CDM Data Quality Validation

Overview and Strategy:

This document outlines the data quality validation processing for populating the CDM data model and defines measures that each domain follows during validation processing. Data quality validation covers several aspects including data content validation, data integrity and data profiling, with the goal of improving data content quality and integrity of the CDM data model.

Performing validation processing of data content builds valuable assets within the consumption layer by aligning predefined business terms and associated mappings. Based on processing many raw source data elements, retrospective analysis can be performed to improve data mappings. This may include a high volume of default ('NI' – No Information) values, which can reveal a new value introduced in the source data for mapping potential and increase business impact.

Extract Transform and Load (ETL) solutions perform reference data validation, comparing value set defined business terms with the source mappings perform during data domain loading. Vocabulary entries are evaluated against each table and column pair defined in the value set reference. Errors identified are stored in data check table for ETL processing to update those errors based on rules defined for that table and column combination being validated.

Management of data quality processing establishes a high confidence level data asset for use within research projects requiring healthcare data.

The processing outlined in this document occurs during population of each site CDM data model and prior to data curation script processing.

Implementation:

Each month the Research Data Warehouse (RDW) loads new data into the atomic warehouse model. After completing the RDW monthly loading, ETL extracts perform loading of the CDM data model using the newly loaded data into a stage environment. The stage environment is a subset of CDM tables that are considered increment data domains. These domains only store the current month of data loaded into the model. This provides a smaller data set for reviewing and performing the value set reference validation.

This processing follows a two step strategy, first step the non-incremental data loading of foundational domains performing updates/inserts. Second step, the incremental processing performing inserts into mainly clinical domains, of newly supplied data from the RDW monthly loading.

The following data validation processing ensures quality, integrity and high level of data accuracy for usage within research projects. Validation rules are built based on the Data Check Exception Summary report from data characterization processing.

- Based on rules definition within the report, ETL processing rules established for detection or mitigation of errors described in each section

- Required data elements with the CDM model
    - ETL processing validates the presence of all required fields within the data model during loading of the staging tables
    - Validation of source data type to CDM data model definition
    - Rules related to each section of the exception summary report
- Orphaned data domains
    - Foreign key constraint validation
    - Validate relationships between data domains ensuring alignment of clinical domains and foundational domains
        - Patient ID – Demographic
        - Encounter ID – Encounter
        - Provider ID – Provider
- Ensuring the existence of foundational data which supports clinical domains
    - Patients exists on the foundational patient table with clinical data
    - Encounter exists for each clinical domain entry (Example: Diagnosis, Encounters, Procedures)
    - Provider exists associated with patient clinical events (Example: Encounters, Medications, Procedures)
- When dates involved, determine acceptable ranges no more than certain years, no clinical dates past patient death date
- Reference data discrepancies between data domain and value set reference table
    - Validate coded data columns with matching business term vocabulary entry
        - Vocabulary term validation according to values by table and column within the value set reference information
        - Default mappings of vocabulary terms not supplied by source system
            - No entry provided by source system
            - No current mapping within the CDM vocabulary values
- Build reference validation table from the value set document associated with current model version
    - Reference table contains validation terms for each table and column in the CDM data model defined with specific values
    - Manage each version of reference information for applying specific rules based on the version being used
- Alignment validation with standard coding terms for columns that have standard term definitions (Example: Units – UCUM for example units in lab results)
- Additionally, dates prior to healthcare data capture abilities or dates future without relevance to the domain dates. Example, appointment dates are valid for future, but collection dates in future would not makes sense for lab result/order domain.
- Once validation errors are identified, extract unique row identifier column for that specific table and place into data check table
- Mitigation ETL processing used pre-defined rules to correct vocabulary mappings
    - Based on table and column characteristics
    - History table used for review patterns and trends of errors
- Consistency patterns for a column in multiple tables

    - ENC_TYPE – Consistent pattern between the Encounter table and the Procedures table
- Track each validation processing statistics on the number of errors identified and what percentage was corrected through automation versus manual intervention
    - Used for trending of data quality validation impact
- Once validation processing is complete and issues are mitigated, the staging data domains are released to the core data model and staging tables are truncated for next processing cycle
- Each monthly error identification stored in data check history table for future reference of trending error patterns

Tools for Validation

- Staging tables for incremental data domains
- Value set reference table containing entries from Values Set Reference File
- ETL processing of validation checking
- Table for detected errors during processing plus history table
- Statistic table of validation outcomes
- PCORnet Value Set Reference file (latest version or all versions available)

Value Set Reference Table

Column Name	Data Type	Sample
CODE	VARCHAR2(100 BYTE)	mg/mL
CODE_DESCRIPTION	VARCHAR2(254 BYTE)	milligram per milliliter
CODE_TYPE	VARCHAR2(25 BYTE)	DOSE_DISP_UNIT
CODE_VERSION	VARCHAR2(25 BYTE)	CDM_5_1_v1.7
TABLE_NAME	VARCHAR2(50 BYTE)	DISPENSING
COLUMN_NAME	VARCHAR2(50 BYTE)	DISPENSE_DOSE_DISP_UNIT
CODE_STATUS	VARCHAR2(10 BYTE)	Inactive
INACTIVE_DATE	DATE	

- Code type column generated categories for multiple rules type per column
- Code status used to apply active rules per version to table columns
- Inactive date to manage temporal rule processing

The value set reference document provided by PCORnet is used for populating the value set reference table. Each version of value set reference is managed within the reference table.

Below are the table layouts for value set error processing used in the ETL processing.

Value Set Error

Column Name	Data Type	Description
TAB_KEY_ID	VARCHAR2(50 BYTE)	Unique ID (Primary Key) of error in table
TABLE_NAME	VARCHAR2(50 BYTE)	Table name error detected
COLUMN_NAME	VARCHAR2(50 BYTE)	Column name error detected

Value Set Error History

Column Name	Data Type	Description
TAB_KEY_ID	VARCHAR2(50 BYTE)	Unique ID (Primary Key) of error in table
TABLE_NAME	VARCHAR2(50 BYTE)	Table name error detected
COLUMN_NAME	VARCHAR2(50 BYTE)	Column name error detected
ERROR_DATE	DATE	

For each data validation run of ETL processing, error statistics are written to table described below for reference on the errors encountered and total number of rows detected.

Data Mart Check Statistics

Column Name	Data Type	Description
RUN_ID	NUMBER(18,0)	Unique run identifier for ETL
RUN_DATE	DATE	
DATA_MART	VARCHAR2(50 BYTE)	The data mart check was performed
RUN_TYPE	VARCHAR2(100 BYTE)	Error identified
TYPE_CNT	NUMBER(18,0)	Number of rows with error

Summary

Many types of source systems provide data to the RDW from the operational healthcare workflow applications. The RDW model manages these disparate sources within each data domain, by utilizing ETL processing to identify the source system and populate the data elements relevant to source within the data model. During data extraction from the RDW, ETL rules applied during the extraction transformation based on specific source system anomalies adhering to CDM model requirements. This approach serves the CDM target model with agnostic source data, but the ability to perform mapping surveillance in the future.

The ETL processing support patterns of extraction from atomic RDW for research projects, requiring several areas of strategies to mitigate data content issues, manage business terms and vocabularies.

The RDW manages reference data for standard code definitions (i.e. ICD, SNOMED, etc…) for use in validation of standard coding and for business terms used in consumption layer. Additional reference data includes data from source systems and mappings for data marts or research projects. This reference data contains codes and business terms for each source system that enables the mapping to local standards, national standards organizations, target data mart vocabularies or data mart to data mart vocabularies (i.e.PCORnet).