Procedures to Reduce the Risk of Respondent Disclosure in a Public-Use Data File: The National Immunization Survey

Meena Khare [1], Michael P. Battaglia[2], and David C. Hoaglin[2]

[1] National Center for Health Statistics, 3311 Toledo Road #3218, Hyattsville, MD 20782, mxk1@cdc.gov;
[2] Abt Associates Inc., 55 Wheeler Street, Cambridge, MA 02138; mike_battaglia@abtassoc.com, Dave_Hoaglin@abtassoc.com

1. Introduction

Rapid advances in information technology have made it feasible to collect, store, transfer, analyze, and disseminate large amounts of data very quickly to a wide range of audiences. Some data files contain sensitive information that can be linked, matched, and merged with exogenous data files. Small cells or subgroups (e.g., with size ≤ 5) of individuals with unique characteristics raise the risk of disclosure. Therefore, agencies and organizations must be extremely careful with the contents of the data files that they release. They must balance the risk of disclosure against the information needed for conducting research or preparing reports, in order to protect the identity of the individuals and the confidentiality of their data. A workshop in January 2002 dealt with "Confidentiality, Disclosure, and Data Access." Papers in a companion book (Doyle *et al.,* 2001) discuss issues related to risk of disclosure and protecting confidentiality, provide guidelines on theory and practical application for statistical agencies, and also discuss methods to control or reduce risk of disclosure with micro-data and tabular data. An American Statistical Association website (ASA, 2003) gives a comprehensive list of references and resources on privacy, confidentiality, and data security.

The National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention (CDC), collects data on vital statistics of the U.S. population and conducts a number of health surveys. Public-use data files (PUFs) containing micro-data for individual respondents are routinely released by the NCHS. Most PUFs contain basic demographic and socioeconomic information on the participants. The combination of these variables and detailed geographic identifiers increases the risk of inadvertently identifying an individual. The risk may be greater in a small geographic area, especially for survey respondents with rare characteristics. The risk of disclosure further increases with the availability of exogenous files that could be covertly matched with a PUF. Section 308(d) of the Public Health Service Act and the Privacy Act of 1974 provide guidelines on legal requirements to protect confidentiality of survey participants and their information (Zarate, 1998; NCHS, 1984).

Since 1994 the National Immunization Survey (NIS) has been collecting immunization data on children aged 19-35 months in the United States. Recently, the NCHS has released a series of annual NIS PUFs, each containing person- and household-level data on approximately 35,000 children and their mothers, along with state and urban area identifiers. The availability of these geographic identifiers facilitates extensive analysis of the rich data from the NIS. One goal is to allow analysis of the NIS data within the 78 Immunization Action Plan (IAP) areas, consisting of the 50 states, the District of Columbia, and 27 large urban areas. Similar to PUFs from other surveys, the NIS PUFs are also constrained by the requirement to protect confidential data and the identity of survey respondents. By combining a number of characteristics of children and mothers (say, in a cross-tabulation), intruders or data analysts might identify a unique respondent in the sample and/or in the U.S. population. The availability of external files and advances in information technology heighten this risk, even though NIS data users agree not to match the NIS data against other files (a warning and penalties are included in the README file that accompanies each PUF). The risks are not theoretical. Commercial databases routinely accumulate large quantities of detailed data on the U.S. population (Doyle *et al*., 2001, Chapter 3), and the computing power to search them is readily available.

This paper describes procedures used in the NIS to reduce the risk of respondent disclosure. A key concern is the availability of state-level exogenous natality files, which could be matched with a PUF. To reduce this risk, we used standard statistical disclosure limitation (SDL) methods (Doyle *et al*., 2001, Chapter 2) and developed an additional data-coarsening step, which involves identifying demographic cells in an exogenous natality file that contain a small number of children in the population. To prevent identification of children in the PUF, we applied a data-recoding procedure to those cells.

2. National Immunization Survey

Sponsored by the CDC, the NIS uses quarterly samples and random-digit dialing to monitor vaccination coverage among children aged 19-35 months in the United States (Zell *et al*., 2000; Smith *et al*., 2001). The NIS collects immunization histories from the parents/guardians of the sampled children (or from the most knowledgeable person in the household) and, with consent, from the children's vaccination provider(s). The NIS also collects demographic and socio-economic information from the household respondents, including geographic information such as city, county, and state of residence. Each PUF contains information on more than 1,000 variables, including child's age, sex, and race/ethnicity, and mother's age, race/ethnicity, and education, family income, and poverty status, along with Census Region and IAP area identifiers. Most of the other variables relate to the child's vaccination history.

Under Section 308(d) of the Public Health Service Act and the Privacy Act of 1974, the NIS staff promises to protect the identity of the respondents and preserve confidentiality of the information collected in the survey. For most research and reporting purposes, data from the NIS are disseminated in tables, published manuscripts, and PUFs on the Internet and on CD-ROMs. Prior to release, all PUFs are submitted for clearance to the NCHS Disclosure Review Board (DRB). The submission describes the methods used to preserve confidentiality and includes a checklist and a wide range of 1-, 2-, 3-, and 4-way tables. Items reviewed in the checklist include:

- Time lag between the data collection and release of the PUF
- Population size of each geographic area (must be at least 100,000 persons)
- Unique demographic characteristics of the respondents
- Unique socio-economic status of the respondents
- Unique identifiers used in the data file and their links (if any) to the unique characteristics of the respondents at geography or socio-demographic levels
- Number of respondents in unique cells in 2-, 3-, or 4-way tables by smallest geography (cell sizes $\leq$ 5 cases)
- A list of exogenous files that contain similar information
- Information on imputation of data items, and the frequency distribution of all variables (i.e., a codebook of the PUF)
- Sample design variables or information (e.g., sampling fractions or probability of selection) that is included or excluded in the Data User's Guide.

The checklist for the NCHS DRB is modeled after the checklist proposed by the Office of Management and Budget (1999). The review and clearance process involves extensive collaboration between the DRB members and the survey representatives prior to preparing the final data file for release. The README file accompanying each PUF contains warnings to data users, including a reminder not to match a PUF with any exogenous file and penalties for using data for other than research purposes (i.e., variables in the PUF should be used only for research and policy making).

3. Procedures Used in the NIS To Maintain Confidentiality

The NIS PUF uses a number of statistical disclosure limitation procedures to preserve data confidentiality and reduce risk of disclosure. They include standard data-coarsening methods such as top- and bottom-coding, dropping sensitive variables or records, collapsing categories of selected variables, and imputation of missing data. A comprehensive list of such SDL methods is included in Doyle *et al*. (2001). It is common in the SDL literature to avoid identifying geographic areas with fewer than 100,000 inhabitants, and in some instances an even higher minimum may be appropriate. Only two IAP areas have population close to (but above) 200,000, and the others have over 500,000 persons. The identification of geographic area was limited to the Census Region, state, and IAP area. The removal of variables such as MSA status, county, city, and ZIP code ensures that we do not identify any subareas within the 78 geographic areas.

We also applied collapsing methods to the basic demographic and socio-economic variables within each IAP area. The use of data recoding for variables such as race/ethnicity and age reduces the chance of being able to identify a child in a race group that is rare within a geographic area. Top-coding of family income ensures that high-income households cannot easily be identified (within the geographic areas). We also omitted all date variables (e.g., interview date, date of birth, and vaccination dates). The deletion of such variables substantially reduces the disclosure risk if someone covertly tries to match the PUF with an exogenous file. The PUF does contain variables that express dates in number of months and number of days from the child's date of birth.

In the next section we discuss methods to reduce risk of matching data for any NIS respondent with the data from an exogenous file. Details are included for an additional data-coarsening step that we developed for the NIS PUF using an exogenous file.

4. Exogenous Files

Despite warnings against such actions and extreme penalties, a data intruder could covertly attempt to match the NIS PUF with an exogenous natality population file. The exogenous population file, in theory, may contain all of the children in the target population. We assume that both the PUF and the exogenous natality population file contain categorical variables A, B, C, and D as well as IAP area of residence. (Variables A, B, C, and D have 3, 4, 2, and 3 categories, respectively.) To reduce the chance of identifying children in the NIS PUF, it is necessary to determine whether cross-classification of variables A, B, C, and D for an IAP area yields cells in which the sample contains most or all of the children in the population. The basic idea is to coarsen data in the small (rare) population cells by applying a technique that distorts data records before the PUF is released. The data-coarsening procedure can be classified as a perturbation technique because we actually change (recode) data values for a small number of cases before the PUF is released. Several methods have been suggested for making the assessment of which cells are at risk for disclosure.

First, one can examine the unweighted cross-tabulation of variables A, B, C, and D within each IAP area. The basic idea is to identify cells with sample counts below a specified minimum value (5 is commonly used). If the sample is self-weighting, such cells suggest that the corresponding number of children in the population may also be small. The strength of that suggestion, however, depends heavily on the sample size. If the sample size is small relative to the population size, the cross-tabulation will tend to contain many cells with small sample counts, even though the corresponding population counts are not small enough to warrant concern.

Second, one can look for cells in the cross-tabulation of variables A, B, C, and D (within each IAP area) whose weighted counts fall below some specified minimum value (again, 5 is commonly used). Cells with small weighted counts indicate that the population size in those cells may be small. Thus this approach has advantages over examining the unweighted cell counts. However, focusing solely on the weighted cell counts also has limitations. If one takes a simple random sample from a list frame, and numerous weight adjustments are *not* required, then the base sampling weight of each sample person (the reciprocal of the selection probability) can be viewed as a number of individuals in the population, though not necessarily in the same cell. The NIS screens a sample of telephone numbers to identify households with age-eligible children and ends up with a sample of children living in telephone households. Ultimately, the final sample weights incorporate an adjustment for multiple voice-use phone lines in the household, several levels of adjustment for unit nonresponse, poststratification to population control totals, adjustments to compensate for the exclusion of nontelephone households, and finally adjustments for children who do not have provider-reported vaccination data. The poststratification within each IAP area is at a more aggregated level than the cross-classification of variables A, B, C, and D. As a result, it is not appropriate to interpret a child's sample weight as indicating that the child represents an exact number of children within a cell in the cross-classification of variables A, B, C, and D.

A third method examines the population counts from the exogenous file in the cells produced by cross-classifying variables A, B, C, and D for each IAP area, to identify any population counts below a specified minimum value (again, 5 is commonly used). The advantage of this method is that it does not rely on the sample to identify small population cells; the assessment is made directly from the population.

A fourth method follows Land's (2002) approach of taking the difference between the population size (from the exogenous file) and the sample size to identify cells in tabulations that may need to be suppressed. This method could be applied to the NIS PUF, using variables A, B, C, and D to form the IAP area tabulations, and specifying a minimum acceptable difference. Land suggests using a difference of 10. In the context of the NIS Land's argument is that it is unlikely that one can identify a child in an IAP area who is not up-to-date on their vaccinations if at least 10 other children with the same demographic characteristics are not in the sample. One limitation of Land's method is that the difference could meet the specified minimum value, but the proportion of the population that is in the sample could be relatively high, say greater than 33%. Table 1 shows examples of this.

Table 1. Examples of cells in which the difference exceeds specified minimum value of 10 but the ratio is high (> 0.33)

Variable and Values				Unweighted Sample Size	Population Size	Difference	Ratio
A	B	C	D				
2	1	3	2	9	24	15	0.38
1	2	1	3	15	30	15	0.50

These examples suggest a fifth method: using the ratio of the sample size to the population size from the exogenous file to identify cells that exceed some minimum specified ratio. That is, the relative size of the sample is a better measure of disclosure risk than a measure based on the difference. For example, if the ratio in a cell were greater than 0.33, the risk of disclosure would be considerably greater than in a cell where the ratio was only 0.05.

Method 3 (population in cell from the exogenous file) does not rely on the size of the sample and was therefore chosen for application to the NIS PUF with 5 children as the minimum cell size. The fifth method, using the ratio of the sample size to the population size, also offers some additional benefits for identifying cells that may be at risk for disclosure. It was therefore decided to identify cells where the population size was 5 or fewer children for data coarsening, and to then supplement this rule by also looking for cells with a ratio greater than 0.33.

The application of Method 3 used the following steps:

1. Within each IAP area we cross-tabulated variables A, B, C, and D in the exogenous file to identify cells with 5 or fewer children in the population.
2. For each of the small cells, the NIS PUF was checked to see whether the sample contained one or more children.
3. For cells identified in Step 2, the value of variable A was recoded to a suitably chosen value for each child in the PUF.
4. After the recoding of variable A in the PUF, the tabulation of variables A, B, C, and D was repeated, and any cells in the exogenous file with 5 or fewer children in the population were identified.
5. For cells identified in Step 4, the value of variable B was recoded to a suitably chosen value for each child in the PUF.
6. This recoding process continued with variables C and D, if necessary, until all cells contained more than 5 children in the population.

The order of the variables in Step 3 through Step 6 is from least to most analytic importance (i.e., less-important variables are recoded first).

If one cross-tabulates variables A, B, C, and D in the NIS PUF and then makes the same tabulation for the exogenous file, the population counts in all of the cells will exceed 5, often by a considerable amount. Thus we never end up with a situation where the population count in a cell is small (e.g., we have no cells where we have 2 children in the sample and 4 children in the exogenous file). This reduces the chance that a data intruder will be able to match an individual child in the NIS PUF with the exogenous file. Table 2 shows an artificial example of the application of Method 3. The five cells listed in Table 2a have population size ranging from 2 to 5 and ratio of unweighted sample size to population size ranging from 0.20 to 0.60. Recoding Variable A from 1 to 2 for the two sample children in the first two rows moves them to cells whose population size is 23 and 14, respectively (Table 2b). Similarly, recoding Variable A from 3 to 2 for the three sample children in the last two rows places them in cells with large enough population size. For the three sample children in the third row, the appropriate recoding changes Variable B from 1 to 2, and the new cell contains 19 children in the population (Table 2c). After the recodings the ratio ranges from 0.18 to 0.35.

Table 2: Artificial example of applying Method 3 (recoded values are shown in bold)

a. Original table

Variable A	Variable B	Variable C	Variable D	Unweighted Sample Size	Population Size	Ratio
1	4	2	3	1	5	0.20
1	3	1	3	1	3	0.33
2	1	2	1	3	5	0.60
3	2	1	2	2	4	0.50
3	1	1	2	1	2	0.50

b. After recoding variable A

Variable A	Variable B	Variable C	Variable D	Unweighted Sample Size	Population Size	Ratio
2	4	2	3	8	23	0.35
2	3	1	3	3	14	0.21
2	1	2	1	3	5	0.60
2	2	1	2	4	17	0.24
2	1	1	2	2	11	0.18

c. After recoding variable B

Variable A	Variable B	Variable C	Variable D	Unweighted Sample Size	Population Size	Ratio
2	4	2	3	8	23	0.35
2	3	1	3	3	14	0.21
2	**2**	2	1	5	19	0.26
2	2	1	2	4	17	0.24
2	1	1	2	2	11	0.18

Other tabulation cells (not shown in Table 2) have population size greater than 5 but ratio of sample size to population size greater than 0.33. Also the recoding of data values (i.e., collapsing of data cells) can itself produce cells with a population size greater than 5 and a ratio that is now greater than 0.33, because of the larger numerator. This occurs in the first row of Table 2: the original ratio is 0.20, and after collapsing the ratio is 0.35.

In one application to the NIS PUF we identified 42 cells with a population size less than or equal to 5. After collapsing on variables A, B, C, and D, all cells contained more than 5 children in the population. However, for eight of the final cells, the ratio of the sample size to the population size was greater than 0.33. Thus, as indicated above, we applied a rule that coarsened cells in which the population size was less than or equal to 5 *or* the ratio was greater than 0.33. We proceeded in a similar fashion, starting with variable A and moving to variable B, etc. The cell collapsing stopped when all cells had more than 5 children in the population and each ratio of sample size to population size was less than 0.33. In the application less than 0.5% of the children in the PUF had one or more variables recoded. The combination of Method 3 and Method 5 produces a PUF where the population size in any cell matched with the exogenous file exceeds 5 and the ratio of sample size to population size is less than or equal to 0.33. This requirement protects against the possibility that someone will match the NIS PUF with an exogenous file and successfully identify individual children.

5. Summary and Conclusion

In releasing public-use files of data on individuals, organizations must often balance data users' desires for more-detailed data against the legal obligation to protect the confidentiality of respondents and control risk of disclosure. Users customarily agree not to match the PUF against other files, and a warning accompanying the PUF reminds them of this obligation, but a higher level of security is essential. Methods of statistical disclosure limitation take systematic steps to protect the identity of respondents and reduce the risk of disclosure for groups of individuals with rare characteristics. By reviewing PUFs prior to their release, the NCHS Disclosure Review Board seeks to ensure that risks of disclosure are acceptably low. In developing PUFs for the National Immunization Survey, we applied several existing disclosure-limitation methods. We also used an exogenous file to identify demographic cells that contain few children in the population; we then applied a data-recoding procedure to cells in which the sample count was small and constituted too high a fraction of the population.

References

American Statistical Association (2003). Privacy, Confidentiality, and Data Security Website, http://gill.amstat.org/comm/CmtePC/

Doyle P., Lane J.I., Theeuwes J.J.M., and Zayatz L.V. (eds.) (2001). *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*. Amsterdam: North-Holland.

Land G. (2002). "Confidentiality Data Release Rules," presented at the Assessment Initiative/NAPHSIS Leadership Institute meeting, January 2002, Minneapolis, MN.

National Center for Health Statistics (l984). *Staff Manual on Confidentiality*, Hyattsville, MD.

Office of Management and Budget (1999). *Checklist on Disclosure Potential of Proposed Data Releases,* prepared by Interagency Confidentiality and Data Access Group: An Interest group of the FCSM. Statistical Policy Office, OMB, July 1999.

Smith P.J., Battaglia M.P., Huggins V.J., Hoaglin D.C., Roden, A.-S., Khare M., Ezzati-Rice T.M., and Wright R.A. (2001). "Overview of the Sampling Design and Statistical Methods Used in the National Immunization Survey," *American Journal of Preventive Medicine*, 20(4S): 17-24.

Zarate A.O. (1998). "Legal, Administrative and Statistical Aspects of Confidentiality Procedures at the National Center for Health Statistics Presentation," paper presented as expert testimony on issues of privacy and confidentiality, for the public meeting on the President's Initiative on Immunization Registries, Atlanta.

Zell E.R., Ezzati-Rice T.M., Battaglia M.P., and Wright R.A. (2000). "National Immunization Survey: The Methodology of a Vaccination Surveillance System," *Public Health Reports*, 115: 65-77.