**Guidelines for Secondary Data Analysis of Existing Data Sets**

The Texas Christian University Institutional Review Board (IRB) recognizes that some research projects involving existing data sets and archives may not meeting 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 document is intended to provide guidance on IRB policies and procedures and to reduce burdens associated with IRB review for investigators whose research involves only the analysis of existing data sets and archives. The IRB acknowledges the

guidance document prepared by the University of Chicago Social and Behavioral Sciences IRB as the model for this Guidance.

Although projects that only involve secondary data analysis do not involve interactions or interventions with humans, they may still require IRB review, because 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.

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

In general, the secondary analysis of existing data does not require IRB review when it does not fall within the regulatory definition or research involving human subjects.

**A. Public Use 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 would not involve human subjects. The IRB recognizes that the analysis of de-identified, publicly available data does not constitute human subject research as defined at 45CFR46.102 and that it does not require IRB review. An IRB review may be required for a research study that relies exclusively on secondary use of anonymous information BUT records data linkage or disseminates results in such a way that it generates identifiable

information.

In addition to being identifiable, 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). For example, a study involving only analysis of the published salaries and benefits of university presidents would not need IRB review since this information is not private, however this should still be submitted in Cayuse for a not human subjects determination letter.

**B. De- Identified Data**

If a dataset has been stripped of all identifying information and there is no way it could 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 subject research, since the data is no longer identifiable. “Identifiable” mean the identity of the subject is known or may be readily ascertained by the investigator or associated with the information. However, this should still be submitted in Cayuse for a not human subjects determination letter.

**Example:** Many student-led research projects involve secondary data analysis of data that belongs to, or was initially collected by, their faculty advisor/PI or another investigator. If the student is provided with 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.

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

1. The investigator and the holder of the key have entered in an agreement prohibiting the release of the key to the investigators under any circumstances, until the individuals are deceased (HHS regulations for human subjects research do not require the IRB 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 investigators, until the individuals are deceased.

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** at http://www.hhs.gov/ohrp/policy/cdebiol.html

***Note:*** If a student is analyzing coded data from a faculty mentor/PI who retains a key, this would be human subjects research, because the faculty mentor/PI is considered an investigator on the student’s protocol, and can 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 study and their work to the original protocol by means of an amendment/modification submission rather than having the student submit a new application of review.

**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, 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 subject.

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

**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 regulations on human subject research protection (45CFR 46.101(2)(b)). However, only one exemption category (Category 4) applies specifically to existing data. If research is found to be exempt, it need not receive full or expedited review. In order to qualify for exemption determination, an application must be submitted in Cayuse for IRB review.

Research involving collection or study of existing data, documents, and records 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 subjects cannot be identified, 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 linked to the subject’s identity). Furthermore, it must not be possible to identify subjects by combining a number of characteristics (e.g., dates 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.

The following do not qualify for exemption: Research involving prisoners and FDA- regulated research.

**Example:** Student A will be given access to data from their faculty advisor’s health survey research project. The data consists of coded survey data responses, and the advisor will train a key that would link the data to identifiers. The student will extract the information they need for their 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.

**When does the secondary use of existing data require expedited or full board review?**

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 the full (convened) IRB, and an application must be submitted in Cayuse for IRB review/approval.

**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 submission.

The IRB may require that informed consent for secondary analysis be obtained from subjects whose data will be assessed.

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 a waiver, the IRB must first be satisfied that the research:

1. Presents minimal risk (no risk of harm, considering probability and magnitude, greater than those ordinarily encountered in daily life or during the performance of routine examinations or test); and
2. The waiver or alteration will not adversely affect the rights and welfare of the subjects; and
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 data use agreement the organization require 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. Research using these data sets requires expedited or full board level review. Note that the data use or restricted use data agreement must be reviewed by The Office Sponsored Programs prior to institutional approval. The IRB will not approve the study until the agreement received approval from OSP. The protocol may be submitted to the IRB at the same time the agreement is submitted to OSP.

**Example:** Student C 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. This student project would require an initial

submission application to be submitted in Cayuse for expedited or full board review by the IRB.

**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

expedited or full board review.

2) Student D 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 an initial

submission application to be submitted in Cayuse for expedited or full board review by the IRB.

**Secondary Data Research Classification**

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

**Projects that are unlikely to be human 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 a 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 research because they involve:**

* Purchasing/obtaining enhanced data sets-data on individuals which may include enough information to potentially identify the individuals. This would include data sets where the owner or vendor requires local IRB approval and a data use agreement prior to allowing access to the 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 most often requires expedited or full committee review.
* 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 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 research must be reviewed and either determined exempt or obtain University IRB approval before the research begins.