Information about identifiers and varying types of datasets regarding identifiers

This document will discuss what data points are considered identifiable, what are considered indirectly identifiable, what are considered de-identified, and what are considered anonymous. 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 NCSU considers identifiable data, indirectly identifiable, limited data, de-identified data, and anonymous data.
- This document will discuss when varying regulations apply to generating identifiable data, receiving identifiable data, and storing identifiable data and what point those regulation “trigger.”
- 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.

Think about identifiable data on a spectrum. Some data are always identifiable, some are identifiable to some PIs as where it may not be to others, some data are indirectly identifiable, and some data are anonymous.

**Always Identifiable Data:** Any information about a living individual that is linked, associated with, or contains the name or any details of the individual that would allow someone to be able to directly or indirectly identify a subject from the information collected.

**Indirectly Identifiable Data:** Indirectly identifiable datasets have two different meanings.
- Identifiable due to researcher expertise/access/role:
  - 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:
    - When the data shared can be indirectly identifiable to anyone with a modicum of effort.
      - 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 data.
- Special Note: sometimes the nature of the data will require the IRB to consider if the data could be identifiable to others outside of the research team, and if so – what is the likelihood of that data being re-identified and what harm could come to the respondents should that data be re-identified. In this case we also take into consideration data handling procedures regarding data protection and security.

**De-identified Datasets and Limited Datasets:**
- **De-identified data:** De-identified data refers to the data set that an NCSU investigator has created. When the NCSU 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 NCSU investigator cannot identify or re-identify respondents from the new data-set.
- **De-identified data with codes:** 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 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.
  - 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. 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.

**Anonymous 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 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.

- 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 research team receiving the dataset can identify an individual from the dataset (with current knowledge, access, or expertise).

<table>
<thead>
<tr>
<th>NCSU: Always Identifiable</th>
<th>NCSU: Indirectly Identifiable Readily Ascertainable due to Expertise, Access, and Role</th>
<th>NCSU: Indirectly Identifiable Due to Content, triangulation of Content, and N size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>IP Address</td>
<td>Birthdate</td>
</tr>
<tr>
<td>Social Security Number</td>
<td>Medical Records</td>
<td>Zip Code</td>
</tr>
<tr>
<td>Physical Address</td>
<td>Health Plan IDs</td>
<td>Enrollment Date</td>
</tr>
<tr>
<td>E-Mail Address</td>
<td>MTurk IDs*</td>
<td>Admission/Discharge Date</td>
</tr>
<tr>
<td>License Plate</td>
<td>Web Address*</td>
<td>Race</td>
</tr>
<tr>
<td>Unity ID</td>
<td>Voice Recording*</td>
<td>Gender/Gender Identity</td>
</tr>
<tr>
<td>Employee ID</td>
<td>Rank/Title*</td>
<td>Years of Service</td>
</tr>
<tr>
<td>Phone/Fax Number</td>
<td>Precinct*</td>
<td>Experiences</td>
</tr>
<tr>
<td>Fingerprint</td>
<td>Mother’s Maiden Name</td>
<td>Veteran Status</td>
</tr>
<tr>
<td>Photo/Video</td>
<td>Biospecimens*</td>
<td>Sex</td>
</tr>
<tr>
<td>GPS*</td>
<td></td>
<td>Sexual Orientation</td>
</tr>
<tr>
<td>Web Address*</td>
<td></td>
<td>Religion</td>
</tr>
<tr>
<td>Usernames*</td>
<td></td>
<td>Ability</td>
</tr>
<tr>
<td>Name of Parents/Guardians</td>
<td></td>
<td>National Origin</td>
</tr>
</tbody>
</table>

* The 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.

Special Note: This is not an all-inclusive list and based on the data you are collecting, some of your data points may be deemed identifiable due to the nature of your research.

For Reference: FERPA and HIPAA

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

The HIPAA and FERPA 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 or FERPA 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).
  - NCSU is a hybrid entity. At NCSU – 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 NCSU person, you may receive data covered by HIPAA through other means (than NCSU 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
  - 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 do 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, 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.