Coded vs. De-Identified; Anonymized vs. Anonymous

Adequately and accurately describing the identifiability of research data and biospecimens throughout the life cycle of a study protocol is an important key to facilitating Institutional Review Board (IRB) review. The manner in which data/specimens are collected (or recorded) and stored has ramifications for whether a research proposal meets the:

- Definition of human subject research (as defined by 45CFR46.102);
- Criteria for exemption determination (as defined by 45CFR46.104); and/or
- Criteria for Institutional Review Board (IRB) approval (as defined by 45CFR46.111 and 21CFR56.111)

Unfortunately, the terms typically used in study protocols to describe how data will be maintained – de-identified, coded, anonymized, and anonymous – are often used interchangeably, when they don't all have the same meaning. This leaves the IRB unsure of what the data looks like and how it will be protected, which can result in additional and potentially frustrating, back-and-forth with the study team.

Coded Data/Specimens

'Coded', means that information that might readily identify a subject has been replaced with a number, letter, symbol or combination thereof (i.e., a code) and there is a separate key linking the code to the subject's identifiable information. The existence of the key means that the data/specimens could be linked back to the individual and therefore makes the data/specimens identifiable, albeit indirectly, to whoever has access to the key. Therefore, the term 'coded' is not synonymous with de-identified, anonymized, nor anonymous, because the information is identifiable to the study team with access to the code. However, for those who do not have access to the key, (e.g. external collaborators) the data is de-identified.

Coding is often used when a study team needs to collect identifiable information in order to meet the aims of the research (e.g., they need to link subject data over time or they need to link subject data between multiple sources). Therefore, as a means of maintaining confidentiality, the data/specimens that are collected are coded. Meaning, the study team assigns each subject a number (or code) and the subject number is used to identify subjects in the dataset and/or to label specimens, while the subject's name and other identifiable information are stored in a separate key, as demonstrated in Tables 1 and 2, respectively, below.

Table 1. Dataset

Subject No.	Data Point 1	Data Point 2
001	7	6
002	8	9
003	5	4

Table 2. Key

Subject No.	Name
001	Superman
002	Batman
003	Wonder Woman

De-Identified & Anonymized Data/Specimens

'De-identified' means the data/specimen cannot be related or attributed to a specific individual, either directly or indirectly, which means there is no reason to believe the data/specimen could be used to identify an

individual. As the federal regulations do not specifically define data elements that could be used to identify subjects, IRBs routinely defer to the 18 identifiers listed in the Health Insurance Portability & Accountability Act (HIPAA), regardless of whether the research contains health information or is conducted within a covered entity. This includes:

- Names;
- Geographic subdivisions smaller than a State, including street address, city, county, precinct, and zip code (except for the initial three digits of a zip code);
- All elements of dates (except year) and ages over 89;
- Telephone and fax numbers;
- E-mail addresses;
- Social security, medical record, health plan beneficiary, account, and certificate/license numbers;
- Vehicle identifiers and serial numbers, including license plate numbers;
- Device identifiers and serial numbers;
- Web universal resource locators (URLs) and internet protocol (IP) address numbers;
- Biometric identifiers, including voice and fingerprints;
- Full-face photos and any comparable images; and
- Any other unique identifying number, characteristic, or code.

De-identified data/specimens are often used in retrospective and secondary analysis research. For example, if a study team conducts a review of existing records (e.g., academic/medical records) and only ***records*** the research data and **does not** record any information that directly or indirectly identifies the subjects, the dataset would be considered de-identified. In this instance, the investigator may have access to or view identifiers during the data collection process but does not record, reference, collect, or link identifiers in the data set.

Another common scenario, is for a study team to obtain data/specimens from a publically available dataset, a tissue/data bank, or from previously conducted research. Whether these data/specimens are considered de-identified hinges on the study team's access to the original source data/specimens (i.e., the original dataset/bank/research) and how the data/specimens for the ***current*** research are provided to the study team.

- If the study team does not have access to the original source data/documents and the data/specimens are provided by a 'gatekeeper' in a manner that is neither directly nor indirectly identifiable, the data/specimen would be considered de-identified.
- If the study team was involved in the original data collection, the gatekeeper engages in the current research study, or the source data/specimens are identifiable, the data/specimens would continue to be considered identifiable. Anytime ***any*** member of the study team has access to identifiable source data/specimens (even if stored separately), the secondary use of the data/specimens would be considered identifiable. A common example of this is conducting secondary research on a dataset pulled from prior research. Even if only one of the study team members has access to the original, identifiable (or coded) data set, the dataset used for the secondary analysis would still be considered identifiable.

Study teams will also often de-identify a previously coded or directly identifiable dataset at the close of prospective research, as a means of further protecting the data. Meaning, all of the identifiers listed above are removed from the data/specimens *and* any code, key, or link that previously identified the data/specimens is destroyed. This is also often referred to as 'anonymized' data/specimens.

To reiterate, to be considered de-identified or anonymized:

- The data/specimens collected cannot include any of the identifiers listed above;

- Link or keys that indirectly identify the data/specimens cannot exist, nor the ability to ***access*** links/keys by any study team members; ***and***
- The data/specimens cannot be re-identified.

Anonymous Data/Specimens

'Anonymous' means the data/specimens were collected without identifiers ***and*** a code/key linking the data/specimens to identifiable information ***never*** existed (so there is no way for the data/specimen to be linked back to a specific individual). For example, data pulled from a publically available database and one-time surveys and specimen collections, where no identifiers are ***ever*** collected or recorded would be considered anonymous. One-time interviews and focus groups might also be considered anonymous provided no identifiers are recorded, only pseudonyms are used (in place of names), and they are not audio-recorded.

Tips for IRB Review

- Carefully consider whether or not you need to collect identifiers to achieve the aims of your research. If so, during your protocol development process (prior to IRB submission) consider the best way to protect the confidentiality of the data you will collect and how you will maintain your data/specimen. All studies are different, so the manner in which you manage your data/specimens will vary. Do not cut and paste from previously approved protocols, assuming that prior plans will 'work' for your current research.
- Do not inaccurately indicate that the data will be de-identified or limit yourself to only collecting de-identified data under the (false!) assumption that doing so will make IRB review easier. Collect the data you need to in order to meet the aims of your research, but do so with a plan to adequately protect the data/specimens.
- Use appropriate and consistent terminology, as described above, in your study protocol. Be aware that the identifiability of the data/specimens collected may change over the course of the study (e.g., once data collection is complete) and accurately providing this clarification within the study protocol will aid the IRB review process.
- Carefully consider context and access while communicating about the identifiability of your data/specimens. Identifiability may vary based on a team member or site's role in the research. For example, an investigator or study coordinator may have access to identifiable information whereas a statistician may only have access to de-identified information. Similarly, a local study team may have access to identifiable information but may only share de-identified data with a collaborating site. Clearly distinguishing access can play a critical role in both the IRB review process and any required research-related agreements. For the purposes of IRB review, the IRB needs to make determinations from a global viewpoint, inclusive of all engaged roles/sites they are reviewing for. Whereas research-related agreements (e.g., data use or material transfer agreements), may be executed for more granular components of research. Using appropriate terminology in consideration of roles/sites and context will further facilitate the process.

ADDITIONAL RESOURCES:

- OHSP Policy 301 RSRB Scope and Authority
- OHSP Policy 404 Criteria for RSRB Approval of Research
- OHSP Policy 501 Levels of RSRB Review
- OHSP Quick Reference Guide: Criteria for Institutional Review Board Approval
- OHRP Guidance on Coded Private Information or Specimens Use in Research

- OHSP Explains… Assurances & Agreements & Reliance