Automation of STCM Review saves from HOURS to DAYS of human time

Presenter: Wai Yi Man

Automation of STCM Vocabularies Review for OMOP CDM

Background: The SOURCE TO CONCEPT MAP (STCM) approach maps local source codes not present in the Athena vocabularies to standard concepts. The vocabularies keep evolving to allow updated and accurate data mapping, and this requires regular manual reviews: each concept ID in an STCM-tailored CSV-formatted vocabulary, which must be checked against Athena, and then updated if necessary. This process must be repeated at each new data release. We have implemented an algorithm to automate this process.

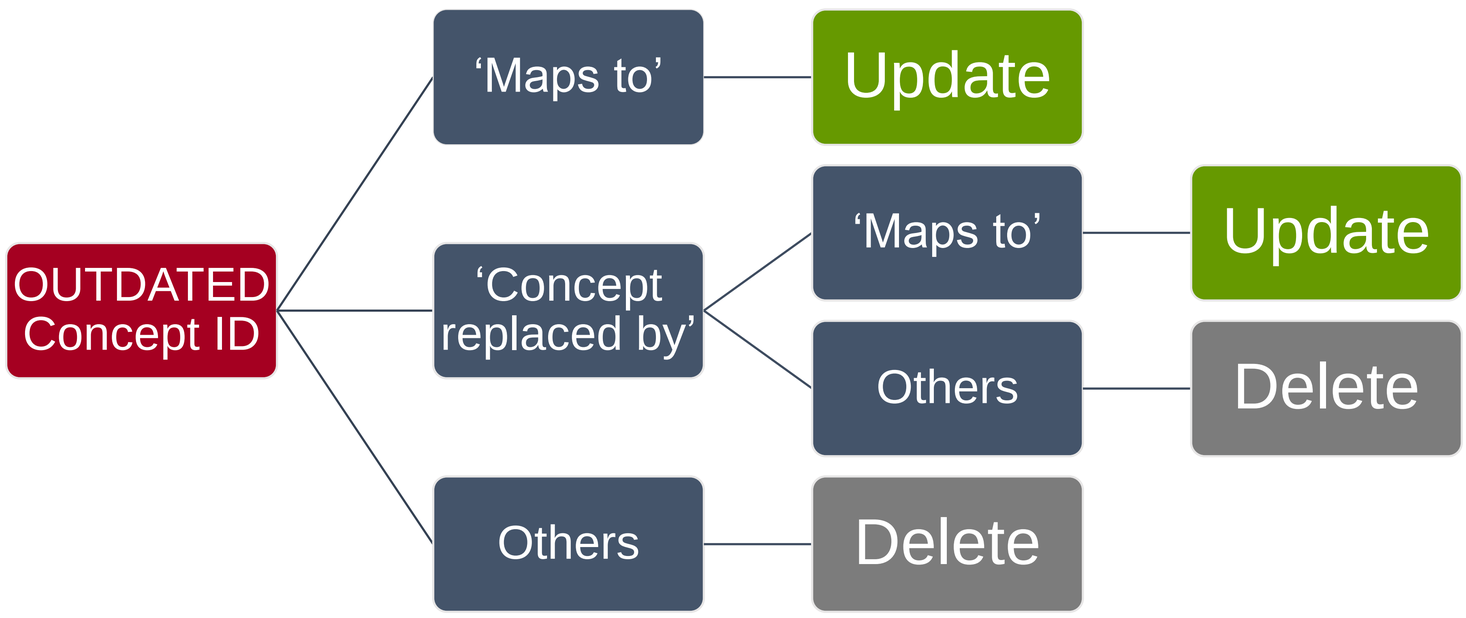
Results:

In the last release of CPRD GOLD* (July 2023 data release) there were over 5,000 concept IDs. Reviewing one concept typically takes between 30 seconds and 2 minutes for one person. By multiplying this by 5,000 concept IDs, the range of saved human hours per CPRD GOLD release extends between 41.7-166.7 hours. Our algorithm is fully automated, and it takes 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 data mapping cycle.

*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 2022, CPRD GOLD, mapped to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) was onboarded in DARWIN EU®. The University of Oxford is committed to onboarding in DARWIN EU® a new OMOP release of CPRD GOLD every six months.

Methods

Before starting the Extract, Transform, and Load (ETL) process, our algorithm, developed in a Python program, searches all outdated (non-standard or/and invalid) target concept IDs in the STCM files, and identifies the new associated standard concept IDs. For each outdated STCM target concept ID, there are three possible scenarios:



1. **The outdated STCM target concept ID is mapped to a new Athena standard concept ID (RELATIONSHIP_ID= 'Maps to') and the STCM target concept ID is updated.**
2. **The outdated STCM target concept ID is replaced by another Athena concept ID (RELATIONSHIP_ID= 'Concept replaced by'), which has then to be mapped.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.** 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 (neither 1 nor 2. A non-standard or/and invalid concept can exist without any relationships with other concepts). The concept is deleted.**

Wai Yi Man, Antonella Delmestri

NDORMS
NUFFIELD DEPARTMENT OF ORTHOPAEDICS, RHEUMATOLOGY AND MUSCULOSKELETAL SCIENCES

UNIVERSITY OF OXFORD

health data sciences