National Center for Health Statistics

Policy on Micro-data Dissemination

July 2002

NCHS Policy on Micro-data Dissemination

Table of Contents

NCHS Policy on Micro-Data Dissemination

This policy addresses when, to whom, and in what form NCHS disseminates data specific to individuals, households, establishments or events–defined as micro-data--and also outlines dissemination procedures.

Introduction

As the nation's principal statistical agency dealing with health, NCHS fulfills its mission through the collection, analysis and dissemination of data on all aspects of the health of the US population. Furthermore, NCHS disseminates the data it collects using a wide range of mechanisms and formats. Although the same general principles apply to all forms of dissemination, this particular policy deals specifically with micro-data produced by NCHS data systems **(**micro-data refers to data files in which each record provides information for the unit of data collection, for example, an individual person; see Terms and Concepts, page 8). Therefore, this policy does not directly address data that are generated from methodological research, that result from administrative monitoring of data collection activities, or that are released in tabular form or in analytic reports. This document outlines guiding principles, relevant terms and concepts, and the practices governing NCHS micro-data dissemination. References are provided for additional information.

Terms and Concepts

Throughout this document, a number of terms and concepts are used that take on specific meanings in the context of a discussion of health statistics. These terms - ranging from "confidentiality" to "dissemination" to "disclosure" - are important to the understanding of NCHS' policy and therefore are defined on pages 8-9.

Guiding Principles

NCHS' authorizing legislation mandates that data be made as widely available as practicable (Section 308(c)). (1) However, the mandate to make data available must be guided by NCHS' role as a federal statistical agency and be balanced against the need to protect respondent confidentiality and to assure data quality.

> *1) Equitability in data dissemination*: In collecting, analyzing, and disseminating data, NCHS adheres to the principles and practices of a federal statistical agency that have evolved to protect the impartiality and credibility of federal statistical efforts. (2-4) NCHS strives for equitable policies and practices on data dissemination, ensuring that federally sponsored and funded data resources are available to all potential users – regardless of organizational affiliation. Data from NCHS are collected and made available in an open environment, with full documentation of methods and reproducible results.

2) Maintenance of confidentiality: The same law that requires that NCHS disseminate data also requires that NCHS safeguard the identity of individuals or establishments included in its data systems. NCHS, under close scrutiny from its Institutional Review Board (IRB), informs respondents as to the potential uses of their data and provides assurances on the protections and security that their data will be accorded. (5) Adhering to the terms of this informed consent – which, in effect, has the force of law – requires that NCHS maintain an overall stewardship role in managing dissemination and security of data that result from this relationship with respondents. A central challenge to NCHS is to implement the dual mandates of making data available while protecting confidentiality. Finding ways to make NCHS data available in sufficient detail for analytic purposes will often mean stretching the limits of data dissemination up to – but not beyond – the point where confidentiality is jeopardized. This dual challenge, the right to privacy vs. the need to know, affects the quantity and quality of virtually all NCHS data products.

3) Sound methodologic practices: As a component of the federal statistical community, NCHS adheres to sound statistical and methodologic practices, including the evaluation of data quality prior to making data available. Sound practices must be applied at all phases of data collection, data transmission, data processing, data analysis, and, finally, data dissemination.

4) Stewardship: NCHS employs an active stewardship to fulfill its obligations to its respondents. For example, to preserve confidentiality, NCHS believes it is not sufficient to rely on the agreements of data users that they will not identify individuals. It is important, rather, that NCHS develop proactive policies and practices that would secure confidentiality. As a federal statistical agency NCHS must demonstrate that it has done all it can feasibly do to maximize data availability, including minimizing the time from data collection to dissemination, to maximize quality of data, and to minimize the risk of disclosure. This includes procedures to safeguard security, careful management and tracking of identifiable data, and the application of statistical tests to proposed disclosures. In addition, this includes the application of judgment to assess the risks of releasing in various forms those data that are viewed as having particularly severe implications for individuals or institutions if there were a disclosure, such as those data that relate to illegal actions or sensitive personal behaviors. Since each data set proposed for release is different, decisions concerning when and how data are to be disseminated must be made on a case-by-case basis, within an overall framework of policies and procedures (specific data dissemination policies for each NCHS data system are forthcoming).

Micro-Data Dissemination Policy

While final decisions about data dissemination can only be made after the data have been collected, processed, and reviewed for unique disclosure issues, strategies for dissemination begin at the time that the data collection activity is being planned. In general, information will be available that could compromise confidentiality with inadvertent disclosure, possibly causing extreme harm to the respondent. In addition, it is important to address how data quality will be evaluated and decisions made concerning when data cannot be disseminated due to failure to meet quality standards. Concerns about procedures for data dissemination in general and in regard to particular data components are addressed during the planning processes, particularly for collaborative activities.

1) *When* micro-data are disseminated

Science and the public good are best served by an open exchange of findings and views. Toward that end, NCHS policy is to disseminate micro-data as soon as possible following data collection, subject only to limits imposed by resources, technology, and data quality. NCHS will not impede the prompt dissemination of micro-data in order to preserve publication rights of its staff, collaborators, or the staff of other organizations.

Prior to final release NCHS will thoroughly evaluate data quality and assure that the data release will preserve the respondents'confidentiality. Expert assistance is often needed to conduct data quality reviews. Procedures for conducting such reviews should be a part of all data planning activities.

Even when micro-data are disseminated as promptly as possible, there are situations where it would be beneficial to release a portion of the micro-data or aggregated data prior to the time when the full set of micro-data can be made available. Such requirements are included in the planning stages or raised as soon as the need for them is apparent. The need for such "early releases" or a staggered release may be raised by NCHS or its collaborators and serves to fulfill important policy and scientific goals.

In keeping with the goal of widespread dissemination of the data collected, once analyses are published in any way and/or once final data files are provided to any collaborator or requestor for analytic purposes, NCHS will provide for more general public access (the form of which would be guided by confidentiality considerations and consistent with informed consent) to those data to ensure that other analysts can reproduce results or reinterpret the data.

2) Content: *what* data are released

NCHS will make micro-data available in the most detailed form possible, subject only to limits imposed by data quality and the need to protect confidentiality.

Selected data will be provided to collaborators for their expert assistance in evaluating data quality when appropriate. These data are in the form of quality assurance/quality control

(QA/QC) data files. Procedures for carrying out this component of the quality control program are developed during the planning stage. The evaluation approach will depend on the nature of the data, past knowledge of the characteristics of the data, availability of expertise, and the specifications in the formal agreement between NCHS and the collaborators.

3) Recipients of NCHS micro-data

NCHS will disseminate the data it has collected as widely as possible, subject only to limits imposed by resources, confidentiality, technology, and data quality.

On occasion NCHS will make available identifiable data to be used only with strict protections for insuring confidentiality. No individual – at NCHS or elsewhere – may claim entitlement to obtain or access identifiable data collected by NCHS by virtue of his or her employment. Access to identifiable data is not determined solely by employment status, organizational affiliation, or financial commitment. More important are the need for the identifiable data, the use to which the data will be put, and the requestor's role and responsibility with respect to the data collection activity. Since any access to identifiable data poses risk, access to such data will be carefully evaluated and tracked after access is granted.

4) Mechanism for data release

The form of NCHS data release varies as determined by disclosure review, ranging from a general public release, to special use files, to more restrictive access in the NCHS data center.

If a micro-data release is addressed in the assurance of confidentiality made to NCHS respondents, NCHS may release potentially identifiable data (generally, but not always, release is to other agencies in the DHHS). However, as stated in the NCHS Staff Manual on Confidentiality,

" ... NCHS would not countenance the transfer of any confidential data to another part of the Department without positive assurance that the data will be used only for the authorized purpose and that the confidentiality of the data will be protected quite as effectively in the other organization as it would be by NCHS itself."(8)

Such positive assurance is stated in an Inter-Agency Agreement, Memorandum of Understanding or other legally enforceable agreement providing details concerning applicable law and government regulation, permissible disclosure, legal responsibilities, treatment and final disposition of confidential records, and the designation of specific persons responsible for the security of such records. Whatever their form, any document developed for the transfer of confidential records from NCHS to another agency must have the written approval of the *Director of NCHS* and be signed by the Director of the Agency receiving the records or other person with broad legal responsibility.

Official agreements that authorize the release of data tapes (either Memoranda of Understanding [MOU] or Inter-Agency Agreements [IAA]) contain provisions for the handling of the data at the conclusion of the research. Generally tapes should not be held outside the NCHS for any longer than is necessary (no more than two years at a time). The researcher/s must return to NCHS all data files (including any copies or backup files) by the date specified in the agreement or provide official confirmation of their destruction. Those needing additional time enter into a new agreement with NCHS. A formal letter to NCHS notifies the Center of final disposition of the tapes.

Research Data Center: If data cannot be released publicly or through special use agreements, NCHS will ensure access through more secure mechanisms, such as Research Data Centers or similar secure access entities (RDC). (7) The continuing demand for analyses that require data with lower levels of geography such as States, counties, and smaller areas, but without confidential identifiers such as names or social security numbers, has been the impetus for the creation of the Research Data Center located at the NCHS headquarters in Hyattsville, Maryland. Designed for the researcher outside of NCHS, this RDC allows access to data that would not be permissible to analyze because of confidentiality/disclosure rules and regulations.

Information that would, if accessed with no restrictions whatsoever, be considered identifiable and not releasable can, under the restricted conditions of RDCs, be subject to statistical manipulation. While information concerning named geographic entities cannot be accessed, data ordered by such units can be analyzed at a level not possible with public use data.

Prospective researchers must submit a research proposal that will be reviewed and approved by a committee whose judgment is based upon the availability of RDC resources, consistency with the mission of NCHS, general scientific soundness, and the feasibility of the project. It is expected that the user will develop the research proposal with the RDC staff to minimize the time required. Although researchers will sign confidentiality agreements, strict confidentiality protocols require that researchers with approved projects must complete their work using the facilities located within an RDC.

NCHS will also continue to invest in new technology and approaches for data access that will simplify and facilitate access to non-public data for users and will explore the possibility of establishing RDCs at additional sites as resources permit.

5) Procedures for data release

Based on the above principles, NCHS has developed the following procedures for micro-data dissemination.

a) *Data quality evaluation*: During and after data collection, NCHS processes (e.g., cleans, codes, edits) and evaluates the quality of the data. Quality control is an intricate aspect of all data collection and processing activities, but final quality evaluation cannot take place until collection and processing are complete. DHHS and NCHS internet web sites contain statements regarding information and data quality standards and how they are employed. (10, 11)

b) *Public release disclosure assessment*: In addition to evaluating data quality, NCHS evaluates the data to determine whether a public release would put the identity of individuals or establishments at risk. This evaluation takes into consideration issues such as 1) the level of detail for which data would be released (particularly as regards geographic specificity, and variables known to be held in common with outside data sources that serve as matching keys to increase the risk of identification); 2) certain variables or combinations of variables that render respondents unique within the sample and which might facilitate their recognition to outsiders; and 3) other linkable data already available outside NCHS, such as those already released from the same or a related survey or information held by others from the same respondent. NCHS is guided by the Privacy Act of 1974 and section 308(d) of the Public Health Service Act (42 U.S.C. 242m) as well as resources from the Federal Committee on Statistical Methodology on disclosure of proposed data releases and disclosure limitation methodology. (1,3)

Determining the risk of disclosing identifiable data is a complex task that involves both empirical statistical analysis and judgment. NCHS has considerable staff expertise on disclosure avoidance and employs a formal Disclosure Review Board (DRB). (6) Considering the extreme sensitivity of much of the data collected, the increased public awareness of and concern for privacy, and NCHS' legal and ethical obligation to fulfill its guarantee of confidentiality, an organized, well-coordinated and statistically sound procedure for establishing acceptable levels of disclosure risk is required. The volume, complexity, and variety of data files requiring review necessitate the implementation of a formal mechanism for the review of files for release outside NCHS. Meetings of the DRB facilitate input on each data file from staff specialists in both sampling and survey statistics.

The Board is chaired by the NCHS Confidentiality Officer and includes representation across NCHS.

The DRB reviews micro-data files for public use, interagency sharing, and other authorized release together with selected tabular materials following procedures established by the Confidentiality Officer.

After informing the Confidentiality Officer of plans to release a public use or other file and scheduling disclosure review, the requesting NCHS program is provided an electronic version of the NCHS Disclosure Potential Checklist.(see Attachment). (8) This document (patterned after one in use by the Census Bureau for many years and adapted for use at NCHS) contains a detailed description of potential problem areas for micro-data files together with suggestions for addressing those problems. (3c) The NCHS program submits the completed checklist together with file documentation, survey background, and other essential documents to the Confidentiality Officer.

Since confidentiality can never be absolutely assured, the risks and benefits of providing access must be weighed against the disclosure risks and sensitivity of the data. Because of the sensitive nature of many of the topics and materials discussed, the meetings and minutes of the Board are

considered confidential. Full details concerning the rationale for the decisions made are, however, shared with the NCHS program and any outside collaborator involved.

c) *Managing access to identifiable data:* Access to identifiable data by NCHS staff, collaborators, other CDC or Federal agency staff, or others permitted by law, is managed by NCHS to ensure strict adherence to confidentiality practices and procedures. The NCHS Staff Manual on Confidentiality (a document distributed to all NCHS employees) states clearly that "each employee of NCHS is responsible for maintaining and protecting at all times the confidential records that are in the employee's presence or under the employee's control. In addition, each employee must at all times follow the principles and obey the laws, rules, and regulations that are cited or referenced in this manual.

"To assure that the employee is fully aware of his responsibilities, each person, on entering employment in NCHS, is given" a detailed statement on nondisclosure and must sign, attesting that he or she has carefully read and understood the stipulations regarding unauthorized disclosure. (8)

Because of the need to limit access to identifiable data based on use and need, NCHS is frequently involved in the evaluation of proposed uses/analyses. While NCHS does not judge the scientific legitimacy of uses to which data collected by NCHS will be put or the analytic methods employed by non-NCHS analysts, requests for access to identifiable data must be evaluated in terms of any potential risk to confidentiality and disclosure. When scientific differences arise, NCHS will seek to engage independent/neutral reviewers before making a decision regarding data release or access.

d) *NCHS Director responsibility*: The Director is the official who has delegated responsibility under Section 306 of the PHS Act, is the signer of letters providing assurances to respondents, and is the responsible official to the IRB. Therefore, the NCHS Director must be in a position to provide stewardship of NCHS identifiable data. Since data release decisions involve judgment, the NCHS Director should have access not only to the views of NCHS program officials but also to those of collaborators and data users. Where there are differences of opinion on the nature of a data release, the Deputy Director, NCHS, will be charged with the responsibility of ensuring that the Director has access to a fair and objective presentation of views, including materials developed by collaborators or other users seeking access to NCHS data.

e) *Report limits of the data*: NCHS will make efforts to fully and clearly outline its judgment as to the limits of NCHS data (e.g., analyses that could be supported by given sample sizes, etc.), and make this available to potential users.

Terms and Concepts

1) **Micro-data:** A data file containing information collected as part of one of the NCHS data systems in which each record provides information at the unit of data collection (e.g., individual persons, households, establishments, or events).

2) **Dissemination of data**: For the purpose of this policy, dissemination refers to any mechanism by which micro-data are made available to users. It includes mechanisms whereby data are released to users as well as those where data are made available without actually being released.

3) **Public Release of Data**: A dissemination mechanism whereby micro-data files are made available to all users using a variety of media (CD-ROM, Internet, diskette, mainframe tapes, etc). The defining characteristic of public release is that any user, including the general public, can be in possession of the micro-data, without the need for special legal status or special arrangements. The files have been edited, documented, and reviewed by the data collection program and undergo a rigorous confidentiality review by the NCHS Confidentiality Officer and the Disclosure Review Board (DRB). (6) The micro-data are judged not to contain identifiable or potentially identifiable information. Users are asked to agree not to try to obtain the identity of respondents. As NCHS does not retain any oversight of the data once released, NCHS' assurance of confidentiality is based on the disclosure review and not on the agreement of the user.

4) **Special Use Agreements**: In some circumstances, data which are not released publicly may be provided by NCHS through a special data use agreement that provides for NCHS oversight over the use of the data. No data which cannot be publicly released will be made available outside NCHS without a data user's signed written agreement to provide such safeguards as are necessary. The agreement must be countersigned by the Director, NCHS, or designee. The circumstances of such agreements are limited by 1) whether the informed consent for the data collection system provided for the data to be used by the recipient; 2) the need for such data (i.e., that the user could not accomplish the analysis with more generally available releases of the data); and 3) the ability of the recipient to provide adequate safeguards as defined by NCHS.

5) **Controlled access to micro-data**: Access to micro-data refers to making data available to users through a mechanism other than public release or special use agreements. In this case, users have access to the micro-data but are not in possession of the data. NCHS exercises more direct supervision of the data use in order to protect confidentiality. For example, users may receive access to micro-data through the NCHS Research Data Center. (7)

6) **Confidential Information**: That information given to NCHS with explicit understanding that it will not be shared with an unauthorized party. (Committee on National Statistics and the Social Science Research Council, Private Lives and Public Policies, National Academy Press, 1993). In the case of NCHS, authorization is secured by means of the informed consent process during which respondents' agreement is obtained concerning which, if any, parties may

have access to identifiable data concerning them. (10, 12)

7) **Identifiable data**: Identifiable information is any tabulation, record, or file which can be used to establish individual or establishment identity, whether directly (using items such as name, address or unique identifying number) or indirectly (by linking data about a respondent with other information that uniquely identifies the individual).

8) **Confidentiality protection**: Removal or suppression of information that could identify a survey respondent to any unauthorized entity. The manner in which data collected by NCHS are to be used and reported is specified in each informed consent statement (see below). Unless explicitly specified in the consent and agreed to by the respondent or other data provider, NCHS protects the confidentiality of all identifying information obtained through its data collection systems.

9) **Informed Consent**: Agreement of the respondent or provider of the data to participate in an NCHS data collection activity after being fully informed of the nature of that activity. 45 CFR 46 (the Common Rule) describes the information that must be provided as part of the informed consent process. (5)

10) **Collaborator**: When applying 308(d), a collaborator or collaborating parties are those with whom NCHS has a formal working relationship at the inception of a survey or project. In most circumstances a formal working arrangement is defined in such documents as a Memorandum of Understanding (MOU) or Inter-Agency Agreement (IAA) but may also be defined in other appropriate instruments. A collaborator must have established a formal working arrangement with NCHS at the initial planning and design stages. Thereafter, the collaborator must have tangible and significant involvement in the planning, design, funding, or execution of the survey or project. A collaborator can be, but is not limited to, other federal agencies, state governments, universities, organizations, colleagues and others working outside NCHS with whom NCHS has a formal working arrangement, as defined in this document. All projects are performed under the auspices of legislatively mandated NCHS programs. Informed consent documents also specifically mention the involvement of one or more collaborators. Collaborators participate fully in data quality assurance/quality control and, accordingly, view micro-data files as part of the eventual data release process.

References

1) Public Health Service Act. Section 306

Current Legislative Authorities of the National Center for Health Statistics, 1999. 20 pp. (PHS) 2000-1303.
http://www.cdc.gov/nchs/data/misc/legis99.pdf

2) Principles and Practices for a Federal Statistical Agency

Committee on National Statistics, Commission on Behavioral and Social Sciences and Education. National Research Council. *Principles and Practices for a Federal Statistical Agency*. Second Edition. Washington, DC: National Academy Press 2001.

3) Federal Committee on Statistical Methodology (FCSM).

http://www.fcsm.gov

3a) Federal Committee on Statistical Methodology. (1978) *Report on Statistical Disclosure and Disclosure-Avoidance Techniques*. Statistical Policy Working Paper 2. Washington, DC: Office of Management and Budget, Office of Information and Regulatory Affairs, Statistical Policy Office. **http://www.fcsm.gov/working-papers/sw2.html**

3b) 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. **http://www.fcsm.gov/working-papers/spwp22.html**

3c) Interagency Confidentiality and Data Access Committee, FCSM (July 1999). *Checklist on Disclosure Potential of Proposed Data Releases.* Washington, DC: Office of Management and Budget, Office of Information and Regulatory Affairs, Statistical Policy Office.

http://www.fcsm.gov/committees/cdac/checklist_799.doc

4) OMB Directive on Statistical Confidentiality

OMB Directive on Statistical Policy.
http://www.whitehouse.gov/omb/inforeg/statpol.html

5) Human Subjects Regulations

Code of Federal Regulations Title 45 Public Welfare Department of Health and Human Services. National Institutes of Health. Office for Protection from Research Risks Part 46.

Protection of Human Subjects.
http://ohrp.osophs.dhhs.gov/humansubjects/guidance/45cfr46.htm

6) Panel on Disclosure Review Boards of Federal Agencies**:** Characteristics, Defining Qualities and Generalizability

http://www.fcsm.gov/committees/cdac/DRB-Panel.pdf

7) NCHS Research Data Center

For more information on the NCHS RDC refer to **http://www.cdc.gov/nchs/r&d/rdc.htm**

8) NCHS Staff Manual on Confidentiality

U.S. Department of Health and Human Services. Centers for Disease Control and Prevention, National Center for Health Statistics. (May 1999) *NCHS Staff Manual on Confidentiality*.

9) Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics. George T. Duncan, Thomas B. Jabine, and Virginia A. de Wolf, Editors; Panel on Confidentiality and Data Access, National Research Council (1993)

10) DHHS Information and Data Quality Standards

http://hhs.gov/infoquality/

11) NCHS Information and Data Quality Standards **http://www.cdc.gov/nchs/about/quality.htm**

12) Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies. Pat Doyle, Julia I. Lane, Jules J. M. Theeuwes, and Laura V. Zayatz. Elsevier (2001)

June 2002 Attachment

NCHS CHECKLIST ON DISCLOSURE POTENTIAL OF DATA

NOTE: Your responses to the questions in this checklist must be treated as strictly confidential.

To differentiate responses from questions, please use an easily distinguishable color. Blue is recommended. If you need more space for an answer, please attach a continuation sheet and identify the number of the question.

Overview of Contents

- **Section 1. General Information:** This asks for basic information about the proposed data release.

- **Section 2. Details on the Microdata File**

Most micro-data files contain data collected from persons or households (referred to as **socio-demographic data**). *Some questions in this section may not be applicable for establishment-based files.*

A major part of this Checklist focuses on geographic information because it is the key factor in permitting identification. While few respondents could likely be identified within a single State, more respondents -- especially those with rare and visible reported characteristics -- could be identified within a county or other small geographic area. In addition to the direct naming of geographic areas, the Checklist elicits geography that may be "implicitly" contained in details concerning sample units or design or variables with a geographic reference.

The risk of inadvertent disclosure is higher in a data set that has both small geographic variables *and* an extensive and detailed set of variables. Certain variables, values for which are very detailed, carry a high risk of identification, for they are very likely to result in a statistical "outlier" or a one-of-a-kind case. For this reason, a number of questions focus on the occurrence of study subjects with extreme or very unusual socio-demographic characteristics.

Interspersed with questions concerning the file under review are suggestions for techniques to reduce or eliminate the disclosure risk presented by extreme or unusual values.

Disclosure risk is also often a function of the quality and quantity of "auxiliary" or "contextual" information (data from sources external to the data being released). Because this information is externally available in a form which may contain names or other identifiable information, it can serve to render a file vulnerable.

In addition, other organizations may have gathered data from the same persons or establishments, using techniques and code structures similar to those employed by NCHS. In such cases, the public availability of such external data may require "coarsening" the data set under review by dropping survey

variables, suppressing information for certain respondents, or collapsing response categories for other variables.

For surveys of establishments, the issues are generally different because such entities are often selected from very skewed populations. For example, in the U.S., there are very few hospitals with 1,000 or more beds, and inadvertent disclosure in a survey of hospitals might be possible using detail on the number of beds and geographic information as large as a Census region.

Final note: Responses to questions in the Checklist are not intended to supply all of the information required by a Disclosure Review Board before a micro-data file or table is released to the public. Some additional questions may need to be answered and/or given special consideration. Those questions will be taken up in detail during the remainder of the disclosure review process.

Section 1. General Information

SURVEY TITLE:_____Date of submission_____

Project Mgr.Name: _____ Div _____Br _____Phone_____

Person completing this Checklist _____Phone_____

Co-sponsoring Agency(ies):_____

Age of Data at Proposed Time of Release:_____(years)

Check the applicable categories below:

[ ] This application is for a single data product.

[ ] This application is for a series of releases with substantially the same content.

(Specify the interval at which future products will be released.)
_____

[ ] This application is for the re-release of an approved product, with the addition of supplemental or previously unreleased data.

(If marked, give the date the original product was submitted)
_____

***(Only those checklist questions for which the answers are now different need be completed*.)**

[ ] Other application(s) will likely follow based on the same survey data set.

Additional materials

The proposed layout and content of the data file.

[ ] The proposed layout and content of the data file are attached.

Section II.. Details of the Micro-data File

1.1 Geographic Information on the File

Identify the variables for geographic identifiers on the file and the minimum population size for such geographical areas. Generally one has to balance the level of survey detail against the level of geography. The greater the amount of detail, the more risk is entailed for lower levels of geography. Similarly, with very high levels of geography, greater detail may be made available.

General Rule: All geographic areas that are identified must have a minimum of 100,000 persons in the sampled area (according to latest Census or Census estimate).

Caution: the figure of 100,000 is not without some risk. For certain target populations, the members of which are found infrequently in a population, a higher number may be desired.

1.1.1 Have you chosen to adopt the above rule or another?

____ 100,000
____ Other; specify and provide rationale
_____
_____
_____

In addition to explicit geographic identifiers on the file, the data items, record identifiers, or file structure may provide additional geographic information by inference. Therefore, steps must be taken to avoid inadvertently identifying geographic areas that do not meet the specified minimum population criteria. Potential problem areas are discussed below.

1.1.2. Primary sampling unit (PSU) or other geographic information are often embedded in control numbers designed for internal use.

How will this problem be avoided on the released file?

___ Control numbers deleted or do not contain geographic information.

___ Control numbers scrambled; describe _____
_____
_____

___ Other; describe_____
_____
_____

1.1.3. Records in many data bases are sequenced so that the first cases are in the lower numbered PSU or county that is first in alphabetic order.

Briefly, describe how the records on this file will be sequenced to avoid such geographic inferences.

_____

_____

1.1.4. Data items that imply specific geography of residence may reveal more than the explicit identifiers displayed. Examples: duration of residence codes revealing State of current residence ("lifetime," "always," years equal to age of respondent); a migration code specifying movement from a metro area to a nonmetro area when metro-nonmetro status has been excluded; residence within X miles of a nuclear reactor or an airport or health care provider when there is only one in an identified geographic area; a housing type that may be unique to a given area; a telephone area code; or latitude and longitude coordinates.

List all items that will be deleted for this reason._____

_____

Identify other geographic-related variables but not identifiers on the file (e.g., center city, noncenter city, metropolitan area, nonmetropolitan area) on the file.

List all items that you think might have geographic significance but could not decide if they should be deleted.

_____

_____

_____

_____

1.1.5. Sampling information also may provide some geographic indicators. For example, certain sampling weights may distinguish between self-representing and nonself-representing PSUs or identify types of areas intentionally oversampled. Also, codes for "second stage units," "Hit number," etc., may be related to geography.

List all sampling information - including that for variance estimation - that will be deleted for confidentiality reasons or subsampling plans to make weights less identifying.

_____

_____

_____

_____

List all other sampling information that you think might have geographic

significance but could not decide if it should be deleted.

_____

_____

_____

_____

1.2. File Contents Presenting an Unusual Risk of Individual Disclosure

The disclosure criteria for public-use micro-data require a review of each file to determine if any of the proposed contents present an unusual risk of individual disclosure. The Disclosure Review Board has identified several measures that can be taken to reduce the possibility of identifying an individual through the characteristics available on a file. The measures are discussed below, and relevant information pertaining to the proposed file is requested to assist the Disclosure Review Board in its review.

1.2.1. Names, addresses, and other unique numeric identifiers such as Social Security, Medicare or Medicaid numbers ***must*** be removed from the file.

1.2.2. High income is a visible characteristic of individuals or households and is considered to be a sensitive item of information. Therefore, each income figure on the file, whether for households, persons, or families, including total income and its individual components, should be "top" coded (i.e., placing cases with extremely high values in a category whose lower limit [e.g., income = $250,000 or more] results in the inclusion of a sufficient number of cases to eliminate "outliers" or unique values). The number of cases included must be relatively large because the range itself (e.g., an income of more than $250,000) may serve, along with other variables, to identify an individual or household.

Top codes for income variables that apply to the total universe (person/households) should include at least ½ of 1 percent of all cases. Note that the strict use of the same criterion could result in changing the cutoff from year to year, which would make things very difficult for data users. One suggested solution would be to change the cutoff only when there has been a substantial change in the upper tail of the distribution. Before making such a change, it is important to take into account how the proposed change will affect time series analyses.

For income variables that apply to subpopulations, top codes should include either 3 percent of the appropriate (non-zero) cases or ½ of 1 percent of all cases, whichever is the *higher top code*. Note that this procedure will result in more cases in non-top coded categories.

While exceptions to this rule are possible, variation from these top code rules should be discussed with the Disclosure Review Board well in advance of the final submission for approval to release a file.

Do all income Top codes satisfy the appropriate rule?

_____ Yes.

_____ No. Specify percent top coded and top code amount and briefly summarize discussions with the Disclosure Review Board.

_____
_____
_____
_____

1.2.3. In addition to continuous variables such as income, certain other characteristics may make an individual more visible than others; for example,

unusual detail on race and/or ethnicity

unusual occupation (e.g., by coding to 3 digits or more of an occupational code),

unusual health condition or cause of death (e.g., by coding to 3 digits or more in the International Classification of Disease codes),

very high age (among all study cases and *in particular*, among special subgroups such as women who have children at a very early or advanced reproductive age), or extreme age *differences* between spouses.

value or purchase price of own property, rent, mortgage amount.

Depending on the geographic detail shown on the file, consideration should be given to top coding (and/or collapsing) these items when they are represented as interval or ordinal variables. The Disclosure Review Board suggests that these top code categories include at least ½ of 1 percent of the total universe (persons/households) represented on the file (weighted counts).

In a few cases, where variables apply only to very small populations, the Disclosure Review Board may consider top code categories including approximately 3 to 5 percent of the appropriate subpopulation. Examples of approved Top codes:

Age--85 years old and over. (Approximately 1.2% of all persons in the 1990 census.)

Value of property--$500,000 or more. (Approximately 0.7% of all units, not just owner-occupied units in the 1990 census.)

Gross Rent (including utilities)--$1,000 or more. Approximately 1.2% of <u>all</u> units, not just renter occupied units in the 1990 census.)

Payments on mortgages - $1,000/month (Approximately 3.0% of all mortgage holders on the 1984 Survey of Income and Program Participation file.)

List all items that will be top coded (or collapsed) and the corresponding Top codes.

_____
_____
_____

List all other items about which you have questions regarding the need to top code.

_____
_____
_____

1.2.4. **Describe any proposed information to be released for the top coded data items (for example, means or medians of the top coded values).**

_____
_____
_____

1.2.5. There are other characteristics that may make a person highly visible, depending upon the geography, that are represented as nonordinal variables and therefore cannot be top coded; for example, codes indicating Foreign or Indian Tribal language spoken; detailed racial identification such as Eskimo, Aleut, Guamanian, or Samoan; detailed ethnic origins, codes for place of prior residence, codes for tenure in the area ("Always," "Lifetime,"), etc. In these cases, the amount of detail on the file may have to be collapsed into larger categories.

List all items that will be collapsed (or deleted) for confidentiality reasons.

_____
_____
_____

List any other items about which you have questions regarding the need to collapse the detail.

_____
_____
_____

1.2.6. Contextual Variables

Contextual or "ecological" variables are those that describe some aspect of an area, such as a state, county, census tract, or block group - percent or frequency of the area's population employed, foreign born, receiving public assistance; number of health facilities; number and specialty of physicians; local government expenditures; measures of air quality; etc.

1.2.6.1. Identify the source(s) of the contextual variables.

_____
_____
_____

1.2.6.2. Identify any contextual variables and the level at which they are coded.

_____
_____
_____

1.2.6.3. List all contextual variables that will be collapsed (or deleted) for confidentiality reasons.

_____
_____
_____

1.2.6.4. List any other contextual items about which you have questions regarding the need to collapse the detail.

_____
_____
_____

1.2.6.5. What is the lowest multiplicity for the set of contextual variables taken in totality (i.e. the lowest number of variables which, in combination, uniquely identify this file)?

_____

1.3. Disclosure Risks Associated with the Ability to Match to External Files

Efforts must be made to reduce the potential for matching micro-data on this file to data on external files, because external files may contain names and addresses, and thus can be used to identify survey respondents. Such matching may be possible if the NCHS file contains highly specific characteristics that are also found on mailing lists or administrative records maintained by other agencies or organizations. For example, the inclusion of exact date of birth or death in conjunction with county or zip code identifiers is unacceptable because these items can be matched to other data bases that contain name and address. Exact dates of events probably could be left on the file if they were recoded (by eliminating day of event). Examples of such external data bases are: voter registration lists; hospital discharge data bases; Federal, State, or local tax records; criminal justice system records; state hunting and fishing license registers; and membership rosters of certain trade associations.

Matching is also highly possible if the sampling frame for a survey comes from a source outside the agency or if the file contains information obtained from other agencies. The agency that provided the sampling frame or the auxiliary information may be able to match survey records to its original records, particularly if the survey records include data from the originating agency's files (e.g., amount of program benefit received, date of entry into program).

1.3.1. External files matchable to proposed file.

1.3.1.1. Are you aware of administrative records, research files, or a mailing list that contains data also included in this proposed file?

___ Yes; Identify. _____
_____
_____

___ No

1.3.1.2. Based on available information, will any data item on the file identify residence in a particular type of institution of which there may be only one in an identified area or for which a system of records could be obtained?

___ Yes; Identify the type of institution.
_____

___ No

1.3.1.3. Were any of the sample cases contained in the proposed file selected from a list provided by an exogenous source?

___ Yes; Identify the source and describe how and by whom sample cases were selected from the list.
_____
_____
_____

___ No

1.3.2. Matching

When an external file related to the proposed file to be released exists, several steps may be taken to reduce the possibility of matching survey data to this file; for example, selected items may be deleted or recoded, or "noise" (i.e., small amounts of random variation) may be introduced into these items.

The Disclosure Review Board cannot specify in advance exactly which steps must be taken to reduce sufficiently the potential for matching. However, it does consider several factors in determining the risk associated with releasing a file when the possibility of matching to external data bases exists; 1) the number of variables available for matching purposes, 2) the resources needed to perform the match, 3)

the age of the data, 4) the accessibility, reliability, and completeness of the external file, and 5) the sensitivity or uniqueness of the data. Some factors that make matching easier are listed below and information is requested on steps that will be taken before the file is released to reduce the matching potential. (NOTE: This information is necessary even if you are not aware of any external files that could be used in matching.)

Matching is easier:

1.3.2.1. ...if any data item or combination of items isolates any small and readily identifiable population subgroup or class. The inclusion of codes that identify very small population segments should be avoided, for example, Indian tribes or detailed occupation groups in combination with highly specific geography. Normally one has to consider more than one variable at a time if that group of variables is likely to appear together on a file or list. For example, age and sex together with country of birth and occupation could permit the disclosure of individual identity.

List all data item(s) proposed for inclusion on the file that isolate a small, readily identifiable population.

_____

_____

_____

List all data item(s) that will be altered (i.e., deleted, recoded, noise added) for this reason.

_____

_____

_____

1.3.2.2. ...if the file includes substantially every member of a population (say $p > 0.5$). Examples: large employers, high-income individuals, doctors, scientists of a specified type, or inmates of certain types of institutions. Additional subsampling frequently is appropriate within certain strata prior to data release.

Identify these populations, if any are on the file, and how they will be subsampled.

_____

_____

_____

1.3.2.3. ...if the file contains any information obtained from records or other sources where that information could serve as a link to an external file that has individual identifiers or detailed geographic information. Examples include the Area Resource Files; CDC STD Files; characteristics from a decennial census; welfare or social security data from a government agency; data from the Centers for Medicare and Medicaid Services; arrest records from a police department; benefits provided to employees such as pensions and health insurance.

List all data item(s) proposed for the file that were not obtained from an interview with the respondent.

_____
_____
_____

List all data item(s) altered or deleted for this reason.

_____
_____
_____

1.3.2.4. ...if the file includes data items frequently used for matching, such as exact date of birth, sex, and race, or if it includes other items that should be identical on both files, such as an exact income amount, real estate taxes or other taxes, or date of entry or termination from a government-sponsored program.

List these data items, if any.

_____
_____
_____

List all data item(s) altered or deleted for this reason.

_____
_____
_____

1.3.2.5. ...if longitudinal data are being collected (i.e., if the data for the same respondents/units will be collected for several different reference periods). Primary concern relates to time series of data items potentially matchable to outside records (e.g., income tax or employment records).

If data are collected from the same respondents more than once, indicate the frequency of interview, length of time any one unit may be in sample, and factors affecting the likelihood of matching a sample unit from one time period to the next.

_____
_____
_____

1.3.2.6. ...if highly specific geography is included on the file, for example, States, MSAs, etc.:

List all geographic identifiers below the level of region.

_____

1.3.2.7. ...if data collected from multiple persons in a household are linked on the released file. Disclosure risks associated with linking of household members are well-known. For example, households can be identified because of significant difference in spouses' ages, atypical number and ages of children, a "unique" multi-racial composition of the household, etc. -- not to mention the fact that one household member, by self-identification, could look up other members' reported information.

(a) Are data collected from multiple persons in a household?
G **Yes.**
G **No.** *If no, skip to 3.3.3.*

(b) If yes, describe the strategy for releasing these data, and indicate whether or not the data from these will be linked. _____

_____

1.3.2.8. **Describe any considerations not previously mentioned that <u>reduce</u> the ability to match this file to external data (e.g., unreliability or natural noise in the data).**
_____
_____
_____

1.3.3. Cross-tabulations to Identify Unique Sets of Characteristics

1.3.3.1. **Were any cross-tabulations performed to identify sets of unique characteristics?**
_____

If no, skip to 1.4.

1.3.3.2. **What were the results?**

_____
_____
_____

1.3.3.3. Will any additional steps be taken to reduce disclosure risk based on these results?

_____
_____
_____

1.4. Noise

1.4.1. Was any noise added to the data? _____

If no, skip to 1.5.

1.4.2. **What procedure(s) was used to add noise to the data? Please give specifics for that procedure (i.e., percent of records affected, distribution of noise, etc.).**

Some possibilities: **random noise**
record swapping
rank swapping
blanking and imputation

_____
_____
_____

1.4.3. Was any attempt made to match back the noise-added data to the original file? _____

If no, skip to 1.5.

1.4.4. How was it done and what was the rate of success in matching?

_____
_____
_____

1.5. Edited data (data values provided by respondents that we have altered) and imputed data (data values that we have created due to non-response) have their own "noise" built in. The processes of editing and imputation decrease the disclosure risk of a file. Please answer the questions in this section if the values are known.

1.5.1. What percent of records contain at least one imputed data item?_____

1.5.2. What percent of all data items were imputed? _____

1.6. Other Issues

1.6.1. Files that include every sample case or cases in strata that are sampled at high rates ($p > 0.5$) are more likely to lead to disclosure than files containing only a subsample of cases. For example, if it were known that a certain individual participated in a particular survey, one could infer that the person's record could be found in the corresponding microdata file, assuming all sample cases were available on that file.

Does this file contain

___ Every case?

___ A subsample of cases (if so, specify the range of sampling rates)?

_____

1.6.2. Project managers should be aware that confidentiality problems may arise if special tabulations are made from an internal version of file, which includes detail omitted from the public use file. For example, the tabulation might provide specific geography not included on the public use file, cross-tabulated by multiple data items on the file. Consult with the Disclosure Review Board if you are planning to release tabulations that make use of detail not available on the public-use file.

1.6.3. Briefly describe the sample design.

1) include a description of any stratification, clustering, and stages, including the identification of the kinds of units sampled at any stage with probability > 0.5.
2) include a comparison and contrast of the proposed sampling units, units of enumeration, and units of analysis in the study.
3) identify the information of the sample design (sampling plan and estimators that will and will not be put in the public domain, including the identity of PSUs).

4) describe how users will estimate sampling variances potentially identifying any proposed "nesting variables" on the proposed file layout or the design of any weights used for replication approaches.

_____
_____
_____
_____
_____
_____
_____
_____
_____

1.6.4. Supplements

Was this information gathered as a supplement to another survey? _____

If no, you are finished with this section of the checklist.

Can this micro-data file be linked to the file produced from the main survey? _____

If yes, what geographic information is on the main file?

_____
_____
_____