# *Common data environment for source vocabularies mapping*

PRESENTER: **Irina Zherko**

## BACKGROUND

The problem of mapping similar codes from different vocabularies that have similar source concepts (e.g., ICD10, ICD10CM) with every new vocabulary and on every refresh called for a more structured semi-automated approach. Commercial datasets, while being specifically designed for each customer, often contain duplicate or similar data that can potentially be combined, mapped together or reused.

## METHODS

The CDE should contain the most complete set of source data, structured and organized into groups by code, source_code_description, as well as the frequency of occurrence of each concept and the group as a whole.

## RESULTS

The algorithm for CDE creating includes:

- Gathering data. For each source_code/source_code_description combination, a flag with the data source (customer) name and/or the name of the dataset in case of combining datasets in one environment as well as respective count of the records for every customer/dataset and overall count must be stored in a separate fields.
- Preparing concept names or descriptions for sorting.
- Additional information per domain adding. In accordance with the requirements of custom mapping, additional fields can be filled for certain domains.
- Defining the groups using 2 fields at once (source_code and cleaned source_code_description) with the help of recursive joins.
- Sorting by group count and concept count for mapping prioritization.
- Mapping to standard concepts.

**A common data environment** approach can streamline source vocabulary processing for **repetitive processes, harmonize and enrich mapping,** speed up the process of **data reuse.**

The **common data environment** should contain the most complete set of source data, organized into groups by **code**, **code_description**, **frequency** of occurrence of **each concept and the whole group.**

Scan QR to link to **script with recursive joins** for grouping using several fields at once

Table 1. Differences in the same concept codes mapping in ICD10 and ICD10CM.

| ICD10/ ICD10CM code | ICD10/ ICD10CM name | target_concept_id for ICD10 | target_name for ICD10 | target_concept_id for ICD10CM | target_name for ICD10CM |
|---|---|---|---|---|---|
| A06.3 | Amoeboma of intestine | 37116456 | Ameboma of large intestine caused by Entamoeba histolytica | 4051447 | Ameboma |
| A32.1 | Listerial meningitis and meningoencephalitis | 4190307 | Inflammatory disease of the central nervous system | 4136344 | Listeria meningitis |
| A32.1 | Listerial meningitis and meningoencephalitis | 438059 | Listeriosis | | |
| E10.6 | Type 1 diabetes mellitus, With other complications | 435216 | Disorder due to type 1 diabetes mellitus | 201254 | Type 1 diabetes mellitus |
| | | | | 442793 | Complication due to diabetes mellitus |

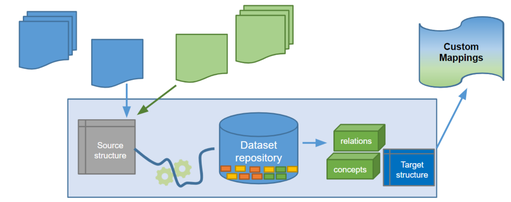


Figure 1. The theoretical scheme of workflow with CDE application

Table 2. Examples of source_code_description transformation

| Original source_code_description | Cleaned source_code_description |
|---|---|
| Photographic diagnosis by simple radiography (a) | photographic diagnosis by simple radiography |
| Steady-state Plasma Glucose 0 min | steadystate glucose min |
| Above the knee deep vein thrombosis age in days | aboveknеethrombosisage |
| BOTH LUNGS: DIFFERENCE IN NUMBER VOXELS ABOVE OR EQUAL TO -200 (PERCENT) | differencenumbervoxelsaboveoreqaltoper cent |

Table 3. The part of the CDE of two custom datasets

There are groups of concepts with identical meaning. Mappings, which need improvements are marked red. Improvements can be easily done using mappings from the rows nearby.

| source_code_description | count dataset1 | count dataset2 | cumulative count | flag | target_concept_code | target_concept_name |
|---|---|---|---|---|---|---|
| stereotactic radiation therapy with a gamma knife | 85595 | | 85595 | dataset1 | 446504008 | Stereotactic destruction of lesion using gamma radiation |
| stereotactic radiotherapy with gamma knife | | 11023 | 11023 | dataset2 | 399315003 | Radionuclide therapy |
| | | | | | | |
| subconjunctival injection procedure | 83201 | | 83201 | dataset1 | 74410004 | Subconjunctival injection |
| subconjunctival injection | | 25238 | 25238 | dataset2 | No matching concept | No matching concept |
| | | | | | | |
| transcutaneous nephrostomy/pyelostomy | 84428 | | 84428 | dataset1 | 39834009 | Nephrostomy |
| percutaneous nephrostomy | | 7553 | 7553 | dataset2 | No matching concept | No matching concept |
| | | | | | | |
| bun | 632454037 | | 632454037 | dataset1 | 105011006 | Blood urea nitrogen measurement |
| bun (urine) | 5159674 | | 5159674 | dataset1 | 84540 | Urea nitrogen, urine |
| urea nitrogen (bun) | | 112719947 | 112719947 | dataset2 | 105011006 | Blood urea nitrogen measurement |
| urea nitrogen [urine] | | 809422 | 809422 | dataset2 | 84540 | Urea nitrogen, urine |

Irina Zherko, MD,
Mikhail Nerovnya, MD,
Michael Kallfelz, MD,
Alexander Davydov, MD

ODYSSEUS DATA SERVICES INC

OHDSI