PURDUE
UNIVERSITY

Existing Public Use Datasets

Guideline and Procedure 101
Issuing Office: Human Research Protection Program
Responsible Officer: Human Protections Administrator
Responsible Office: Human Research Protection Program
Date Issued: April 4, 2011
Most Recently Revised:

I. Summary

This guidance represents the PU HRPP's determination that PU investigators' access to specified "public use" data sets does **not** constitute research with *human subjects* (as it does not involve access to identifiable private information about the persons from/about whom the data were collected) and therefore is not subject to PU IRB review and approval or determination of exemption from PU IRB review.

The data sets to which this guidance is applicable are limited to those identified below. This guidance will be updated, as needed, to include additional data sets that are determined by the PU IRB/HRPP to meet the conditions of this guidance.

Secondary data analysis of publicly available data is a common research method. Increasingly federal agencies supporting research require investigators to make the data they collect publicly available. Additionally, many professional organizations and journals require that research data sets of published works be made accessible to encourage scholarly interpretation and replication of research.

Publicly available data sets stripped of identifiers do not require IRB review under 45 CFR 46. This guideline defines "publicly available" and other terminology to better assist investigators in conducting research.

II. Definitions

1. ***Publicly Available*** means that the general public can obtain the data. Data are not considered "publicly available" if access to the data is limited to researchers.

2. ***Public Use Data Sets*** are data sets prepared by investigators or data suppliers with the intent of making them available for public use. The data available to the public are not individually identified or maintained in a readily identifiable form. Data suppliers may have both (a) publicly available de-identified as well as (b) restricted use data[1] from the same set. Data shared informally among colleagues does not constitute public use data.

[1] Restricted use data are **not** publicly available and are **not** covered under this policy. Restricted use data is defined as files distributed by federal agencies and research organizations upon which use restrictions are imposed. These files generally contain data fields, such as social security numbers, names, or extensive life history markers that might enable an unauthorized user to identify a participant. These files are usually, but not always, accompanied by a data use agreement that details restrictions on use of the data. The restrictions vary, but they typically involve secure (locked) data storage and password protected computers, and forbid the storage of data on computer hard drives that may be accessed through a computer network connection.

3. ***Formal Disclosure Analysis*** is a process in which reasonable attempts are made to identify individual subjects in a data set using the existing data variables, and/or by combining two or more data sets together. The formal disclosure analysis, if successful, proves that subjects cannot be identified using the information existing in the database.

III. Guidelines and Procedures

Public Data Sets That Do Not Require PU IRB Review

1) Research projects involving only the analysis of public use data from the following pre-approved public data sets/repositories will **not** require PU IRB approval or determination of exemption prior to beginning the research.

- Inter-University Consortium for Political and Social Research (ICPSR)
- Better Access to Data for Global Interdisciplinary Research (BADGIR)
- National Center for Health Statistics
- National Center for Education Statistics
- National Child Development Study
- National Election Studies
- Roper Center for Public Opinion Research
- University of Wisconsin-Madison Data and Information Services Center (DISC)
- U.S. Bureau of Census
- U.S. Bureau of Labor Statistics
- The University of Michigan Health and Retirement Study (HRS)
    - Unrestricted data sets only
- Panel Study of Economic Dynamics (PSID)
- Survey of Consumers (SCA)
- Integrated Public Use Microdata Samples – International (IPUMS-i)
- Luxembourg Income Study Project Archive

2) The PU IRBs have determined that data in the above listed data sets have been stripped of identifiers and are publicly available. As a result, research using these data does not meet the definition of "human subjects research" as set out in 45 CFR 46.102, and therefore does not require IRB review and approval or determination of exemption. Investigators whose research solely involves the use of one or more of these public data sets do not need to seek a determination from, or submit an application to, the PU IRB/HRPP except as described in Section 3 below.

3) Research projects that merge public use data sets in such a way that individuals may be identified or which are designed to enhance a public use data set with identifiable or potentially identifiable data are not covered by this guidance and require prior PU IRB review and approval or determination of exemption.

4) If an investigator's use of data is directed by the terms of a data use agreement, where the agreement is more stringent, the terms of the agreement takes precedence over

federal guidance about applying the definition of research with human subjects. For example, a data use agreement could require: (a) IRB review of research that may not meet the federal definition of human subjects research, or (b) that research eligible for exemption or expedited review be reviewed by a convened IRB.

Submitting a Data Set for Consideration

5) Investigators may apply to the PU IRB/HRPP to have a data set registered as a public use data set under this guidance. The data set must be a pre-existing, publicly available data set not yet approved.

6) Data sets that may qualify for inclusion on PU IRB's list of approved data sets include:

   a) Public use data sets posted on the internet that include a responsible use statement or other confidentiality agreement for authors to protect human subjects (for an example, see the ICPSR's responsible use statement[2]).

   b) Public and/or published data sets, accessible without restriction (e.g., a password is not necessary[3]) and containing readily identifiable information such that individuals can reasonably expect this information to be available to the public (e.g., letters to the editor, web logs or blogs, etc.).

   c) Public and/or published data sets, with restrictions to access, that contain data that are presented in aggregate form only (e.g., zip code); thus individuals cannot be identified.

7) Investigators must submit the following information on potentially eligible data sets to the PU IRB/HRPP office prior to conducting research:
   a) Name of data set;
   b) URL of the data set or other information on how to obtain the data set; and
   c) Abstract (maximum one page) describing the content and potential use of the data set.

8) For a public use data set to be listed under this guidance by the PU IRB/HRPP, the data set must adhere to the following rules:

   a) the data set must be publicly available to any person through unlimited access or via a member institution or for a fee;

   b) the original data collection was gathered in anonymous form or on unknown persons, or the original data collection was gathered on identified subjects but the

[2] The Inter-University Consortium for Political and Social Research (ICPSR)'s Responsible User Statement is available online at http://www.icpsr.umich.edu/icpsrweb/ICPSR/irb/index.jsp.

[3] Alternately, if you must register with a site or organization to gain access, the registration for login and password must be without qualifications, i.e., anyone could register with the site.

data set has been stripped of direct identifiers and indirect identifiers that may risk disclosure of subjects' identity; and

c) a formal disclosure analysis was performed to reasonably affirm that identification of individual subjects using variables within the data set cannot occur.

IV. Investigator Responsibility

1. PU investigators whose research project only involves secondary analysis of public use data from the pre-approved public data sets/repositories identified in Section III of this guidance do not need to obtain PU IRB approval or determination of exemption prior to access to the data and do not need to seek a determination from, or submit an application to, the IRB/HRPP, except as described in Section III(3) of this guidance.

2. PU investigators whose research project only involves secondary analysis of public use data from a data set which is not identified in Section III of this guidance may apply to the PU IRB/HRPP to have the data set included in Section III. The process for doing so is outlined in Section III(5-8) above.

3. PU investigators whose research project involves secondary analysis of public use data from one or more of the pre-approved public data sets/repositories identified in Section III of this guidance, and includes additional access to non-public data and/or interaction or intervention with human subjects must submit an application for PU IRB review and approval or exemption request.

4. PU investigators who intend to merge public use data sets or enhance a public use data set with identifiable or potentially identifiable data must submit an application for PU IRB review and approval or exemption request.

5. PU investigators must abide by the conditions of any applicable data use agreements governing the data to be accessed. If use of data is directed by the terms of a data use agreement that requires IRB review of research that may not meet the federal definition of human subjects research, the PU investigator must submit an application for PU IRB review and approval.

V. IRB/HRPP Responsibility

1. The PU IRB Chair or designated PU IRB member will review public use data set registration requests.

2. The PU Chair or designated PU IRB member will consider the following factors when reviewing public use data set registration requests for data originally collected with identifiers:

a. removal of any identifiers of a human subject or of persons named by a human subject;
b. removal of any variables that by definition would serve as surrogates for the identity of a human subject;
c. collapse or combine categories of a variable to remove the possibility of identification due to a human subject being in a small set of persons with specific attributes regarding a variable (e.g., due to the infrequency of subjects in a lower or upper range);
d. collapse or combine variables to provide summary measures to mask what otherwise would be identifiable information;
e. use of statistical methods, where necessary, to add random variation with variables otherwise impossible to mask; and
f. removal of any variables that could be linked to identifiers by secondary users.

3. HRPP staff will provide written notification of the results of the review of the public use data set registration request to investigators.

4. HRPP staff will update this guidance as data sets are approved. Newly approved data sets will be added to Section III, and the updated version of the guidance will be posted on the HRPP website http://www.irb.purdue.edu.

VI. Applicable Regulations and Guidelines

45 CFR 46.102(d)
45 CFR 46.102(f)

National Human Research Protections Advisory Committee (NHRPAC), Recommendations on Public Use Data Files, approved at January 28-29, 2002 NHRPAC meeting.

OHRP Guidance on Engagement of Institutions in Human Subjects Research, October 16, 2008.

OHRP Guidance on Research Involving Coded Private Information or Biological Specimens, October 16, 2008.

VII. Related Documents

100 Determination of Human Subjects Research

Approval

_____Date: _____
Richard D. Mattes, Ph.D.
IRB Chair

_____Date: _____
Peter E. Dunn, Ph.D.
Associate Vice President for Research
Director, University Research Administration