ROSALIND FRANKLIN UNIVERSITY of MEDICINE AND SCIENCE

Institutional Review Board

GUIDELINES: Using Coded or De-identified Private Data

Protection of research participants is a fundamental principle underlying biomedical research and this includes the protection of their private information. Research involving the *initial collection of private data from living individuals* will always require an approved IRB protocol whether or not the data are "coded" subsequently. The secondary use of coded or de-identified data is a different and somewhat more complex issue. Under certain limited conditions, research involving coded or de-identified private information or specimens may be ruled "Not Human Subjects Research" that does not require IRB approval. In other conditions, research involving coded private information or specimens may be ruled "Exempt Research involving human subjects" by the IRB. The investigator is not authorized to make either of these determinations but must seek a determination from the IRB on whether or not proposed research involving coded private information or specimens constitutes human subjects research, or is eligible for exemption.

The IRB has noted that the terms "coded", "anonymous", and "de-identified" to describe data are often used incorrectly by investigators. When an investigator proposes to obtain and use "de-identified data" to conduct studies exempt from IRB oversight, there are very strict conditions that apply to ensure that the identities of research subjects cannot be ascertained with the data (see below). In many cases those strict conditions for de-identification are not met and the data may simply be "coded" and a key to the code exists somewhere. Here are working definitions:

A. Coded Data: Coded data has identifying information (such as name or social security number) that would enable the research team to ascertain the identity of the individual to whom the private information or specimens pertain which is replaced with a "code" (number, letter, symbol, or combination thereof). Note that simply coding the data does not make that data anonymous or de-identified. This is the case because with most coded data, *a key to decipher the code* typically exists, and that key enables linkage of the private information or specimens to individuals. In some cases the investigator may have access to the key; in other cases the investigator may not have any access to the key either because there are written agreements from the provider of the data to never surrender the key, or there are legal constraints on releasing such a key. In other cases, the key to the coded data is on hand but methods to reduce the risk of inadvertent disclosure of private coded information are in place (e.g. The researcher stores the subject's name/other identifiers separately from the research data, or stores the code key separately from the subject's identifiers.) In such cases, the data is fairly secure but is not anonymous or de-identified.

B. Anonymous Data: Data are anonymous if no one, not even the researcher, can connect the data to the individual who provided it. For example, if no identifying information is ever collected from individuals (including direct identifiers such as name, address or student identification number), the data might be anonymous. However, researchers should be aware that collection of indirect identifiers (i.e., information regarding other unique individual characteristics) might make it possible to identify an individual from a pool of subjects. For

example, a study participant who data chart notes them as a member of a minority ethnic group and a particular sex might be identifiable from even a large data pool.

De-Identified Data*:* Data are considered de-identified when all direct or indirect identifiers or codes linking the data to the individual subject's identity are *destroyed*. This means that the following 18 identifiers of the individual or of relatives, employers, or household members of the individual are removed:

1. Names;
2. 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: The geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people; and The initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000).
3. 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;
4. Telephone numbers;
5. Fax numbers;
6. Electronic mail addresses;
7. Social security numbers;
8. Medical record numbers;
9. Health plan beneficiary numbers;
10. Account numbers;
11. Certificate/license numbers;
12. Vehicle identifiers and serial numbers, including license plate numbers;
13. Device identifiers and serial numbers;
14. Web Universal Resource Locators (URLs);
15. Internet Protocol (IP) address numbers;
16. Biometric identifiers, including finger and voice prints;
17. Full face photographic images and any comparable images; and
18. Any other unique identifying number, characteristic, or code (e.g. student ID numbers); and the covered entity does not have actual knowledge that the information could be used alone or in combination with other information to identify an individual who is a subject of the information.

Research involving the use of only "de-identified" data or "anonymous" data may qualify as "not human subjects research" or for an "exempt status" determination from the IRB. Coded data use will typically require an IRB protocol.

See also: Guidelines on Secondary Use of Existing Data