NICHD DATA AND SPECIMEN HUB (DASH)

DASH Codebook User Guide

1. Purpose of this User Guide

This document will serve as a User Guide and provide DASH data submitters instructions on what information to look for when using and filling out the DASH Codebook Template. In addition, the DASH Codebook Template is an essential document that informs the data user about the study, data file(s), variable names and types, labels and definitions, categories, and the corresponding dataset(s) that make up a complete data set.

The purpose of the standardized DASH Codebook is to:

- Catalogue, describe, and categorize each variable submitted to DASH for other researchers to understand and accurately interpret the variables and corresponding values in the study data.
- Standardize metadata associated with variables submitted to DASH to enable cross-study comparison of datasets shared through DASH.
- Facilitate the data submission, quality assurance, archiving, and sharing processes more efficiently in DASH.
- Promote meaningful use of completed studies to allow for additional analyses by the research community.

The sections below in this User Guide detail the steps needed to obtain and document this information. The DASH Codebook Template can be downloaded from the Submission Resources page in DASH.

1.1 Glossary and Index

Below is a list and definition of key terms that appear in this User Guide for quick reference:

- **Data Collection Instrument:** the form or instrument used to collect the data, such as questionnaires, case report forms, etc.
- **Variable Name:** the name assigned to the variable in the data.
- **Variable Label:** a textual description of the variable and what it represents.
- **Variable Label Description:** any additional contextual information regarding the variable or how it was collected or derived.
- **Aliases:** synonyms for any key terms in the variable label to aid in variable search and discovery in DASH
- **Variable Type:** the classification of variables based on the data observations (numeric, coded, categorical, or date/time)
- **Variable Code List:** the associated values and value labels associated with coded variables.
- **Variable Unit:** For numeric variables, the unit of measurement used.
- **Variable Category:** Classification of variables by category according to the Office of the National Coordinator for Health Information Technology (ONC) United States Core Data for

Interoperability (USCDI) standard.

- **Terminology Standard:** An agreed upon approach, to allow for consistent measurement, qualification, or exchange of an object, process, or unit of information.
- **"Other" Standard:** The name of the terminology standard if "Other" is selected as the option for the Terminology Standard field in the Codebook.
- **Additional Standard Information:** Additional information about the Terminology or Other standard used for the variable, such as the standard code, the standard version, or a URL to a specific common data elements (CDE) or coding system.

2. DASH Codebook Template

2.1 Identify Study Variables

All variables from the datasets that will be submitted to DASH must be included in the DASH Codebook template. For example, if there is a DASH submission that contains five datasets with 100 variables each, the codebook should be populated with all 500 variables. If multiple datasets use the same variable name, the variable should have a row in the DASH Codebook for each dataset (it is acceptable for variable names to be repeated across different datasets).

2.2 Template Sheets

The DASH Codebook template is organized into three different sheets: two sheets for submitters to provide dataset and variable information (1 and 2), and one sheet is provided with reference information to support codebook development (3).

1. "Study_Variables" – Complete list of all variables and associated information (i.e., variable name, label/description, type, units, and code-lists)
2. "Code_Lists" – For coded variables, list of numeric codes and what they represent (see Section 2.4.10)
3. "Recommended_Standards" (for reference) – List of common variable categories, associated definitions, and recommended vocabulary standards that can be used for data collection (see Section 2.4.12)

2.3 Entering Variables in the Template

For CSV datasets, or other file types, columns containing variable names (Figure 1) can be copied and pasted into the template vertically (under the "Study_Variables" sheet), using the Paste Transpose option (Figure 2).

Figure 1: Select Columns Containing Variable Names

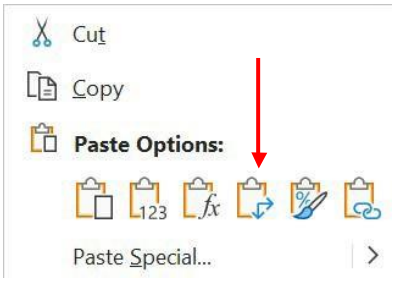

Figure 2. Paste Transpose Icon

Variable Name
STUDY
STUDY_NEW
RACE
AGE

Figure 3: Output

If there is an existing codebook included in the study submission, variable names may also be pasted from there. We suggest reviewing the existing codebook to ensure that all changes made to the datasets when de-identifying and submitting to DASH (variable redactions, recoded date variables), are represented accurately.

If the study includes SAS datasets, we suggest running PROC CONTENTS and exporting the file to Word or Excel. The variable names can then be directly pasted into the Template from the "Variable" column (Figure 4).

The CONTENTS Procedure

Variables in Creation Order					
#	Variable	Type	Len	Format	Label
1	STUDY	Char	5		Sub-study name
2	STUDY_NEW	Char	3		Is this the participant's first research study?
3	RACE	Num	5		Participant's race
4	AGE	Char	10		Participant's age

Figure 4. Example PROC CONTENTS

2.4 Data Preparation

The following section details how to populate each field in the DASH Codebook template for each variable contained in the datasets that will be submitted to DASH.

2.4.1 Dataset Names

Under the "Study_Variables" sheet in the template, populate the first column with the name of the dataset that the corresponding variable comes from. Please ensure the name of each dataset in this column matches the filename that you will submit to DASH.

- For example, if the study is comprised of two datasets, there should be two different dataset names represented in the column (Figure 5).

	Dataset Name*
1	
2	*Enter the name of the dataset that the corresponding variable comes from*
23	BUILD
24	BUILD
25	BUILD
26	BUILD
27	pregnancy_outcomes
28	pregnancy_outcomes
29	pregnancy_outcomes

Figure 5. Multiple Dataset Names Example

If the study only contains one dataset, the dataset name should be consistent throughout the column (Figure 6).

	Dataset Name*
1	
2	*Enter the name of the dataset that the corresponding variable comes from*
3	BUILD
4	BUILD
5	BUILD
6	BUILD
7	BUILD
8	BUILD

Figure 6. Singular Dataset Name Example

2.4.2 Data Collection Instrument Name (Optional)

Populate the second column with the name of the data collection instrument or case report form that corresponds to the variable of the same row. If the variable was from multiple data collection instruments or case report forms, please use a semicolon between the different instrument/form names (Figure 7). This column is optional and is not required for every variable.

	Data Collection Instrument Name
1	
2	*Optional: Enter the name of the form/instrument that the corresponding variable comes from; add semi-colon for multiple forms*
22	MO03-CRF
23	MO03-CRF
24	MO03-CRF
25	MO03-CRF
26	MO03-CRF
27	CMA; CMC; CMD; CBA; CBC; CLA; A09
28	CMA; CMC; CMD; CBA; CBC; CLA; A09
29	S02; A09

Figure 7. Data Collection Instrument Name Example

2.4.3 Variable Names

After including the Variable Names from all the datasets in this column (see Section 2), review all entries to ensure the names follow the subsequent guidelines:

- All variable names should be unique within the dataset and ideally unambiguous (Figure 8).
- They must consist of one string only, consisting of letters and, when useful, numbers and underscores (_). Spaces are not allowed in variables names in most statistical programs, even if data entry programs like Excel or Access will allow this.
- Variables names should be long enough to be meaningful, but short enough to be easy to handle.
- If there are any changes to existing variable names in the template, the changes must also be made to the corresponding variable(s) in the dataset(s).

	Variable Name*
1	
2	*Enter the name of the variable within the dataset*
3	AOPRENAT
4	AOSMOKE
5	AODRUGS
6	AOSURVIV
7	AOMAGNES
8	AOPHENO

Figure 8. Variable Name Example

2.4.4 Variable Labels

Variable Labels are a textual description of a variable or a reference to a question within a data collection instrument. It is important that the variable labels are descriptive enough for future researchers to understand and interpret the contents of the data item (Figure 9).

TIP: Run a spell check after populating this column

	Variable Label*
1	
2	*Provide the textual description of the variable or a reference to a question in the data collection instrument*
4	Smoke
5	Drugs
6	If fetus distress, would obstetrician recommend caesarean
7	Magnesium sulfate
8	Phenobarbital used for IVH prophylaxis
9	Clinical chorioamnionitis
10	Electronic fetal monitor during labor

Figure 9. Variable Label Example

2.4.5 Variable Label Description (Optional)

In the Variable Label Description column, please provide any additional contextual information regarding the variable or how it was collected or derived (Figure 10). This column is optional and is not required for every variable.

	Variable Label*	Variable Label Description
2		
3	*Provide the textual description of the variable or a reference to a question in the data collection instrument*	*Optional: Provide any additional contextual information regarding the variable or how it was collected*
12	Age of mother at delivery	
13	Mother's predominant race	The originally coded variable for ethnicity (AORACE) has been replaced by a 4-level variable (RACE) coded as: Black, White, Hispanic, Other.
30	Gestational age (completed weeks) at pregnancy end using S02F01 and date pregnancy ended from A09	S02F01: project estimated date of delivery; A09: Brief Report of Pregnancy Outcome

Figure 10. Variable Label Description Example

2.4.6 Aliases (Optional)

In the Aliases column, please provide any synonyms for key terms in the Variable Labels to aid with variable search in DASH (Figure 11). This column is optional and is not required for every variable.

	Variable Label*	Aliases
1		
2	*Provide the textual description of the variable or a reference to a question in the data collection instrument*	*Optional: List any synonyms for the Variable Label (Column D); add semi-colon for multiple Aliases*
7	Magnesium sulfate	MgSO4; Magnesium SO4
8	Phenobarbital used for IVH prophylaxis	Intraventricular hemorrhage prophylaxis
9	Clinical chorioamnionitis	Amnionitis; intra-amniotic infection
15	Grade III/IV CNS hemorrhage	Central nervous system hemorrhage

Figure 11. Aliases Example

2.4.7 Variable Type

The Variable Type designates whether the variable contains numeric, character, or coded values (Figure 12). The definitions are as follows:

- Numeric: A numeric variable is a type of variable measured in numbers (e.g., Days, Weeks, Systolic Blood Pressure, Pregnancy Number, Age).
- Coded: Numeric or character variables may also be coded. Codes are an exhaustive and mutually exclusive listing that decodes what values represent within a variable (e.g., 0 = No, 1 = Yes; D = Don't know). See Section 2.4.8 for additional information.
- Date/Time: Date and time variables represent a point in time. Date and time variables should follow the International Organization for Standardization (ISO) formats including YYYY-MM-DD, YYYY-MM, YYYY, or HH:MM:SS.
    - Note: Please refer to the Data and Biospecimen Catalog De-Identification Guidance for information on using de-identified dates.
- Character: Character variables encompass all variables that are not numeric, coded, or date/time. This includes variables with free text responses (e.g., Reason why medicine was not administered, Data entry status).

CAUTION: If pasting Variable Types from a PROC CONTENTS, SAS designates all coded variables as numeric or character. This information must be changed to accurately fill out the template.

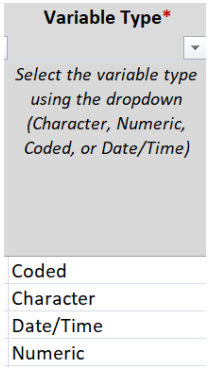

Figure 12. Variable Type Example

2.4.8 Variable Code List (if Coded)

If the Variable Type is Coded, the corresponding Code Name must be included in this column.

For every Code Name, three columns must be populated under the "Code_Lists" sheet (Figure 13) in the template: the Code List Name, the Value, and the Value Label. Descriptions of these columns are as follows:

- Code List Name: The name of the code list. This should match the "Variable Code List (if Coded)" value on the "Study_Variables" sheet of the template.
- Value: An exhaustive list of numeric or character values corresponding to the code (e.g., 1, 2, 03, D, MMN, NR)
- Value Label: The corresponding value labels for each code (e.g., the corresponding labels for the codes: 0 = No; 1 = Yes; M = Missing)

Note: Please include any global indicators in every Code List for which it applies. Global indicators are codes that have a consistent value and label across all code lists. These commonly consist of blanks or missing values (e.g., M – Missing, S – Skip pattern, E – Form Error)

	Code List Name	Value	Value Label
2	Antibiotics	1	No
3	Antibiotics	2	Prophylactic
4	Antibiotics	3	Therapeutic
5	Antibiotics	[Blank]	Missing
6	Antibiotics	.	Missing
7	Antibiotics	M	Missing
8	Antibiotics	-8 or D	Don't know
9	Antibiotics	-7 or R	Refused
10	Antibiotics	S	Skip pattern
11	Antibiotics	E	Form error
12	Blood	1	Arterial
13	Blood	2	Venous
14	Blood	3	Unknown
15	Blood	[Blank]	Missing
16	Blood	.	Missing
17	Blood	M	Missing
18	Blood	-8 or D	Don't know
19	Blood	-7 or R	Refused
20	Blood	S	Skip pattern
21	Blood	E	Form error
22	Brenner	1	Small for gestational age
23	Brenner	2	Average for gestational age

Figure 13. Variable Code List Example

2.4.9 Variable Unit (if Numeric)

If the Variable Type is Numeric, a standardized unit of measurement must be provided to designate the unit of collection for the variable (e.g., milliliters, seconds, weeks, degrees Fahrenheit). Using the drop-down list (Figure 14), select the variable unit based on the Unified Code of Units for Measure (UCUM) code system. The units are represented as "Valid UCUM Code | Descriptive Name" in the drop-down. If you do not see the variable unit in the drop-down list, please select "Other" in this column (see Section 2.4.10).

TIP: For numeric variables without a specific unit, "[arb'U] | Arbitrary Unit" can be used.

Variable Unit* **(if Numeric)** *If numeric (Column G), select the UCUM unit using the dropdown, or select "Other" if none of the units apply; see Recommended_Standards tab for more information*
ng | NanoGram
g/dL | GramsPerDeciLiter
% | Percent
kPa | KiloPascal
[arb'U] | Arbitrary Unit
Other

Figure 14. UCUM Variable Unit Examples

For more information on the UCUM code system, please refer to the "Recommended Standards" tab, which includes a link to the full list of commonly used UCUM codes for healthcare.

2.4.10 Other Variable Unit (if Numeric)

If the Variable Type is Numeric, and you selected "Other" for the UCUM unit (Figure 14), please provide the other unit of measure in this column (Figure 15).

Variable Unit* (if Numeric) *If numeric (Column G), select the UCUM unit using the dropdown, or select "Other" if none of the units apply; see Recommended_Standards tab for more information*	Other Variable Unit* (if Numeric) *If "Other" is selected for the UCUM Unit (Column H), provide the unit of measure.*
Other	Average mg/day
Other	mph
Other	Miles

Figure 15. Other Variable Unit Examples

2.4.11 Value Range Minimum (Optional)

In the Value Range Minimum column, for numeric variables, please specify the minimum permissible value in the range (Figure 16). This column is optional and is not required.

Value Range Mininum *Optional: If numeric (Column G), specify the minimum permissible Value*
10
5
7
9
15
5
10

Figure 16. Value Range Minimum Example

2.4.12 Value Range Maximum (Optional)

In the Value Range Maximum column, for numeric variables, please specify the maximum permissible value in the range (Figure 17). This column is optional and is not required.

Value Range Maximum *Optional: If numeric (Column G), specify the maximum permissible Value*
20
10
11
31
40
9
15

Figure 17. Value Range Maximum Example

2.4.13 Variable Category

In the Variable Category column, use the dropdown list to select the Category that maps to the corresponding variable (Figure 18). Refer to the "Recommended_Standards" sheet for definitions for each category. This tab also includes a detailed list of recommended vocabulary standards that can be used for coding the responses. Utilization of these vocabularies will allow for greater interoperability and aggregation of data across studies and datasets, improving secondary researchers' search results. If there is not an accurate Variable Category for a specific variable, please select "N/A."

Variable Category* *Use dropdown list to select the Category that maps to the variable; see Recommended_Standards tab for more information*
Race
Problems
Problems
Problems
Problems
Medications

Figure 18. Variable Category Definitions

Variables are categorized according to the elements defined in the Office of the National Coordinator for Health Information Technology (ONC) United States Core Data for Interoperability (USCDI) standard (Version 1).

2.4.14 Terminology Standard (Optional)

For each variable, select the appropriate Terminology Standard that it maps onto from the dropdown list of common standards. If the standard used is not on the list, select "Other" and enter the name of the standard in the "Other" Standard column. This field is optional, so it can be left blank.

The dropdown list of standards includes the following:

- Centers for Disease Control and Prevention (CDC) - CVX Vaccine Administered
- CDC Race and Ethnicity Code Set
- Code on Dental Procedure and Nomenclature (CDT)
- Current Procedural Terminology (CPT)
- HL7 Code System Observation Interpretation
- Logical Observation Identifiers Names and Codes (LOINC)
- International Classification of Disease (ICD)
- Medical Dictionary for Regulatory Activities (MedDRA)
- National Drug Code (NDC)
- NIH Common Data Elements
- Office of Management & Budget (OMB) Standards for Race and Ethnicity
- RadLex
- RxNorm
- Systematized Medical Nomenclature for Medicine–Clinical Terminology (SNOMED CT)
- Mondo Disease Ontology (Mondo)
- WHODrug
- Other

The list of common standards was updated as of July 2024. The list of standards will be updated over time in the Codebook.

2.4.15 "Other" Standard (Optional)

For any variable in which "Other" was selected in the Terminology Standard column, list the name of the other standard that was used for that variable.

2.4.16 Additional Standard Information (Optional)

Use this column to include additional information about the Terminology or Other standard that is relevant to the variable. For example, you can indicate the particular code or element for the standard onto which the variable maps.

Variable Category*	Terminology Standard	"Other" Standard	Additional Standard Information
Use dropdown list to select the Category that maps to the variable, or select "N/A" if none of the categories apply; see Recommended_Standards tab for more information	*Provide the name of the standard/code list that the variable was collected from (e.g. ICD, MedDRA, WHODrug, etc.); see Recommended_Standards tab for more information.*	*If selected "other" in column N, please provide the standard/code list used*	*Use to specify additional information about the standard; e.g. the standard code, the standard version, or specific CDEs, coding system URL, etc...*
Laboratory Test	LOINC		Antibody Screen RBC, PEG, 75265-9; https://loinc.org/75265-9
Medications	RxNorm		aspirin[RxCUI 1191]; https://bioportal.bioontology.org/ontologies/RXNORM?p=classes&conceptid=1191
N/A	NIH Common Data Elements		Formula Fed to infant name, NLM, id PQq_xX5JNo0, version 3.1; https://cde.nlm.nih.gov/deView?tinyId=PQq_xX5JNo0
Medications	Other	ATC/DDD Index 2023	https://atcddd.fhi.no/atc_ddd_index/; M01 ANTIINFLAMMATORY AND ANTIRHEUMATIC PRODUCTS

Figure 19. Terminology Standards Examples

Variables are mapped to the existing, published standard using columns: Terminology Standard, "Other" Standard and Additional Standard Information

3. Variable-Level Annotation and Representation in DASH

Once you have completed your study submission to DASH, the variable information in the DASH Codebook template will be programmatically parsed to generate basic statistics for each study dataset. The dataset metadata (e.g., title, description, file format, etc.) and variable information will then be displayed on the respective Dataset Overview pages in DASH for users to explore and identify variables of interest from your study (Figure 20).

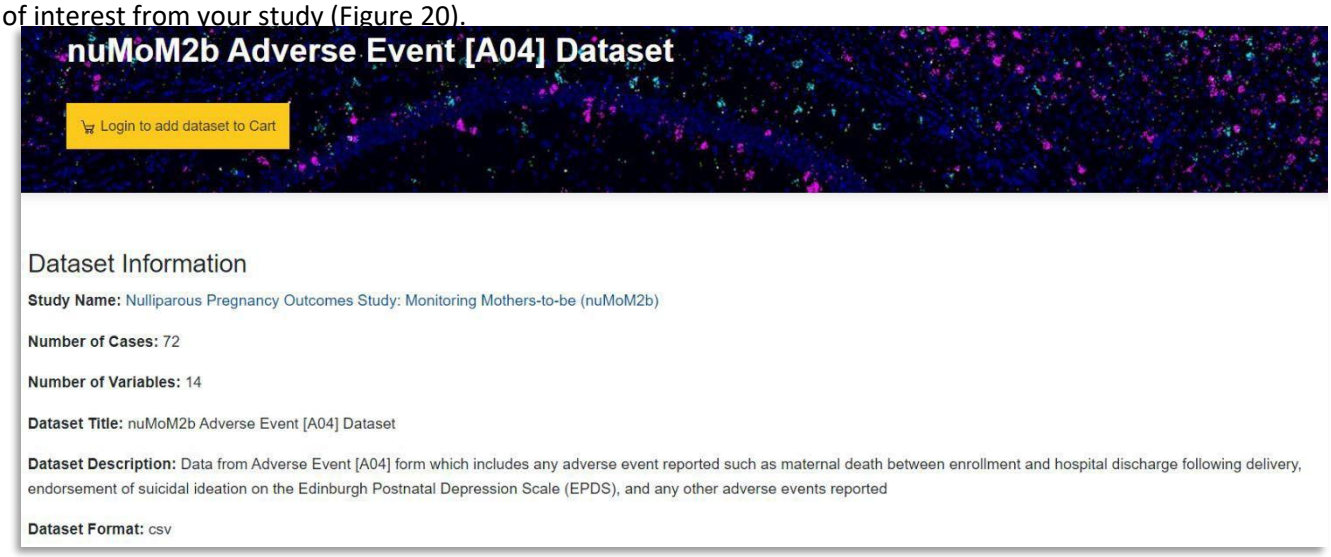

Figure 20. Dataset Overview page Example

Each Dataset Overview page contains a table for each variable type (Coded, Numeric, Character and Date/Time) with the list of variables from that dataset and associated statistics (Figure 21).

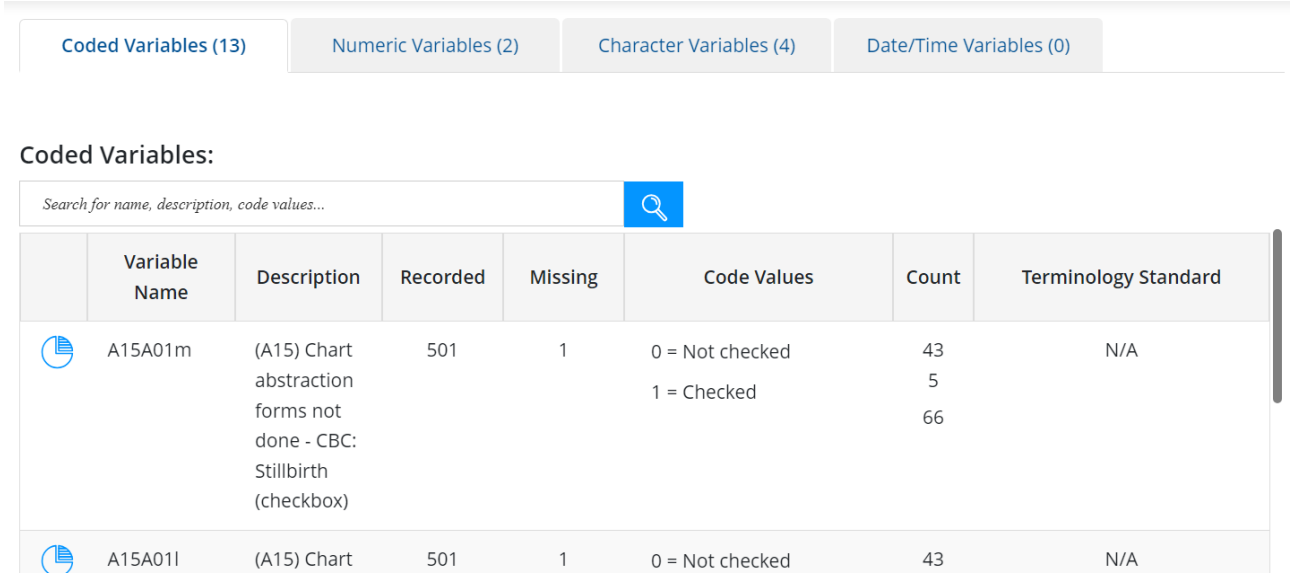

Figure 21. Variable Statistics Example

Additionally, completing the codebook will aid in quality assurance. The codebook will provide validation of the data values, code list values, and types of variables. It will also confirm the accuracy of the number of datasets and variables submitted to DASH.

4. DASH Resources

Additional information is available in the NICHD DASH website under Submission Resources. For new submissions, please reach out to SupportDASH@mail.nih.gov with any questions. For questions about existing submissions, please contact DASHCurator@mail.nih.gov.