Harmonising medical data from a disease registry to the OMOP CDM: pitfalls and lessons learned

Descamps Freija[1], Gesquiere Emma[1], Khan Hamza[2], Parciak Marcel[2], Parciak Tina[2], Peeters Liesbet[2]
[1] edenceHealth NV, [2] UHasselt, Biomedical Research Institute, Data Science Institute; Universitair MS Centrum Hasselt-Pelt
Contact: freija.descamps@edence.health

BACKGROUND

Medical Real World Data (RWD) is an extremely valuable resource for research and the improvement of care, for example in pharmacovigilance and rare disease research. RWD is recorded in a multitude of different source systems (e. g. EHR, registries) by different stakeholders in care, including the patient, and often with the additional use of different terminologies (e. g. SNOMED, ICD10). This makes cross-platform analyses very challenging. The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) is an open-source common data model that standardizes both the structure and semantics of medical RWD [1][2].

We review our steps of the harmonization process for the MS DataConnect dataset and share lessons learned from designing and building the data pipeline.

HARMONIZATION PROCESS

We transformed clinical data of a multiple sclerosis (MS) consortium to the OMOP common data model. Figure 1 outlines all steps included in this process. Figure 2 shows an example of a source variable and its structural and semantic mapping.

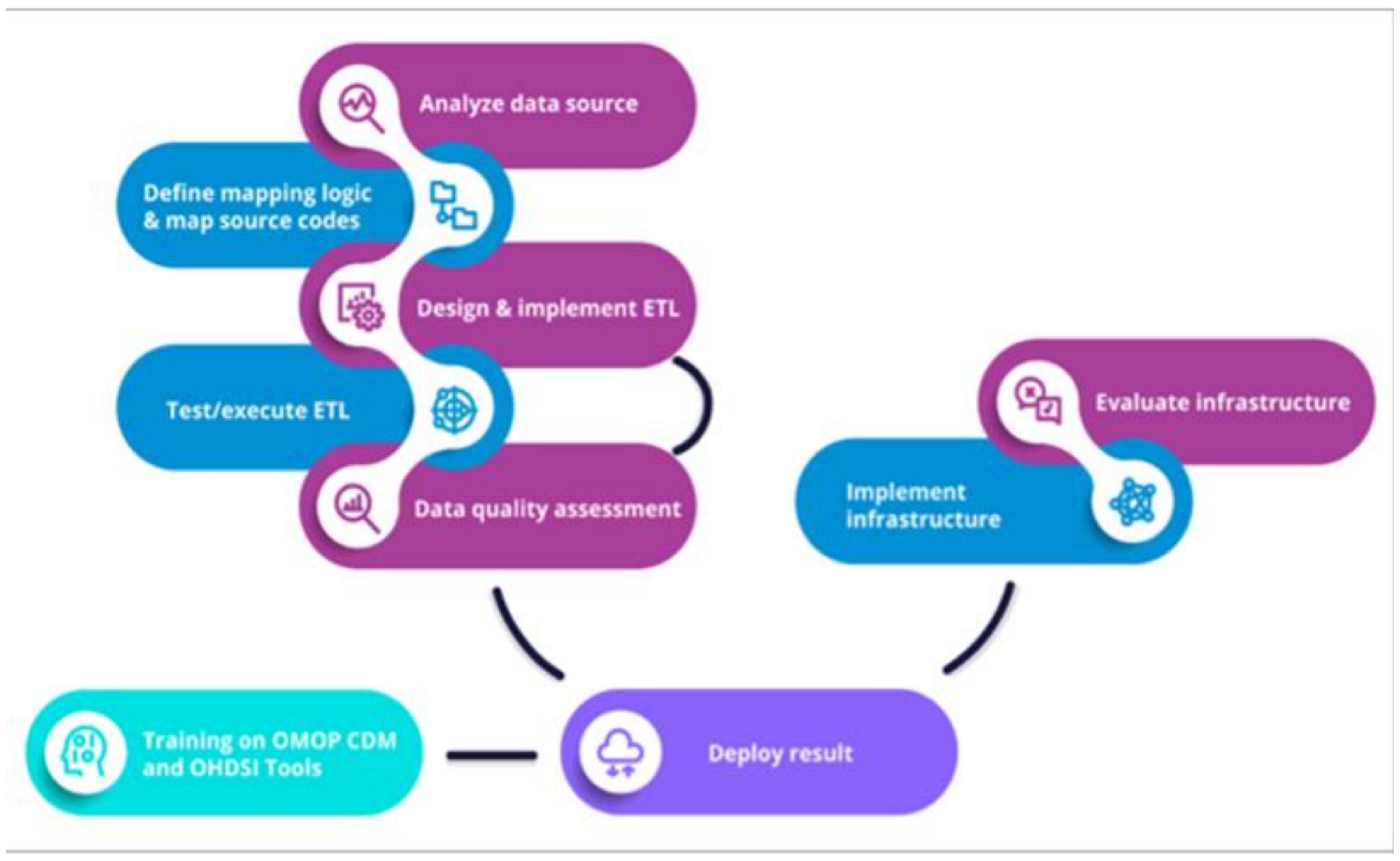
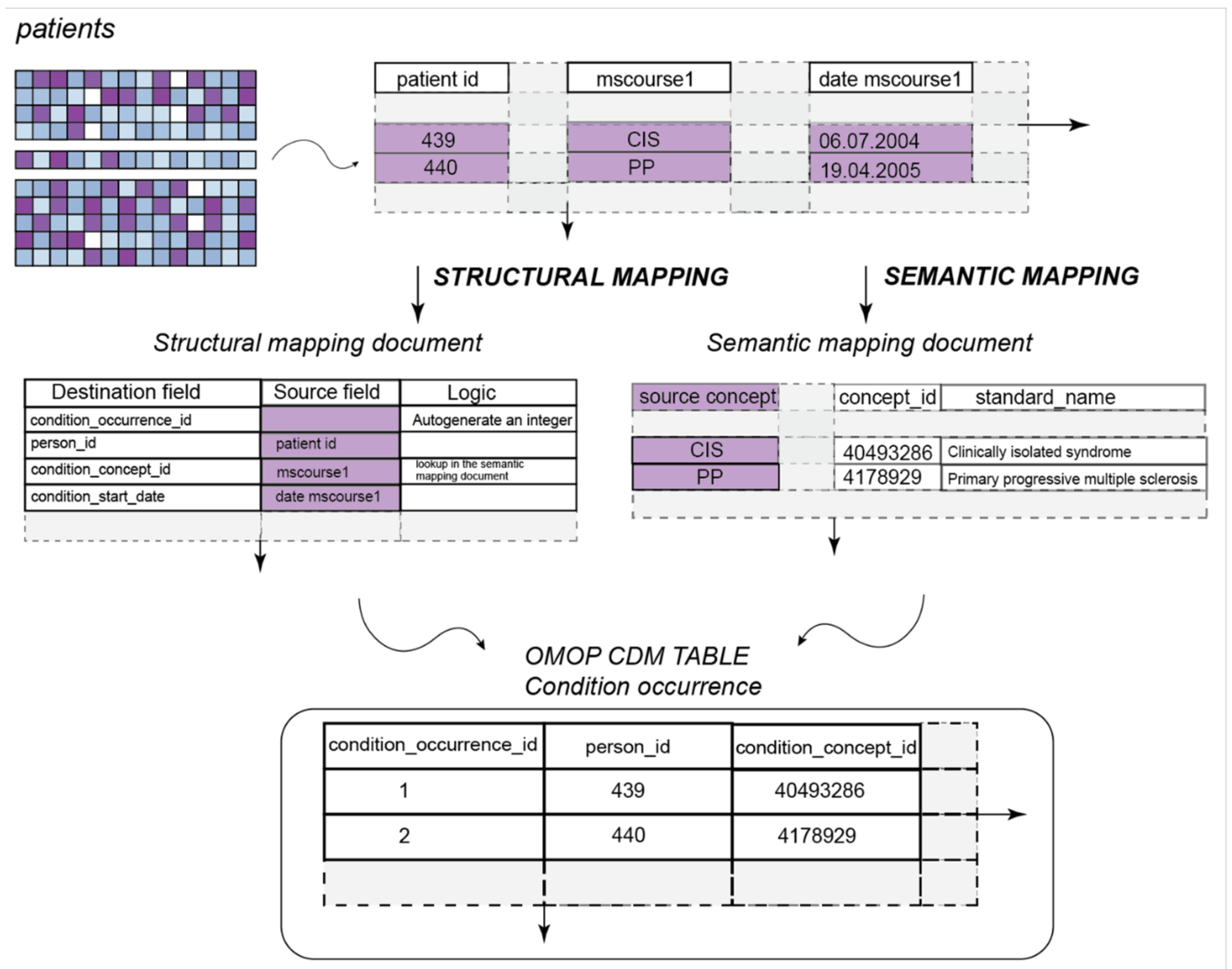


Figure 1. Schematic illustrating the steps in a typical harmonization project.

CHALLENGES

The OMOP common data model lacks features for disease-specific registry data since it is originally designed for electronic health records and claims' data. This became obvious when we dived deep into the details of our clinical data. There are a few concepts missing from the OMOP CDM vocabulary that are disease specific, mandatory variables. Also, since the data emerges from a registry-like source system, the table design of the OMOP CDM was not ideal in some cases. The focus of our data is disease-specific, so its scope is therefore limited relative to a complete EHR. We introduced these challenges to the OHDSI community, confirming the need to develop a joint approach for harmonising similar data.



patient id	mscourse1	date mscourse1
439	CIS	06.07.2004
440	PP	19.04.2005

Structural mapping document

Destination field	Source field	Logic
condition_occurrence_id		Autogenerate an integer
person_id	patient id	
condition_concept_id	mscourse1	lookup in the semantic mapping document
condition_start_date	date mscourse1	

Semantic mapping document

source concept	concept_id	standard_name
CIS	40493286	Clinically isolated syndrome
PP	4178929	Primary progressive multiple sclerosis

OMOP CDM TABLE
Condition occurrence

condition_occurrence_id	person_id	condition_concept_id
1	439	40493286
2	440	4178929

Figure 2. An example showing how a source concept is mapped to the OMOP CDM. The table at the top shows the data in its original format. This data is then mapped structurally (left) where the concept is mapped to the condition occurrence table in the target database and the source fields are matched to the target fields. The right table shows the semantic mapping where the source codes are mapped to a concept in the standard vocabulary.

CONCLUSIONS

Over time we found that changes had to be made to our typical approach to accommodate some of the challenges in this dataset. We decided to drop free-text fields as the work that would go into mapping these would outweigh its limited value. Currently, disease-specific codes are limited in the OMOP standard vocabulary. We advocate to increase the availability of MS-specific as well as other disease-specific codes in the OMOP standard vocabulary in order to enable improved mapping of disease-specific datasets. We also see value in an expansion of the OMOP CDM so it can capture multi-dimensional medical data (time series and imaging data) for which support is currently lacking.

REFERENCES

[1] Overhage, J. M., Ryan, P. B., Reich, C. G., Hartzema, A. G., & Stang, P. E. "Validation of a common data model for active safety surveillance research." *Journal of the American Medical Informatics Association : JAMIA*, *19*(1) (2012): 54–60.

[2] The Book of OHDSI: Observational Health Data Sciences and Informatics. OHDSI, 2019, pp 458. https://ohdsi.github.io/TheBookOfOhdsi/

ACKNOWLEDGMENTS

This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No 806968. The JU receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA. Parts of the resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government.
We thank the medical and administrative personnel of the Noorderhart Rehabilitation and MS Center in Pelt (BE) for the follow-up of the people with MS and corresponding data collection.
Tina Parciak is supported by the Special Research Fund (BOF)/Hasselt University (BOF22OWB01).

Session 2 DATA PREPARATION

1st RADar conference **EmP Engineer meets Physician** 31.05 + 01.06.2022