wcg IRB

Do You Know Me?

The Subtle Distinction Between "Anonymous" and "De-identified" Data in Clinical Research

R. Bert Wilkins, JD, MHA, CIP
Regulatory Chair Director, WCG IRB

INTRODUCTION

There is a subtle, but important distinction between research using anonymous datasets/biological samples and de-identified datasets/biological samples, in that it can change whether the research is considered research involving human subjects (and therefore subject to regulations governing human subject research). For sponsors and clinical researchers, it is important to understand these terms and which status applies to data being collected or data that will be used in research, to ensure that the research is in compliance with the necessary regulations.

DEFINITIONS

Neither the Food and Drug Administration (FDA), which is responsible for protecting the public health by assuring the safety, efficacy, and security of human and veterinary drugs, biological products, medical devices, our nation's food supply, cosmetics, and products that emit radiation[1], nor the Office of Human Research Protections (OHRP), which provides leadership in the protection of the rights, welfare, and wellbeing of subjects involved in research conducted or supported by the U.S. Department of Health and Human Services[2], define anonymous data or de-identified data, nor does there appear to be a consensus definition for these terms. In some instances, de-identified means any identifiers are irrevocably removed from the dataset and there is not a link back to identifiable information. In other instances, de-identified data means any identifiers are irrevocably removed from the dataset but there is a link back to identifiable information. This is often referred to as coded data.

This paper will use the following definitions:

- Anonymous – The dataset does not contain any identifiable information and there is no way to link the information back to identifiable information.
- De-identified – The dataset does not contain any identifiable information, but there is a way to link the information back to identifiable information.

It is also important to make a distinction between datasets that have to comply with the Health Insurance Portability and Accountability Act (HIPAA) and those that do not. Under HIPAA, a dataset is considered anonymous if all 18 identifiers listed at 45 CFR 164.514(b)(2) are removed (see page 6 for a list of the 18 identifiers). If the dataset is

not subject to HIPAA, it is considered de-identified if the identity of the human subjects cannot be readily ascertained. Identity is considered readily ascertainable if the information is publicly available or could be determined from publicly available information. Regardless of whether the dataset is considered anonymous or de-identified, the more unique information that is collected about an individual, the easier it is to identify the individual, even if all of the information by itself is considered de-identified.

HOW DO REGULATIONS APPLY?

The reason it is important to understand the distinction between anonymous data and de-identified data is because research with anonymous data is not considered human subject research and does not need to comply with the federal regulations regarding human subjects research. In contrast, de-identified data is considered human subjects research and does need to comply with the federal regulations for human subjects research known as the Revised Common Rule. The federal regulations define a human subject as:

(e) (1) Human subject means a living individual about whom an investigator (whether professional or student) conducting research:

(i) Obtains information or biospecimens through intervention or interaction with the individual, and uses, studies, or analyzes the information or biospecimens; or

(ii) Obtains, uses, studies, analyzes, or generates identifiable private information or identifiable biospecimens.

(2) Intervention includes both physical procedures by which information or biospecimens are gathered (e.g., venipuncture) and manipulations of the subject or the subject's environment that are performed for research purposes.

(3) Interaction includes communication or interpersonal contact between investigator and subject.

(4) Private information includes information about behavior that occurs in a context in which an individual can reasonably expect that no observation or recording is taking place, and information that has been provided for specific purposes by an individual and that the individual can reasonably expect will not be made public (e.g., a medical record).

(5) Identifiable private information is private information for which the identity of the subject is or may readily be ascertained by the investigator or associated with the information

(6) An identifiable biospecimen is a biospecimen for which the identity of the subject is or may readily be ascertained by the investigator or associated with the biospecimen.

"Research that involves directly interacting or intervening with subjects to collect data is considered to be research involving human subjects, even if no identifiable information is being recorded."

Since even the investigator is not able to re-identify individuals, an anonymous dataset, does not meet the definition of human subjects. On the other hand, a de-identified dataset does meet the definition for human subjects because the investigator is able to readily ascertain the identity of the subjects as there is a link back to the identifiable information.

While it is important to note that most research involving de-identified data will be exempt from Institutional Review Board (IRB) review under the Revised Common Rule, the exemption criteria are narrowly defined and if the study does not meet the exemption criteria it would require IRB review. The most common reason why these research studies would require IRB review is because the dataset is not actually de-identified. Often the dataset will include identifiers such as medical record numbers.

Equally important as knowing what is anonymous data or de-identified data, is knowing what is not. Research with either anonymous data or de-identified data refers to the secondary use of data previously collected for other purposes.

However, any subsequent use of the data collected, would be either anonymous or de-identified depending on whether there is a link back to the identifiable information.

Examples of Research involving anonymous data:

1. An analysis of biological samples (blood, biopsy) that are no longer needed for clinical care and are not labelled with any identifiers.

2. Studies involving data analysis from federal databases and repositories such as the

National Institutes of Health (NIH) database on Genotype and Phenotype (dbGAP) or Medicare claims data where the researcher is given a specific subset of data to use

3. Secondary analysis of hospital satisfaction surveys where the responses are anonymous and the system does not record any identifiable information such as IP address or log in information.
4. An analysis of aggregate data and statistics.

Examples of Research involving de-identified data:

1. Longitudinal chart review studies where data is collected from standard-of-care visits at multiple times and the investigator maintains a link to the medical records.
2. The investigator is conducting a retrospective chart review and replaces identifiable information with a code and keeps a link back to the identifiable information.

Sometimes what seem like minor differences in the conduct of a project can change the type of data and the level of regulatory oversight needed:

- A study looking at specific biomarkers will draw blood from selected donors, pool all of the blood, and not record any identifiable information on the samples. This study is research involving human subjects because there is a direct interaction with the subjects to draw the blood, regardless of the fact that no identifiers are maintained.
- A study looking at specific biomarkers will obtain blood samples from a commercial biobank, where the samples are not labeled with any identifiable information. This study is anonymous because the investigator is not able to identify the donors or link the information back to identifiable information.
- A study looking at specific biomarkers will use blood samples drawn for clinical care, where there is a link back to the medical records. This study is de-identified because the investigator can link the samples back to the identifiable information.

It is not uncommon for investigators to use the terms anonymous and de-identified interchangeably, but the distinction between them can mean the difference in the study needing to comply with the federal regulations regarding human subject research or not. If no one, even the investigator, is able to identify the individuals in a dataset it is not human subjects research. Whereas, if the investigator is able to link back to identifiable information, it will be human subjects research.

The 18 identifiers listed at 45 CFR 164.514(b)(2). Under HIPAA, a dataset is considered de-identified if all 18 of these identifiers are removed:

(2)(i) The following identifiers of the individual or of relatives, employers, or household members of the individual, are removed:

(A) Names;

(B) All geographic subdivisions smaller than a State, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial three digits of a zip code if, according to the current publicly available data from the Bureau of the Census:

(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.

(C) All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and 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;

(D) Telephone numbers;

(E) Fax numbers;

(F) Electronic mail addresses;

(G) Social security numbers;

(H) Medical record numbers;

(I) Health plan beneficiary numbers;

(J) Account numbers;

(K) Certificate/license numbers;

(L) Vehicle identifiers and serial numbers, including license plate numbers;

(M) Device identifiers and serial numbers;

(N) Web Universal Resource Locators (URLs);

(O) Internet Protocol (IP) address numbers;

(P) Biometric identifiers, including finger and voice prints;

(Q) Full face photographic images and any comparable images; and

(R) Any other unique identifying number, characteristic, or code, except as permitted by paragraph (c) of this section; and 45 CFR 164.514(b)(2).

REFERENCES

[1] Food and Drug Administration | USAGov; https://www.usa.gov/federal-agencies/food-and-drug-administration (Accessed 05-18-2021)

[2] About OHRP | HHS.gov; https://www.hhs.gov/ohrp/about-ohrp/index.html (Accessed 05-18-2021)

[3] 45 CFR 46.102(e)

For more information, please visit www.wcgirb.com or follow us on Twitter @wcgirb or LinkedIn.