CHAPTER I - Introduction

A. Subject and Purposes of This Report

Federal agencies and their contractors who release statistical tables or microdata files are often required by law or established policies to protect the confidentiality of individual information. This confidentiality requirement applies to releases of data to the general public; it can also apply to releases to other agencies or even to other units within the same agency. The required protection is achieved by the application of statistical disclosure limitation procedures whose purpose is to ensure that the risk of disclosing confidential information about identifiable persons, businesses or other units will be very small.

During 2004, the Confidentiality and Data Access Committee (CDAC), a special interest committee on data confidentiality and access issues for the Federal Committee on Statistical Methodology (FCSM), revised Statistical Policy Working Paper 22 to incorporate new developments in statistical disclosure limitation methodologies, and to update agency data confidentiality practices and procedures. A description of CDAC and its activities is contained in Appendix D. Statistical Policy Working Paper 22 was written in 1994 by the Subcommittee on Disclosure Limitation Methodology. The 1994 subcommittee's purpose was to review and evaluate statistical disclosure limitation methods used by federal statistical agencies and to develop recommendations for their improvement. A description of that subcommittee is contained in the Cover and Introduction Material of the original 1994 Statistical Policy Working Paper 22.

Legislation passed by Congress after the original release of Statistical Policy Working Paper 22 in 1994 added to the federal agencies' need to protect the confidentiality of the data they collect. The Health Insurance Portability and Protection Act (HIPPA) originally enacted in 1996 had a strong impact on setting requirements for protecting health data. The Confidential Information Protection and Statistical Efficiency Act (CIPSEA) of 2002 created a new mechanism for agencies to protect data confidentiality while at the same time limited the data sharing activity to statistical purposes only. Over this same time period, the interest in federal statistical data within the research and data user community continued to grow. The need for greater data access led to the development of new disclosure avoidance methods so that more data could be released to the public while agencies maintain the protection of respondent information. This revision of Statistical Policy Working Paper 22 updates the discussion of these issues by incorporating current research and new developments in the field.

The goals in revising this report were to:

- describe and evaluate existing disclosure limitation methods for tables and microdata files;
- provide recommendations and guidelines for the selection and use of effective disclosure limitation techniques;
- promote the development, sharing and use of software for the applications of disclosure limitation methods; and
- encourage research to develop improved statistical disclosure limitation methods, for

both tabular as well as public-use microdata files.

Every agency or unit within an agency that releases statistical data should be capable of selecting and applying suitable disclosure limitation procedures to all the data it releases. Each agency should have one or more employees with a clear understanding of the methods and the theory that underlies them. This report is directed primarily at employees of federal agencies and their contractors who are engaged in the collection and dissemination of statistical data, especially those who are directly responsible for the selection and use of disclosure limitation procedures. This report is also useful to employees with similar responsibilities in other organizations that release statistical data, and to data users so that they may better understand and use disclosure protected data products.

B. Some Definitions

In order to clarify the scope of this report, we define and discuss here some key terms that will be used throughout the report.

B.1. Confidentiality and Disclosure

A definition of **confidentiality** was given by the President's Commission on Federal Statistics (1971:222):

> [Confidential should mean that the dissemination] of data in a manner that would allow public identification of the respondent or would in any way be harmful to him is prohibited and that the data are immune from legal process. Duncan et. al., 1993, *Private Lives and Public Policies*, p. 24.

Confidentiality differs from privacy because it applies to business as well as individuals. Privacy is an individual right whereas confidentiality often applies to data on organizations and firms. The second element of this definition, immunity from mandatory disclosure through legal process, is a legal question and is outside the scope of this report.

A second definition is also provided to assist users in understanding this concept.

"Confidentiality pertains to the treatment of information that an individual has disclosed in a relationship of trust and with the expectation that it will not be divulged to others in ways that are inconsistent with the understanding of the original disclosure without permission." IRB Guidebook, Part III.D, Department of Health and Human Services, Office of Human Research Protections.

An agency's need to protect the confidentiality of data it collects is based upon various legislative requirements. Statistical disclosure occurs when released statistical data (either tabular or individual records) reveal confidential information about an individual respondent. This paper is concerned with minimizing the risk of **disclosure** (public identification) of the identity of individual reporting units and information about them.

Section 512 of Title V of the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA) requires all federal agencies to protect data or information acquired by the agency under a pledge of confidentiality for exclusively statistical purposes from being disclosed in identifiable form. Section 502 of CIPSEA defines "**identifiable form**" as any representation of information that permits the identity of the respondent to whom the information applies to be reasonably inferred by either direct or indirect means.

The Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule was implemented on April 14, 2003. This rule obligates most "covered entities," such as Medicare providers, to protect the confidentiality of health care information that they possess. The Privacy Rule subjects the providers of health care information to certain requirements to protect the confidentiality of the data being released. Regardless of the basis used to protect confidentiality, federal statistical agencies as well as some private information organizations involved with health care information, must balance two objectives: to provide useful statistical information to data users, and to assure that the responses of individuals are protected.

The Family Educational Rights and Privacy Act (FERPA) (20 U.S.C. § 1232g; 34 CFR Part 99) was enacted to protect the privacy of student education records. The law applies to all schools that receive funds under an applicable program of the U.S. Department of Education. FERPA gives parents and eligible students (i.e. students over the age of 18 or who attend a school beyond the high school level) certain rights with respect to their education records. Generally, schools must have written permission from the parent or eligible student in order to release any information from a student's education record. However, FERPA allows schools to disclose those records, without consent, to certain designated parties or when specific conditions are present. Schools may also disclose, without consent, "directory" information such as a student's name, address, telephone number, date and place of birth, honors and awards, and dates of attendance. However, schools must tell parents and eligible students about directory information and allow parents and eligible students a reasonable amount of time to request that the school not disclose directory information about them.

The release of statistical data inevitably reveals some information about individual data subjects. Disclosure occurs when confidential information is revealed. Sometimes disclosure can occur based on the released data alone; other times disclosure may result from combining the released data with publicly available information; and sometimes disclosure is possible only through combining the released data with detailed external data sources that may or may not be available to the general public. The accessing and/or linking by the public to electronic data bases creates some degree of risk that disclosure of confidential information may occur even though personal identifiers are removed from a file. At a minimum, each statistical agency must assure that the risk of disclosure from the released data when combined with other relevant publicly available data is very low.

Several different definitions of disclosure and of different types of disclosure risk have been proposed. Duncan et al. (1993: 23-24) provides a definition that distinguishes three types of disclosure:

Disclosure relates to inappropriate attribution of information to a data subject, whether an individual or an organization. Disclosure occurs when a data subject is identified from a released file (**identity disclosure**), sensitive information about a data subject is revealed through the released file (**attribute disclosure**), or the released data make it possible to determine the value of some characteristic of an individual more accurately than otherwise would have been possible (**inferential disclosure**).

Note that each type of disclosure can occur in connection with the release of either tables or microdata. The definitions and implications of these three kinds of disclosure are examined in more detail in the next chapter.

B.2. Tables, Microdata, and On-Line Query Systems

The choice of statistical disclosure limitation methods depends on the nature of the data products whose confidentiality must be protected. Most statistical data are released in the form of tables, microdata files, or through on-line query systems. Tables can be further divided into two categories: tables of frequency (count) data and tables of magnitude data. For either category, data can be presented in the form of numbers, proportions or percentages.

A microdata file consists of individual records, each containing values of variables for a single person, business establishment or other unit. Some microdata files include direct identifiers, such as name, address or Social Security number. Removing any of these identifiers is an obvious first step in preparing for the release of a file for which the confidentiality of individual information must be protected.

Historically, disclosure limitation methods for tables were applied directly to the tables. Methods include redesign of tables, suppression, controlled and random rounding. More recent methods have focused on protecting the microdata underlying the tables using some of the microdata protection techniques. In this way all tables produced from the protected microdata are also protected. This may be done whether there is an intention to release the microdata or not. It is a particularly useful way to protect tables produced from on-line query systems.

B.3. Restricted Data and Restricted Access

The confidentiality of individual information can be protected by restricting the amount of information provided or by adjusting the data in released tables and microdata files (**restricted data**) or by imposing conditions on access to the data products (**restricted access**), or by some combination of these. The number of federal agencies that have implemented restricted access programs have increased during the past ten years and this report provides some references. However, the main trhust of this report is to discuss the disclosure limitation methods that provide confidentiality protection by restricting the data. The fact that this report deals primarily with disclosure limitation procedures that restrict or adjust data content should not be interpreted to mean that restricted access procedures are of less importance. Readers interested in the latter can find detailed information in Duncan et. al., 1993, *Private Lives and Public Policies*,. p. 157 and "Restricted Access Procedures" by the Confidentiality and Data Access Committee (April 2002) at http://www.fcsm.gov/committees/cdac/cdacra9.doc.

As a brief summary, there are four main methods that agencies use to provide restricted access to confidential data: Research Data Centers (RDCs), Remote Access, Research Fellowships and Post Doctoral Programs, and Licensing Agreements. **RDCs** permit use of confidential files in a physically secure environment with specialized equipment. Users agree to terms and conditions governing the access and use of the confidential data. Research products are reviewed by the agency to assure no confidential information is revealed. **Remote access** over secure electronic lines to dedicated computers is a second method. Users can apply statistical techniques to confidential data. The statistical products are reviewed by the agency to assure no confidential data are revealed. **Fellowships and post-doctoral programs** are a third method, and researchers sign agreements to allow them to be treated as agency employees, subject to the same restrictions as employees. Similar to RDC access, researchers may be given limited access and products are reviewed by the agency to make sure no confidential data are released. Fourth, **licensing agreements** permit a researcher to use confidential data offsite, but under highly restricted conditions as spelled out in a legally binding agreement. Arrangements that place restrictions on who has access, at what locations, and for what purposes access is allowed normally require written agreements between agency and users. These agreements usually subject the user to fines, being denied access in the future and/or other penalties for improper disclosure of individual information and other violations of the agreed conditions of use. Users may be subject to external audits conducted by the agency to assure terms of the agreement are being followed. Users in violation may be required to pay fines or be subject to other legal penalties.

Most **public-use** data products are released by statistical agencies to anyone usually without restrictions on use or other conditions, except for payment of fees to purchase publications or data files in electronic form. Both NCHS and NCES require users of public use data files to signify that they will not use the data being made available to them to try to identify an individual respondent. Agencies require that the disclosure risks for public-use data products be very low. In meeting this requirement the application of the disclosure limitation methods described in this document may substantially restrict data content, to the point where the data may no longer be of value for some purposes. The National Center for Education Statistics provides public-use data products that involve access to confidential data. Though these are "Public Use", users must sign agreements assuring that they will maintain the confidentiality of the data. Users may be audited to make sure they are following proper procedures.

C. Organization of the Report

Chapter II, "Statistical Disclosure Limitation Methods: A Primer," provides a simple description and examples of disclosure limitation techniques that may be used to limit the risk of disclosure in releasing tables and microdata.

Chapter III, "Current Federal Statistical Agency Practices," describes disclosure limitation methods used by fourteen (14) major federal statistical agencies and programs. Among the factors that explain variations in agencies' practices are differences in types of data and respondents, different legal requirements and policies for confidentiality protection, different technical personnel and different historical approaches to confidentiality issues.

Chapter IV, "Methods for Tabular Data," provides a systematic and detailed description and evaluation of statistical disclosure limitation methods for tables of frequency and magnitude data. Chapter V, "Methods for Public-Use Microdata Files," describes various statistical disclosure limitation methods used to protect confidentiality in the public release of microdata files. These chapters will be of greatest interest to readers who have direct responsibility for the application of disclosure limitation methods or are doing research to evaluate and improve existing methods or develop new ones.

Due in part to the stimulus provided by previous subcommittee's reports (including Statistical Policy Working Papers 2 and 22), improved methods of disclosure limitation have been developed and used by some agencies over the past 25 years. Based on a review of these methods, guidelines are provided in Chapter VI as recommended practice for all agencies. The development and production of public use microdata files continues to grow and has increased the need to review the possibility of data linkage to external files and the role of identifiers on files.

Three appendices are also included. Appendix A contains technical notes on practices the statistical agencies have found useful in extending primary suppression rules to other common situations. Appendix B is a list of websites and government references on statistical disclosure. Appendix C is a reference list. Appendix D contains a description of CDAC and its accomplishments.

D. Underlying Themes of the Report

Five principal themes underlie the guidelines in Chapter VI:

There are differences between the disclosure limitation requirements that apply to federal agencies. Federal agencies that have specific legislation covering their data collection activities are bound to maintain the confidentiality of all survey responses. Other agencies that do not have specific legislation covering their data collection activities can determine which data may need protection. Nevertheless, agencies that need to protect data should move as far as possible toward the use of a small number of standardized disclosure limitation methods whose effectiveness has been demonstrated.

Statistical disclosure limitation methods have been developed and implemented by individual agencies over the past 40 years. Information and research in this field needs to be shared across all federal agencies. The documentation and the corresponding software used by a statistical agency should then be shared among federal agencies.

Disclosure-limited products should be auditable to determine whether or not they meet the intended data protection objectives of the procedure that was applied. For example, linear programming software can be used to perform disclosure audits for some kinds of tabular data. At the same time, the data utility of the disclosure-limited products should be assessed as part of the evaluation of the applied procedure.

Several agencies have formed disclosure review boards, statistical or review panels, and designated agency confidentiality officers to ensure that appropriate disclosure limitation policies and practices are in place and being properly used. Each agency should centralize its oversight and review of the application of disclosure limitation methods through the development of a standardized list of questions or areas of inquiry. The "Checklist on Disclosure Potential of Proposed Data Releases" by CDAC and located at http://www.fcsm.gov/committees/cdac/resources.html is a useful guide for agencies to structure their review.