Guidance: Secondary analysis of existing data sets

Secondary data analysis includes data that was collected previously for a different purpose. Northeastern University's (NU) Human Research Protection Program (HRPP) recognizes that some research projects involving existing data sets and archives may not meet the definition of "human subjects" research requiring IRB review; some may meet definitions of research that is exempt from the federal regulations at 45 CFR part 46; and some may require IRB review. This guidance details each of these situations and provides examples.

When does secondary use of existing data not require IRB review?

If the information or biospecimens to be used in the secondary research study are not identifiable in any way and there is no way to link the information back to the subjects from whom it was collected, the research study does not meet the definition of human subjects research and does not require IRB review and approval.

Publicly available data sets

Public use data sets are prepared with the intent of making them available for the public. The data available to the public are not individually identifiable and therefore analysis does not constitute human subjects research as defined at 45 CFR 46.102.

> **Example:** portions of U.S. Census data; data from the National Center for Health Services

De-identified data

If a dataset has been stripped of all identifying information and there is no way it could be linked back to the subjects from whom it was originally collected (through a key to a coding system or by other means), its subsequent use by the Principal Investigator or by another researcher would not constitute human subjects research.

> **Example:** Student A is provided with a deidentified, non-coded data set, the use of the data does not constitute research with human subjects because there is no interaction with any individual and no identifiable private information will be used.

Coded data

Secondary analysis of coded private information is not considered to be research involving human subjects and would not require IRB review **IF** the investigator(s) cannot readily ascertain the identity of the individuals to whom the coded private information pertains as a result of one of the following circumstances:

1. The investigators and the holder of the key have entered into an agreement prohibiting the release of the key to the investigators under any circumstances, until the individuals are deceased (HHS regulations for humans subjects research do not require the IRB to review and approve this agreement);
2. There are IRB-approved written policies and operating procedures for a repository or data management center that prohibit the release of the key to the investigator under any circumstances, until the individuals are deceased; or
3. There are other legal requirements prohibiting the release of the key to the investigators, until the individuals are deceased.

> **Example:** Researcher B plans to examine the relationships between attention deficit hyperactivity disorder (ADHD), oppositional defiance disorder, and teen drug abuse using data collected by Agencies I, II, and III that work with "at risk" youth. The data will be coded and the agencies have entered into an agreement prohibiting release of the key to the researcher that could connect the data with identifiers. The use of the data would not constitute research with human subjects.

If the IRB determines that the project does not constitute human subjects research, the IRB will notify the investigator. If the IRB determines that the project does involve human subjects research, the investigator will be asked to submit a protocol for consideration by the IRB.

When is the secondary use of existing data exempt?

Exempt Category #4 considers the "secondary use" of data or specimens that were collected for other purposes. In order to qualify as secondary use, research activities must meet one of the following conditions:

(i) Identifiable information or identifiable biospecimens are publicly available.
(ii) Information is recorded in a way that the identity of the subject cannot be readily ascertained, and the researcher will not attempt to contact or reidentify the participants.

(iii) The research involves only the collection or analysis of protected health information (PHI) from a covered entity, meaning the PHI is already subject to the HIPAA rule.

(iv) Research is conducted by, or on behalf of, a federal department or agency using government-generated or government-collected information obtained for non-research purposes

Further information regarding item (ii): applies in cases where the investigators initially have access to identifiable private information but abstract the data needed for the research in such a way that the information can no longer be connected to the identity of the subjects. This means that the abstracted data set does not include **direct identifiers** (names, social security numbers, addresses, phone numbers, etc.) **or indirect identifiers** (codes or pseudonyms that are linked to the subject's identity). Furthermore, it must not be possible to identify subjects by combining a number of characteristics (e.g., date of birth, gender, position, and place of employment). This is especially relevant in smaller datasets, where the population is confined to a limited subject pool.

Example: Student researcher C will be given access to data from her faculty advisor's health survey research project. The data consists of coded survey responses, and the advisor will retain a key that would link the data to identifiers. The student will extract the information she needs for her project without including any identifying information and without retaining the code. The use of the data does constitute research with human subjects because the initial data set is identifiable (albeit through a coding system); however, it would qualify for exempt status.

Note: The following do not qualify for exemption:

- **research involving prisoners**
- **FDA regulated research**

When does the secondary use of existing data require IRB review?

Research that intends to use identifiable information and/or biospecimens in a secondary research study without further consent from the research participants must be approved by the IRB. In most cases, this will involve expedited review unless the study is of a particularly sensitive nature or uses identifiable information gathered from participants of a protected population (e.g., children). For all such applications, the researcher must request a waiver of

consent from the IRB and must provide a justification for why the secondary research cannot be carried out using non-identifiable information or biospecimens.

Consent

Researchers using data previously collected under another study should consider whether the currently proposed research is a "compatible use" with what subjects agreed to in the original consent form. For non-exempt projects, a consent process description or justification for a waiver must be included in the research protocol.

The IRB may require that informed consent for secondary analysis be obtained from subjects whose data will be accessed. Alternatively, the IRB can consider a request for a waiver of one or more elements of informed consent under 45 CFR 46.116(d). In order to approve such waiver, the IRB must first be satisfied that the research:

1. presents minimal risk (no risks of harm, considering probability and magnitude, greater than those ordinarily encountered in daily life or during the performance of routine examinations or tests).
2. the waiver or alteration will not adversely affect the rights and welfare of the subjects.
3. the research could not practicably be carried out without the waiver or alteration; and
4. whenever appropriate, the subjects will be provided with additional pertinent information after participation.

Restricted Use Data

Certain agencies and research organizations release files to researchers with specific restrictions regarding their use and storage. These restrictions are typically described in a data use or restricted use data agreement the organization requires be signed in order to receive the data. The records frequently contain identifiers or extensive variables that combined might enable identification, even though this is not the intent of the researcher.

> **Example:** Researcher D will be given access to coded mental health assessments from their faculty advisor's research project. The student plans to analyze the data with a code attached to each record, and the advisor will retain a key to the code that would link the data to identifiers. The use of the data does constitute research with human subjects and does not qualify for exempt status since subjects can be identified.

For more information on when analysis of coded data is or is not human subjects research, see the **HHS Office for Human Research Protections Guidance on Research Involving Coded Private Information or Biological Specimens**.

Secondary Data Matrix	
Projects that are unlikely to be human subjects research because they involve only:	Public use data sets such as data from the National Center for Health Statistics—data is available to the public at large and not restricted to researchers. • Data sets from an outside source that have been stripped of all identifying information and of links back to identifiers before being provided to researcher. • Facebook public profiles found from Google searches. • Twitter tweets not in private setting. • Publicly accessible forums or comments sections where users have no expectation of privacy (e.g., New York Times, YouTube, etc.).
Projects that might be human subjects research because they involve:	• Purchasing/obtaining enhanced data sets—data on individuals which may include enough information to potentially identify the individuals. • Receipt of coded data where data holder has code key—depending on whether the data holder only provides data or is a collaborator in the research, and whether an agreement between institutions prohibits receiver from ever receiving identifiers, etc. • Forums or chats where users must register as belonging to a certain group (e.g., cancer survivors) or housed in areas that are not public, e.g., where special passwords are needed to join.
Projects that are human subjects research because they involve:	• Private data sets obtained with identifiers (e.g., traffic violation data with driver's license numbers, survey data with email addresses, medical records with protected health information [PHI], restricted use datasets, etc.). • Stolen, hacked, accidentally released data about individuals—although data may now be publicly available (such as on the surface web or the dark web), the individuals whom the data is about had expectation of privacy, i.e., data will not been hacked, stolen, etc.

Resources

45 CHR Part 46

HHS: Office for Human Research Protections

University of California, Berkeley

University of California, San Francisco