

Guidance and Procedures Number: 42
Title: Research Involving Public Use Data Files
Date of Last Revision: March 5, 2009

I. Introduction

For more information about what constitutes human subjects research, and IRB/OPRS determinations thereof, see [HRPP Guidance & Procedure # 2: Determining When Research Activities Require UCLA IRB/OPRS Review](#).

This guidance represents the UCLA HRPP's determination that UCLA 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 UCLA IRB review and approval or Certification of Exemption from UCLA IRB review.¹

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 which are determined by the UCLA IRB/OPRS to meet the conditions of this guidance.

II. Definitions

- A. **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.
- B. **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 data² from the same data set. Examples of public use data sets include portions of the U.S. Census data, and the Health and Retirement Survey. Data shared informally among colleagues does not constitute public use data.

¹ Research analysis of existing data about *human subjects* requires UCLA IRB approval or certification of exemption from UCLA IRB review prior to UCLA investigators' access to and/or provision of the data. Please see [OPRS/IRB Policy 41: Research Involving Secondary Use of Existing Data](#) for details.

² Restricted use data are *not* publicly available and are *not* covered by this Policy. Restricted use data is defined as special 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. The use 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. These agreements may also limit the types of analyses that are done by the investigator.

- C. **Formal Disclosure Analysis:** A process in which investigators make reasonable attempts to identify individual subjects in a database using the existing data variables, and/or by combining two or more databases together. The formal disclosure analysis, if successful, proves that subjects cannot be identified using the information existing in the database.

III. Use of Public Data Sets That Do Not Require UCLA IRB Review

- A. Research projects involving only the analysis of public use data from the following pre-approved public data sets/repositories will not require UCLA IRB approval or certification of exemption from UCLA IRB review prior to access to the data:
- British Household Panel Survey
 - German Socio-Economic Panel Survey
 - Integrated Public Use Microdata Series - International
 - Inter-University Consortium for Political and Social Research (ICPSR)
 - Luxembourg Income Study Project Archive
 - National Center for Health Statistics
 - National Center for Educational Statistics
 - National Child Development Study
 - National Election Studies
 - UCLA Institute for Social Science Research (ISSR) Data Archive
 - U.S. Bureau of the Census
 - U.S. Bureau of Labor Statistics
 - 1970 British Cohort Study
- B. The UCLA 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”³, and therefore does not require IRB review and approval or certification of exemption from IRB review.** 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 IRB/OPRS, except as described in Section C.
- C. 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 UCLA IRB approval or certification of exemption from UCLA IRB review.
- D. If an investigator's use of data is governed by the terms of a data use agreement, the terms of the agreement take precedence over federal guidance about applying the definition of research with human subjects. A data use agreement could require: (1) IRB review of research that may not meet the federal definition of *human subjects research*, or (2) that research eligible for exemption or expedited review be reviewed by a convened IRB.

³ 45 CFR 46.102

IV. Submitting a Data Set for Registration

- A. Investigators may apply to the UCLA IRB/OPRS to have a data set registered as a public use data set under this guidance. The data set may be a pre-existing publicly available data set not yet approved, or a new data set developed by the investigator which one wishes to make publicly available.
- B. Data sets that may qualify for inclusion on UCLA IRB's list of approved data sources include:
1. 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).⁴
 2. Public and/or published data sets, accessible without restriction (e.g., a password is not necessary⁵), and containing readily identifiable information and where individuals can reasonably expect this information to be available to the public (e.g., letters to the editor, web logs or "blogs", etc.).
 3. 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.
 4. Survey data distributed by UCLA investigators who can certify that:
 - a. the data collection procedures were approved by an IRB that satisfies the Common Rule criteria for an IRB, and
 - b. the data set and documentation as distributed do not contain information that could be used to identify individual research participants.
- C. Investigators must submit the following information on potentially eligible data sets to the UCLA IRB/OPRS office prior to conducting research:
1. Name of data set;
 2. URL of the data set or other information on how to obtain the data set; and
 3. Abstract (maximum one page) describing the content and potential use of the data set.

Information may be submitted to:

UCLA Office for Protection of Research Subjects
Attn: Request for Pre-Approval of Public Use Dataset
11000 Kinross Ave., Suite 102
Los Angeles, CA 90095-1694 (if sending via Campus Mail, use Mailcode 169407)
(310) 825-7122

⁴ The Inter-University Consortium for Political and Social Research (ICPSR)'s Responsible Use Statement is available online at: <http://www.icpsr.umich.edu/org/policies/respuse.html>.

⁵ Alternately, if you must register with a site or organization to gain access, the registration for login and password must be without qualification, i.e., anyone could register with this site.

- D. For a public use data file to be registered under this guidance by the UCLA IRB/OPRS, the data files must adhere to the following rules:
1. the data file must be publicly available to any researcher through unlimited access or via a member institution or for a fee;
 2. 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 data file has been stripped of direct identifiers and indirect identifiers that may risk disclosure of subjects' identity; and
 3. a formal disclosure analysis was performed to reasonably prevent the identification of individual subjects using variables within the database.

V. Investigator Responsibility

- A. UCLA 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 UCLA IRB approval or certification of exemption from UCLA IRB review prior to access to the data and do not need to seek a determination from, or submit an application to, the IRB/OPRS, except as described in Section III(C) of this guidance.
- B. UCLA 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 UCLA IRB/OPRS to have the data set included in Section III. The process for doing so is outlined in Section IV above.
- C. UCLA 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 UCLA IRB review and approval or Certification of Exemption from UCLA IRB review.⁶
- D. UCLA Investigators who intend to merge public use data files or enhance a public use data file with identifiable or potentially identifiable data must submit an application for UCLA IRB review and approval or Certification of Exemption from UCLA IRB review.
- E. UCLA Investigators must abide by the conditions of any applicable data use agreements governing the data to be accessed. If use of data is governed 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 UCLA Investigator must submit an application for UCLA IRB review and approval.

⁶ Please refer to [OPRS/IRB Policy 6: IRB Review Process - Initial Review](#) or [OPRS/IRB Policy 5: OPRS Review Process - Certification of Exemption from IRB Review](#) for details.

- F. UCLA Investigators who propose to provide public use data files are responsible for having the public use data files appropriately reviewed by the UCLA IRB before making the data available to the public.

VI. IRB/OPRS Responsibility

- A. The UCLA IRB Chair or designated UCLA IRB member will review public use data set registration requests.
- B. The UCLA IRB Chair or designated UCLA IRB member will consider the following factors when reviewing public use data set registration requests for data originally collected with identifiers:
 - 1. removal of any identifiers of a human subject or of persons named by a human subject
 - 2. removal of any variables that by definition would serve as surrogates for the identity of a human subject
 - 3. 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)
 - 4. collapse or combine variables to provide summary measures to mask what otherwise would be identifiable information
 - 5. use of statistical methods, where necessary, to add random variation with variables otherwise impossible to mask
 - 6. removal of any variables that could be linked to identifiers by secondary users.
- C. OPRS staff will provide written notification of the results of the review of the public use data set registration request to investigators.
- D. OPRS 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 OPRS website (<http://www.oprs.ucla.edu>).

Regulations:

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

References:

National Human Research Protections Advisory Committee (NHRPAC), Recommendations on Public Use Data Files, approved at January 28-29, 2002 NHRPAC meeting.
(<http://www.hhs.gov/ohrp/nhrpac/documents/dataltr.pdf>)

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.