RESEARCH INVOLVING THE SECONDARY USE OF EXISTING DATA

This document provides guidance to investigators conducting research involving the secondary use of existing data. Should you need additional assistance please contact the Office for Protection of Human Subjects (OPHS) at 510-642-7461 or at ophs@berkeley.edu.

Table of Contents:

A. Scope

This guidance applies *only* to activities that involve the secondary analysis of existing data, such as medical records, student records, data collected from previous studies, audio/video recordings, etc. that were initially collected for another purpose.

Though such projects do not involve interactions or interventions with humans, they may still require CPHS/OPHS review, since the definition of "human subject" at 45 CFR 46.102(f) includes *living* individuals about whom an investigator *obtains identifiable private information for research purposes*.

Data analysis activities that meet the definition of research with human subjects may qualify for an exemption or require expedited or even full committee review. Any such project must receive CPHS approval or a determination of exemption *before* the investigator accesses the data.

B. When does the secondary use of existing data *not* require review?

In general, the secondary analysis of existing data does not require CPHS/OPHS review when it does not fall within the regulatory definition of research involving human subjects, as referenced above.

Note: Although the definition of a human subject includes only living individuals, thereby excluding decedents, there are cases in which the health information of the deceased and death data files may require CPHS review. See What Needs CPHS/OPHS Review for more details.

Public data: Public use data sets (such as portions of U.S. Census data, data from the National Center for Educational Statistics, National Center for Health Statistics, etc.) are data sets prepared with the intent of making them available for the public. The data available to the public are not individually identifiable and therefore their analysis would not involve human subjects.

In addition to being identifiable, the existing data must include "private information" in order to constitute research involving human subjects. Private information is defined as information which has been provided for specific purposes by an individual and which the individual can reasonably expect will not be made public (e.g., a medical or school record). Information that contains identifiers and can be accessed freely by the public (without special permission or application) is not "private" and the research therefore does not therefore involve human subjects. For example, a study involving only

analysis of the published salaries and benefits of public university presidents would not need CPHS/OPHS review since this information is not private.

De-identified data: If the dataset has been stripped of all identifying information and there is no way that it could be linked back to the subjects from whom it was originally collected (through a key to a coding system or by any other means), its subsequent use by the PI or another investigator would not constitute human subjects research, since it is no longer identifiable. Identifiable means the identity of the subject is known or may be readily ascertained by the investigator or associated with the information. In general, information is considered to be identifiable when it can be linked to specific individuals by the investigator(s) either directly or indirectly through coding systems, or when characteristics of the information obtained are such that by their nature a reasonably knowledgeable person could ascertain the identities of individuals. Therefore, even though a dataset may have been stripped of direct identifiers (names, addresses, student ID numbers, etc.), it may still be possible to identify an individual through a combination of other characteristics (e.g., age, gender, ethnicity, and place of employment).

Example: Many student research projects involve secondary analysis of data that belongs to, or was initially collected by, their faculty advisor or another investigator. If the student is provided with a *de-identified, 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. The project does not therefore require CPHS/OPHS review.

Coded data: Secondary analysis of coded private information is not considered to be research involving human subjects and would not require CPHS/OPHS review if the investigator(s) cannot readily ascertain the identity of the individual(s) 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 (DHHS 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.

Note: *If a student is analyzing <u>coded data</u> from a faculty advisor/sponsor who retains a key, this would be human subjects research, because the faculty advisor is considered an investigator on the student's protocol, and <u>can</u> readily ascertain the identity of the subjects since he/she holds the key to the coded data. If the student's work fits within the scope of the initial protocol from which the dataset originates, the faculty advisor (or investigator who holds the dataset) may wish to consider adding the student and his/her work to the original protocol by means of an amendment application rather than having the student submit a new application for review.*

Example: Researcher A 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 and does not require CPHS/OPHS review.

C. When is the secondary use of existing data exempt?

There are six categories of research activities involving human subjects that may be exempt from the requirements of the Federal Policy for the Protection of Human Subjects (45 CFR 46), and one UCB-specific policy (Category 70). Among them, either Category 4 or Category 70 may apply to secondary data analysis, if the corresponding criteria are met. If research is found to be exempt, it need not receive full or subcommittee (expedited) review. In order to qualify for an exempt determination, an eProtocol application must be submitted to OPHS for review.

Category 4: Research involving secondary data analysis of data, documents, and biospecimens can be exempted under Category 4 of the federal regulations if: (i) the sources of such data are publicly available; or (ii) the information is recorded by the investigator in such a manner that the resulting dataset contains no information that can identify subjects, directly or through identifiers linked to the subjects.

The latter condition of this category 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. Data must be de-identified before any analysis is conducted in order to qualify under exempt category 4.

Category 70: Research involving secondary analysis of identifiable private data, documents, and biospecimens may be exempted under Category 70 if: disclosure of the data outside of the research does not have the potential to place the subject at risk of criminal or civil liability, be damaging to the subject's financial standing, employability, insurability, or reputation, or be stigmatizing in any other way. Exempt determinations under Category 70 are subject to OPHS/CPHS discretion.

The following do not qualify for exemption: Research involving prisoners except for research aimed at involving a broader subject population that only incidentally includes prisoners, research involving protected health information from UCB HIPAA-covered entities (University Health Services (including its health care services on behalf of Intercollegiate Athletics) and the Optometry Clinic)), and FDA-regulated research.

Examples:

1) A researcher conducts a study of treatment outcomes for a certain drug that involves the review of patient charts at a non-UCB medical facility. The researcher records patient age, sex, diagnosis, and treatment outcome in such a way that the information cannot be linked back to the patient. This project could qualify for an exemption under Category 4 because, while the information extracted from patient charts is private, it is not identifiable.

2) Student B 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 exemption under Category 4 because the student will record the data she needs for analysis without identifiers and without a code.

3) Student C will obtain a private data set from UCLA for analysis at UCB for a new project. The data set is not available to the public and includes subject home addresses. The data set is not subject to HIPAA. This secondary analysis could qualify for exemption under Category 70 if the identifiable private information could not harm subjects if disclosed (subject to OPHS/CPHS discretion).

D. When is the secondary use of existing data non-exempt?

If secondary analysis of existing data *does* involve research with human subjects and does not qualify for exempt status as explained above, the project must be reviewed either through expedited procedures or by a full (convened) Board, and a non-exempt eProtocol application must be submitted for CPHS review.

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

1. the research presents no more than minimal risk of harm to the subjects; and
2. the research could not practicably be carried out without the waiver or alteration; and
3. if the research involves using identifiable private information or identifiable biospecimens, the research could not practicably be carried out without using such information or biospecimens in an identifiable format;* and
4. the waiver or alteration will not adversely affect the rights and welfare of the subjects; and
5. whenever appropriate, the subjects or legally authorized representatives will be provided with additional pertinent information after participation.

* If an individual was asked to provide broad consent for the storage, maintenance, and secondary research use of identifiable private information or identifiable biospecimens, and refused to consent, an IRB cannot waive consent for the storage, maintenance, or secondary research use of the identifiable private information or identifiable biospecimens.

See CPHS Informed Consent Guidelines for more details regarding informed consent and waivers.

"Restricted Use Data": Certain agencies and research organizations release files to researchers with specific restrictions regarding their use and storage. The records frequently contain identifiers or extensive variables that combined might enable identification, even though this is not the intent of the researcher. Research using these data sets often requires non-exempt level review, per the data holder's request.

Examples:

1) Student D will be given access to coded mental health assessments from his 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 and disclosure of mental health assessments outside of the research could potentially be damaging to subjects. This student project would require an application to be submitted for non-exempt review by the CPHS. **Note:** *As previously noted, if the student's work fits within the scope of the initial protocol from which the dataset originates, the faculty advisor (or investigator who holds the dataset) may wish to consider adding the student and his/her work to the original protocol by means of an*

amendment application rather than having the student submit a new application for non-exempt review.

2) Student E is applying to the National Center for Health Statistics for use of data from the National Health and Nutrition Examination Survey that includes geographic identifiers and date of examination. The analysis of this restricted use data would require non-exempt review by CPHS.

Secondary Data Matrix

When is secondary data (e.g., medical records, purchased data, data from the Internet, etc.) considered human subjects research? Research involving secondary data analysis is considered human subjects research when data about individuals is both private and identifiable.

Examples	
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.). Researchers who are unsure whether their project fits under this category should contact OPHS (ophs@berkeley.edu) for consultation.
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. Researchers should contact OPHS (ophs@berkeley.edu) for consultation.
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., that the data would not be hacked, stolen, etc. Human subjects research must be reviewed and either determined exempt or obtain CPHS approval before the research begins.