC: 2003. Available at: http://nces.ed.gov/statprog/2002/std5_3.asp (accessed February 1, 2013).
- NHIS (2010). National Health Interview Survey Description document. Available at: ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHIS/2010/srvydesc.pdf (accessed February 1, 2013). Division of Health Interview Statistics, National Center for Health Statistics. Centers for Disease Control and Prevention. U.S. Department of Health and Human Services.
- NHIS (2011). National Health Interview Survey website link, <http://www.cdc.gov/nchs/nhis.htm> (accessed February 1, 2013). National Center for Health Statistics, Centers for Disease Control and Prevention. U.S. Department of Health and Human Services.
- ONS (2006). Review of the Dissemination of Health Statistics: Confidentiality Guidance. Health and Care. London, UK: Office for National Statistics.

- Ramanayake and Zayatz (2010). Balancing Disclosure Risk with Data Quality. Research Report Series (Statistics #2010-04). Statistical Research Division, U.S. Census Bureau, Washington, D.C.
- Reiter, J. (2003). Model Diagnostics for Remote-Access Regression Servers. *Statistics and Computing*, 13, 371–380.
- Shlomo, N. and Skinner, C. (2010). Assessing the Protection Provided by Misclassification-Based Disclosure Limitation Methods for Survey Microdata. *Annals of Applied Statistics*, 4, 1291–1310.
- Simard, M. (2011). Real Time Remote Access at Statistics Canada: Development, Challenges, and Issues. Proceedings of the 58th World Statistics Congress of the International Statistical Institute. Available at: isi2011.congressplanner.eu/pdfs/650096.pdf (accessed February 1, 2013).
- Skinner, C.J. and Shlomo, N. (2008). Assessing Identification Risk in Survey Microdata Using Log-linear Models. *Journal of American Statistical Association*, 103, 989–1001.
- Statistics Netherlands (2008). μ -ARGUS version 4.2 User's Manual. Available at: <http://neon.vb.cbs.nl/casc/Software/MuManual4.2.pdf> (accessed February 1, 2013).
- Sweeney, L. (2002). Achieving k-anonymity Privacy Protection Using Generalization and Suppression. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10, 571–588.
- Tam, S. (2011). On-line Access of Micro-Data at the Australian Bureau of Statistics – Challenges and Future Directions. Proceedings of the 58th World Statistics Congress of the International Statistical Institute. Available at: isi2011.congressplanner.eu/pdfs/650030.pdf (accessed February 1, 2013).
- UCLA (2012). AskCHIS website: <http://ask.chis.ucla.edu/main/default.asp> (Accessed February 1, 2013). University of California, Los Angeles Center for Health Policy Research.
- Woo, M., Reiter, J., Oganian, A., and Karr, A. (2009). Global Measures of Data Utility for Microdata Masked for Disclosure Limitation. *The Journal of Privacy and Confidentiality*, 1, 111–124.

Received January 2012

Revised March 2012

Accepted November 2012