Wayne State University
Institutional Review Board (IRB)

WSU IRB Administration Office
87 East Canfield, Second Floor
Detroit, MI, 48201
313-577-1628
irb.wayne.edu

Collecting Data for Human Participant (Subject) Research

Recommended procedures for data collection, protecting the confidentiality of research participants and preventing a breach of confidentiality

The most common potential risk associated with Human Participant Research is a breach of confidentiality. Confidentiality refers to how private information provided by individuals will be protected. The IRB looks closely at the details of the data being collected: what is being collected, and how and where the data is being accessed, collected, stored, shared and destroyed. One of the most common forms of non-compliance with IRB policy is related to the protection of identifiable research data. The information in this guidance document will provide you with some tips to effectively obtain the data you need to conduct your research, and protect it in a way that protects your research participants' confidentiality.

During the process of soliciting informed consent from research participants, you must clearly and completely describe to them how you plan to manage, use, and potentially share the information they provide and confirm that they have understood your plans. When private/sensitive identifying information is collected as part of the research process, you should de-identify the data before sharing with other parties whenever possible.

What is a Breach of Confidentiality?*

Investigators are responsible for the confidentiality of participant information collected during the course of a study. A breach of confidentiality is an unanticipated problem that must be reported to the IRB. Additional requirements apply if the breach involves Protected Health Information (PHI) covered under HIPAA regulations. Examples of breaches of confidentiality include, but are not limited to, the following:

- Lost or stolen laptops storing participant information
- Lost or stolen USB/thumb drives with unencrypted participant information
- Accessing PHI without a *business* need to know
- Any unencrypted PHI sent outside of the responsible institution/covered entity
- Faxes sent to the wrong fax machine outside of the research site
- Improper disposal of paper containing PHI i.e., not shredding documents
- Information delivered to the wrong participant using the postal service, courier, or other delivery method

**This information was inspired by guidance from the University of Utah IRB with their permission.*

Tips for reducing the risk of a Breach of Confidentiality:

There is one risk in human participant research that exists in every study and that is a breach of confidentiality. Every research protocol/proposal should address this risk with a detailed description of the measures that will be taken to mitigate this risk. Recommended methods to reduce the risk of a breach of confidentiality include:

- Never share identifiable data with anyone outside of the IRB approved research team without a data use agreement (when PHI is involved) or a written agreement regarding the terms of use,
- Store all research data on an encrypted server (as data hackers have become more sophisticated, storing data on a password protected computer is no longer an adequate protection measure),
    - **Note:** *"One Drive" is a cloud-based encrypted server that is available to all Wayne State University faculty, students and staff. See important information about One Drive security below.*
- Collect only the minimum data necessary for your data analysis,
- If your research requires the collection of identifiable data, then code the data as described below and store the master list in a different location from the coded data,
    - Limit the accessibility of the master list among members of the research team to only those who will need to de-code.
- Recorded interviews and focus group discussions should be transcribed to omit all identifiable information.
    - Recordings must be securely stored until transcribed and immediately destroyed after transcription is complete unless maintenance of the original is required (e.g., by a sponsor).

What is Identifiable Data?

Any collection of data that includes direct personal identifiers or any combination of data sets that would make it possible for the researcher to identify the participant. With advance planning, you can minimize the collection of such identifying information (for example, by not asking a participant to state their name during a recorded interview).

Note: Identifiable data subject to HIPAA regulations must only be stored in an environment that his HIPAA compliant.

Note: Identifiable data is defined differently by the various regulations that apply to the research (Health Information Portability and Affordability Act (HIPAA), Family Educational Rights and Privacy Act (FERPA), Common Rule 45 CFR 46, and NIH Certificate of Confidentiality) It is the principal investigator's responsibility to be knowledgeable of other definitions that may apply based upon where research participants reside or where data is being managed (e.g., GDPR), who is funding the research, etc. See the following policies & guidance for additional information:

- IRB Policy- 6-7 Additional Requirements for Research Involving Other Federal Agencies
- IRB Policy- 10-1 HIPAA Requirements in Research
- IRB Guidance- Data Use Agreements & Limited Data Sets- *Applying the HIPAA Privacy Rule to Research*

Relevant HIPAA Definitions:

Individually Identifiable Health Information:

A subset of health information, including demographic information collected from an individual, and: 1) is created or received by a health care provider, health plan, employer,

WSU IRB Administration Office
87 East Canfield, Second Floor
Detroit, MI, 48201
313-577-1628
irb.wayne.edu

http://www.irb.wayne.edu/hipaa.html or health care clearinghouse, and 2) relates to the past, present, or future physical or mental health or condition of an individual; the provision of health care to an individual; or the past, present, or future payment for the provision of health care to an individual, and a) identifies the individual, or b) with respect to which there is a reasonable basis to believe the information can be used to identify the individual.

Protected Health Information (PHI):

Individually identifiable health information that is transmitted by electronic media, maintained in electronic media, or is transmitted or maintained in any other form or medium (e.g., paper records). PHI is comprised of 18 elements of identifiable data. These elements are provided in the table below.

Use:

The sharing, employment, application, utilization, examination, or analysis of individually identifiable health information for research purposes within the health entity that maintains the information.

Direct and Indirect Identifiers:

A person's or an organization's identity can be disclosed through direct and indirect identifiers. These identifiers may be found in the data or their documentation.

Direct identifiers:

Information that is sufficient, on its own, to disclose the identity of a research participant or organization.

Examples:

- name,
- address,
- date of birth
- telephone number,
- email address
- picture
- Social Security Number
- Any other number that could be tied directly to an individual (e.g., driver's license number, medical record number, account numbers)

WSU IRB Administration Office
87 East Canfield, Second Floor
Detroit, MI, 48201
313-577-1628
irb.wayne.edu

HIPAA Privacy Rule 18 Elements of Identifiable Data:

1.	Names	**10.**	Account numbers
2.	*All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP, Code, and their equivalent geographical codes, except for the initial three digits of a ZIP code if, according to the current publicly available data from the Bureau of the Census	**11.**	Certificate/license numbers
3.	*All elements of dates (except year) for dates directly related to an individual including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of 90 or older.	**12.**	Vehicle identifiers and serial numbers, including license plate numbers
4.	Telephone numbers	**13.**	Device identifiers and serial numbers
5.	Facsimile numbers	**14.**	Web universal resource locators (URL's)
6.	Electronic mail addresses	**15.**	Internet protocol (IP) address numbers
7.	Social security numbers	**16.**	Biometric identifiers, including fingerprints and voiceprints
8.	Medical record numbers	**17.**	Full-face photographic images and any comparable images
9.	Health plan beneficiary numbers	**18.**	Any other unique identifying number, characteristic, or code, unless otherwise permitted by the Privacy Rule for re-identification
** Limited data set requires the removal of all elements with the exception of parts of #2 and all of #4.*			

(Department of Health and Human Services, 2003, p.10)

Indirect identifiers:

Demographic and contextual information that creates a risk of possible disclosure in combination with other information about a participant or organization (either collected as part of your project, or available elsewhere).

Examples:

- institutional affiliations,
- occupation,
- geographic region,
- unique values or characteristics (outliers)

WSU IRB Administration Office
87 East Canfield, Second Floor
Detroit, MI, 48201
313-577-1628
irb.wayne.edu

What is Coded Data?

Data is considered coded when direct personal identifiers have been removed (e.g., from data or specimens) and replaced with words, letters, figures, symbols, or a combination of these (not derived from or related to the personal information) for purposes of protecting the identity of the source(s). When data is coded the original identifiers are retained in such a way that the de-identified data can still be traced back to the source(s).

Coded data is considered identifiable data because it is possible to de-code the data thereby allowing the researcher to ascertain the identity of the participant.

Example of a Coded Data Set:

	Partial -Identifiers		Other Variables	
ID	**Gender**	**Year of Birth**	**Lab Test**	**Lab Result**
1	Male	1950	Triglycerides	141
2	Male	1967	Creatine Kinase	86
3	Female	1978	Monocytes	14
4	Male	1981	Hematocrit	44.1
5	Female	1976	BUN/Creatinine Ratio	15
6	Male	1953	B-type Natriuretic Peptide	128.5

Example Data Set Master List:

ID	Name
1	John Smith
2	William May
3	Denise Brown
4	Wayne Green
5	Lisa Cooper
6	Daniel Miller

The Minimum Necessary Standard:

When planning a research proposal/protocol, carefully think through the data you will need, and collect the minimum necessary in order to conduct the research. In many cases, identifiable data sets are collected when it is not necessary and does not add value to the research data. Limit use of these direct identifiers to only what you absolutely need to conduct the research.

What is De-Identified Data?:

Under the Common Rule (45 CFR 46) a data-set is "de-identified" only when no one could "re-identify" the data: not the recipients, nor the data provider, nor anyone else. If the data were "coded," any "key to the code" must be destroyed to "de-identify" the data-set.

It is important to recognize that the ease of re-identification has increased with advances in technology and algorithms and the pure volume of publicly accessible data available for cross-reference.

A biospecimen that does not have an identifier directly tied to the specimen is no longer de-identified when the specimen undergoes whole genome sequencing.

A de-Identified data-set is a data-set that meets both of the following criteria:

- Does not identify any individual that is a subject of the data.
- Does not provide any reasonable basis for identifying any individual that is a subject of the data.

Tips for De-identifying Identifiable Data:

One way of reducing risk in human participant research is by keeping the data you collect confidential. You can guard against inadvertently disclosing participants' identities through targeted and thoughtful de-identification – removing direct identifiers and reducing the precision of indirect identifiers.

Researchers who collect the data needing to be de-identified and who are very familiar with their research context are best positioned to perform this delicate task. The data must be de-identified in a way that makes it impossible for any individual, including individuals who are intimately familiar with the research, to re-identify participants.

Develop uniform de-identification rules at the beginning of your project and follow them consistently throughout the project; this is particularly important if working as part of a team. These de-identification rules should be included in your research protocol/proposal.

Include a general description of the measures that will be taken to maintain the confidentiality of the study participant's data in the informed consent form, information sheet or oral consent transcript as applicable.

Here are some steps you can take to de-identify research data:

- Remove direct identifiers
- Reduce precision/detail of direct and/or indirect identifiers through aggregation
- Instead of using date of birth use the year, decade, or date of service
- Research involving interviews or focus group conversations that are recorded:
    - Transcribe the recording- omitting any unnecessary information and any information given that could identify the participant, or any other individual discussed in the interview or focus group
    - Securely destroy the recording as soon as it has been transcribed.

WSU IRB Administration Office
87 East Canfield, Second Floor
Detroit, MI, 48201
313-577-1628
irb.wayne.edu

- Instead of using the name of a participant's town, or individual place name use the region or urban/rural location.
- Generalize meaning of detailed variables
    - Instead of naming the participant's specific professional position use the name of their occupation or area of expertise
    - Restrict upper or lower ranges of a data set to hide outliers
    - Group income or age into broader categories
    - A 72-year old → grouped in "people in their 70s" or "senior citizens"
- Combine variables
- Maintain a master log of all replacements, aggregations, or removals made and keep it in a secure location separate from the de-identified data files (see examples of coded data above).

Tips for Maintaining Compliance with Data Collection Procedures:

1. Think through the process for the collection and use of research data and prepare a standard operating procedure (SOP) that can be easily followed. Make sure this SOP addresses the following:
    a. Detailed methods for handling all forms of data (i.e., paper signed consent forms, electronic Excel spreadsheets, audio recordings, etc.)
    b. Storage of all forms of data including separate storage of data with identifiers such as master lists, and signed consent forms.
    c. Detailed methods for sharing data outside of your institution.
    d. Detailed methods for destroying data. (We recommend destroying identifiers as soon as possible. All other research data must be retained for 3 years after the study is closed by the IRB)
        i. Research involving PHI that is subject HIPAA regulations must be retained for 7 years after the close of the study.
        ii. Investigational clinical trials involving drugs or devices subject to FDA regulations must retain records for 2 years following the date a marketing application is approved for the drug for the indication for which it is being investigated; or, if no application is to be filed or if the application is not approved for such indication, until 2 years after the investigation is discontinued and FDA is notified.
    e. Accessing identifiable data
2. Train research staff on the data gathering process.
3. Re-assess the SOP and make necessary corrections along the way to ensure compliance. Any change made needs to be submitted to the IRB as an amendment. Those changes can be implemented once IRB approval has been granted.
4. Submit an amendment to the IRB if any changes to the data collection process need to be made. Do not implement any changes until after you are notified of the IRB's approval of the amendment.
5. NIH funded research must adhere to the terms of the Certificate of Confidentiality (COC) if applicable.

a. COC's protect "covered information." Covered information includes names or any information, documents, or biospecimens containing identifiable, sensitive information related to a research participant.
b. If there is at least a very small risk that information, documents, or biospecimens can be combined with other available data sources to determine the identity of an individual, then they are also protected by the Certificate.

When the Collection of Data/Bio-specimens for Research does NOT Require IRB Approval:

Research that falls within the scope of the IRB's oversight must meet the regulatory definition of a human subject (participant) **and** research.

Data/bio-specimens used for research does not meet the regulatory definition of a human subject (participant) when all of the following conditions apply:

1. The data/bio-specimens being requested has not been collected specifically for the proposed research
2. The data/bio-specimens will be provided to the investigator by a third party that has no connection to the proposed research.
3. The research does not include the use of biospecimens to evaluate the safety or effectiveness of a medical device (e.g., diagnostic).
4. The party supplying the investigator with the requested data/bio-specimens will remove all information that could potentially identify the human subjects (participants) before the investigator receives the data/bio-specimens.
    a. The data or planned analyses do not include genomic data NIH considers individual level human genomic data to be "identifiable, sensitive information." For more information, see https://grants.nih.gov/grants/guide/notice-files/NOT-OD-17-109.html
        i. "Currently available and emerging technologies make the re-identification of specific individuals from raw genomic data increasingly feasible. For example, some research has demonstrated that data and other information in publicly accessible resources can be compared with genotypic or phenotypic information obtained from other sources to re-identify the individual who is the source of the data" (NIH Grants and Funding, n.d.)

The IRB advises investigators to complete the Human Participant Research Determination tool and send it to irbquestions@wayne.edu for IRB confirmation that IRB approval is not required before the research begins. If the IRB finds that the collection of data/bio-specimens does not involve human subjects (participants), a memo documenting this review and determination will be provided for the investigator's records. The Human Participant Research Determination tool is available on the IRB Forms and Submission Requirements webpage.

*Using One Drive:

One Drive is an encrypted data storage and collaboration platform that is available to all WSU staff, students and faculty. There are a few things to know about using One Drive:

WSU IRB Administration Office
87 East Canfield, Second Floor
Detroit, MI, 48201
313-577-1628
irb.wayne.edu

One Drive and HIPAA:
One Drive is not HIPAA compliant, therefore identifiable PHI that is subject to HIPAA regulations cannot be stored in One Drive. The IRB recommends investigators check with the covered entity's privacy office for their preferred HIPAA complaint data storage and collaboration methods.

Access Control in One Drive:
Access in One Drive is given out by the owner of the data (or whoever puts it there first). When setting up a One Drive file for research, it is important to ensure that data stored in One Drive is only accessible to individuals who are approved by the IRB to have access to that data.

Resources:

- Revised Common Rule: 45 CFR 46
- HIPAA Privacy Rule: 45 CFR Part 160
- FDA 21 CFR 50.312.62
- NIH Certificates of Confidentiality FAQ's
- FERPA Definition of Personally Identifiable Information