DISCLOSURE AVOIDANCE TECHNIQUES FOR THE 1994 NEHIS DATA PRODUCTS

David W. Chapman, Klemm Analysis Group, and Christopher L. Moriarity, Nat'l Center for Health Statistics
David W. Chapman, KAG, 1785 Massachusetts Ave., NW, 5th Floor, Washington, DC 20036

KEY WORDS: Confidentiality, Public Use Files, Establishment Surveys

1. Introduction

Maintaining the confidentiality of survey respondents is a fundamental objective in any government survey. It is also a basic goal to make data available to the public that are useful for research purposes. Since these two important goals are conflicting, it might be assumed that compromises between the two must be made in designing data products for users. (For a discussion of the tension between confidentiality and data access, see Duncan, et al. (1993), Chapter 1.) However, protecting confidentiality is an absolute goal that must not be compromised, and takes priority over the objective of providing data useful to researchers. As stated in the Manual on Confidentiality of the National Center for Health Statistics (NCHS) (1984, p. 5), data collected by NCHS surveys are protected by Section 308(d) of the Public Health Service Act which states that information collected in surveys "may not be published or released in any manner in which the establishment or person supplying the information or described in it is identifiable unless such establishment or person has consented."

The purpose of this document is to describe the issues involved in protecting the confidentiality of respondents to the National Employer Health Insurance Survey (NEHIS), and to recommend specific disclosure avoidance measures to be used in developing NEHIS data products. The 1994 NEHIS was a national survey of business establishments (i.e., individual business locations) and government agencies that collected detailed information on health insurance that employers provided for their employees in 1993. The basic design of the survey was a stratified random sample with states as the major stratifier because of the importance in NEHIS of producing state estimates. The information collected included the names and types of health insurance plans offered (if any), the number of employees eligible for insurance, the number of enrollees in various plans, specific coverage characteristics of plans, the costs of insurance for both employers and employees, and claims paid in 1993. Although data collection has been completed, no data will be published until 1996. As part of the preparation for releasing NEHIS data, we are beginning to develop the disclosure avoidance methods that will be used.

By disclosure avoidance methods, we refer specifically to measures taken to avoid the release of any survey results, as published tables or public use files (microdata), that disclose the responses of any specific survey participant. The terms "protecting confidentiality" and "disclosure avoidance" will be used interchangeably. The major goal in developing public use products will be to provide data that are valuable to users and that are adequately protected in terms of disclosure avoidance.

Disclosure avoidance is generally more difficult for an establishment survey like NEHIS, as compared to a household survey, because there are fewer establishments than households and there is greater size variation among establishments. Specifically, the size distribution among establishments is highly skewed, with most establishments having only a few employees, but a relatively small number of establishments having large numbers of employees. Protecting confidentiality for the larger establishments is a special concern because they are often highly visible.

For establishment surveys, disclosure avoidance is especially difficult for public use files (microdata). In fact, as reported in a major recent report on disclosure limitation by the Office of Management and Budget (1994, p. 20), "there are virtually no public use microdata files released for establishment data." This is because the amount of data that would have to be suppressed to guarantee confidentiality would make the microdata files of marginal value for research purposes. In a phone discussion in December, 1994, with Brian Greenberg of the Census Bureau, who is a member of the Census Bureau's Microdata Review Panel, Dr. Greenberg said that the Census Bureau concluded that it would not be possible, because of confidentiality concerns, to release any useful microdata files for the Business Census or any of the major Business surveys. Therefore, the challenge of releasing useful data to the public from NEHIS, especially in terms of public use files, is a difficult one.

Disclosure avoidance for either published tables or microdata can be broken down into two major aspects:

(1) Identifying circumstances in a data product that jeopardize the confidentiality of respondents, and

(2) Modifying the survey data, or the presentation of data, in some way to avoid disclosure.

We discuss these two aspects of disclosure avoidance for NEHIS in the following sections, for both published tables and public use files. We address risk detection and corresponding adjustments in Section 2 for published tables and in Section 3 for public use files. We summarize various approaches and make and specific recommendations for NEHIS in each section.

2. Disclosure Avoidance for Published Tables

There are a number of methods that various federal agencies have used to identify cells in proposed published tables for which confidentiality is jeopardized, and corresponding methods to mask these cells. Statistical Policy Working Paper 22, prepared by a subcommittee appointed by the Office of Management and Budget (1994), provides descriptions of many of these methods. The brief summaries of the procedures given in this section are abstracted from that report. [For simplicity, Statistical Policy Working Paper 22 hereafter will be referenced as OMB (1994).]

As discussed by OMB (1994, p. 10), considerable confidentiality protection is achieved in published tables whenever a sample survey is used instead of a complete census. With cell entries consisting of weighted up sample responses, rather than straight sums of unweighted responses, it is especially difficult to identify specific respondents. However, even though NEHIS is a sample survey, we cannot assume that confidentiality is automatically provided in published tables. In fact, there were some strata for which the establishments were selected with very high probabilities. For these strata, the sample is close to a census and the confidentiality protection associated with the selection of a sample is diminished.

A discussion of methods used to detect and mask sensitive cells is provided in Subsections 2.1 and 2.2, followed by recommendations for NEHIS in Subsection 2.3.

2.1 Procedures Used to Identify Sensitive Cells

Published tables can be classified into two basic types: frequencies (or percentages) and magnitudes (i.e., totals or means). For frequency tables, the most common rule for identifying sensitive cells is the **Threshold Rule**. This is simply a rule that identifies a cell as sensitive if it does not contain at least n respondents, where n is taken most often to be 3. Some agencies use a higher number like 5 or 10. For NCHS surveys, n is taken to be 3. [See p. 16 of the NCHS Staff Manual on Confidentiality, hereafter referred to as NCHS (1984).] In addition to the Threshold Rule, NCHS defines a cell as sensitive if it is the only non-empty cell in a row (or column), regardless of the number of entries the cell has (see NCHS, 1984, p. 16).

For tables of magnitudes, there are several rules that are used to identify sensitive cells, involving the level of dominance of one or more establishments in terms of the cell estimate. A straightforward rule used by many agencies is the **(n,k) Rule**, which identifies a cell as sensitive if n or fewer respondents account for k percent or more of the cell total. The most common version of this rule, which is the version used by NCHS, takes n=1 and k=60% (see NCHS, 1984, p. 16).

There are two more complex rules used by some agencies to identify sensitive cells for magnitude tables: the **p-Percent Rule** and the **pq Rule**. Both of these rules identify a cell as sensitive if lower or upper bounds for the largest reported value of a survey variable can be derived to be within p-percent of the actual value by a "coalition" of c respondents, where c is usually taken to be 1 or 2. For additional details of these rules, see OMB (1994), Chapter 4.

2.2 Procedures Used to Treat Sensitive Cells

For cells that are identified as sensitive, a number of methods have been developed to treat them. The most obvious approach is to suppress these cells (**primary suppressions**). In such cases, other cells have to be suppressed (**complementary suppressions**) so that the cell or cells suppressed cannot be derived from the marginal frequencies. Based on the set of primary suppressions identified, the selection of a corresponding minimum set of complementary suppressions needed to protect the primary cell suppressions can be complex, requiring linear programming methods (see Cox, 1980).

Of course, cell suppression diminishes the value of the tables for data users. Other alternatives have been developed which do not require cell suppression. One method is to collapse some of the rows or columns so that the revised table has no sensitive cells. Although it eliminates the need to suppress cells, this method also diminishes the value of the tables because of the combining of two or more categories of a variable.

Two other methods for protecting sensitive cells are **random rounding** and **controlled rounding**. Both involve rounding off the cell frequencies (e.g., to the nearest multiple of 5 or 10) in order to mask or disguise the data. Random rounding is more straightforward, but can provide cell frequencies that do not add to the original marginal frequencies. Controlled rounding forces cell frequencies to add to the original marginal frequencies, but requires linear programming procedures (see Cox and Ernst, 1982).

Another approach that can be applied to frequency tables, or to magnitude tables, has been developed by

the Census Bureau. It is referred to as a **confidentiality edit** and involves "switching" or "swapping" of survey responses between sets of respondents in different geographic areas that have similar demographic characteristics.

A major weakness of the rounding procedures and the confidentiality edit approach is that, although they may mask the true responses, they may give the appearance of allowing disclosure. Even though the survey documentation can state that the data have been modified to protect confidentiality, the perception of disclosure could cause some confusion and distrust among survey respondents.

A different approach that has been used by some agencies to deal with sensitive cells is to ask respondents to release the government from its promise of confidentiality. Although there are obvious advantages to this approach, it would be time consuming and awkward to implement, and it may have adverse effects on future requests of respondents to participate in NEHIS. Furthermore, if only a portion of the respondents give permission to release NCHS from its confidentiality pledge, there would still be the need to protect the confidentiality of the other respondents.

2.3 Recommendations for NEHIS

Because of the confidentiality protection that is provided because NEHIS is a sample survey, rather than a census, we recommend that only minimal (though important) checks be made to avoid disclosure in tables. This approach is consistent with other federal agencies. For example, OMB (1994, p. 30) states that the Census Bureau reports that "For economic magnitude data most surveys do not need disclosure analysis."

To identify sensitive cells in tables, it should be sufficient to apply the rules discussed above from the NCHS Staff Manual on Confidentiality. Specifically, for frequency tables, any cell with less than three respondents would be defined as potentially sensitive (Threshold Rule). We recommend that any cell of this type be examined to see how many cases there are in the cell in the entire sampling frame. If there are at least four frame cases in the cell, we recommend that this cell not be considered sensitive. In addition, if a cell is the only non-empty cell in a row or column in a frequency table, we recommend that it be identified as potentially sensitive. Any such cell would be checked to see if there are other cases in the row or column in the sampling frame. If so, we recommend that the cell not be considered sensitive.

For magnitude tables, we recommend that the (n,k) Rule be applied, with n=1 and k=60%. Specifically, any cell for which one respondent provides 60% or more of the value of the cell total would be defined as potentially sensitive. In such cases, we recommend that a final decision on the sensitivity of the cell be made based on the weight of the respondent. For example, if the weight is 1, so that the respondent's unweighted value accounts for 60% or more of the weighted cell total, the cell should be treated as sensitive. However, if the weight is 10 or more, then the unweighted value accounts for 6% or less of the weighted cell total. In such a case, the cell should not be treated as sensitive. A specific rule of thumb that we recommend is to classify a cell as sensitive only if the unweighted contribution of a single respondent exceeds 30% of the cell estimate.

The approach we recommend to mask any cells identified as sensitive is the method of cell collapsing. This would involve a judgment as to whether to collapse the corresponding row or column categories to remove the sensitive cell, and which rows or columns to collapse. Although there is some loss of information when response categories are combined, it is a straightforward approach and does not give the appearance that disclosures are revealed.

3. Disclosure Avoidance for Microdata Files

Although a public use file (PUF) of individual survey records can be a valuable tool to researchers, such a file poses a considerable threat to the confidentiality of survey respondents. With the availability of data from many outside sources, there is the potential for matching PUF records to other data files. The objective in developing a PUF is to limit the risk of disclosure to an acceptable level while still providing useful data for researchers.

In Subsections 3.1 and 3.2, we summarize methods that agencies have used to identify sensitive records and corresponding methods to protect these records. Recommendations for disclosure avoidance methods for NEHIS will be given in Subsection 3.3.

3.1 Identification of Sensitive Cases for Microdata Files

In general, federal agencies have not been able to use objective methods for identifying sensitive microdata records. As reported by Jabine (1993, p. 436), the major releasers of public use files have established procedures for reviewing these files which, unlike those for published tables, "do not rely on parameter-driven rules. Instead, they require judgements by reviewers who take into account factors such as: the availability of external files with comparable data, the resources that might be needed by

an 'attacker' to identify individual units, the sensitivity of individual data items, the expected number of unique records in the file, the proportion of the study population included in the sample and the expected amount of error in the data."

There are two main sources of disclosure risk for public use files (OMB, 1994, p. 62). The first is the existence of high visibility records: records for respondents with unique characteristics. For an establishment survey like NEHIS, there is considerable potential for high visibility records, because of the high skewness of the establishment size distribution. Generally, the high visibility records would be those corresponding to very large establishments (in terms of the number of employees) or establishments that belong to very large firms. In addition, an establishment could have high visibility if it were the only one of a specific type [i.e., standard industrial classification (SIC) code] in a given state or region.

The other main source of disclosure risk is the potential for matching the PUF with other external files that are available. For NEHIS, an obvious risk is a match of the private sector sample with the national establishment file available commercially from Dun and Bradstreet (D&B), since an earlier version of this file was used as the private sector sampling frame for NEHIS. As a minimum measure to protect confidentiality, basic identifiers from the D&B file must be deleted from the public use file.

There are several other files, in addition to the D&B file, that could be matched to the NEHIS PUF. These would include both government and private files. The government files would include those maintained by the Bureau of the Census, the Bureau of Labor Statistics, and the Internal Revenue Service. Some of the private organizations that have files developed from health care surveys that could be matched to the NEHIS include the Health Insurance Association of America, the Foster Higgins Company, the Robert Wood Johnson Foundation, and others.

Although they have not been widely accepted, several mathematical measures of risk for microdata have been proposed (OMB, 1994, pp. 64-65). These methods involve estimating various probabilities associated with disclosure: for example, the probability that the respondent for whom the "intruder" is looking for is contained on both the PUF and on some other file available for matching. It is certainly possible that one or more of these mathematical measures could be useful in terms of assessing the risk associated with matching the NEHIS PUF to another available file. However, the time and resources may not be available to investigate and apply these approaches.

3.2. Procedures Used to Treat Sensitive Cases in Microdata Files

The first step that federal agencies take to avoid disclosure of responses from microdata records is to suppress all of the basic identification information, such as establishment name and address. In addition to suppressing the address, agencies also suppress additional geographic information (e.g., county and state). Jabine (1993, p. 436) reports that "The Census Bureau and National Center for Health Statistics specify that no geographic codes for areas with a population of less than 100,000 can be included in public use data sets." With establishment microdata, a much higher cutoff would presumably be required, depending on the other characteristics (e.g., SIC code and size measures) that are included in the public use file.

In addition to suppressing geographic information, agencies must consider suppressing other variables that can be used to identify a specific respondent. For an establishment survey, variables of this type include SIC code and size measures. An alternative to suppression, discussed later in this subsection, is to combine categories of such variables to prevent identification of high visibility cases. The decision as to the suppression of geographic identifiers and other file characteristics requires careful examination of the structure of the proposed microdata file and other files available for matching.

OMB (1994, p. 63) reports that one approach to protecting confidentiality in public use files is to provide only a sample of the population. With only a sample of establishments available in a public use file, an "intruder" would have difficulty matching records to another source (e.g., the D&B file) which contains all, or almost all, of the establishment records in the population. In addition, because a sample survey generally has both unit and item nonresponse, and imputed responses for some missing items, it is more difficult to match to another source.

In addition to suppressing data and using sampling, several methods are used by agencies to modify the reported data to help protect confidentiality in its public use files. One of these methods is to recode continuous variables (e.g., number of employees or premium amounts) into class intervals (categories). This method may include using "top codes" or "bottom codes" for values of a highly visible variable which puts together all responses greater (or less) than a specific threshold chosen to guarantee an adequate number of respondents in the top (or bottom) category of a variable. A related method is to combine outcome categories into fewer categories. Although some detail of information is lost with these approaches, the basic magnitudes of the

variables are preserved.

In addition, there are several methods which actually alter the reported values to protect confidentiality. One of these methods is to add a random component or "noise" to the responses. Other methods include "rounding" responses to an adjacent round number and "blurring" reported values. With blurring, reported values are replaced by "average" values computed across a group of respondents. A final method of altering the responses is the method of "swapping," or switching responses, discussed in Subsection 2.2. Some additional details and related references for these and related methods are provided by OMB (1994, pp. 66-67).

Methods of altering the reported data (also referred to as **disturbing** the data) have three basic weaknesses. First, they introduce error into the data which will reduce the precision of estimates. Second, some of these methods require considerable time and resources to develop and apply. Third, the altered data may not appear to be sufficiently masked to protect confidentiality. Even though the perception of disclosure may be false, it could cause some confusion and hard feelings among survey respondents.

3.3 Recommendations for Disclosure Avoidance for Public Use Files

The first step that should be taken to protect confidentiality is to suppress name, identification numbers, and basic geographic identifiers from each record. Several other measures still need to be taken to identify and treat highly visible cases, and to prevent the matching of the NEHIS PUF to other data sources.

To identify highly visible records for NEHIS in terms of size or type, we recommend that basic tabulations be generated of the number of establishments in a geographic area by size categories, by major SIC groups, and by size by SIC group cross-classifications. This may give some idea as to the extent of high visibility cases for a PUF for alternative geographic identifiers.

As more variables are included on the PUF, identifying highly visible records becomes more complex. As a general approach, we propose to identify a set of variables suggested for inclusion in the PUF. From these, define a subset, S, which are basic variables that appear to have potential for defining high visibility cases. These variables would include geographic area, any size measures (e.g., number of employees in the establishment or firm), SIC group, public or private, and others. We recommend that subject matter experts be used to help identify the subset S since it plays a critical role in the procedure. Assuming that all of the variables in S are defined as categorical (and this can certainly be done), they can be used to define a multi-way cross-classification table. If a cell of this table has only one or two establishments in it, these establishments are highly visible.

Depending on the number and type of high visibility cases identified, we recommend that the PUF be revised, as needed, to eliminate such cases. This would be done by either suppressing one or more of the variables in S or, preferably, by collapsing some of the categories in one or more of the variables in S. In some cases, the collapsing of categories could be equivalent to using top codes or bottom codes, discussed in Section 3.2. It is recommended that subject matter experts assist in the process of collapsing categories, at least in terms of establishing priorities.

In terms of preventing a matching of the NEHIS PUF with other available files, we recommend that subject matter experts be consulted to identify available files, in addition to the D&B file, that could be used to match to NEHIS. Some of the alternate files that should be investigated in this regard were noted in Section 3.1.

Once the files are identified that potentially could be matched to NEHIS, we recommend that the variables in each of these files be obtained and compared to the variables intended for inclusion in the NEHIS PUF. Of course, the more variables that two files have in common, the better the chances are that records can be linked. Regardless of the number of variables that two files have in common, detailed comparisons of the records from the two files would be required to be sure that matches are possible between them.

This type of extensive analysis would not be practical to make for NEHIS, except perhaps for comparing the NEHIS PUF and D&B file. However, we recommend that at least some basic comparisons between NEHIS and alternative sources be made. These would include comparisons of the target populations, the variables and corresponding categories included, the level of sampling involved, and the accessibility of each alternate file.

To the extent that it seems necessary to avoid the possibility of matching the NEHIS PUF to other data sources, we recommend that the categories of PUF variables be collapsed. It is also possible that some variables would have to be deleted from the NEHIS PUF to prevent matches. We highly recommend that subject matter experts assist with these comparisons.

In developing the PUF for NEHIS, there is a fundamental question of whether or not state identifiers should be included. A primary goal of the 1994 NEHIS is to provide baseline data to help evaluate the impact of health care reform. Since many of the health care

reform initiatives are generated at the state level, it would be valuable to provide a PUF with state identifiers. However, there is concern about confidentiality protection if state identifiers are included. It may be that for some of the smaller states there is only one very large establishment, or only one establishment in a major SIC group.

As a result of the uncertainty as to whether state identifiers should be included on the file, the following three options for a PUF for NEHIS are being considered:

(1) National PUF with no state identifiers

(2) National and state-level PUF with state identifiers

(3) Separate national and state-level PUFs

Although there would still be considerable effort involved in preparing the file and minimizing the risk of disclosure, the first option would be the easiest and safest of the three to produce. In terms of geographic identifiers for this option, we recommend that nothing below Census Region be included. It is anticipated that Regions may be large enough so that many other characteristics, such as size and SIC group, could be included on the PUF without jeopardizing respondent confidentiality.

The most problematic of the three options would be the development of a single PUF with state identifiers. It is possible that if state identifiers are included on the PUF, very little information of value can be provided without jeopardizing the confidentiality of NEHIS respondents. However, it is recommended that this option be given the first priority in the development of a PUF because of its potential value to the research community. That is, a PUF with state identifiers should be created if it could include enough valuable microdata without compromising NCHS's confidentiality pledge.

The potential advantage to providing both national and state-level PUFs (Option 3) is that users that need state-level data would have some limited information available from the state-level file while users who only need national data would have considerably more data available on the national file. There are two major problems with this approach. First, it would involve more time and resources to prepare than either of the single-file approaches discussed above. Second, there is the additional risk that the two files could be matched to each other which would probably lead to numerous disclosures.

In order to prevent matching the two files, no continuous variables would be allowed on the state-level file, since such variables could provide a fairly detailed match between respondents in the two public use files. Even with these precautions there would still be the concern of matching the two files on the basis of the respondent weights. Therefore, methods would have to be developed to prevent the two files from being matched on that basis. Possible procedures include modifying the respondent weights in one of the two files in some way, such as rounding the weights or adding "noise" to the weights.

Finally, if it is determined that a useful PUF cannot be developed with state identifiers, it is suggested that a contractor be hired to serve as a clearinghouse to provide desired analyses for states. The contractor would be funded by the states and would be authorized to work with the microdata records.

References

Cox, L.H. (1980), "Suppression Methodology and Statistical Disclosure Control," Journal of the American Statistical Association, Vol. 75, pp. 377-385.

Cox, L.H. and Ernst, L.R. (1982), "Controlled Rounding," INFOR, Vol. 20, No. 4, pp. 423-432. Reprinted: Some Recent Advances in the Theory, Computation and Application of Network Flow Methods. University of Toronto Press, 1983, pp. 139-148.

Duncan, G.T., Jabine, T.B., and de Wolf, V.A., eds. (1993). Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics. Panel on Confidentiality and Data Access, Committee on National Statistics. Washington, DC: National Academic Press.

Jabine, T.B. (1993), "Statistical Disclosure Limitation Practices of United States Statistical Agencies," Journal of Official Statistics, Vol. 9, No. 2, pp. 427-454.

National Center for Health Statistics (September, 1984), NCHS Staff Manual on Confidentiality, U.S. Department of Health and Human Services, Public Health Service, Hyattsville, MD.

Office of Management and Budget (May, 1994). Statistical Policy Working Paper 22: Report on Statistical Disclosure Limitation Methodology. Subcommittee on Disclosure Limitation Methodology, Federal Committee on Statistical Methodology.