ClinEpiDB dataset formatting guidelines for data providers

Overview

Data from clinical and epidemiology studies is loaded into ClinEpiDB from files provided by members of the study team. Case report forms, data dictionaries, and study protocols are also used to inform how data is loaded. Below, we've collated information on data collection, data cleaning, and data transfer practices that will minimize the amount of time it takes to load a study.

Data collection

- Specify units of measurement on case report forms (CRFs) (e.g. Temperature in celsius)
- Indicate expected values on the CRFs with number of lines and decimal points
- Use consistent data formatting and coding across CRFs and studies
    - Use an unambiguous date format for all date variables like DD-Mon-YYYY (ex. 14-Mar-2018), which specifies all 4 digits of the year and the first 3 characters of the month so you can never confuse the day, month, or year with each other
    - Use codes consistently across binary variables (e.g. Yes = 1, No = 0)
- Limit free text fields since they will require cleaning later
- Use indicator questions (e.g. "Was a blood sample taken?"), which can indicate for other variables like "Red blood cell count (/mL)" whether data is missing or just not applicable
- Use standardized coding for missing data:
    - UNK for parts of dates that are missing
    - Not applicable = -88
    - Unknown/missing = -99
- Resources:
    - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4170533/
    - https://www.cdisc.org/standards/foundational/cdash/cdash-20#Bookmark23

Poorly designed	Well designed
Date of visit:_____	Date of visit: _ _-_ _ _-_ _ _ _ (DD-Mon-YYYY)
Blood pressure: _____	Blood pressure: _ _ _/_ _ _ (mmHg)
Temperature: _____	Temperature: _ _._ (˚C)

Data cleaning

- Never change the original files. Save new, cleaned files separately using a standardized naming convention and document all steps of data cleaning in a 'lab notebook'
- Avoid using formatting as data (e.g. highlighting data in an excel sheet)
- Remove/fix data that is impossible
    - A 2-year old is 180 cm tall → remove outlier data points and describe standards used for determining outliers
    - A temperature of 101°C → transform values from Fahrenheit to Celsius and describe transformation as well as standards used to identify values recorded in incorrect units
- ClinEpiDB displays data as provided, so clean free text fields as needed
    - Standardize spellings
    - Create value codes based on free text answers

Free text field for "Other diagnosis"	Coded value
DIABETES MELLITUS	Diabetes mellitus
DIABETES	
DIABETIS MELLITUS	
DM	

- Ensure coding is consistent across dataset (e.g. binary variables always coded 0 = No, 1 = Yes)
- Ensure value formatting is consistent across dataset (e.g. dates always in DD-Mon-YYYY)
- Minimize 'missing' data
    - Use a standardized code like 'NA' or '-88' for data that is not conditionally required
    - Use a standardized code like '-99' for missing data
    - Ex. "Number of live births" should be '-88' for males and '-99' for females of childbearing age where no data is available

sex	num_livebirth
M	
F	4
M	
F	
F	0
F	1

→

sex	num_livebirth
M	-88
F	4
M	-88
F	-99
F	0
F	1

- Make sure variable names are consistent
    - Variables should have distinct names when they contain different data
    - Variables should have the same name (same spelling and capitalization) when the same data is collected across distinct data collection phases or subsequent studies done by the same research group
    - Avoid duplicating variables across multiple files except for variables like Participant ID that are required to link data across files to each other
    - Duplicate variables that must appear in multiple files should be named the same (e.g. don't use "participant_ID" in one file and "participantID" in another file)
- Ensure unique identifiers like Participant ID and Observation ID are truly unique
    - Ex: A file of enrollment data should not have multiple rows with the same Participant ID and different demographic data
- Resource: https://www.cdc.gov/globalhealth/healthprotection/fetp/training_modules/10/managing-data_pw_final_09252013.pdf

Data transfer

Get a Box folder link for uploading files from the ClinEpiDB study team and upload the following:

- Data files
    - Comma or tab-delimited files preferred (.csv and .txt) to prevent any chance of losing or altering data during file conversion
    - Remove sensitive/identifying information (e.g. participant names, street addresses, and rare diagnoses). The ClinEpiDB team can also remove this, but prefer to minimize access to files with this type of information
- Blank case report forms
- Data dictionaries
    - Ensure variables match those found in the data files and include:
        - Variable names
        - Variable descriptions or labels (including units where relevant and collection timepoints for longitudinal data)
        - Variable types (i.e. numeric, text, date)
        - Response options/codes (e.g. 1 = Male, 2 = Female)
    - Use the data dictionary to indicate any variables you don't want us to load
- Protocol (general study enrollment and data collection information)
- Key publications
- Use cases
    - What questions did you ask with your data?
    - What questions do you want to commonly ask of your data?