**IRB Guidance: Identifiability**

This guidance addresses what information makes data identifiable and what information needs to be removed from the data to de-identify it.

Study teams often have questions about what makes data identifiable. This guidance discusses what it means for data to be identifiable under the Common Rule (45 CFR 46) and the Health Insurance Portability and Accountability Act (HIPAA). The guidance also describes what it means for a data set to be coded or de-identified/anonymous.

**Identifiability under the Common Rule**

An identifier includes any information that could be used to link research data with an individual subject.

- The Common Rule defines "individually identifiable" to mean that the identity of the subject is, or may be, readily ascertained by the investigator or associated with the information.
- A data set may be identifiable under the Common Rule if it contains: initials, address, zip code, phone number, gender, age, birth date, occupation, employer, racial or ethnic group, type of biopsy performed, date sample taken, diagnosis, primary care physician, referring physician, and genealogy.
- Age, ethnicity/race, and gender may be identifiers under the Common Rule if fewer than five individuals possess a particular cluster of traits.
- Data may be identifiable if any combination of variables could potentially identify a subject.
- Some of the identifiers listed above become less problematic if the sample size is large enough that the potential identifiers could describe several individuals and thus cannot be linked to only one person. Conversely, if the sample size is small, the potential to identify an individual may increase, even in the absence of direct identifiers.

The IRB will take into consideration indirect identifiers and sample size when determining if a data set is truly de-identified. Indirect identifiers are not direct identifiers but are characteristics, which depending on the size of your population may become identifiers either on their own or when combined with each other. Examples of indirect identifiers may include:

- Names or other identifiers of the individual's relatives, employers, or household members
- Direct quotes taken from websites or social media sites that if published, could identify/be traced back to a participant
- Medical conditions, hospitalizations, and accidents
- Job titles, number of years with an employer, education, and income
- Gender, race, ethnicity, age, marital status, household composition, number of children, place of birth, etc.
- Dates such as marriage, divorce, graduation, arrest, crime, trial, or conviction
- Non-randomly assigned ID numbers (e.g., assigning the first participant ID #1 or making codes based on personal characteristics such as birthdates)

**Identifiability under HIPAA**

The HIPAA Privacy Rule regulation specifies 18 identifiers, listed below, most of which are demographic. Inclusion of even one of the following identifiers makes a data set identifiable. However, there are levels of identifiability. The following are considered limited identifiers: date of birth, date of death, dates of clinical service, and age over age 89. The remaining identifiers in the list below are considered to be direct identifiers. If the data set contains any limited identifiers, but none of the direct identifiers, it is considered a limited data set and not a de-identified data set.

- Names
- All geographic subdivisions smaller than a state
- All elements of dates (except year) that are directly related to an individual, including birth date, admission date, discharge date, death date, and all ages over 89
- All elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older
- Telephone numbers
- Vehicle identifiers and serial numbers, including license plate numbers
- Fax numbers
- Device identifiers and serial numbers
- Email addresses
- Web Universal Resource Locators (URLs)
- Social security numbers
- Internet Protocol (IP) addresses
- Medical record numbers
- Biometric identifiers, including finger and voice prints
- Health plan beneficiary numbers
- Full-face photographs and any comparable images (including video)
- Account numbers
- Any other unique identifying number, characteristic, or code
- Certificate/license numbers

**Coded data**

This phrase refers to data from which all direct subject identifiers have been removed, but each record has its own study ID or code that is linked to identifiable information such as name or medical record number. The linking file must be separate from the coded data set. This linking file may be held by someone on the study team (e.g., the PI) or it could be held by someone outside of the study team (e.g., a researcher at another institution). Under HIPAA, a coded data set may include limited identifiers. Of note, the code itself may not contain identifiers such as subject initials or medical record number.

**Anonymous/De-identified data**

This refers to data that have been stripped of all subject identifiers and that have no indirect links to subject identifiers. If the data are subject to HIPAA, the data must be stripped of the 18 direct identifiers listed above. This means there can be no data points that are considered limited

identifiers under HIPAA, (e.g., geographic area smaller than a state, elements of dates, and age over 89). If the data set contains any limited identifiers, it is considered a limited data set under HIPAA. If the data includes an indirect link to subject identifiers (e.g., via coded ID numbers), then the data are considered by the IRB to be coded, not de-identified.
Please note that data can be considered de-identified under the Common Rule but not the HIPAA Privacy Rule (e.g., limited data sets), and vice versa (e.g., no HIPAA identifiers are included but the combination of data points could make subjects identifiable).