American Institutes for Research®

Child Trends

The Evaluation Data Coordination Project

INFORMATIONAL PAPERS

PREPARED FOR:

DEPARTMENT OF HEALTH AND HUMAN SERVICES
ADMINISTRATION FOR CHILDREN AND FAMILIES
OFFICE OF PLANNING, RESEARCH AND EVALUATION

FROM:

THE AMERICAN INSTITUTES FOR RESEARCH
1000 THOMAS JEFFERSON STREET, N.W.
WASHINGTON, DC 20007

AND

CHILD TRENDS, INC.
4301 CONNECTICUT AVENUE, N.W.
SUITE 100
WASHINGTON, DC 20008

May 6, 2004

GUIDELINES FOR THE RELEASE OF PUBLIC-USE DATA

This document is one of two products created under the auspices of the Evaluation Data Coordination Project (EDCP)—options documents providing measures for a number of constructs and informational papers describing guidelines for data reporting and the release of public-use data. The EDCP's overarching goal is to develop common measures of constructs and reporting formats for selected evaluation projects to facilitate and improve the quality of potential future secondary analyses and cross-project syntheses. This document provides guidelines for releasing public-use data to facilitate meta-analysis, secondary analysis, and cross-product synthesis; for protecting the confidentiality of study participants when releasing data sets; and for addressing issues relevant to creating data sets from administrative data and pragmatic issues, such as documentation. The evaluators of the nine U.S. Department of Health and Human Services/Administration for Children and Families (DHHS/ACF) studies are the primary audience; however, the information is applicable to a wide range of research and, therefore, to a broader audience. (For more information about the EDCP and the nine ACF studies, refer to the Options Document.)

This document is divided into six parts: legal references, guidelines for the release of public use data, additional suggestions for preparing documents for public use, recommendations for creating data sets from administrative data, recommendations for creating linked data sets, and additional references on data confidentiality.

I. LEGAL REFERENCES

To protect the confidentiality of data that contain individually identifiable information, researchers must be aware of relevant legal regulations and must monitor the confidentiality of individually identifiable information in their daily activities and in the release of information to the public.

Much of the research involving human subjects in the United States operates under the rule "Federal Policy for the Protection of Human Subjects" (known as the "Common Rule"; Title 45 Code of Federal Regulations [CFR] Part 46, Subpart A)[1] and/or the Food and Drug Administration's (FDA) Protection of Human Subjects Regulations (Title 21 CFT Parts 50 and 56).[2] Both the Common Rule and FDA regulations are intended to ensure the privacy of human subjects and the confidentiality of information about the subjects, with the Common Rule applicable to human subjects research conducted or supported by DHHS and the FDA regulations applicable to research involving products regulated by FDA.

Building on these existing federal human subject protection regulations, HHS established the "Standards for Privacy of Individually Identifiable Health Information," referred to as the "Privacy Rule" in 2000 in response to the mandate of the Health Insurance Portability and Accountability Act of 1996 (HIPAA; Public Law 10-191).[3] Unlike the Common Rule and the FDA regulations that apply to most

[1] Available at http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.htm
[2] Available at http://www.fda.gov/oc/ohrt/irbs/appendixb.html
[3] The Privacy Rule took effect on April 14, 2001, with a compliance date of April 14, 2003. The full text of the Privacy Rule is available at the HIPAA Privacy Web site of the Office of Civil Rights: http://www.hhs.gov/ocr/hipaa

federally funded and to some privately funded research, the Privacy Rule applies more broadly to HIPAA-defined covered entities. Regardless, the Privacy Rule guidelines outlined in Part II. can serve as a general model for researchers in other disciplines as well.

II. GUIDELINES FOR THE RELEASE OF PUBLIC-USE DATA

Guideline 1. Research Use/Disclosure With De-identified Health Information. There are no restrictions on the use or disclosure of de-identified health information, which neither identifies nor provides a reasonable basis to identify an individual. There are two ways to de-identify information: 1) a formal determination by a qualified statistician or 2) the removal of specified identifiers of the individual and of the individual's relatives, household members, and employers, which is deemed adequate only if the researcher has no actual knowledge that the remaining information could be used to identify the individual.

Guideline 2. Research Use/Disclosure With Individual Authorization. Researchers may use or disclose protected health information for research purposes when a research participant authorizes in writing the use or disclosure of information about himself or herself. Such an authorization may be combined with consent to participate in the research or with any other legal permission related to the research study.

Guideline 3. Research Use/Disclosure Without Authorization. To use or disclose protected health information without authorization by the research participant, the researcher must obtain one of the following unless the use or disclosure is for any of the 12 national priority purposes:[4]

- Documented Institutional Review Board or Privacy Board approval for an alteration or waiver of research participants' authorization for use/disclosure of information about them for research purposes.

- Representation from the researchers, either in writing or orally, that the use or disclosure of the protected health information is solely to prepare a research protocol or for similar purposes preparatory to research and representation that protected health information for which access is sought is necessary for the research purpose.

- A data-use agreement for limited data sets, according to which a limited data set that excludes certain specified direct identifiers (e.g., social security numbers, addresses, phone numbers, names, case numbers) of the individual or of relatives, employers, or household members of the individual may be used or disclosed for research, public health, or health care operations.

Guideline 4 Accounting for Research Disclosures. Upon individual request, the researcher should be able to provide an accounting of certain disclosures of protected health information about the individual. This accounting must include disclosures of protected health information that occurred during the 6 years prior to the individual's request or since the compliance date (whichever is sooner) and contain specified information about each closure. Exempt from this requirement are two types of

[4]See OCR Privacy Rule Summary for a list of the 12 national priority purposes: http://www.hhs.gov/ocr/privacysummary.pdf.

disclosures: research disclosures made pursuant to an individual's authorization and disclosures of the limited data set to researchers with a data-use agreement.

III. ADDITIONAL SUGGESTIONS FOR PREPARING DOCUMENTS FOR PUBLIC USE

The previous guidelines and recommendations address the legal, technological, and ethical issues associated with public-use data; provide essential protection for the privacy of human subjects involved in research that generated the data for public use; and ensure the ethical use of the data. A number of additional issues should be considered for the release of public-use data. Before releasing the data, for instance, the researcher may want inspect or even extensively analyze (i.e., conduct a disclosure analysis) the data to ensure figuring out the identity of a unit (e.g., organization or individual) by looking at the pattern of data would be difficult. For example, all direct identifiers must be removed from all data files being prepared for public use and distribution, but there are no definitive rules about including indirect identifiers. The DHHS/ACF and BRI Consulting suggest that a good general rule to follow is that any cell of a cross-tabulation of identifying variables should have at least 5–10 cases. Data contributors may consult the Inter-University Consortium for Political and Social Research "Guide to Social Science Data Preparation and Archiving"[5] for more detailed guidelines and recommendations.

Further, documentation on the data collection methodology and the quality of the data should accompany the data file. Specifically, the documentation maintained by principal investigators should include detailed information on file structure, format, and content of the data sets. If the study includes more than a single data file, DHHS/ACF and BRI Consulting recommend that the documentation describe the relationship between the data files and the procedures for linking records in the files to one another, including information on the variable or variables that can be used to uniquely identify individuals. In addition, descriptions of the meaning of each variable in the data file(s) and the valid codes for each variable should be included with the frequency distribution, including the number of missing variables for all categorical variables, when possible. In general, blanks should not be used as missing data codes; if missing values have been assigned, the data file must contain a variable that indicates which values are imputed.

Finally, a codebook, either in a text format or in an electronic format, should also accompany the data file. The codebook should specify the names and provide a description of every variable in the data file, including an explanation of how they were created. The codebook should also identify the range of valid codes and their meaning for each variable.

IV. CREATING DATA SETS FROM ADMINISTRATIVE DATA

The rules and regulations for creating data sets from administrative data follow the same basic principles as mentioned above; however, preparing these data sets for public release presents a number of unique challenges because of the complexity and evolving nature of administrative data systems. Given that child care policy research and policy research in general have begun to increasingly rely on the valuable information administrative data can provide, briefly addressing the issues specific to using administrative data to create research data sets is useful. DHHS/ACF and the BRI Consulting Group recommend that 1) researchers take great care to thoroughly document every variable in the research data files that are made publicly available and 2) investigators submitting administrative data include both a comprehensive description of the policy context and program guidelines for future users of the data and a description of the data context. To protect subject confidentiality in administrative data, many of the

[5] Available at http://www.icpsr.umich.edu/access/dpm.html

guidelines are identical to the ones discussed in part II. However, it is important to consider that, unlike voluntary survey respondents or interviewees, most of the subjects in administrative data files are unaware that the data collected on them is being used for research purposes.

V. CREATING LONGITUDINAL LINKED DATA SETS

Like the use of administrative data, research opportunities have further been expanded by technological advances allowing for the development of longitudinal data sets linked to health, economics, contextual geographic, and employer information. However, such technological advances have not only improved the range and depth of data collected, but also led to new methods for identifying individuals from the information made available. According to the report of the National Academy of Sciences workshop (2000), researchers must seriously consider the resolution of technical, legal, and ethical issues associated with the construction of longitudinal files that link survey, administrative, and contextual data. Researchers must weigh these issues and decide on one of two approaches that will best serve data users and maintain sufficient levels of subject confidentiality. The first approach entails physically restricting access to the data; the second approach involves altering the data to allow for safe broader public access (refer to guidelines in part II.)[6]

VI. ADDITIONAL RESOURCES ON DATA CONFIDENTIALITY

For additional resources on data confidentiality, visit the following URLs:

http://www.amstat.org/comm/cmtepc
http://aspe.hhs.gov/datacncl/privacy/
http://www.elsevier.com/inca/publications/store/6/2/2/1/2/9/index.htt
http://www.aspe.hhs.gov/hsp/leavers99/datafiles/intro.htm

For a comprehensive (100+ articles and documents) reference list related to ensuring data confidentiality assembled by ORC Macro, visit http://www.aspe.hhs.gov/hsp/leavers99/datafiles/app_c.pdf. The articles mostly address restricting access and restricting content.

For resources on federal statistical agency policies related to data release, visit

http://nces.ed.gov/statprog/Standards.asp
http://www.cdc.gov/nchs/products/elec_prods/intro/relpolic.htm

[6]View the full workshop report, "Improving Access to and Confidentiality of Research Data" at http://www.nap.edu/openbook/0309071801/html/1.html