

DEFINING DATA

in Human Subjects Research



■ **OVERVIEW:**

In Human Subjects Research, the protection of private, personal data is both an ethical obligation and a regulatory requirement.

It is essential for researchers to understand the distinctions between the different types of data to ensure compliance with IRB regulations, data protection laws, and funding agencies.

■ The 4 Types of Data ■

Data gathered or used in human subjects research studies can be regarded as:

Confidential

Anonymous

Coded

De-identified

The following guidance document defines each of these 4 terms as they relate to an individual's participation in HSR, and the way that their data are collected and maintained for analysis.

See pages 4-6 for information about how to code and de-identify HSR data.

1. CONFIDENTIAL DATA



Confidential data is Personally Identifiable Information that is provided by research participants to investigators under an expectation of privacy. Confidential data can be used to identify an individual—either directly or indirectly.

Personally Identifiable Information (PII) includes any and all of the 18 identifiers listed below:

 Names	 Social Security Numbers	 Vehicle Identifiers
 Geographic Elements	 Medical Record Numbers	 Website URLs
 Dates	 Health Plan Beneficiary Numbers	 IP Addresses
 Phone Numbers	 Account Numbers	 Biometric Elements
 Fax Numbers	 Certificate/License Numbers	 Facial images
 Email Addresses	 Device Identifiers & Serial Numbers	 Any other unique identifiers

SDSU investigators must inform the IRB if they will access and/or record any of the 18 PII identifiers by listing them on the IRB Initial Application under the “Identifiers” table.

7.0 Identifiers

7.1 Select any **PII** or **PHI** identifiers you will be **accessing** (i.e. using) or **recording** (i.e. collecting) among the following:

PII or PHI Identifier	Accessing & Recording
☐ Patient/Participant Names	☒ Accessing ☐ Recording

“PHI” is Personal/Protected Health Information that relates to the health of an individual.
PHI is a specific type of PII.

2. ANONYMOUS DATA

Data is anonymous when it has not at ANY point been connected to an individual through direct identifiers (i.e., name, email address), OR indirect identifiers (age, race) that can be triangulated to ascertain the identity of an individual.

If you directly access or record any identifiers through your study methods, your study is NOT anonymous.

EXAMPLE:

Peter's Study

Peter designed a survey for college seniors on Qualtrics. The survey is open to individuals who identify as 18 years of age or older, are college seniors, and live in the United States. Participants' Qualtrics accounts will not be connected to their survey responses, and Peter has enabled "Anonymize Responses", so IP Addresses and location data will not be collected.



- His survey is multiple choice, and includes Likert-style scales like this:

A.

When I am with other people, I feel included:

- 1 = Disagree
- 2 = Disagree a little
- 3 = Neither agree or disagree
- 4 = Agree a little
- 5 = Agree

...and demographic questions like this:

B.

What is your age?

- 18 - 25 years
- 26 - 35 years
- 36 - 45 years
- 46 - 55 years
- 56 - 65 years
- 65 + years

Question A does not collect any information that could directly identify an individual.

Question B is a basic demographic question, and even if Peter asks about age, gender, or other demographics, the information combined would not allow for an individual to be identified, because he cannot determine where the responses came from, or who they are associated with, as the participants did not sign up for the survey using their email address, phone number, name, or any other identifiable information.

3. CODED DATA

Data is considered coded if:

1. Identifying information (such as name or address) that would enable the investigator to readily ascertain the identity of the individual to whom the private information or specimens belong has been replaced with a number, letter, symbol, or combination thereof (i.e. the code);

and,

2. A key to decipher the code exists, enabling linkage of the identifying information to the private information (like survey responses) or specimens (like blood).

Coding HSR data allows investigators to discuss a study without revealing their study participants' identities. All identifying information should be stored separately from coded study data, and no one outside of the study team should be able to link a participant's identity to coded data.

4. DE-IDENTIFIED DATA

Data is considered de-identified if:

All direct or indirect identifiers or codes linking the data to the individual participant's identity are destroyed or broken, so that the investigator no longer has the ability to ascertain the identity of the participant.

De-identification can occur by removing the code from the dataset or destroying the file link.

After de-identification, even members of the study team cannot link the data to the participant. De-identification of data is often used in research or public health reporting where privacy is critical and re-identification is not needed.

HOW-TO: Data Coding & De-Identification

Moana's Study

Moana has just finished interviewing participants for her study. Study participants reached out to her using their **email address** to express interest. The interviews were **audio-recorded** over Zoom. Moana has also recorded each participant's **name**, and has their audio files stored in her SDSU Google Drive.



Moana keeps the list of her study participants linked to their contact information in a Google Sheet.

This is the original, **identifiable** data:

PARTICIPANT NAME	EMAIL ADDRESS	LINK TO AUDIO FILE
Jake Young	jake@email.com	www.googledrive-folder1
Brenda Perez	brenda@email.com	www.googledrive-folder2
Kai Lee	kai@email.com	www.googledrive-folder3

- Each study participant is linked by their **name** to their own **email address**, and to their interview **audio recording**.



IN THE EVENT OF A BREACH OF CONFIDENTIALITY...



...the hacker would know **WHO** participated in the study, **HOW** to contact them, and **WHAT** the participant shared in the interview. Depending on the subject matter of the study, the unauthorized access to this information could be harmful to the participant. This is why it is important to have robust risk management procedures to ensure your study data is secure and protected.

See SDSU Sensitive Data Storage Best Practices for more information.

1. Coding

- **After Moana transcribes each interview**, she destroys the audio file, gives each participant a numerical code, and creates a **KEY**, stored **separately** from the original sheet.

THE CODE:

PARTICIPANT	LINK TO INTERVIEW TRANSCRIPT
P001	www.googledocs/folderA
P002	www.googledocs/folderB
P003	www.googledocs/folderC

The KEY to the CODE:



CODE	PARTICIPANT NAME	EMAIL ADDRESS
P001	Jake Young	jake@email.com
P002	Brenda Perez	brenda@email.com
P003	Kai Lee	kai@email.com

At this point, Moana's data is considered to be **CODED**. Even though Jake's name was changed to **P001** on the original sheet, Moana can still reference her **KEY** to ascertain his identity and **link** him to his interview transcript.

2. De-Identification

- **Once Moana has drawn her conclusions and finished analyzing the transcripts** she no longer needs to keep any of her participants' identifiable information on file.

Moana will **de-identify** her data by destroying the **KEY to the CODE**.

Now, the only data Moana has left is this:

PARTICIPANT	LINK TO INTERVIEW TRANSCRIPT
P001	www.googledocs/folderA
P002	www.googledocs/folderB
P003	www.googledocs/folderC

Moana's data is now considered **de-identified**, because even though there is no way for her or for anyone on the street to identify any of her participants, Moana had access to the identifiers prior to de-identification, so her data is **NOT** considered anonymous.