The Value of Early Disclosure Risk Decisions

JoAnne McFarland O'Rourke[1]

Inter-university Consortium for Political and Social Research (ICPSR)
Institute for Social Research
University of Michigan

November 15, 2011

Abstract:

When data are made available to others to analyze for their purposes, steps must be taken to ensure confidentiality, that is to prevent the identities of the persons or institutions that were studied are not disclosed and cannot be deduced. Disclosure risk analysis is conducted in order to create a public-use file (PUF) from confidential, or restricted-use, data. Based on this analysis of disclosure risks, statistical disclosure limitation (SDL) methodologies are applied to the data to create the PUF.

The public-use file (PUF) is the only version of the microdata to which most researchers ever have access and the version from which much of the utility of the data is extracted. Therefore, decisions made to create the PUF, in terms of variable changes (e.g., deletions, recodes) and the selection of statistical disclosure limitation (SDL) methods (e.g., data swapping, imputation collapsing categories) are very important and must match the key intended purposes of the data collection and the disclosure risk.

Typically, decisions regarding disclosure risk are made after data collection is completed. This article will describe a new model for conducting disclosure risk analysis for the creation of PUFs that moves decisions regarding disclosure risk to the beginning of the survey research process. Early thinking and decision-making regarding disclosure risk can lead to a more analytically useful PUF and the most optimal set of data products that can be developed (tables, maps, online analysis, and so on, in addition to the PUF). Efficiencies created between the various stages of the research process by the model will shorten the time between data collection and data release, thus increasing the value of the shared data to secondary analysts and to science.

Key words: Disclosure analysis, disclosure risk, statistical disclosure control, disclosure limitation, public-use file, restricted-use data, survey data.

[1] Contact JoAnne McFarland O'Rourke at jmcfar@umich.edu.

Description of the current process

Standard practice for creating new public-use files (PUFs) is for disclosure risk analysis (DRA) to be conducted at the end of the survey research process, after data are collected. This requires that data cleaning and editing be completed. By this time, several publications may also be have been completed and released, which may pose dangers of deductive disclosure of identities.

Often, different organizations or groups within the same organization are involved in the various steps of survey design, data collection, disclosure analysis, PUF preparation, and data dissemination. This can create specialization, or efficiency, in the process. One issue that creates inefficiency and disconnection, however, is thinking of each survey step as unrelated to the other steps in terms of disclosure risk and the final PUF content.

Prior to the late 1970's, PUF creation included the removal of direct identifiers such as name, record identifiers, and social security numbers. Subsequently, several factors led to the development of disclosure analysis as a field and science, including advancements in computing technology; increasingly inexpensive data storage; the expansion of survey research and correspondingly, an increase in the availability of data about individuals and organizations; public demand for data; and the desire and ability to link data across systems. These factors combined to progressively increase disclosure risk and led to the development of SDL techniques to reduce such risk (see Fienberg, 1985).

Disclosure risk and analysis has not previously been fully considered until after data collection is completed due in large part to disclosure analysis being relatively new to social research. Therefore, this step has been added to the end of the process, even when a PUF is planned at the time of proposal development. Similarly, though PUF creation may be an ultimate goal, the survey model has remained substantially linear, to match the steps of questionnaire design, data collection, data cleaning and editing, and data release.

Description of Disclosure Analysis

Disclosure analysis is now typically considered a fundamental step in the protection human subjects. For the social and behavioral sciences, disclosure analysis extends and ensures the promises made during informed consent procedures. For the medical sciences, disclosure analysis provides a route to data when other provisions in the Health Insurance Portability and Accountability Act (HIPAA), such as the removal of the 18 variables specified by HIPAA to create a deidentified dataset, are insufficient.

Disclosure analysis involves the careful examination of indirect identifiers that pose the risk re-identification of a respondent (O'Rourke, 2003) and publicly available databases that could be used to link data, and thus enable one to deduce identities.

Optimally, a disclosure analysis begins by answering the question "What are the key analytic uses of this data collection?" Based on this assessment, decisions are made to modify the data in order to create the PUF. Modifications may include combining categories or removing variables altogether due to their sensitivity and therefore, the risk of respondent re-identification. Reasons that variables are removed altogether from a file include that they are highly sensitive and potentially identifying (e.g., sexual orientation) or that they are both sensitive and had a low or moderate response rate (i.e., low utility). Examples of SDL techniques applied to files to protect the data include coarsening (U.S. Bureau of the Census, 2003), data swapping (Feinberg and McIntyre, 2005), imputation [citation], multiple imputation [Drechsler and Reiter, 2010], and microaggregation [citation]. For an overview of disclosure risks as well as methods used to protect against risks, see Federal Committee on Statistical Methodology (2005).

Regarding the impact of decisions made regarding the public-use data file, O'Rourke, et. al (2006) state:

> "The public-use version of the data is very important because it is likely to be the only one to which most researchers, policy analysts, teaching faculty, and students will ever have access. Hence, it is the version from which much of the utility of the data is extracted and often it effectively becomes the historical record of the data collection. Large national studies containing thousands of variables are often not, nor are they necessarily intended to be, very thoroughly analyzed prior to their public release. At most, a detailed report or series of tables are sometimes released ahead of the microdata. The data are subsequently used for research, policy, and teaching purposes in the years after their release, and even decades later for comparative analysis. For these reasons, great care must be taken to create a public-use version of a data collection that truly does balance utility and risk. Those creating public use files must ensure they have identified and retained intended uses of the data and yet accurately defined and fully addressed risk."

The following is an example of a demographic variable – whether the respondent served in the military –that might be collapsed by eliminating detail such as period of military service:

- Period or place of military service
    - World War II
    - Between WWII and Korea
    - Korea
    - Between Korea and Vietnam
    - Vietnam
    - Between Vietnam and Gulf War
    - Gulf War
    - Iraq
    - Afghanistan
    - Other

For the PUF, the detail for military service, due to disclosure risk, might be collapsed to:

- "Ever Served"
    - Yes
    - No

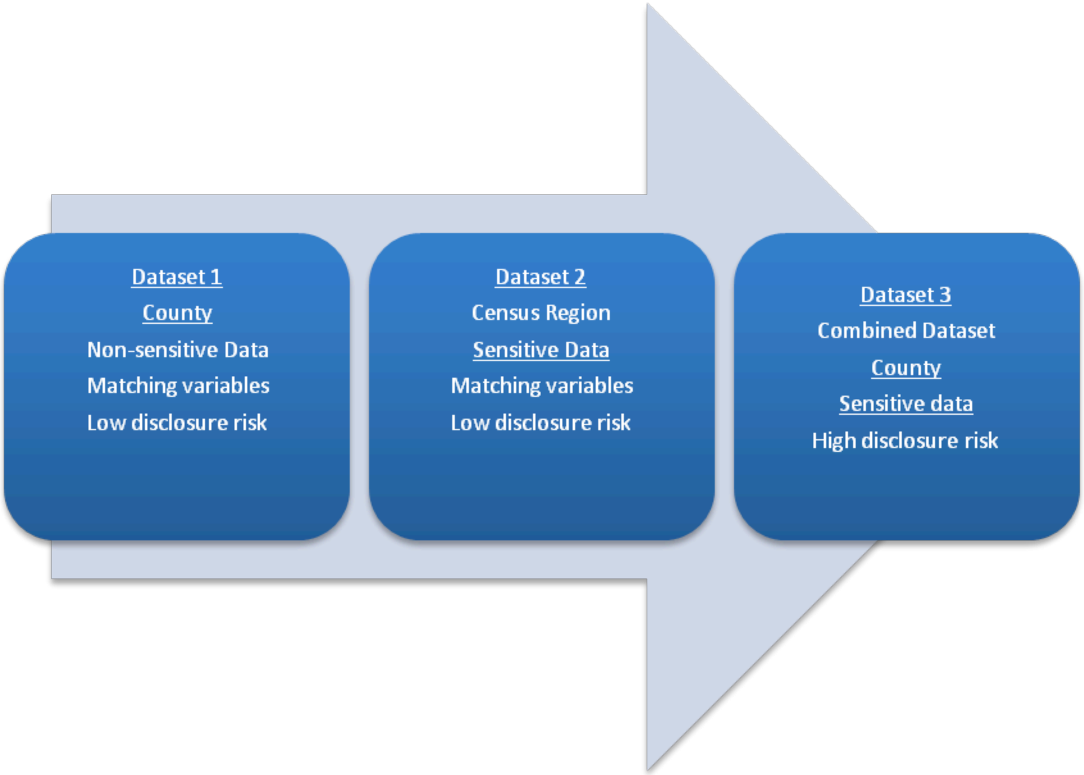
Of note is that future uses of data are difficult to determine at the time of the disclosure analysis. For example, data about veterans with service prior to Iraq and Afghanistan can be used for comparative analysis to veterans with contemporary military service. These comparisons might include the nature of differences in the impact of military service, as well as duration and severity of impacts. However, the specificity regarding period or place of service is often removed from PUFs. Prior to the Iraq and Afghanistan wars, the significance of this type of analysis was perhaps not anticipated. When a particular and salient need arises, it is difficult for the data community to quickly respond, such as determining existing data that could provide answers to current questions. Moreover, when data are removed from the PUF, unless researchers are familiar with the original data collection or take the extra step to examine the questionnaire and other survey instruments, if they are publicly available, they will not realize that certain questions and response categories were asked of respondents but that those variables were withheld from the PUF.

Other examples of sensitive variables that may be altered include (1) lesbian, gay, bisexual, and transgender (LGBT) codes that are collapsed to a single, dichotomous LGBT category, or depending on frequencies and disclosure risk, removed from the file (2) race codes such as white, black, American Indian, Asian, Pacific Islander, and other race that are recoded to white, black, and other (3) ethnicity codes such as Mexican, Cuban, Chicano, Mexican American, Other Spanish that are transformed to a single, dichotomous Hispanic code and (4) continuous years of education, training, and other schooling that is coarsened to broad categories. These alterations to a datafile have different implications, depending on the type of desired statistical analysis (e.g., cross-tabulations, means, regression) and content of the analysis (e.g., differences among sub-populations).

Often, a single variable or small group of variables alone do not pose a risk. It is when the variables are combined with geographic variables that disclosure risk becomes of concern. This is because, typically, to identify a respondent, one begins by narrowing the search to a geographic location. For example, detailed health or mental health diagnostic codes alone may not be problematic but when they are combined with detailed geography, such as county or primary metropolitan statistical area (PMSA), as well as demographic data such as age, race, ethnicity, marital status, and number of children it becomes easier to isolate cases. This problem is magnified by the collection of data for multiple persons per household. Geographic data may also be problematic when common variables are available for data linkage across two datasets, as shown in Figure 1. Each file may have low disclosure risk but when combined using common variables between the files, they create a superset of variables with high disclosure risk due to the geographic specificity and variable sensitivity of the superset.

Figure 1.

Data linkage between non-sensitive and sensitive data with varying geographic specificity and disclosure risk.



Depending on the type of risk, disclosure analysis may examine risk, both within the sample and within the population. For characteristics that can be known, or observed, such as demographics, Census and other publicly available data can be used to determine population denominators. This usually allows for a more robust PUF.

A record may be unique in the sample but not in the population. For example, a record in a study may include a pregnant Hispanic veteran in Milwaukee and this may be the only record with these characteristics in this geographic location. However, given the timeframe of the study and using Census data, it could be determined that there were many pregnant Hispanic veterans in Milwaukee and therefore, that the single record in the sample was not a disclosure risk. Therefore, the record could be determined to be of low (acceptable) risk because it was selected from among many similar records within the population yet be the only record of its kind within the sample.

Conversely, a study may sample a high proportion of the population under examination (e.g., people in a rural region affected by a health condition). Study participants may be known to each other (e.g., most participants were treated at the same specialty health facility), thus creating high disclosure risk. Certain sampling designs, such as snowball designs, create unique disclosure risks. With snowball designs, study participants are asked whether they know others with characteristics that fit the study's definitional criteria. The The fact that respondents may know each other must be factored

Uneven treatment of data

Alterations to the data may disproportionately affect certain sub-populations regarding the release of variables and response codes on the PUF. The uneven treatment of variables is due to some variables making records stand out in a data collection and thereby, increasing disclosure risk for respondents. This may include any sensitive variables, such detailed race and ethnicity. Geographic data are frequently coarsened due to disclosure risk. However, these types of data provide the nuanced understanding of the broader topic under study in which researchers are often the most interested.

A particular problem with coarsening race and ethnicity codes is that disclosure protections are often applied to all records, though risk may be localized. This can lead to detailed race and ethnicity codes being grouped together (e.g., white, black, other). However, when disclosure risk is considered in a nuanced manner, more detailed data can typically be retained on the PUF. This provides researchers the ability to fully analyze and explain the data and distinctions within it. The blunting of codes impacts every type of research (e.g., health disparities) due to with withholding of geographic data, race and ethnicity codes and other types of demographic data. This issue is perhaps best understood with an example.

When considering risk for unique records in a national file, the risk for Asian Americans will appear different depending on whether one considers the distribution of this group as a whole or at a detailed level. As shown in Table 1, the racial category "Asian alone" represent 4.8 percent of the U.S. population. As shown in Table 2, three sub-groups within the "Asian alone" category include about half (46%) of the "Asian alone" population, as follows: Asian Indian (.9%), Filipino (.8%), and Vietnamese (.5%). Given the cultural differences within these sub-groups, retaining the detail for race / ethnicity could provide a great enhanced public-use file, if it is possible to do so.

Table 1. Race Alone or in Combination (Census Table AT-P5, 2010, Summary File 1)

Subject	Number	Percent
Total population (all races)	308,745,538	100.00%
WHITE		
White alone or in combination [1]	231,040,398	74.83%
White alone	223,553,265	72.41%
White in combination	7,487,133	2.43%
Not White alone or in combination	77,705,140	25.17%
BLACK OR AFRICAN AMERICAN		
Black or African American alone or in combination [1]	42,020,743	13.61%
Black or African American alone	38,929,319	12.61%
Black or African American in combination	3,091,424	1.00%
Not Black or African American alone or in combination	266,724,795	86.39%
AMERICAN INDIAN AND ALASKA NATIVE		
American Indian and Alaska Native alone or in combination [1]	5,220,579	1.69%
American Indian and Alaska Native alone	2,932,248	0.95%
American Indian and Alaska Native in combination	2,288,331	0.74%
Not American Indian and Alaska Native alone or in combination	303,524,959	98.31%
ASIAN		
Asian alone or in combination [1]	17,320,856	5.61%
Asian alone	**14,674,252**	**4.75%**
Asian in combination	2,646,604	0.86%
Not Asian alone or in combination	291,424,682	94.39%
NATIVE HAWAIIAN AND OTHER PACIFIC ISLANDER		
Native Hawaiian and Other Pacific Islander alone or in combination [1]	1,225,195	0.40%
Native Hawaiian and Other Pacific Islander alone	540,013	0.17%
Native Hawaiian and Other Pacific Islander in combination	685,182	0.22%
Not Native Hawaiian and Other Pacific Islander alone or in combination	307,520,343	99.60%
SOME OTHER RACE		
Some Other Race alone or in combination [1]	21,748,084	7.04%
Some Other Race alone	19,107,368	6.19%
Some Other Race in combination	2,640,716	0.86%
Not Some Other Race alone or in combination	286,997,454	92.96%

X Not applicable.

[1] The race concept "alone or in combination" includes people who reported a single race alone (e.g., Asian) and people who reported that race in combination with one or more of the other race groups (i.e., White, Black or African American, American Indian and Alaska Native, Native Hawaiian and Other Pacific Islander, and Some Other Race). The "alone or in combination" concept, therefore, represents the maximum number of people who reported as that race group, either alone, or in combination with another race(s). The sum of the six individual race "alone or in combination" categories may add to more than the total population because people who reported more than one race are tallied in each race category.

Source: U.S. Census Bureau, 2010 Census.
Summary File 1, Tables P3 and P6.

NOTE: For information on confidentiality protection, nonsampling error, and definitions, see http://www.census.gov/prod/cen2010/doc/sf1.pdf.

TABLE 2. Asian Alone by Selected Groups

	United States	PERCENT U.S.	PERCENT ASIAN
	Estimate	POPULATION	POPULATION
TOTAL U.S. POPULATION	308,745,538	100.0%	--
ASIAN ALONE	14,727,806	4.8%	100.0%
Asian Indian	**2,765,155**	**0.9%**	**18.8%**
Bangladeshi	115,037	0.0%	0.8%
Cambodian	264,080	0.1%	1.8%
Chinese , except Taiwanese	3,291,388	1.1%	22.3%
Filipino	**2,512,686**	**0.8%**	**17.1%**
Hmong	245,807	0.1%	1.7%
Indonesian	77,104	0.0%	0.5%
Japanese	774,104	0.3%	5.3%
Korean	1,456,076	0.5%	9.9%
Laotian	210,571	0.1%	1.4%
Malaysian	20,438	0.0%	0.1%
Pakistani	356,939	0.1%	2.4%
Sri Lankan	40,285	0.0%	0.3%
Taiwanese	165,524	0.1%	1.1%
Thai	177,445	0.1%	1.2%
Vietnamese	**1,625,365**	**0.5%**	**11.0%**
Other Asian	496,039	0.2%	3.4%
Other Asian, not specified	133,763	0.0%	0.9%

2010 American Community Survey 1-Year
Estimates, ACS Table BO2006
Universe: Total Asian alone population

Although the American Community Survey (ACS) produces population, demographic and housing unit estimates, for 2010, the 2010 Census provides the official counts of the population and housing units for the nation, states, counties, cities and towns.

Total includes people who reported Asian only, regardless of whether they reported one or more detailed Asian groups.

Other Asian. Includes people who provided a response of another Asian group (such as Burmese); and includes people who provided multiple Asian responses.

Other Asian, not specified. Includes people who answered the "Other Asian" response category and did not provide a specific group; and includes people who provided only a generic term such as "Asian."

While the 2010 American Community Survey (ACS) data generally reflect the December 2009 Office of Management and Budget (OMB) definitions of metropolitan and micropolitan statistical areas; in certain instances the names, codes, and boundaries of the principal cities shown in ACS tables may differ from the OMB definitions due to differences in the effective dates of the geographic entities.

Estimates of urban and rural population, housing units, and characteristics reflect boundaries of urban areas defined based on Census 2000 data. Boundaries for urban areas have not been updated since Census 2000. As a result, data for urban and rural areas from the ACS do not necessarily reflect the results of ongoing urbanization.

Source: U.S. Census Bureau, 2010 American Community Survey

Explanation of Symbols:

1. An '**' entry in the margin of error column indicates that either no sample observations or too few sample observations were available to compute a standard error and thus the margin of error. A statistical test is not appropriate.
2. An '-' entry in the estimate column indicates that either no sample observations or too few sample observations were available to compute an estimate, or a ratio of medians cannot be calculated because one or both of the median estimates falls in the lowest interval or upper interval of an open-ended distribution.
3. An '-' following a median estimate means the median falls in the lowest interval of an open-ended distribution.
4. An '+' following a median estimate means the median falls in the upper interval of an open-ended distribution.
5. An '***' entry in the margin of error column indicates that the median falls in the lowest interval or upper interval of an open-ended distribution. A statistical test is not appropriate.
6. An '*****' entry in the margin of error column indicates that the estimate is controlled. A statistical test for sampling variability is not appropriate.
7. An 'N' entry in the estimate and margin of error columns indicates that data for this geographic area cannot be displayed because the number of sample cases is too small.
8. An '(X)' means that the estimate is not applicable or not available.

Depending on how the altered variables are treated for the PUF, certain types of analyses may be precluded. For example, categorization prevents the ability to use measures of central tendency, such as means, whereas data swapping and imputation preserve this ability.

While restricted-use versions of data are available in some cases, the requirements and application procedures for accessing them are often stringent. Restricted-use data are available via licensing agreements and in some cases, online data analysis and virtual data systems, such as those provided by ICPSR and the National Organization for Research at the University of Chicago (NORC). Most data collections do not have both public-use and restricted-use versions available. Typically, the only file version available to researchers other than the original investigator is the PUF.[2] Of ICPSR's 7,567 data collections, 6,573 collections (87%) are public-use only; 800 (10.5%) are restricted-use (entire data collection is restricted) and 194 (2.5%) have both restricted- and public-use versions of the data.

A New Model

A model for conducting disclosure analysis that moves disclosure risk considerations to the beginning of the survey design, rather than waiting until data collection is finished, will create efficiencies throughout the survey process, as well as provide for a more robust set of data products to be released. Moving this analysis to the beginning of the survey design optimizes the development and release of data products because it forces planning and decision-making to a stage in the survey process prior to any data or information release. Therefore, decisions are made prior to any type of release that could prevent the most optimal set of data products from being developed.

Any information released regarding a data collection must be factored into a disclosure analysis, even when the error is mitigated (i.e., an unintended release of map showing the primary sampling units (PSUs), or data collection sites, that is eventually removed). The release must continue to be considered because it is unknown how many people read, downloaded, copied, forwarded, or saved the information. Some types of data releases cannot be mitigated, such a table of sample sizes by PSU published in a professional journal. Publishing PSU maps and sample sizes by geographic area typically lend little to analysis but rather are used for descriptive purposes. Both of these practices increase disclosure risk by pinpointing geographic areas and sample size characteristics based on geography, usually without analytic benefit. It is critical to consider all disclosure risks and the benefit of the information prior to publishing any data products based on geography.

The disclosure analysis must take into account that any information publicly released prior to the disclosure analysis can be combined with the PUF or other data products and potentially

[2] Some data can only be released in restricted-use format. Examples include surveys or variables that geographically pinpoint respondents or include enough detail about respondents, along with geographic information, to create high disclosure risk. One such example might include surveys of disaster survivors.

put respondents at risk. For these reasons, it is optimal to consider the disclosure risk resulting from all data products that will be publicly released from a given data collection at the beginning of the survey, proposal, or request for proposal (RFP) development. Data products may include publications, maps, tables, estimates, online analysis files, and public-use datafiles. If longitudinal follow-ups are planned, even if they are not funded at the time of the disclosure analysis but are desired and might be conducted, the risk presented by these additional data can be taken into consideration using this model. The model will facilitate the release of the longitudinal data once they are collected, without penalty for having released earlier waves of data.

This model consists of three steps: (1) Identify disclosure risks early (2) Adjust survey design, including the sample size, if desired (3) Identify preliminary statistical disclosure limitation (SDL) methodologies.

Identify disclosure risks (early)

Once the questionnaire and survey design take shape, disclosure analysis can begin because disclosure analysis starts with a few fundamental determinations, including:

- Key analytic uses
- Sensitive variables
- Unique risks
- Inherent disclosure protections (e.g., self-report data, recall error)[3]

Even preliminary data are not required for the above determinations to be made and for disclosure risk assessment to begin. Next, a preliminary report regarding risk can be completed, taking all data collection and planned releases into account. After the data are collected, final determinations regarding risk and SDL techniques can be agreed upon and the PUF created.

Adjust survey design

During the design stage, needed samples for given variables are determined for estimation and analysis. Examples of these variables are age, race, and gender. Expected distributions for these sub-groups by key analysis variables are examined to ensure sufficient samples during data collection.

If questions regarding disclosure risk are also considered at this point, the survey could be of greater utility and resources better directed. For example, disclosure risk is often created by indirect variables, taken together, along with geographic variables that create unique records.

[3] Survey questions or techniques requiring self-report and recall provide inherent disclosure protections because they add a measure of error to the data. Respondents are known to under- or over-report behavioral data, depending on social desirability (self-report data) (e.g., how many religious services did you attend in the last 3 months?) (Citation). Also, when asking respondents to recall an event of an earlier period, there will be a range of accuracy in responses (recall error) (Citation).

Examples of indirect variables include characteristics about a person that can be known, such as age, gender, ethnicity, education, number of children, and marital status. Such records may be unique in the sample and also within the population.

SDL methods that tend to provide a more robust PUF, and that can work well with disclosure risk created by unique records, include data swapping and multiple imputation. These methods, compared to methods such as coarsening categories, allow a broader array of analyses for the PUF. However, data swapping and multiple imputation require a suitable number of like records for the method to be implemented. Reiter and Drechsler (2010) have suggested *Sampling with Synthesis* whereby multiple imputation is used as an SDL technique with sensitive data. The authors use Census data to demonstrate this technique. This is a good method if the sample size is sufficient to support the imputations.

Under the new model, questions such as sample size and the disclosure risk created by record and variable uniqueness are considered during the design stage. If it is determined that a given sample size (or methodology) is required in order for a sensitive variable or response category to remain on the PUF, and this is important to the funder or investigator, the sampling plan can be adjusted. That is, the sample can be adjusted to ensure that enough respondents populate additional cells based on a preliminary analysis of disclosure risk.

On the other hand, if it is determined that a sufficient sample size will not be achieved to allow the sensitive variable or response code to be released in a PUF, or this will be too costly, the funder or investigator may decide to funnel resources in other ways. For a repeated survey, given questions or response categories can be used in one year (with an increased sample) and alternative questions asked in a different year.

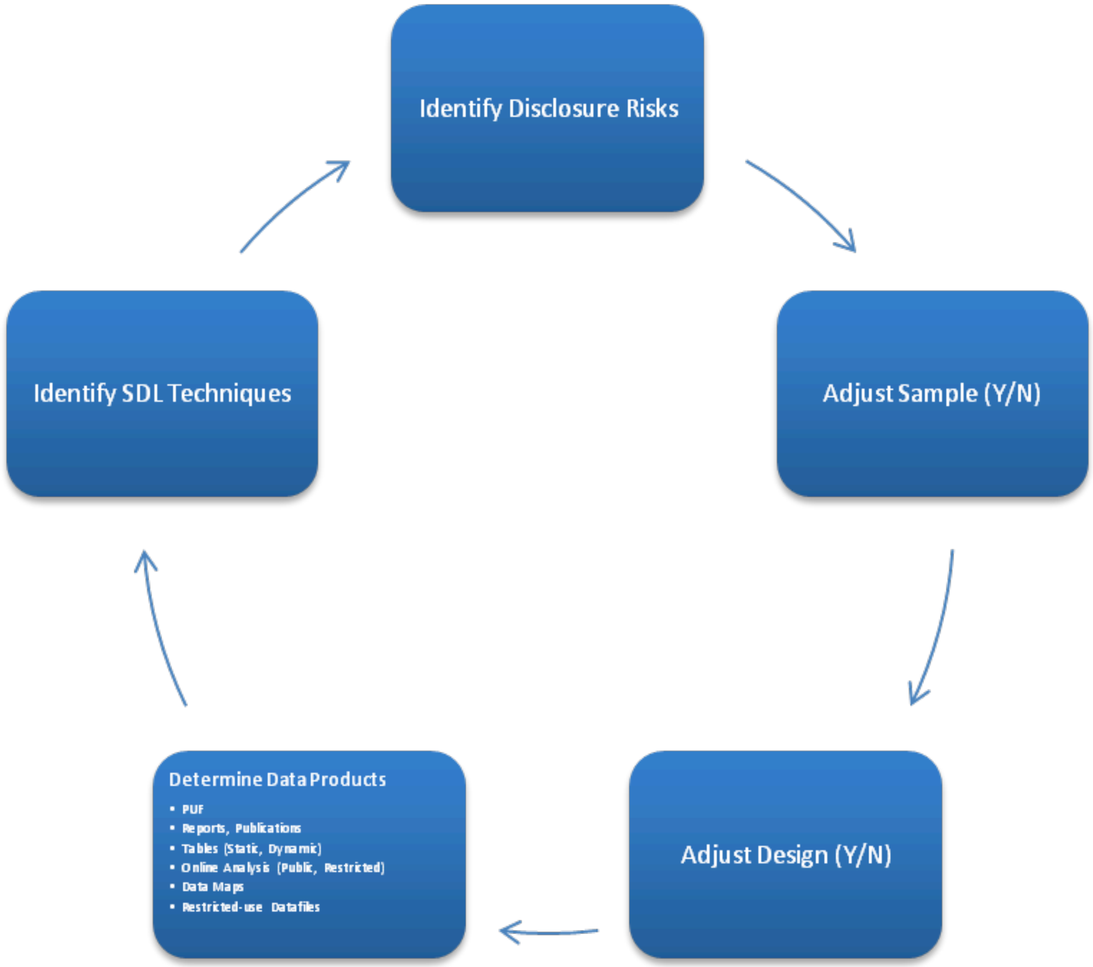
If the survey proceeds and the sensitive variable will not be released on the PUF and the variable is important for analysis (e.g., 9/11 variables, period of military service) plans can be made from the beginning of the survey for how to release the data, such as through a restricted online data analysis system or data portal. At a minimum, more fully developed dissemination plans than are currently envisioned at the outset of data collection can be constructed.

Identify SDL techniques

The last step in the early disclosure analysis process is to identify the SDL techniques that are likely to be used with the final data file. For example, a data swapping technique can be tested with early data to determine whether it appears that matches will be found for records to be swapped within the planned parameters. This allows for the method(s) to be tested with early data and if necessary, adjusted prior the close of data collection and cleaning. Algorithms for SDL and testing could, in fact, flow from the cleaning and editing processes. This will further shorten the time to data release. Identifying the SDL methods early will help ensure that the most efficient methods will be used based on the data type and key uses of the data.

Figure 2 shows the steps in the model for early identification of disclosure risks, including delineating the products that will be released. Planning publications, tables, and other types of releases helps to both ensure human subjects protection and that the most utility will be realized from the data. Considering all types of releases from the beginning of the survey process will minimize inadvertent errors.

Figure 2. Model for Early Identification of Disclosure Risks.



Best practices

Implications for best practice are multi-tiered, as they impact several aspects of the survey process. Moving the initial disclosure assessment to the survey design stage may require survey sample and design modifications based on the disclosure risk that is revealed. Examples of these possible modifications will include whether or not to:

- Increase the sample in order to retain certain questions on the PUF or utilize a given SDL (e.g., data swapping vs. coarsening);
- Eliminate certain questions from the survey (e.g., highly detailed ethnicity) because they will not be retained on the PUF and redirect resources accordingly; conceivably and depending on other risks and issues, this could lead to a suggested overall decrease in sample size;
- Ask certain questions less frequently (e.g., biennially) in repeated surveys to better utilize resources, depending on the analytic goals of the survey and the importance of the PUF relative to other goals.

To fully achieve the goals of the model, cooperation is required by investigators, data collectors, funding agencies, and data distributors. Better planning is also required to build funding for consulting regarding disclosure risk analysis into proposals, and even Requests for Proposals (RFPs), and also for making time for the work described in the front-end of the design process and then again, though more modestly, at the data release stage. This goes beyond planning for data dissemination alone and requires incorporating strategies for the PUF and data dissemination from the beginning of the survey process. If distribution will involve a new organization, it will also be necessary to bring that organization into discussions early.

Including disclosure risk as part of the planning process may or may not change the questions asked in a survey. Federal law mandates some questions in surveys sponsored by government agencies, such as those that produce certain components of economic forecasts and national estimates of crime victimization or child abuse. However, at a minimum, considering disclosure risk at the beginning of the survey process will make dissemination more efficient and allow for a smoother and faster data release. These changes will not only help create a more useful PUF but will help agencies, data collection organizations, investigators, and data disseminators create the strongest data dissemination plans possible, which will ultimately, speed the release of data products, and help ensure better human subjects protections, and more timely use of survey results.

To address the need for researchers and others who will want to know about modifications to data, changes from the original, restricted-use version should be documented in codebooks or other permanent texts accompanying a data collection. In this way, users can electronically search changes. The changes based on disclosure protections should summarize adjustments affecting key analyses so that researchers are aware of these modifications. Confidential changes to the data based on the disclosure protection that would put the plan at risk (e.g., the p-value for records that are exchanged with a data swapping technique) can be preserved in a restricted-use manner in order to retain the fidelity of the disclosure protection procedures.

Research agenda

As the model is put into practice and information is shared via conferences and publications, refinements can be made. Distinctions and particular challenges for different types of data will emerge. For example, longitudinal data carry increased risk because they have more data points over a period of time about the same subject. It is particularly important to ensure that longitudinal data are considered at the time of disclosure analysis and the early stages of survey design due to (a) the desire to release early waves of data and publications prior to all data collection being completed and (b) changes that may be made to later waves of the survey. Efficiencies that were gained from the model will also be important to share and discuss so that these can be capitalized.

The increasing availability of, and demand for, image data, such as biomedical scans, and video data underscores the need to consider disclosure risk early. These types of data, compared to numerically coded data, have more risk because they more uniquely identify individuals, and they do so with a single “variable” (image), and they are highly sensitive.

Moving disclosure considerations to the forefront of the survey process will heighten awareness regarding disclosure risk, as well as force consideration of all data releases for a given data collection. The result will enhance the ability to identify disclosure risks for all data products.

Educational implications

A basic and continuing educational implication is better preparation of graduate students in disclosure risk and disclosure analysis, including disclosure protection methods. The concept of disclosure risk and researcher responsibilities needs to be integrated into graduate education. Another implication is post-graduate training in disclosure analysis and SDL methods. Training is particularly important for non-statisticians so that social scientists engaged in human research, not only become more aware of the issues of disclosure risk but also understand steps to take to mitigate risks and where to turn for help.

References

Federal Committee on Statistical Methodology (2005). Statistical Policy Working Paper 22, Second Version, Report on Statistical Disclosure Limitation Methodology. Statistical and Science Policy Office of Information and Regulatory Affairs, Office of Management and Budget, Washington, D.C.

Fienberg, S.E., Martin, M.E., and Straf, M.L. (1985). Report of the Committee on National Statistics. In S.E. Feinberg, M.E. Martin, and M.L. Straf (Eds.), Sharing Research Data. Washington, D.C.: National Academy Press.

Fienberg, S.E. and J. McIntyre. (2005). Data Swapping: Variations on a Theme by Dalenius and Reiss, Journal of Official Statistics, 21, 309-323.

Health Insurance Portability and Accountability Act (HIPAA) of 2002, Federal Register …

O'Rourke, JM, Roehrig, S, Heeringa, SG, Reed, BG, Birdsall, WC, Overcashier, M., Zidar, K. (2006). Solving Problems of Disclosure Risk While Retaining Key Analytic Uses of Publicly Released Microdata, Journal of Empirical Research on Human Research Ethics, 1(3), 63-84.

O'Rourke, J.M. (Fall, 2003). Disclosure Analysis at ICPSR. ICPSR Bulletin, Volume XXIV, No. 1. http://www.icpsr.umich.edu/files/ICPSR/org/publications/bulletin/2003-Q3.pdf.

Reiter, J.P. and Drechsler, J. (December, 2010). Sampling With Synthesis: A New Approach for Releasing Public Use Census Microdata, Journal of the American Statistical Association, 105 (492), 1347-1357.

U.S. Bureau of the Census. (2003). 2000 Census of Population and Housing, Public Use Microdata Sample, United States: Technical Documentation. U.S. Census Bureau.

Zayatz, L. (2002). SDA in the 2000 U.S. Decennial Census. In Inference Control in Statistical Databases. Vol. 2316, Pages 193-202.