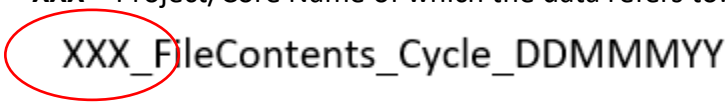
CLEAR DMAC General Guidance for Data Formatting

Purpose: The purpose of this guidance is to provide a clear and concise set of best practices for formatting data when submitting to the DMAC group. Following this guidance will ensure that the submitted data is organized, standardized, and readily usable by the DMAC

File Naming Convention

XXX_FileContents_Cycle_DDMMMYY

- **XXX =** Project/Core Name of which the data refers to:



  - B01 – Development of VOC Exposure in Zebrafish
  - B02 – Impact of BTEX chemicals exposures
  - B03 – Epidemiology Study of VOCs
  - E01 - Building above ground Strategies
  - E02 – Integrated IoT sensing and Edge Computing
  - CEC – Community Engagement Core

  - The middle character for the Project Name is a ZERO
  - Note: if data may be applied to multiple project/cores – contact the DMAC (DMAC@hfhs.org) for an assigned three digit Project/Core Code (XXX)
  - Add an underscore after the 3 project/core characters

- **File Content**: File contents should indicate the type of dataset that is being submitted:

  - The name should be limited to 9 characters or less
  - Use camel case only to separate words in the file name. Camel case is an initial capital used for the first letter of a word forming the second element of a closed compound, e.g. LabResults

    Examples of File Content Naming: Demog, Genetic, Geography, etc.
  - Add an underscore after the File Content

- **Cycle**: Single number 0-9:

  - The cycle number indicates the frequency of which the data will be collected and submitted for the specified dataset
  - If this is a single data set cycle, meaning the same data is collected at a different time interval, the cycle should be equal to 0

- If the same data will be collected at a different time interval, use the appropriate cycle number, 1-9, that corresponds to the given submission cycle
- If the data submission is found to have errors and is corrected and resubmitted on the same day, use the same cycle number (and submission date) as the erroneous data set will be overwritten in the processing file. Note that *all* submitted data is saved in an archive file.
- Pre-defined structure data sets will typically have a pre-determined number of cycles
- Free-style data sets will typically be a single data set (i.e., cycle = 0)
  - See Appendix A: Data Structures for additional information
- Add an underscore after the cycle digit

- **DDMMMYY**: Date the file is uploaded to the DMAC:

  XXX_FileContents_Cycle_DDMMMYY

  - **DD = 2 digit numeric day of submission** (01, 02, etc.)
  - **MMM = 3 character month of submission:**
    - JAN = January
    - FEB = February
    - MAR = March
    - APR = April
    - MAY = May
    - JUN = June
    - JUL = July
    - AUG = August
    - SEP = September
    - OCT = October
    - NOV = November
    - DEC = December
  - **YY = 2 digit year (i.e., 2023 = 23)**

Example: B01_Demog_0_08AUG23

- **Naming Convention for submitting formatting templates:**

  STR_XXX_specificinfo_X_DDMMMYY

  - Use the file naming convention as described above, however, add STR followed by an underscore to the beginning of the filename:

Example: STR_B01_Demog_0_08AUG23

General Data Rules

- **Variable Naming**:
  - Use descriptive and meaningful variable names that reflect the nature of the data they represent
    Example, *Age, Height, Income, IndexDate*
  - Avoid using abbreviations or acronyms unless they are widely recognized and understood in the field
  - Variable *name* length should be limited to 20 characters, preferably. If a variable name exceeds 35 characters, the name will be truncated
  - Use underscores (_) or camel case (e.g., Variable_Name or VariableName) to separate multiple words within a variable name.
  - Do not begin a variable name with a number

- **Variable Lengths & Formatting:**
  - Determine the appropriate length for each variable. Choose lengths that are sufficient to accommodate the largest expected value without wasting excessive storage space. Otherwise, try to minimize the length of your data.

    Example, if the maximum age expected is 100, the age variable should have a length of 3
  - The maximum variable length is 251
  - *Consider using categorical data to classify data into specific categories to account for variability*

    Example: 1 = Wayne County, 2 = Oakland County
  - Follow established standards and consistent formats throughout the dataset for consistency and ease of analysis

  - **Numeric Data:**
    - Numeric data should be stored as numbers rather than text, whenever feasible
    - Numeric data: Use appropriate numeric formats based on the precision/accuracy required

      Example: For the variable, "Temperature", the length should equal 5 and the format should equal 9.2 This means the variable can accommodate numbers with up to 5 digits, including the decimal portion. The format "9.2" specifies that the variable will have a total of 5 digits (including the decimal point) and 2 decimal places

  - **Dates:**
    - Dates should be stored using a standardized format:

      Most commonly: MM-DD-YYYY

  - **Character/Text Data:**
    - Character data: Use appropriate text formats, ensuring consistency and avoiding excessive lengths or variability

      For instance, the variable, "Name" should have a text format with a limit to 50 characters

    - Comments may be up to 251 characters, however, free text should be used sparingly as this is not easily analyzable
    - Avoid using special characters or symbols in variable values unless necessary

- **Missing Values:**
  - Decide on a standard for missing values and ensure consistency across variables
  - Missing numerical data should be period/dot, “.”
  - If a character data is not applicable, this should be consistently marked as NA (without a slash “/”)

- **Data Validation:**

  - Validate the data for accuracy, consistency, and adherence to predefined criteria
  - Check for outliers, data integrity issues, and logical inconsistencies
  - Consider providing a data dictionary or codebook to the DMAC that describes the variables, their definitions, and any permissible value ranges. (See Appendix B: Sample Data Dictionary)

  Example: Check that variable "Age" values fall within a reasonable range (e.g., 0 to 100).

- **Documentation:**
  - Wherever possible, provide documentation that explains the purpose of the dataset, the data collection methods, and any assumptions made during data generation
  - When applicable, provide instructions on how to interpret the data and any necessary transformations or calculations
  - When applicable, document any limitations or potential biases in the dataset

- **Submission Guidelines:**
  - Follow the guidelines provided by the DMAC for file naming conventions, formatting and submission requirements
  - Organize the data tables appropriately, ensuring clear labeling and structure

When in doubt, reach out to the DMAC Shared Mailbox: DMAC@hfhs.org

Appendix A: Pre-Determined Data Structures vs. Free-Style Data Structures

General Data Table Descriptions for Data Submitted to the DMAC:

- **Pre-determined data table** (i.e., VDW formats from EMR data). A pre-determined data table is a table that has already been created with a set of data points prior to its use.

- **Free-style data table** with conventional data elements (i.e., Microsoft Excel spreadsheets). A free-style data table is a table that does not have a pre-determined structure or format. Unlike a pre-determined data table, a free-style data table is not bound by any predefined rules or restrictions.

Appendix B: Sample Data Dictionary from a Virtual Data Warehouse (VDW) Encounters Table

CLEAR DMAC General Guidance for Data Formatting

Variable Name	Definition	Type(Len)	Values	Implementation Guidelines	Labels
StudyID	**Unique study ID assigned to the patient**	**char(8)**	**First two characters of the StudyID should be your HCSRN site code**	****This variable is not normally included in the VDW table. To be imputed locally**	**Study ID**
ENCTYPE	Refer to the ENCTYPE variable on the ENCOUNTER table for definition, type, length, and value set. This variable's redundancy is to improve querying performance.	char(2)			**Encounter Type**
ENC_ID	Foreign key to the ENCOUNTER table uniquely identifying the encounter.	char(73)	Unique to each encounter at each site.	Length may vary by site	**Encounter Identifier**
PROVIDER *	**Refer to the PROVIDER variable on the ENCOUNTER table for definition, type, length, and value set. This variable's redundancy is to improve querying performance.**	**char(50)**		*** OPTIONAL (if ENC_ID is populated)**	**Provider**
DIAGPROVIDER *	**Identifies the provider that made the diagnosis. If unknown, set the value equal to the PROVIDER varaible.**	**char(50)**		*** OPTIONAL (if ENC_ID is populated)**	**Diagnosing Provider**
PROVIDER_SPECIALTY	**The provider's specialty.**	**char(3)**	**See Appendix C - SPECIALTY for value list**	****This variable is not normally included in the VDW table. To be merged in based on 'Provider' code in the PROVIDER look-up table**	**Provider Speciality**
DIAG_SPECIALTY	**The diagnosis provider's specialty**	**char(3)**	**See Appendix C - SPECIALTY for value list**	****This variable is not normally included in the VDW table. To be merged in based on 'Provider' code in the PROVIDER look-up table**	**Diagnosing Provider Speciality**
ADATE	Refer to the ADATE variable on the ENCOUNTER table for definition, type, length, and value set. This variable's redundancy is to improve querying performance.	num(4)			**Admission Date**
DX	The diagnosis made. For ICD diagnosis coding, include decimal points in the value.	char(20)	ICD diagnosis values are set by CMS, others may be defined locally	Length may vary by site	**ICD**
DX_CODETYPE	The coding set used in the DX variable for this record.	char(2)	07='ICD-7-CM' (including 'ICD-7') 08='ICD-8-CM' (including 'ICD-8') 09='ICD-9-CM' (including 'ICD-9') 10='ICD-10-CM' (including 'ICD-10') 11='ICD-11-CM' (including 'ICD-11') OT='Other'		**Coding Set for ICD**
ORIGDX	The diagnosis code as reported in source data without standardization or cleaning.	char(20)			**Original Diagnosis Code**
PRINCIPAL_DX	For hospital admissions, whether this diagnosis is the principal discharge diagnosis of the encounter. The principal diagnosis indicates the main reason why the patient was admitted to the hospital for care and the value on which a DRG is assigned.	char(1)	P = Principal diagnosis N = Not principal diagnosis X = Unknown or not classifiable	Assigned after discharge after review by the medical record department, the principal diagnosis is main reason why the patient was admitted to the hospital for care. This is the diagnosis on which the DRG is based. Note that the principal diagnosis is very different from the admitting diagnosis which is assigned at the beginning of the stay. For example, if a patient was admitted to a hospital with an admitting diagnosis of chest pain which was later diagnosed as a heart attack during the stay, the principal diagnosis would be heart attack.	**Principal Diagnosis**
PRIMARY_DX	Whether this diagnosis is the primary diagnosis of the encounter. The primary diagnosis is the most serious or resource intensive diagnosis and is the primary reason for the procedures being rendered.	char(1)	P = Primary diagnosisS = Secondary diagnosisX = Unknown or not classifiable	Primary diagnosis is the illness or injury that was the most serious/severe/life-threatening and/or resource intensive. From a claims perspective, it is the main reason for a provider's services being rendered (and billed/paid for). Multiple primary diagnoses are allowed if the final/last professional claim can't be determined using the criteria above or if the primary diagnosis was a local combination code that has to be put into multiple records to have values within a standard coding system.	**Primary Diagnosis**
ENDPULLYR	Last year of data included in the dataset	num(8)	SAS Date	****This variable is not normally included in the VDW table. To be imputed locally**	**End Pull**