# THE FCSM CONFIDENTIALITY AND DATA ACCESS COMMITTEE

Alvan O. Zarate, Jacob Bournazian and Virginia de Wolf

Paper presented at the FCSM Statistical Policy Seminar: Integrating Federal Statistical Information and Processes November 8-9, 2000

## ABSTRACT

The Confidentiality and Data Access Committee (CDAC) operates as an interagency "Interest Group" of the Federal Committee on Statistical Methodology (FCSM) to promote cooperation and sharing of information concerning data access issues and statistical disclosure methods among federal agencies. CDAC has developed several products relating to the confidentiality review of data intended for general use and restricted access to confidential data, for example a "Checklist on the Disclosure Potential of Proposed Data Releases" to assist agencies in preparing tabulated and micro data for public release. Other projects underway include a brochure describing confidentiality, data protection and data access procedures among federal agencies as well as the development of "auditing" software that would assess the degree of protection afforded tables that contain confidential information. In addition, members of the group have developed workshops and training modules for special presentations at professional meetings and elsewhere, served on ad hoc disclosure review panels, and contributed expertise to the development of privacy regulations.

**Key words: Disclosure limitation, Restricted access, Publicly available data**

## Introduction

An increasing number of electronic data products are released to the public by federal agencies both as separate products such as CDROMs as well as via the Internet. At the same time , the proliferation of publicly available electronic data bases and advances in computer technology have heightened concerns among federal statistical agencies about the protection traditionally used for public -use microdata as well as tabular data. These developments have also stimulated the consideration of whether (and how) to provide selective and controlled access to microdata files containing identifiable information collected under an assurance of confidentiality.

Faced with these issues a group of federal statisticians initiated informal contact late in 1995, and in early 1996, the first meeting of the Interagency Confidentiality and Data Access Group (ICDAG) took place. Since renamed the Confidentiality and Data Access Committee (CDAC), it serves as a special interest committee on data access and confidentiality issues for the Federal Committee on Statistical Methodology (FCSM) and is comprised of staff members from federal agencies who work in the "confidentiality area". These include those who work at statistical agencies as well as staff of nearly any agency confronted by problems dealing with privacy, confidentiality, or statistical disclosure. Nearly every federal department in the Executive Branch is represented, with staff of the Bureau of the Census, the Department of Health and Human Services, the Bureau of Labor Statistics, the Department of Justice, and the Department of

Energy among those participating in CDAC activities.

The principal goal of the committee has been to operate as a forum where members share information and ideas on disclosure limitation methodology, and discuss problems as well as solutions to issues concerning confidentiality and data access. In doing so, the group has attempted to provide a mutually supportive environment in which individuals can ask questions and seek advice across agency boundaries on issues concerning data access and confidentiality. In order to encourage the open communication of ideas, only federal employees may become members and meetings are restricted to members and invited guests. The closed meetings promote increased cooperation and sharing among agencies by serving as a safe environment in which to discuss sensitive topics such as disclosure limitation methodology and data access.

Of the many Statistical Policy Working Papers (SPWP) produced by the FCSM, two have received special attention as references for those concerned with disclosure limitation methods, SPWP #2, "Report on Statistical Disclosure and Disclosure-Avoidance Techniques" (1978) and, SPWP #22, "Report on Statistical Disclosure Limitation Methodology" (1994).

The 1994 report has proven to be so influential that a few words concerning it are warranted. The product of an ad hoc FCSM interagency subcommittee, an important contribution of the report is the chapter described as a "Primer" on statistical disclosure limitation. This chapter is especially valuable to those new to the field or who are interested in a nontechnical treatment of essential concepts and techniques. In addition to a description of current agency practices, detailed discussions of methodology for both tabular data and microdata files are provided. The report concludes with a list of recommendations for disclosure limitation practices and a research "agenda". An annotated bibliography is a very helpful appendix.

**CDAC as a "Forum"**

The same principles established in SPWP#22, are embodied in CDAC's very first product, The Checklist on Disclosure Potential of Proposed Data Releases. Before discussing the checklist in some detail, it will be useful to briefly describe the principal activities of CDAC.

CDAC holds meeting four times each year, in January, April, July and October. At these meetings topics such as the following have been discussed:

- Updates on proposed legislation and regulations, internet sites, new journal and publications; conferences, workshops, etc.;
- The use of checklists (e.g., disclosure risk);
- Agency practices concerning flexiplace;
- Computer software for disclosure limitation;
- Issues related to the release of data for public use - consent, probability of re-identification, restricted access, use restrictions, addition of "noise" to public use files;
- Licensing as a means of restricted access;
- Development of research data centers;

- Links with professional societies and other groups (Privacy and Confidentiality Committee of the American Statistical Association, Privacy Committee of the Department of Health and Human Services);
- Remote data access;
- Case studies of disclosure problems and solutions;
- Disclosure review boards and ad hoc panels;
- Auditing of licensees and others authorized off-site users of nonpublic use data;
- CDAC tutorials on disclosure limitation methods.

During discussions at these meetings, it became apparent that a need existed for certain types of documents and that the skills and resources needed to develop these documents were available to this group as a whole. Among these documents were the need for some agreed upon, standardized way of reviewing the many considerations attendant to an adequate disclosure risk review of both microdata files and tabular material. Those who were accustomed to producing such files were in need of a mechanism to insure that all necessary elements had been taken into account for the increasing number of electronic data products being produced. At another extreme, for those who were considering the release of a file for the first time, the expertise was rarely present and new staff unfamiliar with disclosure risk review were in need of reference materials and background information. For these and other situations, it was thought that a standardized list of considerations based upon accepted principles of statistical disclosure limitation could represent a solid platform from which their own particular needs could be crafted.

**The Checklist**

Based on documents used by the Bureau of Census, the "Checklist on Disclosure Potential of Proposed Data Releases" (which we will refer to as "the Checklist") consists of a series of questions that are designed to assist in determining the suitability for release of microdata files and tabular data collected from individuals and organizations under an assurance of confidentiality. These questions generally include definitions of terms used (a limited glossary is appended), concrete examples of problems to be avoided or dealt with, and suggested techniques for disclosure limitation. The Checklist's introductory section includes a statement concerning the uses of the Checklist, an overview of its contents, suggestions as to who should fill it out, and how the Checklist can be useful not only to statisticians, but to non-statisticians interested in learning more about statistical disclosure limitation. Next comes a cover sheet that elicits basic information about the proposed data release (a single release, part of a series, reference period, related releases, etc.) and then follows with three main sections. Section 3 pertains to microdata files that contain information from individuals or establishments, while Section 4 and 5 refer to tabular data from individuals and establishments, respectively. We will review, briefly, the contents of those sections:

Microdata: A major part of this section of the Checklist focuses on geographic information because it is the key factor in permitting inadvertent identification. In a demographic survey, few respondents could likely be identified if located within a single State, but more respondents -- especially those with rare and visible reported characteristics -- could be identified if located within a county or other geographic area with 100,000 or fewer persons. An interesting feature of the Checklist is that it draws the user's attention to the variety of ways in which geography may be derived or inferred from items not intended to provided

geographic detail - e.g., record numbers that reflect sequencing by state, county, primary sampling unit, etc. or information on proximity to unique sites.

After a warning concerning direct identifiers, detailed attention is given to certain variables which, when provided in detailed form, increase the ease of matching with external files: income, race, occupation, health conditions, age, and rent/mortgage. The discussion of these variables contains suggestions for the masking or reduction of detail that can cause problems, such as re-coding and top/bottom coding. In this context, the kinds of external files that might be used by an intruder are described. Because many microdata files are enhanced with data from other sources (other research files, administrative data bases), the Checklist highlights the problems that arise when files are enriched in this way. To assist in the evaluation of disclosure risk, questions are posed concerning both "natural" and added sources of statistical error or "noise" in the file. Finally, information concerning details of the sample design that would be part of the release, or which might already have been released is elicited. Such information often contains details helpful to an intruder and must also be considered.

Tabular Data from Persons or Households: In contrast to microdata files in which data for each individual is presented separately, tabular data generally present person or household data in aggregated form. In some cases, however, a record can appear in such a way that reveals information concerning an individual respondent or class of respondents and an unintended disclosure can occur. This section address such situations.

Consideration is first given to the nature of the data upon which tabulations are based; whether they are from a sample or a complete count, whether some groups are sampled with certainty, the public availability of sample weights employed, the number of tabular dimensions, levels of geography, whether a preliminary or final release is being considered, and addition of data from external sources.

The Checklist then moves on to disclosure limitation methods in frequency count data (where numbers or percents are shown) and magnitude data (the aggregate of a "quantity of interest").
The distinction is important because some methods apply only to one or the other, but not both.

Methods described include primary and complementary cell suppression (with discussion of criteria for cell sensitivity), key item suppression, auditing of suppression patterns, and the addition of controlled statistical perturbation ("noise"). A final section alerts the Checklist user to the need for agency coordination when more than one agency will disseminate data.

Tabular Data from Establishments or Other Types of Organizations: As with data from persons or households, tables can be of two types - tables of frequency count data, such as a table of the number of establishments within the manufacturing sector by industrial classification group, and tables of magnitude data, such as a table that presents the total value of shipments for those establishments in the same cells. Also, as set forth in Section 2 of the Checklist, different statistical disclosure limitation methods can be used depending on the type of data being presented.

This section overlaps considerably with the section dealing with tabular data for persons and households.

While many of the disclosure limitation techniques are the same, however, their use is described in this section with examples drawn from establishment surveys. Where establishment data require a different approach (e.g., multiple locations of a single "firm") appropriate questions are raised. In addition, because establishments are often selected from very skewed populations and because there is already a great amount of information concerning them already available to the public, avoiding disclosure from published data is often very difficult. For example, in the U.S., it is well known there are only a handful or so of hospitals with 1,000 or more beds and inadvertent disclosure in a survey of hospitals might well be possible using detail on the number of beds within a geographic unit as large as a Census region.

Although it is rare that public use establishment level data files are released, the Checklist is an appropriate place to comprehensively review the issues associated with the public release of establishment level data files.

**Limitations of the Checklist**

Responses to questions in the Checklist are not intended to supply all the information that might be required before a microdata file or table is released to the public. Some additional questions may need to be answered and/or given special consideration. Nonetheless, if files and tabular material are reviewed with the aid of the Checklist early enough, the need for time-consuming and costly re-programming of the data to be released can be avoided. This allows additional time for coordination with collaborators and other potential users.

**Restricted Access**

Frequently, the results of the Checklist or other considerations, will mean that a file cannot be released for public use, yet there is an acute need in the research community for detailed data. In those instances, some agencies have developed access procedures that permit access with restrictions as to who, for what purpose, and where[1]. Two of these, in particular, have been the subject of intensive CDAC discussions - licensing and research data centers. Such was the interest in licensing procedures that one member took the initiative to update Jabine's work in this area. That work has been consolidated in a paper entitled "Data Licensing Agreements at U.S. Government Agencies and Research Organizations"[2].

**Licensing**: Procedures for five federal agencies and two private research organization are included in the report. For example, the National Center for Education Statistics licenses their researchers to use, at their university or research center, data sets that contain more detailed information than the "standard"

---

[1]See Jabine, 1993 for a description of a broad array of restricted access procedures used by U.S. statistical agencies.

[2] Authored by Paul Massell and Laura Zayatz, this paper was presented at a special contributed session at the International Conference on Establishment Surveys II, Buffalo, NY, June 17-21, 2000

public-use microdata file. Under licensing, the researcher must sign an agreement with the agency which permits the installation of the restricted data on their computer in return for meeting the agency's conditions relating to maintaining confidentiality of the data.

The following are common themes in licensing agreements used at the entities described in this report:

- Demonstration of a need for detailed data;
- Designation of those who have access;
- Statement of legal provision;
- Data security and enforcement/provision for inspection;
- Restrictions on use (prohibition against linking with other files);
- Restrictions on release of research results/adherence to agency policies;
- Return/Destruction of data provided;
- Costs.

An additional project CDAC is considering undertaking is the development of a Checklist that Federal agencies may use to decide whether (and how) to provide restricted access to microdata files.

**Research Data Centers:** The Census Bureau's and NCHS Research Data Centers permit access to qualified researchers under highly restricted conditions. The essential characteristics of these centers are:

- review of research protocol;
- formal agreement covering work to be done, data used, and types of output;
- in-house files without identifiers;
- limitations on types of analysis;
- no outside (linkable) data brought in by researcher;
- dedicated computers;
- disclosure review of output;
- inspection of material removed from site;
- physical presence of agency staff.

The research data center at the National Center for Health Statistics provides for on-site as well as remote access.

**Brochure**

In view of the many requests CDAC members were receiving from persons inquiring about confidentiality procedures, it was felt that some easy to read and concisely crafted document ought to be developed to respond to these requests and to distribute at strategic sites. Accordingly, a document (which we will refer to as the "Brochure") was drafted entitled "Confidentiality and Data Access Issues for Statistical Data" for people unfamiliar with these issues. The public is generally unaware that some statistical agencies are bound to preserve the confidentiality of their survey responses while other agencies may withhold release

of the data under the Freedom of Information Act (FOIA) only if formally exempted. The Brochure provides a comprehensive review of the disclosure limitation and restricted access policies and procedures that various government agencies follow to preserve the confidentiality of their reported data. Covered are confidentiality statutes and policies, data protection, statistical disclosure limitation for tabular and microdata, and mechanisms for restricted access. References to helpful internet sites, and annotated references are appended.

**Suppression Audit Program**

Prior to publishing confidential data in tabular format, a statistical agency must review the table to make sure that cells do not contain values which disclose sensitive, confidential information. Cell suppression is a common technique for protecting the confidentiality of survey responses that are used to generate frequency count and aggregate magnitude data. The suppression of these sensitive cells, called primary suppressions, does not always insure against disclosure. The value of suppressed cells is sometimes derivable from other values shown in the table when the total is published along a row or column. To guard against this kind of disclosure, it may be necessary to suppress additional cells through a process referred to as complementary suppression. The resulting suppression patterns become complex in tables that contain by up to 5 dimensions of the data.. Software exists which determines necessary complementary suppression cells to prevent derivation of primary suppressed cells using various mathematical formulas. However, no existing software that (a) uses generally available computer language and (b) produces low cost, easy, reliable, and efficient results. Without an automated system which is also easily modifiable, agencies currently need to spend considerable resources to perform an audit of the suppressions for all tables in a publication.

In 1999, CDAC launched an interagency project to develop a suppression audit software for general use. Seven agencies are currently participating in the funding and development of project specifications for a user friendly suppression audit program, with four of them contributing test data sets. The auditing system software being developed is written in SAS and it stores data and parameters in SAS data sets. A user needs version 8.1 of SAS which includes the Operations Research module to run this software program.

The first phase of the project involved developing a methodology for importing tabular cell data into a processing system that could be used by any agency. This included specification of the program code; output data sets, and descriptions. The SAS import routine reads and converts a CSV (comma separated value - ASCII) file into a SAS data set. The CSV file must contain at least four types of records and may contain other types for proper application processing. Because some tables contain independent rounding of cells within a table, the user needs to input specific epsilon factors which indicate the range by which the incoming data are allowed to vary around the published values. The user also needs to specify the hierarchy of the data within a dimension if the data follow some order or rank The program checks for internal consistency of the data and verifies that all cells along a row or column sum to the marginals before beginning the LP module.

Parameters are coded in the input file which state the protection range for the cells of a table. The user has the option of specifying either a percent or absolute value to define the protection range. Also, the user

must elect to use different protection ranges for each cell or a global protection range for all cells in the table. The user has the option of specifying a different protection range for primary and complementary cells, however, if different protection ranges are used for these two classes of suppressed cells, the primary cells must have a wider protection range because those cells present direct disclosure risks. If user doesn't select different protection ranges for primary and complementary cells, the system defaults to using the same protection range for both primary and complementary cells in a table.

The program provides a mathematical solution for tables containing up to 5 dimensions for the data - i.e. each variable in a table is a dimension. The user also has the option of specifying the mathematical inter-table relationships including the number of relationships and components that are related between tables and the number of tables involved in the inter-table relationships. The audit system permits the use of alternate optimizers and the program is designed to run using PROC LP, PROC NET FLOW, and PROC INTPOINT..

The program outputs a data sets which shows the values from the original import file with the protection range column values for each cell. The integration of the input modules with an output file that displays the protection range for each cell in a table provides an agency with an audit of the suppression quality of the primary and complementary suppressions used to protect the confidentiality of the published data.

**Short Courses and Tutorials**

In an effort to respond to expressed needs for an introduction to, and overview of, issues and practices in the field of statistical disclosure limitation, a team of CDAC members have, since 1997, given day-long short courses in a variety of setting. The courses have been given in conjunction with the Washington Statistical Society, the Continuing Education Program of the American Statistical Association, at the American Public Health Association annual meetings, the CDC/ATSDR Symposium on Statistical Methods, and at the Survey Research Center of the University of Michigan. Involving 4-5 members of CDAC, the short courses have covered the following topics:

- Legal Issues;
- Informed Consent;
- Statistical Disclosure Limitation (SDL) Techniques for Microdata;
- Applications of SDL for Microdata;
- SDL for Tables;
- Checklist on Disclosure Potential;
- Restricted Access Procedures.

**Provision of Expertise to Regulatory Development**

Recent years have witnessed a sharp increase in legislation directed at the privacy of medical records. Much of this legislation has developed concepts of "identifiability" that would require a statistical assessment. It has only been with the need to develop specific regulations to implement the Health Insurance Portability Act of 1997, however, that this need became explicit. Fortunately, work already

done by CDAC was known and became available to serve as a resource. The future implementation of these regulations will doubtless reflect the influence of many others in this field (among other, members of the American Statistical Association's Privacy and Confidentiality Committee have offered their expertise), but for now, DHHS staff developing portions of the regulations dealing with statistical issues have drawn upon the FCSM SWP #22 and the CDAC Checklist as a references to indicate the procedures necessary to evaluate questions of identifiability. It is noteworthy that when questions arose as to the utility of certain approaches, empirical information bearing upon them was easily obtained through contacts built up within CDAC.

**Summary and Future Plans**

CDAC has several projects underway for developing common procedures and methodologies for federal statistical agencies to use when providing access to confidential data and in the dissemination of data suitable for public use. The Checklist is not a "fixed" document and agencies are encouraged to adapt it to suit their particular needs. Of course, the Checklist will be modified as disclosure limitation methods improve and as new problems emerge. The Brochure on data confidentiality will be a useful information tool for informing others of federal practices and resources in disclosure limitation. The development of an audit suppression software will also be an easily modifiable program for an agency to use to audit the quality of the suppressions for the data being released.

Recently, a project to bring together reports on major mechanisms for granting restricted access to confidential data was initiated. When completed, the report will provide helpful information concerning data licensing agreements, fellowships and post doctoral programs, and research data centers.

All of these information products will be available through the Internet. CDAC will continue to function as a resource for federal agencies and develop information products which are relevant to current and emerging issues relating to data confidentiality, privacy, and access.

**New Web Site**

The Checklist, the Brochure, and other information products can be accessed and downloaded from CDACs website using the url ***http://www.fcsm.gov/cdac/index.html.*** The CDAC website also includes a statement of the committee's purpose and duties, the responsibilities of members, and CDAC contact persons. Other information products such as the Statistical Policy Working Papers Series are available through the web site *http://www.fcsm.gov*.

**References and Resources**

de Wolf, V.A. (1997). The "Interagency Confidentiality and Data Access Group," *American Statistical Association, 1997 Proceedings of the Government Statistics Section and Social Statistics Section*, 323-328.

Eurostat. (1996). *Manual on Disclosure Control Methods*. (Catalogue #: CA-94-96-283-EN-C). Luxembourg: Eurostat.

Evans, T., Zayatz, L., & Slanta, J. (August 1996). "Using Noise for Disclosure Limitation of Tabular Data," *Proceedings of the 1996 Annual Research Conference and Technology Interchange*. Washington, DC: U.S. Department of Commerce, Bureau of the Census, 65-86.

Federal Committee on Statistical Methodology. (May 1978). *Report on Statistical Disclosure and Disclosure-Avoidance Techniques*. (Statistical Policy Working Paper 2). Washington, DC: U.S. Department of Commerce, Office of Federal Statistical Policy and Standards.

Federal Committee on Statistical Methodology. (May 1994). *Report on Statistical Disclosure Limitation Methodology*. (Statistical Policy Working Paper 22). Washington, DC: Office of Management and Budget, Office of Information and Regulatory Affairs, Statistical Policy Office.

Jabine, T. B. (1993). "Procedures for Restricted Access," *Journal of Official Statistics*, 9(2), 537-589.

Kim, J.J. & Winkler, W.E. (1995). "Masking Microdata Files," *American Statistical Association, 1995 Proceedings of the Section on Survey Research Methods*, 114-119.