CHAPTER VI – Recommended Practices For Federal Agencies

A. Introduction

Based on its review of current agency practices and relevant research, the Confidentiality and Data Access Committee (CDAC), a subcommittee of the FCSM, developed a set of recommendations for disclosure limitation practices. The implementation of these practices by federal agencies will result in an overall increase in disclosure protection and will improve the understanding and ease of use of federal disclosure-limited data products. Sometimes the methods used to reduce the risk of disclosure make the data unsuitable for statistical analysis (for example, as mentioned in Chapter V, recoding can cause problems for users of time series data when top-codes are changed from one period to the next). In deciding what statistical procedures to use, agencies also need to consider the usefulness of the resulting data product to data users.

The first set of recommendations in Section B.1 is general and pertains to both tables and microdata. Section B.2 describes CDAC recommendations for tables of frequency data. Recommendations 7 to 11 in Section B.3 pertain to tables of magnitude data. Lastly, Recommendations 12 and 13 in Section B.4 pertain to microdata.

B. Recommendations

B.1. General Recommendations for Tables and Microdata

Recommendation 1: Seek Advice from Respondents and Data Users. In order to plan and evaluate disclosure limitation policies and procedures, agencies should consult with both respondents and data users. Agencies should seek a better understanding of how respondents feel about data disclosure risks, data sharing across agencies, the availability of matching to external administrative data files, and data protections under CIPSEA and non-CIPSEA surveys.

Similarly, agencies should consult data users on issues relating to: balancing the risk of disclosure against the loss in data utility; increasing the availability of public use microdata files; the need for restricted data access procedures so that researchers may access microdata in a controlled and safe environment, and the development of on-line public use data base query systems through the Internet. Other issues that affect data utility include whether users would prefer disclosure limitation methods that modify, replace, or adjust the data in some manner rather than methods that suppress data.

Recommendation 2: Standardize and Centralize Agency Review of Disclosure-Limited Data Products. It is important that disclosure limitation policies and procedures of individual agencies be internally consistent. Results of disclosure limitation procedures should be reviewed. Agencies should standardize the review process by adopting standards and/or guidelines on protecting data confidentiality. The "Checklist on Disclosure Potential of Proposed Data Releases" available at http://www.fcsm.gov/committees/cdac/ should be used as a guide for this review process. The checklist should be modified to suit the agency's data release policy and procedures. Agencies should also centralize responsibility for this review in the organizational

structure through mechanisms such as disclosure review boards (permanent or ad hoc), or a confidentiality officer, review panel, or group of staff knowledgeable and experienced in the area of disclosure limitation procedures and data confidentiality protection.

CDAC recommends that agencies become familiar with external databases that are available to users for matching to agency data products. They should evaluate any proposed data release both in terms of disclosure risks internal to the variables and values inside the file and in terms of external risks of disclosure from potential matching to external files. In agencies with small or single programs for microdata release, this may be assigned to a single individual knowledgeable in statistical disclosure limitation methods and agency confidentiality policy. In agencies with multiple or large programs, a review panel should be formed with responsibility to review each microdata file proposed for release and determine whether it is suitable for release. Review panels should be: as broadly representative of agency programs as is practicable; knowledgeable about disclosure limitation methods for microdata; prepared to recommend and facilitate the use of disclosure limitation methodologies by program managers, and should be empowered by their agency to verify that disclosure limitation techniques have been properly applied.

Tabular data products of agencies should also be reviewed. Disclosure limitation and suppression should be an auditable and replicable process. (Disclosure limitation for microdata is not currently at the stage where a similar approach is feasible.) There are administrative efficiencies for centralizing the review of both micro data files and table files. Depending upon institutional size, programs, and culture, an agency should combine the review of microdata and tables in a single individual, review panel or office.

Recommendation 3: Share Software and Methodology Across the Government.
Federal agencies should share software products for disclosure limitation and record linkage, as well as methodological and technical advances. Specifically, CDAC should continue to make software for disclosure limitation methodologies and documentation available from its website to the federal agencies and public for their use. Software should be written in a common processing language that is easily modifiable with clear documentation.

As advances are made in software for statistical disclosure limitation and record linkage by academia, government, and private businesses, CDAC should evaluate these new methodologies and software, and provide guidance to the federal agencies on the practical and appropriate applications for their use. CDAC has available on its website at http://www.fcsm.gov/committees/cdac/ software which performs primary and complementary suppression, and suppression auditing software which reviews and generates a report indicating the extent of the protection applied from the suppression pattern used for a table.

Recommendation 4: Formal Interagency Cooperation is Needed for Data Sharing. Sharing data files between agencies requires formalized agreements between agencies in order to safeguard data confidentiality protections and meet an agency's legal obligations for collecting and publishing information. The release of identical or similar data by different agencies or groups within agencies (either from identical or similar data sets) and the availability to match to

external files are other factors that contribute to the need for interagency cooperation. Interagency panels or teams may be needed to plan and review data sharing activities between agencies. Interagency cooperation on reviewing overlapping data sets and the use of identical disclosure limitation procedures is encouraged. Agencies should expand the shared use of research data centers as a method for increasing access to confidential data by researchers. Agencies may also consider requesting representatives from other agencies that have more experience with releasing public use micro data files to serve on disclosure review boards so that knowledge and experience across agencies may be shared.

Recommendation 5: Use Consistent Practices. Agencies should strive to employ disclosure limitation methods in standard ways and be consistent in defining categories in different data products and over time. They should standardize variable definitions internally to the extent it meets the agency's program needs and common definitions between agencies should be developed where possible. Such practices will improve data access by the public and make it easier to implement disclosure limitation methodologies. Examples include using consistent schemes for combining categories, establishing standardized practices for similar data such as categorizing or top-coding variables like age or income, and moving towards standardized application of minimum geographic size limitations for household data. Software should be developed, made broadly available and used to implement these methods to assure both consistency and correct implementation.

B.2. Tables of Frequency Count Data

Recommendation 6: Research is Needed to Compare and Evaluate Methods. There has been considerable research into disclosure limitation methods for tables of frequency data. The most common method used is suppression. Besides suppression, other well-developed methods that are available include controlled rounding, controlled tabular adjustment, and applying data perturbation methods prior to tabulation. Additional research is needed to apply these methods to different types of data and compare and evaluate these different methods in terms of data protection and usefulness of the resulting data product. If suppression is used, the guidelines listed in Recommendations 9 and 10 also apply to tables of frequency data.

B.3. Tables of Magnitude Data

Recommendation 7: Use Only Subadditive Disclosure Rules For Identifying Sensitive Cells. Agencies should develop and apply operational linear sensitivity rules (See Chapter 4) to identify and then protect primary disclosure cells. Disclosure rules that have the mathematical property of **subadditivity** provide assurance that a cell formed by the combination of two non-sensitive cells remains non-sensitive. Agencies should employ only subadditive primary disclosure rules. The p-percent, pq, N, and (n, k) rules are all subadditive. **Primary disclosure cells** must be protected using disclosure limitation techniques.

Recommendation 8: The p-Percent or pq-Ambiguity Rules are Preferred. The p-percent and pq-ambiguity rules are recommended because the use of a single (n, k) rule is inconsistent in the amount of information allowed to be derived about respondents (see Chapter IV). The p-percent and pq rules do provide consistent protection to all respondents. In particular, the pq rule should be used if an agency can quantify the extent that data users already know something about respondent values. If, however, an agency feels that respondents need additional protection from close competitors within the same cells, they might use the p-percent or pq rule in conjunction with an (n, k) rule. When using only the (n, k) rule, a sequence of (n, k) rules is better than a single set of parameters. An example of a sequence of (n, k) rules is (1,75) and (2,85). When a combination of (n, k) rules is applied, a cell is sensitive if it violates either rule.

Recommendation 9: Do Not Reveal Suppression Parameters. To facilitate releasing as much information as possible at acceptable levels of disclosure risk, agencies are encouraged to make public the kind of rule they are using (e.g. a p-percent rule) but they should not make public the specific value(s) of the disclosure limitation rule (e.g., the precise value of "p" in the p-percent rule) since such knowledge can reduce disclosure protection. (See Chapter 4 Section B.4 for an illustration of how knowledge of both the rule and the parameter value can enable the user to infer the value of the suppressed cell.) The value of the parameters used for statistical disclosure limitation can depend on programmatic considerations such as the sensitivity of the data to be released.

Recommendation 10: Redesign Tables, Apply Cell Suppression, Controlled Tabular Adjustment, or Perturbation Methods to Microdata Prior to Tabulation There are four methods of limiting disclosure in tables of magnitude data. First, for single tables or sets of tables that are not related hierarchically, agencies may limit disclosure by combining rows and/or columns. Second, for more complicated tables, cell suppression may be used to limit disclosure. Third, controlled tabular adjustment may be applied to protect sensitive cells after tabulation. Fourth, sensitive cells may be protected prior to tabulation by applying some perturbation method that adds noise to the underlying microdata.

Suppression is widely used by the federal agencies. Cell suppression removes from publication (suppresses) all cells that represent disclosure, together with other, nondisclosure cells that could be used to recalculate or narrowly estimate the primary, sensitive disclosure cells. Zero cells are often easily identified and should not be used as complementary suppressions. The suppression patterns should be audited to check whether the algorithms that select the complementary suppression pattern permit estimation of the suppressed cell values within “too close” of a range. Suppression methods should provide protection with minimum data loss as measured by an appropriate criterion such as minimum number of suppressed cells or minimum total value suppressed. If the information loss from cell suppression undermines the utility of the data, other methods may more useful.

Controlled tabular adjustment applied to tables and perturbation methods applied to microdata prior to tabulation eliminate the information loss associated with suppression. One cautionary note is that both methodologies may not provide sufficient protection to a cell that has one

respondent or a cell that is dominated by one respondent. There may also be some inferential loss in information from changing the data. The interrelationship between tables also needs to be checked to minimize any adjustments to cells in other tables or set of tables should be reviewed to check if any of the table(s)' analytical properties have been distorted or limited. These recommended practices also apply if suppression is used for tables of frequency count data.

Recommendation 11: If Applying Cell Suppression, Auditing of Tabular Data is a Necessity. Tables where suppression is applied to protect sensitive cells should be audited to assure that the values in suppressed cells may not be derived by manipulating row and column equations. This recommendation applies to both tables of frequency data and magnitude data.

B.4. Microdata

Recommendation 12: Remove Direct Identifiers and Limit Other Identifying Information From Microdata Files. The challenge of applying statistical disclosure methods to microdata is to thwart the identification of a respondent from data appearing on a record while allowing release of the maximum amount of data. The ability to match variables from external files generates additional disclosure risks that expand the list of variables on a file that need to be reviewed. The first step to protect the respondent's confidentiality is to remove from the microdata file all **direct identifying information** such as name, social security number, exact address, or date of birth. Certain univariate information such as occupation or precise geographic location can also be identifying. Other univariate information such as a very high income or presence of a rare disease can serve both to identify a respondent and disclose confidential data. These data should also be removed or protected. Agencies should also continue to identify univariate data that tend to facilitate identification or represent disclosure, and set limits on how this information is reported. For example, the Census Bureau presents geographic information only for areas of 100,000 or more persons. Income and other information may be top-coded to a predetermined value such as the 99th percentile of the distribution. Lastly, appropriate distributions and cross tabulations should be examined to ensure that individuals are not directly identified. Circumstances can vary widely between agencies or within an agency between microdata files.

After direct identifiers have been removed, a file may still remain identifiable, if sufficient data are left on the file with which to match with information from an external source that also contains names or other direct identifiers. For this reason, agencies should perform re-identification studies and attempt to match variables on the released files to external files outside of the agency.

Recommendation 13: Agencies Need to Share Information on Assessing Disclosure Risks. Agencies need to share information on what external files that are available to a user for matching to agency data products. Information on external files should be updated and widely circulated among the statistical agencies so that disclosure review boards, confidential officers, and other ad-hoc disclosure review boards can properly assess the disclosure risk from a proposed data release.

GLOSSARY

Attribute disclosure – A disclosure that reveals sensitive information about a data subject.

Audit – Check a proposed suppression pattern to make sure sensitive cells are adequately protected.

Bottom-coded – Replacing values below a certain number or percentile ranking with the same value.

Complementary suppression – Withholding non-sensitive cells from release in order to protect other sensitive cells from disclosing confidential information.

Confidential Information – information reported under an expectation that the information will not be released in a manner that allows public identification of the respondent or causes some harm to a respondent.

Disclosure – revealing information that relates to the identity of a data subject, or some sensitive information about a data subject through the release of either tables or microdata.

Frequency count data – Data that show the number of units of analysis in a cell.

Hierarchy – A series of items organized or classified according to rank or order; especially a ranked classification schema used to structure a table or microdata file such as NAICS codes.

High risk – information that has a high probability of being used to either identify a respondent or reveal confidential information about the respondent.

Identifiable form – Any representation of information that permits the identity of the respondent to whom the information applies to the reasonably inferred by either indirect or indirect means.

Inferential disclosure – A disclosure that makes it possible to determine the value of some characteristic of any individual more accurately than otherwise would have been possible.

Identity disclosure – A disclosure that identifies a data subject.

Informed consent – Written permission from a respondent to publish sensitive cell values. It is has the effect of acting as a waiver of the promise to protect sensitive cells and specific authorization or consent to the agency for public releasing he confidential information.

Intruder - An outside user who attempts to link a respondent to a microdata record.

Linear sensitivity measure – A rule that indicates how close a respondent's data may be estimated from a published cell value.

Magnitude data – Data that show the aggregate of a "quantity of interest" that applies to units of analysis in the cell.

Primary suppression rules – A linear combination of respondent level data that is used to determine whether a given table cell could reveal individual respondent information.

Primary suppression – Withholding from publication any cells that are identified as being by a primary suppression rule.

Public-use – Data products that are released by statistical agencies to anyone without restrictions on use or other conditions, except for payment of fees to purchase data in electronic form.

Restricted Data – Adjusting data in released tables and microdata files or limiting the amount of information released.

Restricted Access – Imposing terms and conditions on users' access to the data products.

Sample – A set of records or data elements drawn from a population and used to estimate the characteristics of a population.

Sensitive – A classification of a cell value established by using a primary suppression rule.

Suppression – Withholding information in selected table cells from release.

Subaddivity – The property that the union of two non-sensitive cells is also non-sensitive.

Tabular Data – Data presented in tables.

Three-dimensional table – A table containing aggregate cell values over three variables.

Top-coded – Replacing values above a certain percentile ranking with the same value.

Two-dimensional table – A table containing aggregate cell values over two variables.