NIH *Eunice Kennedy Shriver* National Institute of Child Health and Human Development | DASH Data and Specimen Hub

NICHD DASH Data and Biospecimen Catalog De-Identification Guidance

This document provides general guidance on the de-identification of study data and biospecimen catalogs[i] that will be submitted to DASH. It is important, however, to understand that each dataset and biospecimen catalog is unique and may require a customized de-identification approach to modifying data while retaining the scientific value and preserving as much data as possible. The Principal Investigator (PI) or data submitter should consider consulting with NICHD DASH prior to beginning data de-identification to ensure that the data submitted is consistently de-identified and retains as much information as possible to maximize the value of the data, safeguard replicability, promote meaningful secondary use, and accommodate a wide range of research plans proposed by the scientific research community.

The general process when de-identifying data for submission to DASH is:

- Determine variables to de-identify
- Perform de-identification
- Document de-identification procedures

Each step of the de-identification process is described in detail below.

1. Determine Variables to De-Identify

All study data and biospecimen catalog variables should be carefully reviewed for personally identifiable information (PII) and protected health information (PHI) as defined in the "Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule"

- Determine which variables include identifiers listed in the HIPAA guidance, such as names, specific locations, addresses and zip codes, Social Security Numbers (SSN), date and place of birth, etc. Note that an original participant ID is considered an identifier and will have to be re-coded during the de-identification process; both the study data and biospecimen catalog will have the same de-identified random unique ID.
- Review free text fields to ensure that the identifiers referenced above are not included. To retain as much data as possible and preserve the meaningfulness of the study data, only affected cell(s) should be de-identified as opposed to redacting all data within a variable. For example, redactions can be denoted by a symbol such as changing a cell from “Participant diagnosed with shingles on June 10, 2008” to “Participant diagnosed with shingles on <<>> 2008”.
- Examine the overall study data for potential identifiers to protect the privacy of study participants. We strongly encourage consulting with NICHD DASH representatives prior to implementing modifications to the study data and/or biospecimen catalog to help determine de-identification actions needed.

2. Perform De-Identification

After the de-identification needs have been determined, decide on how to best de-identify the study data and/or biospecimen catalog such that they are compliant with the standards outlined in the HIPAA Privacy Rule / Safe Harbor Method[ii]. General guidance on how to de-identify variables is outlined below. Again, each dataset and/or biospecimen catalog is unique and may warrant a specific, tailored approach.

Participant Identifiers

Original participant IDs should be recoded into new, randomly-generated, unique IDs within the study. A reference file with the original IDs and the corresponding recoded IDs should not be submitted to DASH and instead be maintained by the PI or data submitter.

In addition, multiple data quality checks for participant identifiers should be performed to ensure that IDs are correct, consistent, and free from errors / typos. Furthermore, the participant identifiers should fully match between the study data and biospecimen catalog.

Site Codes

Site codes should be anonymized and not linked to a specific location smaller than a state per the HIPAA Privacy Rule / Safe Harbor Method. For example, original site IDs with identifying information embedded can be recoded into new site IDs by first assigning random numbers to the sites, then rank ordering the sites according to the random numbers, and then assigning a new site ID based on their rank.

Birth Dates

Birth dates cannot be included. Instead, participant age or days since a reference date can be computed as a new variable and birth dates can be removed.

Other Dates

Other dates, such as when a certain procedure or test was performed, cannot be included. These full dates could be modified to retain the year, but the month and day must be removed. Consider replacing such dates with age of the participant at the time of the event, or compute the number of days since a reference point. For date conversions, it is best to keep the unit of measurement (e.g., days, months, years) as consistent as possible within and across datasets.

Free Text Fields

Information contained within free text fields will need to be carefully reviewed and individually edited to de-identify information as described above.

"Do's" and "Don'ts" of Data De-identification

A few examples of what to do or not to do during data de-identification are summarized in the table below.

Table 1: Examples of De-identification "Do's" and "Don'ts"

Do...	Do not...
Remove any data containing the 18 HIPAA identifiers (e.g., names, dates (except years), locations smaller than a State)	Delete data outside of the 18 HIPAA identifiers; recommend consulting with the NICHD DASH team beforehand to help determine if additional de-identification measures are needed
De-identify affected cell(s) as needed, especially for open text responses to allow for qualitative data analysis and to complement quantitative analyses	Delete all open text responses (e.g., explanation fields; comments; description of symptoms, medical history, adverse events, etc.)
Adhere to NICHD DASH policies, governance, and de-identification guidelines to retain as much data as possible and preserve replicability and meaningfulness of the study data	Remove or change responses to missing/blank across visits, data collection instruments, or within a variable if number of responses are less than an arbitrary number (e.g., 15)
Remove participants that did not consent to having their data shared	Selectively remove participants and their corresponding data (e.g., participants that died during the study)
Preserve data/variable structures and data collection context to allow for data to be segmented and detailed analyses performed	Change the data/variable structure from continuous to categorical (e.g., changing "number of pregnancies" to categories of "1 pregnancy" and "2 or more pregnancies")

3. Document De-Identification Procedures

As de-identification is performed, be sure to clearly document all steps taken to de-identify the data. Document all variables that contained potentially identifiable data such as PII / PHI and how the information was de-identified. Submitted study data and/or biospecimen catalogs should be accompanied by a reference document summarizing the methods and steps taken to de-identify the data for traceability, replicability, and accountability. In addition, information regarding the construction of derived variables should also be included.

For inquiries on the NICHD DASH Data and Biospecimen Catalog De-Identification Guidance, please send an email to supportdash@mail.nih.gov.

[i] Biospecimen catalog submission functionality will be available in DASH soon. For any questions regarding biospecimen catalog submissions, please contact supportdash@mail.nih.gov.

[ii] The identities of research participants cannot be readily ascertained or otherwise associated with the data by DASH or secondary data users (Common Rule); and the following data elements have been removed (HIPAA Privacy Rule / Safe Harbor Method).

- Names
- 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: 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 is changed to 000

- 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
- Telephone numbers
- Fax numbers
- Electronic mail addresses
- Social security numbers
- Medical record numbers
- Health plan beneficiary numbers
- Account numbers
- Certificate / license numbers
- Vehicle identifiers and serial numbers, including license plate numbers
- Device identifiers and serial numbers
- Web Universal Resource Locators (URLs)
- Internet Protocol (IP) address numbers
- Biometric identifiers, including finger and voice prints
- Full face photographic images and any comparable images
- 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.