eMERGE Data Dictionary/Data Validation Tool

Accessing the Tool

The Data Dictionary/Data Validation tool is available to all registered PheKB users. Once you have logged in, you will be able to access the tool under the "Tools" menu. There are sample data dictionaries and data files (as well as corrected versions) for use in exploring the tool.

Using the Tool

The validation tool allows you to verify a data dictionary file for adherence to eMERGE and dbGaP standards, and optionally check a data file if it adheres to the data dictionary. Since the data file is dependent on the data dictionary, it is recommended that you verify the data dictionary first.

Select the data dictionary file that you would like to check. Please note that the validation tool currently only accepts CSV files. If you have a data dictionary saved as an Excel document, you will need to convert it first.

After you choose and validate the file, you will get a report detailing which issues (if any) were found. Results are categorized as "warnings" and "errors". While it is recommended that you address all warnings and errors, in some cases it is understandable that warnings may be ignored.

The sample report shown in Figure 1 highlights the type of responses you might see. In this example file, the "SOURCE ID" column is named "SRCID". There are two warnings that appear which alert us of this issue – the fact that the "SOURCE ID" column is missing, and that the fourth column named "SRCID" is not a standard column. The validation tool is only able to make certain recommendations, so you may need to do some investigation to identify the underlying issue.

In addition to warnings, we also see one error in particular for the list of values for an encoded type. In this sample file, the list of values was incorrectly written as:

Female; Male; Missing; Not Assessed;

The system expects that values are represented as a semicolon-delimited list, and notes that each entry should be in the format "val=Description". Correcting this error would mean we need to change the VALUES column for this variable to be:

Female=C46110;Male=C46109;.=Missing;NA=Not Assessed

As you correct errors and warnings, you will resubmit the data dictionary until everything shows up as valid. If you have a data file you would like to also validate, you will follow the same process. Data file validation requires a data dictionary, so specify that first. Then select your data file (also a CSV file) and submit it. The process to review and revise the data file will be the same.

Figure 1 – Sample report from the Data Validation tool