NICHD DATA AND SPECIMEN HUB (DASH)

Data De-Identification and Coding Guidance

1 INTRODUCTION

Protection of research participants is a fundamental principle underlying biomedical research. NICHD is committed to responsible stewardship of data throughout the research process; such stewardship is essential to protecting the interests of study participants and to maintaining public trust in biomedical research.

To ensure that the identities of research subjects cannot be readily ascertained with the data, NICHD DASH will store only data that are without identifiers and coded. Specifically, before submitting the data to NICHD DASH, investigators must:

a. Strip the data of individually identifying information according to:
   i. the standards set forth in the HHS Regulations for the Protection of Human Subjects and related guidance (which covers individually identifiable private information), and
   ii. the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule (which covers protected health information); and
b. Assign to the de-identified data random, unique codes.

NICHD will not hold direct identifiers to individuals whose data are stored within NICHD DASH, nor will NICHD have access to the link between the data code and identifiers that may reside with the primary investigators and institutions for particular studies.

Data must be de-identified according to the following criteria:

- The identities of research subjects cannot be readily ascertained or otherwise associated by NICHD DASH staff or secondary data users (45 C.F.R. 46.102(f));
- The 18 identifiers[i] enumerated at section 45 C.F.R. 164.514(b)(2) (the HIPAA Privacy Rule) are removed; and
- The submitting institution has no knowledge that the remaining information could be used alone or in combination with other information to identify subjects.

2 DATA DE-IDENTIFICATION AND CODING GUIDANCE

This document provides general guidance on the de-identification and coding of study data that will be submitted to NICHD DASH. References utilized in the preparation of this Guidance are located at the end of the document.

The following considerations should be taken into account when defining the approach for de-identifying and coding study data that is to be submitted to NICHD DASH.

a. "Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule" represents a *minimum* set of actions that should be performed on datasets to de-identify participant information.

b. Codes linking the new and original data should be retained by the PI and should not be submitted to NICHD.

c. It is extremely important to approach each dataset uniquely as each research study may require slightly different application of the masking and de-identification process that is described below.

Below is a five-step outline of the process for de-identification and coding of data:

Step 1: CATEGORIZE EACH VARIABLE
Step 2: RECEIVE IRB GUIDANCE
Step 3: PERFORM REDACTION, MASKING, or DE-IDENTIFICATION
Step 4: REVIEW
Step 5: DOCUMENT

Data submitters should review the guidelines below and submit questions to NICHD DASH prior to data preparation.

2.1 Step 1: CATEGORIZE EACH VARIABLE

Categorize variables associated with study data according to the "Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule."

Categories include fields that:

a. Must be completely redacted. This includes obvious identifiers such as names, SSN, date and place of birth, etc.

b. Do not need to be fully redacted but must be obscured. These data types correspond to addresses that must be reduced to 3-digit zip codes or dates that should be altered to obscure the actual date while preserving the period of time between events. It can also refer to creating new variables that are derivations of identifiable variables. For example, weight and height measurements can be transformed into Body Mass Index (BMI) measures, which are much less identifiable.

c. Require additional review.
    1. Quasi-identifiers, possible rare and/or potentially identifiable attributes require further review by the IRB or another governing body.
    2. Free text or comments. These fields will need to be reviewed carefully to determine identifiable or quasi-identifiable information.

3. Geographic information fields other than zip codes need to be reviewed to determine whether or not they should be de-identified and which method should be used.
4. Variables representing values with rare occurrences within the dataset, such as one occurrence of an ethnic group, or identifiable measurements, e.g., exceptionally large weight or height, if not pertinent to the scientific value of the study.

2.2 Step 2: RECEIVE IRB GUIDANCE, AS APPROPRIATE

Obtain specific IRB guidance on how potentially identifiable data should be categorized, what should be done with potentially identifiable variables, and which methods should be used for masking and converting variables. It is the responsibility of the PI to contact the IRB and receive appropriate guidance, as each study may have different specific rules by which data is de-identified. In some circumstances, removing certain variables may diminish the scientific value of the study data and the IRB can make a decision on the masking of such data or how the actual values can be made available by special requests.

2.3 Step 3: PERFORM REDACTION, MASKING OR DE-IDENTIFICATION

Perform the necessary redaction, masking or de-identification process on the data associated with each potentially identifiable variable. For example:

a. Ensure that records are coded with a unique identifier, with no reference to recruitment center, geographic location, participant name, etc.

b. Study event dates should be expressed as the number of days since a "reference" event. The time intervals between events expressed in number of days since a "reference date" should be the same as time intervals between the original dates that were in the dataset. For example, if date of birth is the reference date, subsequent study event dates can be expressed as "days since birth" and actual date of birth redacted in the final dataset.

c. In the example above, if analysis of temporal trends is essential to the scientific value of the study, year of birth may be used, with study event dates still expressed as "days since birth".

d. Rules for handling partial dates for reference events or subsequent study events should be established and consistent throughout the dataset. Examples would be to consider the 15th of the month the event date when the day is missing and July 1 if both month and day are missing.

e. To avoid identification of an individual with a rare clinical or demographic characteristic, small sample sizes (less than 5) should not be included *as is* in the dataset, but could be pooled with other small samples or categorized as "Other". Another approach would be to eliminate the variable from the dataset altogether.

Note: Data submitted to NICHD DASH should have all 18 identifiers enumerated in Section 45 CFR 164.514(b) (2) (the HIPAA Privacy Rule) removed; however, quasi-identifiers and sensitive data may remain if they are directly associated with the outcome goal of the research study. For example, the IRB associated with the submitter's institution, may decide that a study dataset maintain participant information on HIV status, if the study is to correlate a specific measurement with HIV treatment.

2.4 Step 4: REVIEW

Review the study data in accordance with the HIPAA rules (Section 45 CFR 164.514(b)(2)) and IRB guidance, paying particular attention to ensure that there are no residual participant identifiers present in the dataset.

2.5 Step 5: DOCUMENT

Documentation should occur throughout the data preparation process. Submitted data should be accompanied by a reference document summarizing the steps taken to de-identify the dataset along with the following descriptions:

a. Redaction, masking or de-identification methods used.

b. Derivation of derived variables (Note: often the primary variables are deleted and only the derived variables remain).

c. Cut-off threshold for recoded variables, such as "fewer than 5 participants with this value".

d. Descriptions of recoded variables. For example: If there are fewer than 5 college graduates in a study, then explain that high school and college graduate responses were grouped together (recoded) as high school/college graduate.

3 NICHD REVIEW AND APPROVAL PROCESS

All study data submitted to NICHD DASH will be reviewed by NICHD to verify that data has been appropriately de-identified before it is made available to other investigators. NICHD will inform the PI if identifiable data is discovered and request that the PI review the data further and determine the best approach to de-identification as outlined in this Guidance.

Study data submitted to NICHD DASH will not be freely and openly accessible. All requests for access to study data archived in NICHD DASH will be reviewed for approval by the NICHD DASH Data Access Committee composed of program staff from NICHD. Certain studies may require additional approvals (such as a research network Steering Committee) before access is granted to the study data in NICHD DASH.

4 REFERENCES

a. www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html
b. http://www.nhlbi.nih.gov/research/funding/human-subjects/set-preparation-guidelines
c. Arbuckle, Luk and Eman, Khaled El; *Anonymizing Health Data: Case Studies and Methods to Get You Started*; December 23, 2013; O'Reilly Media, Inc.; ISBN: 978-1449363079
d. Eman, Khaled El; *Guide to the De-Identification of Personal Health Information*; May 6, 2013; Taylor and Francis Group, LLC; ISBN: 978-1466579064.

ENDNOTES

[i] The identities of research subjects cannot be readily ascertained or otherwise associated with the data by the repository staff or secondary data users (Common Rule); and the following data elements have been removed (HIPAA Privacy Rule).

1. Names.
2. All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP Code, and their equivalent geographical codes, except for the initial three digits of a ZIP Code if, according to the current publicly available data from the Bureau of the Census: a. The geographic unit formed by combining all ZIP Codes with the same three initial digits contains more than 20,000 people. b. The initial three digits of a ZIP Code for all such geographic units containing 20,000 or fewer people are 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. Facsimile 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) addresses numbers.
16. Biometric identifiers, including fingerprints and voiceprints.
17. Full-face photographic images and any comparable images.
18. Any other unique identifying number, characteristic, or code, unless otherwise permitted by the Privacy Rule for re-identification

In addition, the submitting institution should have no actual knowledge that the remaining information could be used alone or in combination with other information to identify the individual who is the subject of the information.