# Automation of STCM Vocabularies Review in OMOP CDM

**Wai Yi Man[1], Antonella Delmestri[1]**
**[1]Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences (NDORMS), University of Oxford, UK**

## Background

Clinical Practice Research Datalink (CPRD) GOLD is a well-established structured UK primary care data source with over twenty million patients and more than ten billion events in total in recent releases[1]. In 2022, CPRD GOLD, mapped to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) and represented by the University of Oxford, became part of the DARWIN EU®[2] Network to provide real-world evidence across Europe on diseases and medicines' use, risks and benefits. The University of Oxford is committed to onboarding in DARWIN EU® a new OMOP release of CPRD GOLD (CDM_GOLD) every six months.

In CPRD GOLD, some clinical data, like measurement units and drug routes, are recorded as descriptions instead of being coded. Thus, it is not feasible to directly adopt the OHDSI standardised vocabularies to map these data to the OMOP CDM. Using the OMOP CDM *SOURCE_TO_CONCEPT_MAP* (STCM) table, we have added STCM-tailored vocabularies to expand the mapping information. These STCM vocabularies are comma-separated values (CSV) files containing mapping information between local source codes in CPRD GOLD and OHDSI standard concepts[3,4]. However, regular manual reviews and updates on the STCM files are required as the OHDSI vocabularies keep changing to facilitate better content for the mapped data.

## Methods

The original approach to update the STCM vocabularies requires manually reviewing (i.e., browsing codes in Athena[5], the OHDSI vocabularies repository, and then verifying the concept IDs in the CSV files) and updating, when necessary, STCM CSV files, before loading them into the database. We have implemented an algorithm to automate this process.

Before the Extract, Transform, and Load (ETL) process, we ran a Python program[6] to prepare a database with source data, OHDSI standardised and STCM vocabularies, primary keys, indexes, and constraints. Then, the program searches all outdated (i.e. non-standard or/and invalid) STCM concepts and the new associated standard *concept IDs*. For each outdated STCM concept, there are three possible scenarios, illustrated in Figure 1:

1. The outdated STCM concept maps to an OHDSI standard concept (i.e., RELATIONSHIP_ID= 'Maps to'). The mapping is updated to the new standard OHDSI concept.
2. The outdated STCM concept is replaced by another OHDSI concept (i.e., RELATIONSHIP_ID= 'Concept replaced by'), which has then to be mapped. The mapping goes under an extra check as the replaced concept may be invalid or not standard. Likewise, the mapped concept is updated if a ‘Maps to’ concept exists, or deleted if no ‘Maps to’ concept is found.
   In addition, duplication results from scenarios 1 and 2 are removed as ‘Maps to’ and ‘Concept replaced by’ relationships are not mutually exclusive for a concept.
3. Others (i.e., neither 1 nor 2. A non-standard or/and invalid concept can exist without any relationships with other concepts.). The concept is deleted.

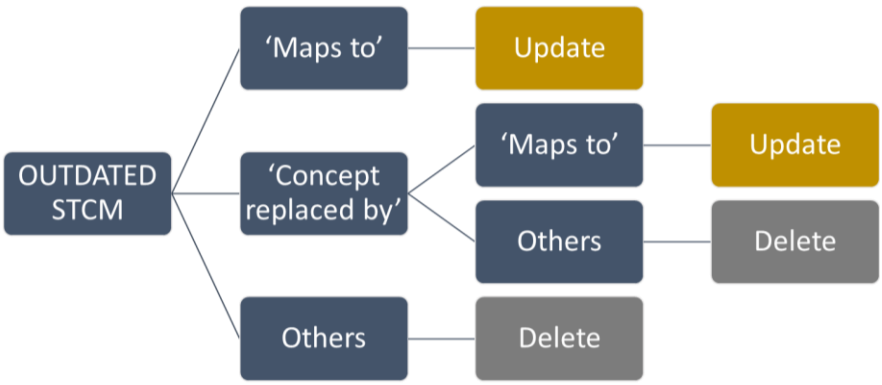


Figure 1: Three possible scenarios for outdated STCM concepts

After updating one STCM vocabulary in the database, the program will generate a new STCM CSV file, replacing the outdated one for the next ETL process.

**Results**

In the original (i.e., manual) approach, each concept ID in a STCM file, must be checked in Athena, and then updated if necessary. This process must be repeated at each new release for each concept ID, which was over 5,000 in the last release of CPRD GOLD (July 2023 data release). To review one concept, it is reasonable to assume that one person can take between 30 seconds and 2 minutes. By multiplying this by 5,000 concept IDs, the range of saved human hours per CDM_GOLD release is [41.7-166.7]. On the other hand, our algorithm is fully automated, may take from seconds to minutes to review around ~5,000 concept IDs. This automation reduces the overall processing time, and removes any potential human errors, producing higher quality data and a more robust cycle of CDM_GOLD onboarding.

**Conclusion**

The automation reduces 99% of the processing time of the STCM vocabularies review, eliminates any manual work and human errors, enhances data quality in CDM_GOLD data, and streamlines the onboarding of CDM_GOLD in DARWIN EU®.

**References**

1. *CPRD GOLD July 2023 dataset*. Available at: https://cprd.com/cprd-gold-july-2023-dataset (Accessed: 4 March 2024).
2. *Data Analysis and real-world interrogation network (Darwin EU) Data Analysis and Real World Interrogation Network (DARWIN EU) | European Medicines Agency*. Available at: https://www.ema.europa.eu/en/about-us/how-we-work/big-data/data-analysis-real-world-interrogation-network-darwin-eu (Accessed: 19 February 2024).
3. Clair B. and Erica V. (2019) '6.3.1 Usagi'. Observational Health Data Sciences and Informatics *The Book of OHDSI*. Italy: Independently published, pp. 86-93.
4. Observational Health Data Sciences and Informatics (OHDSI) *Data model conventions*, *dataModelConventions.knit*. Available at: https://ohdsi.github.io/CommonDataModel/dataModelConventions.html#Source_to_Concept_Map (Accessed: 31 January 2024).
5. *Athena*. Available at: https://athena.ohdsi.org/ (Accessed: 27 February 2024).
6. *Oxford-pharmacoepi/etl_ndorms*, *GitHub*. Available at: https://github.com/oxford-pharmacoepi/etl_ndorms (Accessed: 26 February 2024).