Source vocabulary mapping, typical pitfalls, solutions and quality assurance

Oleg Zhuk, MD[1], Gregory Klebanov, MSc[1], Christian Reich, MD[2], Alexander Davydov, MD[1]

[1]Odysseus Data Services, Inc., Cambridge, MA, USA; [2]IQVIA, Cambridge, MA, USA

Contact: oleg.zhuk@odysseusinc.com

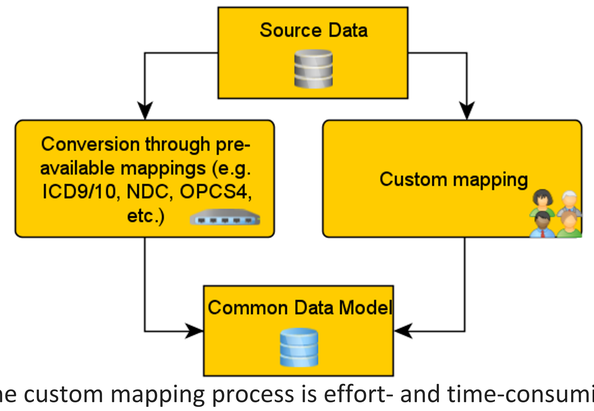
Background

To get all the benefits from the OHDSI ecosystem, users have to perform an extract, transform, and load (ETL) process to walk from the native data to the OMOP Common Data Model (CDM) format. In ETL data is restructured to the CDM format, and mapping from source to the Standard terminology occurs. There are two approaches to the conversion of the source codes to Standard concepts being performed subsequently.

The first option, a vocabulary mapping, includes script-driven conversion through pre-available concept relationships. This direct approach is possible when source data is based on Standard terminology (e.g. SNOMED, LOINC, RxNorm) or the respective vocabulary has already been added to OMOP and mapped (e.g. ICD9/10, NDC, OPCS4). These mappings, either provided by some of the terminology systems (UMLS, ICD10CM, NDC) or developed by OHDSI internally (ICD10, ICD9Proc, international drug vocabularies), are being constantly maintained by the OHDSI vocabulary team.

Another option, a custom mapping, is required when a conversion through existing crosswalks doesn't provide the desired data coverage or scientific use cases are laying outside. The amount of such uncovered data varies among the data sources being highest for oncology data, clinical trials, specific disease registries and sets using proprietary coding systems or uncovered vocabularies (IMO, MedDRA, WHODrug, etc.).

Figure 1. Extract, Transform, Load (ETL) process



The custom mapping process is effort- and time-consuming and can lead to various types of errors that are usually hard to detect and eliminate, but significantly affect the result of the conversion and further study.

There is one more option to deal with source codes that are extremely important and can't be mapped without information loss – turning them into Standard concepts and assigning 2bil+ concept_id. Even though a custom vocabulary (including concept_relationship and concept_ancestor tables) might be built, these concepts will not be available for other community members and cohort building from the perspective of network studies becomes challenging.

Table 1. *QA/QC tests, typical pitfalls and solutions.*

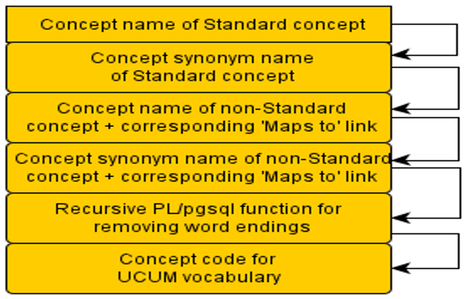
Test	Pitfall	Solution
Source Testing		
Check source_code uniqueness	Data extraction error Source_code definition error	Use another field or even concatenation of some fields as a source_code
Check the length of fields	CDM constraints violation	Use another field or cut to the needed length. Extend CDM field length
Unit Testing (to check general mapping consistency and accordance with CDM specifications/constrains)		
Check if a target concept has the same values in the concept table and are valid and standard	Each ETL conversion runs on a specific OMOP CDM Vocabulary version. Re-run on updated Vocabulary version may result in some concepts are not Valid/Standard anymore Manual mapping mistakes can lead to corrupted target_concept_id and semantic evaluation of a different concept source_code_description | relationship_id | target_concept_id | target_concept_name | target_vocabulary_id | target_domain_id | valid_end_date Kleefstra syndrome | Maps to | 44805996 | Kleefstra syndrome | Snomed | Condition | 1/30/2018 Kleefstra syndrome | Maps to | 37110119 | Kleefstra syndrome | Snomed | Condition | 12/31/2099	Use the same Vocabulary version in mapping and ETL Amend corrupted target_concept_id
'Maps to' mapping to abnormal Domain/Concept class	Each CDM table and field has its purpose so the list of possible concept Domains/Classes is predefined. E.g., mapping to Unit, Meas Value, Specimen Domains must not be used if ETL rules are adjusted to event_concept_id fields source_code_description | relationship_id | target_concept_id | target_concept_name | target_vocabulary_id | target_domain_id | target_concept_class_id Asthma | Maps to | 45877009 | Asthma | LOINC | Meas Value | Answer Asthma | Maps to | 317009 | Asthma | Snomed | Condition | Clinical Finding	Amend mapping. Change ETL rules
Value ambiguous mapping (2 Observation/Measurement concepts for 1 value or vice versa)	This leads to duplication of records in CDM source_code_description | relationship_id | target_concept_id | target_concept_name | target_vocabulary_id | target_domain_id | target_concept_class_id Allergy to house dust | Maps to | 4304626 | House dust RAST | Snomed | Measurement | Procedure Allergy to house dust | Maps to value | 9191 | Positive | Snomed | Meas Value | Qualifier Value Allergy to house dust | Maps to | 4048168 | Allergy to house dust | Snomed | Observation | Clinical Finding Allergy to house dust | Maps to | 4048168 | Allergy to house dust | Snomed | Observation | Clinical Finding	Amend the multiple mapping. Skip if duplication is required
'Maps to value' without 'Maps to'	Value_as_concept_id field cannot be populated if event_concept_id is not defined source_code_description | relationship_id | target_concept_id | target_concept_name | target_vocabulary_id | target_domain_id | target_concept_class_id Allergy to phytosterols | Maps to value | 19044812 | Phytosterols | RxNorm | Drug | Ingredient Allergy to phytosterols | Maps to | 4169307 | Allergy to substance | Snomed | Observation | Clinical Finding Allergy to phytosterols | Maps to value | 19044812 | Phytosterols | RxNorm | Drug | Ingredient	Add 'Maps to' relationship mapping
Used vocabularies	Most vocabularies are used only in certain circumstances Some vocabularies are license-required source_code_description | relationship_id | target_concept_id | target_concept_name | target_vocabulary_id | target_domain_id | target_concept_class_id implant /abutment supported fixed denture for partially edentulous arch - mandibular | Maps to | 944898 | implant /abutment supported fixed denture for partially edentulous arch - mandibular | CDT | Observation | CDT implant /abutment supported fixed denture for partially edentulous arch - mandibular | Maps to | 40663172 | IMPLANT/ABUTMENT SUPPORTED FIXED DENTURE FOR PARTIALLY EDENTULOUS ARCH | HCPCS | Device | HCPCS	Amend mapping Confirm the license
Semantic Testing		
Concepts that can't be used without 'Maps to value' link	Includes 'History of', 'Disease suspected' and other concepts that don't make sense without 'Maps to value' source_code_description | relationship_id | target_concept_id | target_concept_name | target_vocabulary_id | target_domain_id | target_concept_class_id Family history of viral pneumonia | Maps to | 4083519 | Family history of disorder | Snomed | Observation | Context-dependent Family history of viral pneumonia | Maps to | 4083519 | Family history of disorder | Snomed | Observation | Context-dependent Family history of viral pneumonia | Maps to value | 261326 | Viral pneumonia | Snomed | Condition | Clinical Finding	Add 'Maps to value' mapping. Remove unnecessary mappings
Multiple mapping	Review multiple mappings (only 'Maps to' / 'Maps to' with value separately)	Approve or remove unnecessary mappings
Key terms loss/misuse	Acute, recurrent, suspected, chronic, left/right and other attributes might be lost or misused source_code_description | relationship_id | target_concept_id | target_concept_name | target_vocabulary_id | target_domain_id | concept_class_id Acute arthritis of left knee joint | Maps to | 4159739 | Arthritis of knee | Snomed | Condition | Clinical Finding Acute arthritis of left knee joint | Maps to | 759891 | Arthritis of left knee | Snomed | Condition | Clinical Finding Acute arthritis of left knee joint | Maps to | 4000634 | Acute arthritis | Snomed | Condition | Clinical Finding	Add/amend mapping
USAGI reverse mapping	Run USAGI using target concepts as source and source as vocabulary. Review codes mapped to different targets and having the lowest scores	Amend mapping

Results

Custom mapping algorithm implies:

1. Source data analysis and source code definition (depending on the context, it might be not trivial, e.g. concatenation of several fields, etc).
2. Source vocabulary clean up by excluding junk and meaningless source terms.
3. Source consistency testing (table 1).
4. Automated term match in Standardized vocabularies - subsequent matching by following pathways:

Figure 2. Pathway to match Standard terminology



5. Automated term match in previously mapped concept pool: (a) source_code_description; (b) concept_synonym_name - source_code_description translated to English if the source has a different language.
6. USAGI (**https://github.com/OHDSI/Usagi**) automated mapping and manual review based on match scores.
7. Manual mapping of the rest portion (with the usage of Athena and JavaScript UX/UI extension for automation of copying by one click).
8. Mapping QA/QC tests (table 1). Also manual cross-review by another expert to provide the best quality and reduce the number of human errors.

Methods

In order to recognize, amend and eliminate typical mapping pitfall we present the system of QA tests implemented in SQL scripts. Also, a formalized custom mapping algorithm is recommended to make the process more efficient.

Conclusions

The QA/QC tests and algorithm that were developed may improve mapping accuracy and effectiveness of the process. We recommend implementing both automated tests and those that require further expert review. The current mapping rate achieved with the help of the provided mapping algorithm is around 50%. We believe that further improvements are possible with the implementation of Natural language processing (NLP) and an extensive increase in the number of collected crosswalks.