Assessing the quality of mapping a data set to the OMOP Common Data Model

Michel Van Speybroeck¹, Lars Halvorsen², Peter Rijnbeek ³
¹Janssen Data Sciences, Beerse, Belgium, ²Erasmus MC, Rotterdam, The Netherlands

Background

Within the IMI project European Medical Information Framework (EMIF), 10 databases were mapped to the OMOP Common Data Model. The Health Improvement Network (THIN) was one of the databases. THIN contains the data of 711 GP Practices in the UK, covering 15.6 million subjects.

As part of the workflow THIN is working together with the participating providers to input good quality data in line with the standards as identified in the General Medical Services (GMS) contract and Quality and Outcomes Framework (QoF).

Within EMIF, we were specifically interested in how to specify the quality of the mapping to an OMOP Common Data Model. Measuring the quality of the mapping process is a critical feature to enable reproducible and trustworthy evidence generation.

Methods

THIN v1701 was mapped to OMOP V5.2 using the open source version of Pentaho v7.1. A total of 21 Pentaho scripts were developed to map 6 input tables to 15 different OMOP tables. Given the sheer volume of data (in total over 6 billion rows), the scripts had to be optimized to keep processing times to acceptable levels. This was achieved by reducing expensive lookups between tables and enabling –where appropriate– parallel processing in Pentaho.

Following the mapping process, different assessments were made to measure the quality of the mapping process. A distinction was made between the ‘structural’ mapping quality and the ‘semantic’ mapping quality. While the first one measures the extent to which data elements in the source are mapped at all to the target system and to the right entity, the second element – the semantic mapping quality- tries to quantify if the variables can be mapped to a standard OMOP concept and if there is a potential loss of granularity in this process.

The quality assessment included the following steps:
1) Reviewing the log of the mapping script execution (structural mapping quality)
2) Manual Review of Achilles and Achilleshee results (structural and semantic mapping quality)
3) Summary level queries per OMOP entity on completeness of mappings (semantic mapping)
4) Specific queries on drug_exposure and condition_occurrence comparing input data vs OMOP mapped data (structural mapping)

The quality assessment was also used to drive a gradual improvement in the mapping, so when results were deemed unacceptable, scripts (and/or the semantic mappings) were adapted / expanded to obtain a higher mapping quality and the whole cycle was repeated.

Results

During the execution of the mapping scripts, the generated logs were summarized to keep track of number of processes records, processing times and the mapped entities (see fig 1).

![Fig 1: Screenshot of the script execution tracker](image)

As the table structure in the THIN source doesn’t map one-on-one to the target OMOP table schema, one particular use of the tracker was to understand the relative contribution that each source entity made to particular target tables. In a second step, the counts and possible discrepancies in counts between source and target was used to assess potential deficiencies in the mappings.

Achilles and Achilleshee output were also reviewed as part of the quality assessment and the identified errors were used to gradually improve the mapping scripts.

Summary level queries executed per OMOP entity were executed and results are represented in fig 2.

![Fig 2: Mapping overview per OMOP entity (all results are against CDM tables)](image)

By applying the repetitive cycle of improving scripts and mappings, almost all records that were uploaded to OMOP could be mapped using standard concepts. For drugs, despite the fact that only 62% of the codes were mapped, this still resulted in 98% of the records that could be mapped. It has to be noted in addition, that since the mapping was done, a more comprehensive coverage of local drug codes is available in OMOP, which would result in a significantly higher percentage of mapped codes.

The last step assessed specifically the full count comparison or a given code in the source system vs a count on the standard concept in OMOP. Results are given below for drug_occurrence.

![Figure 3](image)

Figure 3 shows that there is indeed a difference in count between source and OMOP CDM. As indicated in figure 2, this discrepancy is only marginally caused by the absence of a standard concepts (counting for about 2% of the losses). The main contributing factor was the explicit business rule to only include records which had at source a flag indicating an ‘acceptable record’.

Conclusions

The assessment indicates that based on absolute counts, a difference can be found between the source system and OMOP. This difference is based on explicit design decisions. For the semantic mappings, a near complete mapping to standard concepts could be obtained. However, this is following a repetitive cycle of assessment, adjustment and re-execution of the mapping logic. In addition, it’s imperative that consistent design decisions are made when mapping data sources i.e. mapping all records or only mapping records that adhere to certain established criteria.