**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](image-url)