Information about identifiers and varying types of datasets regarding identifiers

This document discusses what data points are considered identifiable, what are considered indirectly identifiable, and what the terms deidentified and anonymous mean in varying contexts. This is what the IRB considered identifiable, indirectly identifiable, deidentified, or anonymous. These recommendations take into consideration NC State OIT guidance, varying laws, and the IRB regulations regarding participant protections. Datasets include data that are information and/or biospecimens.

- This document will discuss the identifiers from HIPAA, FERPA, and finally the IRB *Final Rule* and what NC State University considers identifiable data, indirectly identifiable, de-identified data, and anonymous data.
- This document will discuss when varying regulations apply to research that generates identifiable data, receives identifiable data, and stores identifiable data including when varying regulations "trigger" and are subsequently applied to research.
- This document should serve as a guide for you when discussing your data collection, transfer, and storage of data regarding the identifiable nature of the data.

Identifiable data is on a spectrum. Some data are always identifiable, some are identifiable to some PIs as where the data may not be identifiable to others and some data are anonymous.

Some factors affecting the nature of identifiability

- Directly Identifiable, Indirectly Identifiable, Anonymous
- International, Federal, State, and Local Laws (Such as GDPR, FERPA, HIPAA)
- Researcher Role, Access, Expertise, Type of Analysis
- Breadth of N (International, U.S.A, State, County, School, Club, Class)
- Researcher Knowledge of Participants (if the researcher knows who the people taking the survey are)
- N Size in correlation with Indirect IDs collected
- Type of analysis completed
- Additional Modes of Data Collection or pairing data with other data

Always Identifiable Human Subjects Data:

Any information *about a living individual* that is directly linked, associated with, or contains the name or any other direct identifiers from the individual. Please see the table below for examples.

Anonymous Human Subjects Data:

Any information *about a living individual* that was collected in a manner that identifiers were never associated with the information and that no one was ever able to identify an individual from whom the information was collected. Subjects' identities are unknown to the investigator, not requested, not recorded and not given. There is no possible way that the researcher, research team or anyone else could possibly link the data to the participant (including in publication).

- When you generate anonymous data – you as the researcher/research team should not collect any type of identifier that would ever allow you to be able to identify a participant/respondent. There should be no way that you or anyone on the research team can know who said/did what.
- When you receive anonymous datasets – the original dataset may have IDs on them, but when that dataset is shared with you – there are absolutely no identifiers on the dataset and no one on the NC State research team or their collaborators can identify an individual from the dataset (with current knowledge, access, expertise, or triangulation of data).

Indirectly Identifiable Human Subjects Data:

Indirectly identifiable datasets have two different meanings.

- Identifiable due to researcher expertise/access/role. Please see the table below for examples.
  - The identifiers are considered “readily ascertainable” to the researchers/PIs due to their expertise, access to related information and technologies, and their roles outside of the research.
  - When a respondent can be indirectly identified from a dataset due to the expertise of the PI/researcher, due to the role of the PI/researcher, and/or due to the access the PI/researcher has due to their position outside of the research.
  - For example, you are an admissions officer by profession and you want to use admissions data for research purposes. When you have the data-set, you have the access needed due to your role, to re-identify the data, even if it has been de-identified.
- Identifiable due to data content, triangulation of data, and N size. Please see the table below for examples.
  - When the data shared can be indirectly identifiable to anyone with a modicum of effort (the dataset or reported data).
  - For example, compiling multiple indirect identifiers to be able to directly identify someone. One could potentially use someone’s race, gender, years of experience, and rank to be able to identify a participant.
  - For example, when the content of the data reveals details of an experience that someone could identify the respondent based on details shared within the dataset or reported data.

De-identified Human Subjects Data and HIPAA Limited Datasets:

- **De-identified data:** De-identified data refers to the dataset that an NC State University investigator has created. When the NC State University investigator generates identifiable data and then when appropriate, they have removed all direct and indirect identifiers from the data

set. They at one time had access to identifiable/indirectly identifiable data, but they have processed the data in such a way that the data no longer has IDs associated with it and the NC State University investigator cannot identify or re-identify respondents from the newly cleaned data-set.

- **De-identified data with codes/master list:** Identifiers have been removed from the dataset but can readily be found through the use of a master list that is accessible to the investigator.
    - The link that cross-references the subject's identity with the code should be stored in a separate location from the data and should be securely protected.

- **Limited Dataset:** A limited data set is a type of dataset specifically termed by HIPAA and only refers to HIPAA covered entities and their release of a data set.
    - A limited data set excludes 16 categories of direct identifiers and may be used or disclosed, for purposes of research, public health, or health care operations, without obtaining either an individual's Authorization or a waiver or an alteration of Authorization for its use and disclosure with a Data Use Agreement (DUA) or Business Associate Agreement (BAA) that is initiated by the data provider.
    - A limited data set allows retention of specific elements of identifying private information: geographic subdivisions, town, city, state, ZIP code, dates, age. Limited data sets are not considered to be de-identified information.
    - Usually when a Limited Dataset is shared, an agreement between institutions/organizations must be in place (such as a DUA or a BAA). A data use agreement is the means by which covered entities obtain satisfactory assurances that the recipient of the limited data set will use or disclose the PHI in the data set only for specified purposes.
    - A limited data set is often treated as "yellow" data at NC State uncles otherwise obligated to treat it differently (such as through a contract)

Examples of Data Considered Identifiers

NC State IRB considers the following table as direct or indirect identifiers when other laws do not apply. This is NOT an all inclusive list.

NC State University: Directly Identifiable Digital and Civil Identities	**NC State University: Indirectly Identifiable** Readily Ascertainable due to Expertise, Access, and Role	**NC State University: Indirectly Identifiable** Due to Content, triangulation of Content, and N size
Name	IP Address and some URLs	Code
Social Security Number	Medical Record Number	Enrollment Date
Physical Address	Health Plan IDs	Admission/Discharge Date
License Plate	MTurk IDs*	Race
NC State Unity ID	Online Panel IDs	Gender/Gender Identity
Phone/Fax Number	Voice Recording*	Years of Service
Photo/Video of Faces	Rank/Title*	Personal Experiences
GPS (precise geographic location)	Precinct*	Veteran Status
Name of Parents/Guardians	Mother's Maiden Name	Sex assigned at birth
Government/Organization Identifier (ex: Campus ID, Badge Number, Licensure ID, Conference ID)	Genomic Data (or analysis where genomic sequencing occurs) where an individual can be readily re-identified	Photos containing screenshots, personal items, tattoos, birthmarks, skin patterns
Digital Identifiers (ex: E-Mail Address, Usernames and Profile Names, Personal Web Address)	Student Artifacts such as specific projects, papers, or presentations	Unique Content of Information Shared such as unique stories
	Device IDs/Serial Numbers	Ability
	Account Numbers	National Origin
	Biometrics (ex: iris scans, fingerprint)	Religion
	Biospecimens*	Age (including full date of birth)
	SONA ID*	Ethnicity
		Sexual Orientation

indication of a star means that there are possibilities for that type of data to move category due to context of research and researcher. Context can make it either directly identifiable or indirectly identifiable. Some data sets that include indirect identifiers can be considered NHSR if an individual cannot be readily re-identified through triangulation, researcher role/access, master list, or researcher expertise. This is usually when the dataset is composed of a larger N, or there are fewer demographics, or the researcher (and research team) didn't interact with any of the participants.

https://privacyruleandresearch.nih.gov/pr_06.asp https://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html

Updated 2/18/25

FERPA and HIPAA Identifiers

18 HIPAA Identifiers that comprise Personally Identifiable Information (PII)

PII may be used alone or with other sources to identify an individual. PII in conjunction with medical records (including payments for medical care) becomes Protected Health Information (PHI).

1. Name (including initials)
2. Address (all geographic subdivisions smaller than state: street address, city, county, zip code)
3. All elements (except years) of dates related to an individual (including birthdate, admission date, discharge date, date of death, and exact age if over 89)
4. Telephone numbers
5. Fax number
6. Email address
7. Social Security Number
8. Medical record number
9. Health plan beneficiary number
10. Account number
11. Certificate or license number
12. Any vehicle identifiers, including license plate
13. Device identifiers and serial numbers
14. Web URL
15. Internet Protocol (IP) Address
16. Finger or voice print
17. Photographic image - Photographic images are not limited to images of the face
18. Any other characteristic that could uniquely identify the individual

A data set containing any of these identifiers, or parts of the identifier, is considered "identified"

HIPAA – Limited Data Set

A Limited Data Set must omit all of the HIPAA Identifiers in the left-hand column except for the following:

1. City, state, zip code
2. Dates of admission, discharge, service, date of birth, date of death
3. Ages in years, months or days or hours

To re-iterate: initials are always considered PHI/PII

HIPAA – De-identified Data

All of the 18 HIPAA Identifiers in the left-hand column must be removed in order for a data set to be considered de-identified with caveats for the following:

1. All geographic subdivisions smaller than a state, except for the initial three digits of the ZIP code: (1) The geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people; and (2) The initial three digits of a ZIP code for all such geographic units containing 20,000 or fewer people is changed to 000;
2. Ages in years and for those older than 89, all ages must be aggregated into a single category of 90 or older

FERPA – Personally Identifiable Information

In the context of FERPA, PII includes, but is not limited to:

1. Student's name
2. The name of the student's parent(s) or other family members
3. Address of the student or student's family
4. Student's personal identifiers, such as:
   a. Social Security Number;
   b. Student number; or
   c. Biometric record (i.e. Finger or voice print)
5. Student's other indirect identifiers, such as:
   a. Birthdate;
   b. Place of birth; or
   c. Mother's maiden name
6. Other information that, alone or in combination, is linked or linkable to a specific student that would allow a reasonable person in the school community, who does not have personal knowledge of the relevant circumstances, to identify the student with reasonable certainty
7. Information requested by a person who the educational agency or institution reasonably believes knows the identity of the student to whom the education record relates

http://sites.nationalacademies.org/PGA/fdp/PGA_170894

At NC State FERPA data is often treated as "yellow" data with a few exceptions.
At NC State identifiable HIPAA data is treated as "red" data" and HIPAA data that can be considered a "Limited Data Set" is often treated as yellow.
Other research activities, applicable laws, sponsors, or contracts/agreements may also influence the data sensitivity level consideration

https://privacyruleandresearch.nih.gov/pr_06.asp https://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html
Updated 2/18/25

GDPR Identifiers

https://www.gdpreu.org/the-regulation/key-concepts/personal-data/

Sensitive data is a set of special categories that should be handled with extra security. These special categories are:

- Ethnic or racial origin.
- Political opinions.
- Cultural or social identity.
- Philosophical or religious beliefs;
- Trade union memberships.
- Genetic data.
- Biometric data (that can be used to uniquely identify someone).

Personal data is any information relating to an *"identified or identifiable natural person."*When most people think of personal data, they think of phone numbers and addresses; however, personal data covers a range of identifiers.

- Name and surname.
- Email address.
- Phone number.
- Home address.
- Date of birth.
- Race.
- Gender.
- Political opinions.
- Credit card numbers.
- Data held by a hospital or doctor.
- Photograph where an individual is identifiable.
- Identification card number.
- A cookie ID.
- Internet Protocol (IP) address
- Location data (for example, the location data from a mobile phone).
- The advertising identifier of your phone.

GDPR does not cover:

- Information about someone who is deceased.
- Properly anonymized data.
- Information about public authorities and companies.

A person can be identified if they are distinguishable from another individual. The GDPR asks companies to consider:

- If they can identify an individual person just by looking at the data they are processing.
- That you don't need a name to identify a person, it could be a combination of other pieces of data that act as the identifier.
- How they assess the data they are processing and if another could feasibly use it to identify a person.
- Whether there is a future likelihood that the data could be used to identify someone.
- The data content and whether it's about the person or what they do.
- The reason they are processing the data.
- The possible effects on the person from the data processing.

Pseudonymization is when data is masked by replacing any identified or identifiable information with artificial identifiers. Although it can be a great way to protect the security and privacy of personal data – pseudonymization is limited. Even though pseudonymous data will not identify a person directly, they can be indirectly identified relatively easily. Some examples of this type of personal data include:

- An internet user name, such as a name used to post to an online discussion forum.
- Any social networking data, such as a person's friend list and login information.
- Internet user-generated data – data that is knowingly generated by an individual, such as discussion forum posts, internet searches, and personal data that they input into their social networking profiles.
- RFID codes (radio frequency identification)- RFID chips will usually include an identifiable unique number, which individualizes any property to which it is attached and can therefore be used to identify someone.
- Unique identification numbers on personal devices. For example, Mac addresses, IP address, Bluetooth number, International Mobile Equipment Identity (IMEI) number, or Near Field Communication number.

The HIPAA, FERPA, and GDPR regulations dictate what data points are considered identifiable and they "kick in" when the entity that generates, stores, and releases the data are subject to the HIPAA, FERPA, or the GDPR regulations.

HIPAA is a federal law that sets a standard for the protection of medical records and personal health information. HIPAA must be followed by groups considered to be *covered entities*. A covered entity is a) health plans, b) health care clearinghouses, and c) health care providers who transmit any health information in connection with transactions for which HHS has adopted standards.

- HIPAA "kicks in" at the beginning of the process of data generation and continues throughout the generation, storage, and transfer of data. When the data is generated, stored, and shared – the HIPAA covered entity must follow the HIPAA regulations including identifiable datasets and limited datasets (see above for definition of datasets).
    - NC State University is *a hybrid entity.* At NC State University – the HIPAA covered entities are Student Health Services, the Counseling Center, Sports Medicine, and the Psychology Clinic within the College of Education. This means that when these departments and units generate information about their clients/students, that their information is protected by the regulations of HIPAA and the information must be treated accordingly.
        - If this entity wants to share data – they must do so in accordance with HIPAA regulations, meaning it is subject to certain transfer and storage requirements.
        - If you receive data from a HIPAA covered entity, depending on the type of data you get, they will likely put an agreement in place regarding the treatment of the data.
- However as an NC State University person, you may receive data covered by HIPAA through other means (than NC State University entities).
    - If that is the case, then the entity giving you the data – are subject to follow the regulations and they are the ones responsible for deciding if and how the data can be shared and if they need for you to follow certain agreements.

FERPA (The Family Educational Rights and Privacy Act) is a Federal law that protects the privacy of student education records. The law applies to all schools that receive funds under an applicable program of the U.S. Department of Education. FERPA gives parents certain rights with respect to their children's education records. These rights transfer to the student when he or she reaches the age of 18 or attends a school beyond the high school level. Students to whom the rights have transferred are "eligible students."

- FERPA "kicks in" at the beginning of the process of data generation and continues throughout the generation, storage, and transfer of data. When the data is generated, stored, and shared – the educational institution must follow the FERPA regulations.
    - Data protected by FERPA includes:
        - Any information from the students education record o Data not protected by FERPA includes:
        - Directory information. such as a student's name, address, telephone number, date and place of birth, honors and awards, and dates of attendance. However, schools must tell parents and eligible students about directory information and allow parents and eligible students a reasonable amount of time to request that the school not disclose directory information about them.
    - If an educational institution wants to share data covered by FERPA – they must decide when it is appropriate to do so (criteria below) and then do so in accordance with FERPA regulations.
- FERPA allows schools to disclose those records, without consent from an individual, to the following parties or under the following conditions:
    - School officials with legitimate educational interest; Other schools to which a student is transferring; Specified officials for audit or evaluation purposes; Appropriate parties in connection with financial aid to a student;
    - Organizations conducting certain studies for or on behalf of the school; Accrediting organizations; To comply with a judicial order or lawfully issued subpoena; Appropriate officials in cases of health and safety emergencies; and State and local authorities, within a juvenile justice system, pursuant to specific State law.

GDPR (The General Data Protection regulation) is a is a European law that establishes protections for the privacy and security of personal data about individuals located in the European Economic Area (EEA). The law establishes circumstances under which it is lawful to collect, use, disclose or process personal data. The law also establishes certain rights of individuals in the EEA in certain cases, including rights to access, amendment, and erasure (or the right to be forgotten), requires appropriate security measures for personal data, and requires notification to certain authorities in the event of a

breach of personal data. The GDPR is not about citizenship, but instead is about anyone who is physically within the EEA at the time their data is accessed, generated, or collected.

- The GDPR "kicks in" when
    - The GDPR is not about citizenship, but instead is about anyone who is physically within the EEA at the time their data is accessed, generated, or collected.
    - An entity is physically located in the EEA and they are giving or receiving personal data
    - An entity not based in the EEA offers goods or services (regardless of payment) to people physically in the EEA
    - An entity outside of the EEA monitors behavior from people physically within the EEA
    - When any information about people physically located in the EEA is accessed, generated, and collected for research purposes.
    - NC State collaborates with an entity in the EEA and has access to fellow researcher's personal and employment data, even if the research conducted does not involve EEA entities.
    - The GDPR does not apply to anonymous data

- Data protected by the GDPR includes:
    - Personal data means any information relating to an identified or identifiable natural person.
    - An identifiable natural person is one who can be identified, directly or indirectly, by reference to an identifier.
    - In practice, this includes all data which are or can be assigned to a person in any kind of way.
    - Since the definition includes "any information," the term "personal data" should be as broadly interpreted as possible.