Identifiability Guidance

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, de-identified, and 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, gender may be identifiers under the Common Rule if fewer than 5 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 so 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.

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 under HIPAA: geographic area smaller than a state, elements of dates (date of birth, date of death, dates of clinical service), and age over age 89. The remaining identifiers in the bullet list 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 under HIPAA.

- names (this includes parts of names, such as initials)
- geographic subdivisions smaller than a state
- all elements of dates (except year) related to an individual (including dates of admission, discharge, birth, death and, for individuals over 89 years old, the year of birth must not be used)
- telephone numbers
- FAX numbers

Source: University of Wisconsin Health Sciences IRB

- electronic mail addresses
- Social Security numbers (this includes parts of SSNs, e.g. last 4 digits, and scrambled SSNs)
- medical record numbers
- health plan beneficiary numbers
- account numbers
- certificate/license numbers
- vehicle identifiers and serial numbers including license plates
- device identifiers and serial numbers
- web URLs
- internet protocol addresses
- biometric identifiers (including finger and voice prints)
- full face photos and comparable images

Coded data

This refers to data which have been stripped of all direct subject identifiers, but in this case each record has its own study ID or code, which 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). A coded data set may include limited identifiers under HIPAA. Of note, the code itself may not contain identifiers such as subject initials or medical record number.

De-identified data

This refers to data which have been stripped of all subject identifiers, including all 18 HIPAA identifiers. This means that there can be no data points that are considered limited identifiers under HIPAA, i.e. geographic area smaller than a state, elements of dates (date of birth, date of death, dates of clinical service), and age over age 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 is 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).

Anonymous data

Essentially the same thing as de-identified data, this refers to data which have been stripped of all subject identifiers and which have no indirect links to subject identifiers. There should be no limited identifiers in an anonymous data set.

Source: University of Wisconsin Health Sciences IRB