Date of Edition: 11/16/2001

Confidentiality and Data Access Issues Among Federal Agencies

Confidentiality and Data Access Committee
and
Federal Committee on Statistical Methodology

Federal agencies publish data to meet the needs of data users and to fulfill legislative mandates. Various federal laws require certain agencies to collect and disseminate data to the public and other governmental organizations. An agency must also protect the confidentiality of the data providers as required by applicable law. This brochure describes some examples of data protections used by federal agencies - legal sanctions, removal of personal identifiers from data sets, the application of statistical procedures to published information, certificates of confidentiality, institutional and disclosure review boards, and restricted data access (research data centers, remote access, special employee status and data licensing).

Key references and websites are appended.

BRIEF DESCRIPTION OF SURVEYS AND CENSUSES

The federal government collects and processes a wide variety of surveys and censuses. Data are collected from individuals and business entities and are validated and edited before being published.

Examples include the Economic Census from the Bureau of the Census in the Department of Commerce; the Consumer Expenditure Survey from the Bureau of Labor Statistics in the Department of Labor; the Annual Survey of U.S. Direct Investment Abroad from the Bureau of Economic Analysis in the Department of Commerce; the Weekly Retail Motor Gasoline Price Survey from the Energy Information Administration in the Department of Energy; and the Medicare Current Beneficiaries Survey from the Health Care Financing Administration in the Department of Health and Human Services.

Congress passed legislation authorizing these surveys and censuses to assist the President and Congress in formulating various public policies. These laws, and the regulations derived from them, restrict the use of data for statistical purposes only. They also require that statistical agencies collecting the data preserve the confidentiality of the respondent's identity. Examples of such laws are 7 U.S.C. 2276 covering the National Agricultural Statistics Service, and 13 U.S.C. 9 covering the U.S. Census Bureau.

CONFIDENTIALITY PROTECTION AND THE NEED FOR DATA

Agency Confidentiality Protection: An agency's need to protect the confidentiality of data it collects must be considered along with other legislative requirements imposed on an agency concerning public disclosure of information. Data collection activities by a few federal agencies are covered by specific legislation which require the agency to maintain the confidentiality of all survey responses. The great majority of federal agencies, however, do not have specific confidentiality legislation governing their data collection activity.

All agencies are subject to the general requirements of both the Freedom of Information Act (FOIA), 5 U.S.C. section 552, and the Privacy Act, 5 U.S.C. section 552a. The FOIA requires agencies to make available to the public all information obtained in the course of conducting business, unless the information satisfies one or more of nine exceptions. Several exceptions are based on the need to protect individual or enterprise information. For example, an agency may withhold information contained in "personnel and medical and similar files" if the disclosure "would constitute a clearly unwarranted invasion of privacy." 5 U.S.C. 552(b)(6). Similarly, an agency may refuse to disclose survey responses from business surveys if the release could cause competitive harm to a respondent. 5 U.S.C. 552(b)(4). The agency is always required to balance the potential benefit of the public's need to know particular information against the privacy interests that would be protected by the exception. In contrast, the Privacy Act prohibits federal agencies from disclosing information about individuals held in their files, unless the agency has obtained consent from the individual or the disclosure is authorized by law, regulation or a "routine use" established by the agency when collecting the information. The Privacy Act also requires agencies to disclose, upon written request by an individual, all records held in their systems of records concerning that individual to correct their data, and to give the individual an accounting of disclosures made from those systems, except as exempted by law.

Professional Standards: Data collection activities in the federal government may be influenced by standards and principles followed by organizations outside of the government. Various professional societies, such as the American Statistical Association, the American Sociological Association, and the American Association for Public Opinion Research, have adopted a set of ethical principles that address privacy, confidentiality, and informed consent. These principles provide guidance in protecting confidentiality both for those directly involved in the data collection and for secondary users of statistical data. Members and nonmembers alike can benefit by studying one or more of these professional society codes.

The International Statistical Institute (ISI) adopted a Declaration of Professional Ethics in 1986 (International Statistical Review 54(2):227-242), that provides a detailed discussion on data collectors' obligations to their subjects. It recognizes a difference between voluntary surveys, in which the informed consent of participants is required, and mandatory surveys, such as the U.S. Census of Population and Housing, in which the selected participants are required to disclose the requested information. In mandatory surveys, the data collector has an

obligation to notify participants about the purpose(s) for collecting the data and any other intended uses for the reported data. It also notes that essentially the same obligations exist whether the survey participants are individuals or organizations.

Data users working with nonpublic microdata files that contain values of variables for a single person or establishment also have an obligation to protect the confidentiality of information about individuals and organizations. They should:

- Restrict access to the original data set to only those persons who are permitted access under the agreed conditions and ensure that such persons scrupulously follow those conditions of use.
- Make no attempt to identify particular individuals or other units whose data are considered confidential.
- Notify the organization providing the data set if one or more individuals or other units are inadvertently identified so that it may take appropriate action.

PROTECTING DATA DURING COLLECTION AND PROCESSING

Several different approaches may be taken to ensure that information provided will not (or cannot) be released to others without a respondent's permission.

Anonymization: After the data collection has been completed, responses to selected questions and sometimes even to an entire survey are stripped of names and other personal identifiers. By limiting agency use and analysis to these anonymized files, it greatly reduces the likelihood that agency employees can identify individual respondents. The original files, containing names, addresses, etc. -- or a crosswalk file, containing the true identifiers and their corresponding synthetic personal identifying number -- may be used for certain authorized purposes and are kept under highly secure conditions.

Agency-Specific Regulations: Most agencies follow specific policies and procedures for safeguarding the confidentiality of the survey responses they collect. These measures can include employee acknowledgment letters, employee non-disclosure affidavits or pledges, records management procedures, employee security training, password protection on data systems and encryption of data files.

Institutional Review Boards: Most Government-funded research involving participation of identifiable living subjects or private information about them must be approved by a panel that examines the potential for harm to subjects through the research. Resulting loss of privacy is one type of harm that researchers must satisfactorily address before approval is granted. Such approval is necessary before research can begin.

Certificates of Confidentiality: Individuals and agencies conducting research on particularly sensitive topics may apply to the Department of Health and Human Services for a government guarantee of immunity from judicial subpoena or request by other authorities. This is common for research involving sexual attitudes, preferences and practices, controlled substances, illegal conduct and mental health and other topics where public disclosure of identities and the information they provide could cause great

harm to respondents. It is not necessary for a researcher to be federally funded, but the research protocol must have been approved by an Institutional Review Board. Once approved by DHHS, investigators can assure respondents absolute confidentiality.

DATA ACCESS AND RELEASE WHILE PROTECTING CONFIDENTIALITY

Administrative and Statistical Use of Data: The type and level of protection that an agency applies to the data it collects depends on whether the information is used for statistical or administrative purposes.

Statistical use involves the description, estimation, or analysis of survey responses without regard to the identities of specific respondents. Such uses typically generate an aggregate description of a group of persons or establishments, and relationships between characteristics of individuals, establishments, or other units. Published data are expressed with summary figures such as means and coefficients of variation.

Administrative use includes the use of survey responses in identifiable form, such as determining whether a person is eligible for a license, privilege, right, grant, or benefit or whether such a person's conduct was or is in accordance with law. The purpose can be regulatory, program administration, legislative, or judicial.

Confidentiality and Disclosure: When data providers are given assurances of confidentiality by a federal agency, the agency is committed legally and ethically to abide by its confidentiality pledge. The inadvertent public disclosure of confidential information constitutes a violation of the agency's confidentiality pledge to its survey respondents. Care must be taken to ensure that adequate disclosure limitation procedures have been used to protect the data.

Types of Data: Statistical data are generally of two types - aggregate estimates from survey responses -- like those shown in tables, charts, and graphs; and microdata that refer to individual units. Three data formats commonly used to present the data are tables of frequency counts, tables of aggregate magnitude data, and microdata.

Frequency count tables are one-, two-, or higher-dimensional tables comprising counts of the number of respondents with specified characteristics. For example, a two-dimensional frequency count table may have rows corresponding to employment sectors (industry, academia, nonprofit, government, military) and columns corresponding to income categories (in increments of $10,000). An individual table cell at the intersection of a given row and a given column would indicate the number of residents of a certain county employed in the corresponding sector with salary in the corresponding range.

Using this example, such a tabulation could result in a disclosure of confidential information if (a) one sector fell into only one income category (in which case income could be known within more precise limits than without the information in the table, or (b) only 2 cases of any sector fell into the same income category (permitting the conclusion on the part of anyone privy to the information about one of the cases, to know the income of the other).

Tables of *aggregate magnitude data* are

analogous to frequency count tables in that they are defined by cross-classification of one, two, or more categories. However, the cells contain aggregate values, over the covered respondents, of a quantity of interest. For example, a two-dimensional table defined by employment sector and race based on income would contain total incomes in each sector-by-race cell. If cell sizes are very low, other publicly available information could be used to derive more precise income information and, therefore, result in a disclosure.

For confidentiality purposes, it is important to note that even though tabular data may be presented as a collection of two-dimensional tables, these tables may be interrelated or comprise higher-dimensional tables. For example, the suppression patterns in previously published tables may need to be reviewed to prevent any inadvertent disclosure in a currently published table. The interrelationships between sets of tables must be considered to adequately provide confidentiality protection for tabular data.

Microdata files consist of individual records that contain values of variables for a single person, business establishment or other economic unit. Public-use microdata are microdata files with personal identifiers removed that are released to the public for research and analytical purposes after being subjected to procedures to limit the risk of disclosure. Public-use microdata files are popular because the user has the detailed information underlying the published tabular data, and users are free to tabulate data according to their specific needs. The flexibility in generating tabulations together with the highly detailed information in the files permit the user to develop tabulations that carry high disclosure risk. Special care must be taken to ensure that information contained in the microdata file cannot be linked to records for the same person or business in another data set.

An emerging avenue of data release is tabulations from an on-line query to a statistical database using the Internet. Data users may create their own tabulations by customized queries to the database. The "Wonder" system used by the Centers for Disease Control and Prevention is an example of such a system, where users may access various public health files electronically, using the Internet, to create their own tabulations without seeing the underlying data. Another example is the American Fact Finder implemented by the U.S. Census Bureau.

Statistical Disclosure Restraints on Public Tabular Data: Cell suppression is a common technique for protecting frequency count and aggregate magnitude data from disclosure. Before publication, tabular cell data are classified as either sensitive or non-sensitive to disclosing survey respondent-level data. If the cells are classified as sensitive, then the cells are called primary suppressions and are withheld from publication. The suppression of primary cells alone, however, does not always protect the sensitive data. When a table contains cells that represent the sum of either a row or column, the original value of a primary cell may be determined exactly or within a narrow range through subtraction. When this occurs, it is necessary to suppress additional cells, called complementary suppressions, to protect the primary suppressed cells from disclosure.

The extent that cells are suppressed in a table may be analyzed as an optimization problem.

After a set of primary suppression cells is identified, a set of complementary suppressions may be necessary to ensure a certain level of protection at minimal cost to the table for containing suppressed cells. Most optimization programs either minimize the sum of the values of the suppressed data or the number of suppressed cells.

Other methods for protecting frequency count data include data swapping and rounding. For more information on disclosure limitation techniques, see Statistical Policy Working Paper 22 (1994) in the References and Resources section.

Statistical Disclosure Restrictions on Public Microdata Sets: In surveys that collect a large number of variables, it is possible that a number of respondents are unique in the sample based on some combination of characteristic variables and some may also be unique in the population. The likelihood that an unscrupulous user could identify a sample case is a function both of the availability of individually identified external information that can be used to match against some of the variables in the survey and of the probability of combinations of real data in the survey.

There are two common types of procedures that are applied to data contained in public microdata sets. First, there are methods intended to reduce the number of unusual cases directly by reducing the variation within the data, including such methods as rounding, top- and bottom-coding, collapsing of categorical responses, and data suppression. Second, there are a variety of techniques that increase the uncertainty associated with reported data such as swapping of values, adding predetermined random noise to the data, and performing other more structured randomization of the data.

RESTRICTED ACCESS

Agencies are aware that some data simply cannot be made available in an unrestricted manner. Applying statistical disclosure limitation procedures to preserve the confidentiality of the published date would likely make the information useless for research purposes. Valuable detailed information can still be made available to a qualified user under certain conditions.

Restricted Sites: Restricted access sites are secure sites where researchers may go to access confidential data. These sites can be located either at the statistical agency or at another approved location. An approved user may access the data at these restricted sites for approved statistical purposes only.

Restricted access sites have several characteristics. Access to the physical office is limited to those persons authorized to enter, the computer network is not accessible from outside the site, and the computer systems meet the agency's security criteria.

Users at restricted access sites are typically researchers from universities or other research organizations, but may also include other federal agencies. Depending on such circumstances as the nature of the research project and the agency's legislative authority, an agency may require the user to swear to protect the data, subject to penalties for wrongful disclosures before granting access to the data. Often, prospective users are required to submit detailed research proposals that the agency reviews. If the proposal is approved, a

contract is written that specifies the work to be done, the data to be used, and the type of output to be released. Users are given access only to the data they need for their projects and are not allowed to take confidential data out of the secure facility. An agency employee is stationed at the site to provide security, to perform disclosure analysis on any results users may wish to remove from the site, and to provide consultation on the use of the data, including references to agency experts.

Remote Access: Some agencies provide users remote electronic access to certain data files through the Internet. This type of system shares many of the characteristics of the restricted access sites except that computational programs are reviewed before application and prohibited tabulations are suppressed. Statistical output is subjected to disclosure review. A record is maintained of the tabulations provided in order to keep track of released data for purposes of ensuring against disclosures of complementary suppressed cells.

Sworn employees and fellows: Some agencies have authority to employ the researchers as temporary staff members to perform authorized work. These employees are referred to as having "Special Sworn Status." The agency can appoint these individuals under special circumstances such as, (1) the individual is employed by an organization with which the agency is engaged in joint work, and the individual has specialized knowledge that can aid that joint work; (2) the individual is employed by an organization performing work for the agency under contract or provides information to the agency for statistical purposes; or (3) the individual is employed when federal law requires someone to audit or inspect agency operations. People who are granted Special Sworn Status swear to protect the data and are subject to penalties for wrongful disclosure.

Several U.S. agencies have fellowship programs that recruit highly qualified researchers to carry out research that both aids the agency's mission and provides benefits to the researchers' fields. These individuals carry out their research at the agency under restricted access, and the agency grants them Special Sworn Status.

Licensing of Data Users*:* Some agencies have authority to license data users to access respondent-level data. Licensees are required to sign disclosure protection agreements with the agencies prior to having access to the data. Licensees may install the restricted data on their computers in return for adhering to the agency's conditions relating to maintaining confidentiality of the data. Agencies are careful to permit restricted access only for those uses that have already been carefully described to the respondent or information provider. Statistical agencies, for example, rarely collect information for anything but statistical or research uses and when they do, they are careful to inform respondents of the purpose of the data collection and the planned uses that will be made of the data. In turn, anyone granted access to these data must agree, in writing, not to use them for any other purpose other than that for which the data were collected and for which they are being given access. Before applicants are granted access, they are required to submit detailed descriptions of research plans and to sign legally binding document(s) agreeing to use the data only for the uses described.

Agency-Specific Statutes: Organizations such as the Census Bureau, the National Agricultural Statistical Service, the National Center for Education Statistics, and the National Center for Health Statistics collect information under laws specifically established for agency collection activities. These statutes provide no discretion (except the consent granted by study subjects themselves) for an agency to decide whether to release confidential information.

Disclosure Review Boards. Some agencies have established special panels called Disclosure Review Boards to review data releases before they are made public. These boards review microdata files and tables to determine if releasing the information to the public would conflict with the agency's confidentiality policies. Over time, these boards develop substantial expertise and experience concerning their agency's practices and confidentiality issues regarding public releases of data. Other agencies which receive requests for microdata files on a less frequent basis may use ad hoc panels comprised of existing agency staff or staff from other agencies to assess the confidentiality risk of the data.

REFERENCES AND RESOURCES

Publications:

American Statistical Association, Committee on Privacy and Confidentiality (1998). *Surveys & Privacy*, (2nd Edition). Alexandria, VA: American Statistical Association.

Intended for the general public, readers are told what to expect when they are asked to participate in a survey with respect to: assurances of confidentiality (what they mean and how they are implemented), how an individual's survey responses are used to compile statistics, and factors to consider (as well as what information should be provided) before participating in a mail, phone or face-to face interview.

Association of Public Data Users , *Of Significance* (2000), Volume 2 Number 1.

A compilation of articles on data confidentiality by academics and professionals within and outside of the U.S. federal government and in European statistical offices. Topics cover recent activities in health and medical statistics, access to education statistics, research data centers, issues in the production of electronic data products for public use, and more.

Duncan, G., Jabine, T.B. and de Wolf, V. (1993), *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*, Editors; Panel on Confidentiality and Data Access, National Research Council, Washington, D.C.: National Academy Press.

This works provide insight into the tension between the production of useful government statistics and the protection of individual and organizational privacy. Comprehensive and well-handled treatment of how the federal statistical system functions (basic principles,

concepts, problems), its legal underpinnings, the treatment of data subjects, expectations of data users (including their legal and ethical responsibilities), restricted data access, special consideration of data on establishments, and internal agency staffing and organization of information management activities. The report concludes with recommendations on each of the areas treated.

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

A comprehensive treatment of major techniques used in statistical disclosure limitation. Treatment of techniques for tabular data (macrostatistics), as well as micro-data (microstatistics). Includes sections on random perturbation. Intended for those with statistical backgrounds, though non-statisticians may also find it useful.

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.

Update of previous report on same topic, the report begins with 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. Following a description of federal agency practices as of the early 1990s, the report presents detailed discussions of disclosure limitation methodology for both tabular data and microdata files. The report concludes with a list of recommendations and a research "agenda." Contains extensive annotated bibliography.

International Seminar on Statistical Confidentiality Proceedings, November 28-30, 1994, (1995), (Catalogue #C4-88-95-6124-EN-C). Brussels, Luxembourg.

Twenty-eight papers presented at an international conference organized by major European statistical organizations and offices. Papers deal with legislative and administrative aspects of confidentiality, mathematical and computing aspects of confidentiality (including software for disclosure limitation), and organizational aspects of data security, access and protection.

Jabine, T. B. (1993). *Procedures for Restricted Access*, Journal of Official Statistics, Vol. 9, No. 2, p. 537-589.

A summary of restricted access procedures used by U.S. statistical agencies to make data available to other agencies, other organizations and individuals. Topics include criteria for release without restriction; access to special classes of employees such as contractors, grantees, researchers and other agencies; location and mode of access; conditions of future release; security measures; written agreements; penalties. Provides numerous examples of restricted access as well as failures to grant access. Appendices provide forms used by several agencies in

granting restricted access.

Journal of Official Statistics, *Confidentiality and Data Access.* Vol. 9, No. 2, 1993.

The entire issue is devoted to a wide variety of topics, such as informed consent, measures of disclosure risk and harm, statistical disclosure limitation (techniques and practices), database systems, confidentiality legislation, and restricted access. Useful for those new to the field, as well as advanced statisticians.

National Center for Health Statistics *Panel on Disclosure Review Boards of Federal Agencies: Characteristics, Defining Qualities and Generalizability*. NCHS Technical Workshop Report, (forthcoming).

Presentations on disclosure review boards of four federal statistical agencies concerning: principal objectives; types of data reviewed and documentation of the review process; intended use of data; organization of the review board; decision making process; and, statistical disclosure methods applied to the data. Agencies represented are Census Bureau, Bureau of Labor Statistics, National Center for Education Statistics, and National Center for Health Statistics.

Willenborg, L., & de Waal, T. (1995). *Statistical Disclosure Control in Practice.* (Lecture Notes in Statistics 111). NY: Springer-Verlag, Inc.

A basic text on statistical disclosure that will help those new to the field but that also contains useful material for statisticians. Coverage similar to the Eurostat manual. Draws extensively on European data (particularly the Netherlands and Great Britain) for case studies.

Zarate, A.O. , Bournazian, J., de Wolf, V. The FCSM Confidentiality and Data Access Committee. Paper presented at the *FCSM Statistical Policy Seminar: Integrating Federal Statistical Information and Processes November 8-9, 2000*

Description of the origin, activities and plans of the Committee on Data Access and Confidentiality, a group of federal statisticians that meets periodically to discuss common interests and problems related to confidentiality, statistical disclosure limitation and restricted data access. Presents examples of topics discussed and cooperatively produced papers and documents that address common problems. Notable "products" of this group include a Checklist for the review of disclosure risk in tables and electronic files and a series of tutorials covering privacy law, informed consent, statistical disclosure limitation techniques and restricted access procedures.

Web Sites:

Federal Committee on Statistical Methodology, http://www.fcsm.gov

See especially the Statistical Policy Working Papers # 2 and 22 (cited above)

Confidentiality and Data Access Committee, http://www.fcsm.gov/cdac

Background information on organization

and purpose of committee together with description of important documents and activities.

American Statistical Association (ASA), Committee on Privacy and Confidentiality (P&C) http://www.amstat.org/comm/ then go to the P&C Committee's home page.

Background information on the Committee, definitions of key terms; disclosure limitation methods, restricted access, confidentiality sessions at professional meetings, conferences and events of interest, legislation and governmental policies, respondent issues, and Committee activities and services.

Checklist on Disclosure Potential of Proposed Data Releases. http://www.fcsm.gov/docs/checklist_799.doc

A useful and informative series of considerations for the preparation of microdata and tabular data for release to the public in "checklist" format.

Statistics Netherlands' Statistical Disclosure Control Project: http://www.cbs.nl/sdc

This project carries out methodological research in statistical disclosure control (SDC) and develops specialized SDC software (ARGUS). Statistical offices and universities from three European countries (Netherlands, Italy, United Kingdom) participate. The Website contains detailed descriptions of software for tabular and microdata and related articles on SDC.