The National Human Subjects Protection Advisory Committee (NHRPAC) approved the following recommendations on Public Use Data Files at the January 28-29, 2002 Committee meeting.

The content of this document does not represent the official views or policies of the Office for Human Research Protections (OHRP) nor of the Department of Health and Human Services (HHS). The content represents solely the advice and views of the National Human Research Protections Advisory Committee that were provided to the Secretary of HHS, the Assistant Secretary for Health, and OHRP for their consideration.

Recommendations on Public Use Data Files

These recommendations are intended to advise Institutional Review Boards (IRB) on the use of publicly available data by investigators at their institutions. They are also intended to aid IRBs in their review of protocols to create public use data files.

Background

In the social and behavioral sciences, data sets are often created with the intent of making them available for multi-users such as research analysts and other relevant users. Some data resources like the U.S. Census are federal statistical data collections; others like the General Social Survey or the Panel Study of Income Dynamics are collected in academic settings by research teams. Within the social and behavioral sciences, there is a culture of data sharing and data access to advance the public good through verification of findings, testing of rival hypotheses, and asking important new questions. Federal funding agencies, professional organizations, and scholarly journals encourage investigators to de-identify data files and make them publicly accessible to other users for secondary analysis. In other research disciplines, there may also be publicly available data files developed for similar purposes. These recommendations are applicable to all public use data files.

Definition

Public use data files are data files 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.

Features of Public Use Data Files

Public use data files are data files that have been reviewed under the jurisdiction of an IRB with the intent of making them available for public use. In the case of Federal statistical data collections, the Federal government has responsibility for that review.

Public use data files fall outside of the C~~c~~ode of Federal Regulations for the Protection of Human Subjects, once they have been appropriately classified as public use data files.

Producers and suppliers of public use data files are responsible for having public use data files appropriately reviewed by IRBs before making them available to the public.

Once data files are certified by an IRB as public use data files, no further review by subsequent IRBs is required.

Responsibility of Users of Public Use Data Files

Users of public use data files do not need to obtain IRB approval to use such files or seek a determination that the use of the public data files meets the criteria for being exempt from IRB review. However, users must follow any institutional policy in place for recording that they are using data files classified for public use.

Users seeking to merge public use data files or enhance a public use data file with identifiable or potentially identifiable data need to obtain IRB review and approval.

Approval of Data as Public Use Files by IRBs

Protocols submitted by investigators or research teams may include plans for making data publicly available. In reviewing data files for public use, IRBs should determine that either:

(a) the original data collection was gathered in anonymous form or on unknown persons or

(b) 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.

When IRBs are satisfied that a protocol involves the collection of data in anonymous form or on unknown persons, these data should be classified as public use data files. When IRBs are asked to authorize public data files from data originally collected with identifiers, the following factors should be considered by the IRB to be certain the data files has been effectively de-identified for analysis by secondary users.

(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 (for example, 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

(f) removal of any variables that could be linked to identifiers by secondary users

Inventory of Public Use Data Files

Institutions are urged to post a list of approved publicly available data sets, including data files reviewed at that institution that qualify for public availability. Such a listing would permit IRBs and investigators with ready access to information regarding data sets that have already been approved as a public use data file by a qualified IRB (at that institution or any other institution). In the social and behavioral sciences, such a list should include, but not be limited to, public use data files available through the following:

(a) Inter-University Consortium for Political and Social Research (ICPSR)

(b) U.S. Bureau of the Census

(c) National Center for Health Statistics

(d) National Center for Educational Statistics

(e) Bureau of Labor Statistics