Lessons from mapping cancer information from European hospitals to ICD-O-3 conditions in OMOP

Lars Halvorsen[a], Olivier Bouissou[b], Elisabeth Ross[b], Stelios Theophanous[c], Joëlle Thonnard[d], Piers Mahon[e]

a) edenceHealth NV, Kontich, Belgium, b) Oslo University Hospital, Oslo, Norway, c) Leeds Teaching Hospitals NHS Trust, Leeds, United Kingdom, d) Clinique universitaires Saint-Luc, Brussels, Belgium, e) IQVIA, London, UK

Background

European hospitals in the DigiONE network[1] are converting their cancer data to the OMOP common data model (CDM) and face a recurring conversion challenge in mapping cancer conditions, as their source data is typically: (i) primary cancer type, typically in ICD10/9, (ii) histological data, and (iii) staging with no structured location data.

Cancer OMOP often depends on ICD-O-3 codes that use (a) location (topography), (b) histology (morphology) and (c) staging. Although the hospitals mapped their cancer diagnoses to ICD-O-3 or SNOMED using topography and morphology information from pathology reports, differences in terminologies used at source presented different challenges for different hospitals. All hospitals also faced similar challenges in mapping a subset of ICD-O-3 diagnosis codes, as a sizeable set of these codes have not yet been defined in the OHDSI vocabularies[2]. With a planned update to the ICD-O-3 vocabulary in an upcoming OHDSI vocabulary release, one objective of the following analysis was to identify a common subset of missing ICD-O-3 condition code mappings for inclusion in the next vocabulary update.

Here we compare how three DigiONE hospitals approached this mapping challenge; Cliniques universitaires Saint-Luc (Brussels, Belgium)[3], Leeds Teaching Hospitals NHS Trust (Leeds, UK)[4], and Oslo University Hospital (Oslo, Norway)[5]. In the discussion below, the three hospitals are referred to as Hospital 1 through 3, in random order.

Methods

The three hospitals have different paths for their mappings to ICD-O-3 condition concepts, and therefore undertook somewhat different approaches:

<u>Hospital 1:</u>

ICD10 cancer diagnosis codes were mapped to standard SNOMED concepts using the OHDSI vocabulary and stored in the *condition_occurrence* table, and ICD-O-3 morphology codes were mapped to standard and included in the *observation* table. To create ICD-O-3 condition codes, the ICD-O-3 topography codes were derived from the ICD10 codes, where possible, using guidance from SEER[6].

For solid cancers, there were some missing morphology codes in the source data. Some ICD10 codes could not be converted to ICD-O-3 topography codes (C22.2, C22.3, C22.4, C22.7, C22.9), while in other cases there was a one-to-many relationship between the ICD10 diagnosis and ICD-O-3 topography codes, and these were therefore not included.

For haematological cancers, the mappings were created based on the ICD-O-3 morphology codes alone, with no topography code (e.g. 8140/3-NULL [*Neoplasm defined only by histology: Adenocarcinoma, NOS*]).

Hospital 2:

In addition to ICD10 cancer diagnosis codes, which were mapped to standard SNOMED concepts using the OHDSI vocabulary, pathology results are available coded in a local vocabulary. The local codes specify a T code (topography) and an M code (morphology).

The local vocabulary is based on an older SNOMED version, with most of the T codes having a corresponding SNOMED concept code available. In most cases, the M codes corresponded to an equivalent ICD-O-3 morphology code. However, there were cases where either no valid SNOMED or ICD-O-3 condition reference could be found for a local code, so a three-stage process was developed:

1. Use the T code provided as SNOMED concept and convert the M code from non-standard ICD-O-3 morphology concept to standard SNOMED morph abnormality concept using the concept_*relationship* table in the OHDSI vocabulary. Search the vocabulary, using concept_*relationship*, for a standard SNOMED concept which has (only) these T and M SNOMED concepts defined as *Has finding site* and *Has associated morphology* relationships, respectively.
2. If (1) fails, use the same approach, but for the ICD-O-3 path; convert the T code to a corresponding ICD-O-3 topography code, combine this with the ICD-O-3 morphology representation of the M code to find the ICD-O-3 Condition concept. If this is a non-standard concept, use the concept_relationship table to find the corresponding standard (SNOMED) concept.
3. If both steps fail, attempt a manual mapping to a standard condition concept based on the information provided in the local T and M codes.

In all cases above, the standard concept for the individual T and M codes were stored in the *observation* table.

Hospital 3:

Both ICD-O-3 topography and ICD-O-3 morphology codes were available in the source system, which enabled direct construction of the ICD-O-3 condition codes.

The ICD-O-3 topography and ICD-O-3 morphology concepts were included as individual records in the *observation* table, while the constructed ICD-O-3 condition codes were used to look up the standard condition concept.

ICD-O-3 condition concepts which exist in the OHDSI vocabulary as standard concepts were added to the *condition_occurrence* table. For non-standard ICD-O-3 condition concepts, the *concept_relationship* table was used to look up the standard (SNOMED) concept, which was added to the *condition_occurrence* table instead. If neither of these were possible, the ICD-O-3 code was captured for further follow-up.

Results

The main results from the mapping exercise are presented in Table 1 and Figure 1:

Solid and haematological cancers		**Hospital 1**	**Hospital 2**	**Hospital 3**
OHDSI vocabulary version		v5.0 29-FEB-24	v5.0 23-JAN-23	v5.0 29-FEB-24
Source	ICD-O-3 source data	ICD-10 diagnosis codes (*converted to ICD-O-3 topography*) + ICD-O-3 morphology	Local pathology vocabulary for topography and morphology	ICD-O-3 topography + ICD-O-3 morphology
	# unique source codes	3588	2979	1823
Mapping Result	# standard ICD-O-3 Condition concept ID	2413	1524	1308
	# non-standard ICD-O-3 Condition concept IDs mapped to standard (SNOMED) concept IDs	706	568	338
	# ICD-O-3 Condition codes with no concept ID in vocabulary	469	887	177

Table 1: Results of mapping to ICD-O-3 Condition concepts at the three DigiONE hospitals

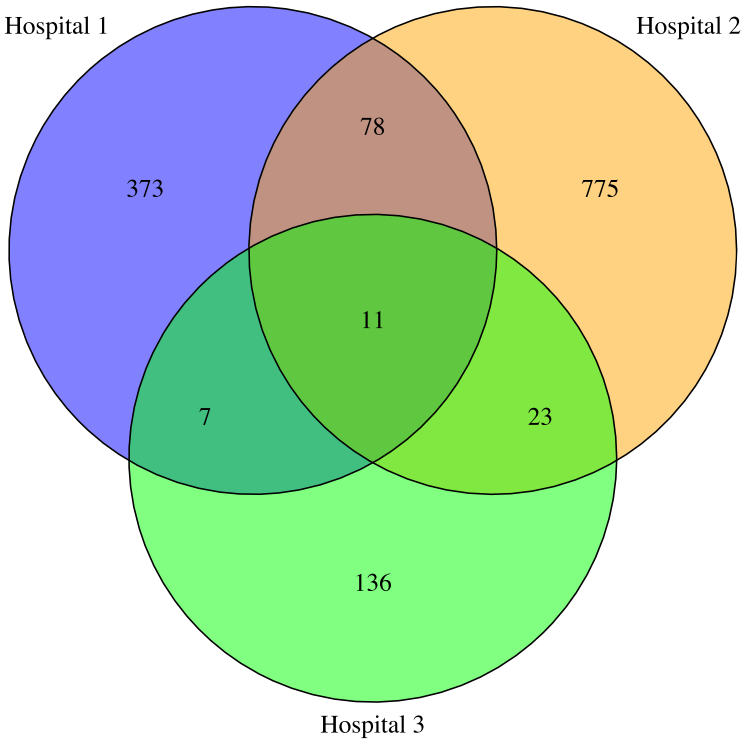

Figure 1: Venn diagram illustrating the overlap between the sets of missing ICD-O-3 Condition codes between the 3 hospitals

Conclusion

Two different vocabulary versions were in use at the three hospitals at the time of the transformation to OMOP CDM. As there were no recorded changes to the included ICD-O-3 concepts between these two vocabulary versions, this has no impact on the comparisons here.

All three hospitals ended up with ICD-O-3 condition codes which do not exist in the OHDSI vocabularies yet. There were 119 such codes which appeared in at least two of the three hospitals; missing codes that appeared at only one hospital were excluded from the final list to censor possible invalid topography and morphology combinations. Of these missing codes, 53 have already been defined for inclusion in a future OHDSI vocabulary release, while 66 of these codes had not yet been included. As a result, the DigiONE hospitals will further evaluate these findings and work with the Oncology WG – Vocabulary subgroup to include an additional set of ICD-O-3 condition codes considered as relevant. This would positively impact future studies within the DigiONE network and other OMOP CDM-based oncology research networks.

References

1. Mahon, P., Chatzitheofilou, I., Dekker, A. et al. A federated learning system for precision oncology in Europe: DigiONE. Nat Med 30, 334–337 (2024). https://doi.org/10.1038/s41591-023-02715-8
2. Prinsen P, van Swieten M, Attanasio C, Enerly E, Larønningen S, Rapiti E, Fournier E, Künzli P, Schnell M, Marcic D, Couffignal S, Backes C. Mapping of ICD-O-3 codes to standard concepts: Lessons learned from four population-based cancer registries. *OHDSI Europe Symposium 2023*. https://ohdsi-europe.org/images/symposium-2023/posters/poster_OHDSI_symposium_ICDO3_Peter_Prinsen_-_Peter_Prinsen.pdf
3. Thonnard J, Calay F, Timmermans A, Van Marcke C. An OMOP Cancer Data warehouse for data Quality Control, collaborative Real World Evidence Studies and data submission to the Belgian Cancer Registry. *OHDSI Europe Symposium 2023*. https://www.ohdsi-europe.org/images/symposium-2023/posters/Digione_OHDSI_poster_june2023_final_-_Joelle_Thonnard.pdf
4. Theophanous S, Zucker K, Hick L, Bolton E, Riaz M, Fenton H, Corkett J, Cheeseman S, Hall G. OMOP for oncology data: A single-centre and network perspective. *OHDSI Europe Symposium 2023*. https://ohdsi-europe.org/images/symposium-2023/posters/Theophanous_OMOPforOncologyData_poster_-_Stelios_Theophanous.pdf
5. Bouissou O, Small Hanto IK, Ross E. Oslo University Hospital Participation in a European Cancer OMOP Network - DigiONE. *OHDSI Europe Symposium 2023*. https://ohdsi-europe.org/images/symposium-2023/posters/Bouissou_OUS_OMOP_Network_-_elisabeth_rosshotmailcom.pdf
6. SEER, ICD Conversion Programs (2024). https://seer.cancer.gov/tools/conversion/

Appendix 1

ICD-O-3 Condition codes, used at 2 or more of the three hospitals, which are not yet in the standardized OHDSI vocabularies.

8000/3-C44.2*	8070/3-C38.4*	8070/6-C67.9	8140/3-C41.9	8246/3-C76.3*	8800/3-C76.4*
8000/9-C34.9	8070/3-C49.1	8070/6-C76.3	8140/6-C17.0	8265/3-C50.9*	8801/3-C76.5
8000/9-C71.9	8070/3-C49.2	8072/3-C44.7*	8140/6-C17.1	8323/3-C07.9	8805/3-C76.5
8000/9-C76.3	8070/3-C49.6	8076/3-C44.2*	8140/6-C18.0	8323/3-C18.1	8811/3-C76.4
8004/3-C44.9	8070/3-C49.9*	8076/3-C44.3*	8140/6-C18.7	8430/3-C41.1*	8851/3-C34.9
8010/3-C38.3*	8070/3-C76.2*	8076/3-C44.5*	8140/6-C19.9	8441/3-C20.9	8851/3-C76.0
8010/3-C38.4*	8070/3-C76.4*	8076/3-C44.6*	8140/6-C20.9*	8460/3-C76.3*	8851/3-C76.2*
8010/3-C44.1*	8070/3-C76.5*	8076/3-C44.7*	8140/6-C22.1	8461/3-C55.9	8890/3-C11.9*
8010/3-C44.2*	8070/6-C02.9	8078/3-C44.9*	8140/6-C25.2	8480/6-C18.0	8890/3-C41.2*
8010/3-C76.2*	8070/6-C06.9	8083/3-C44.7*	8140/6-C26.9	8480/6-C18.1*	8890/3-C76.4
8010/3-C76.4	8070/6-C10.9	8083/3-C44.9	8140/6-C34.1*	8480/6-C76.2	8936/3-C49.6
8010/6-C40.9	8070/6-C14.0	8090/3-C69.6*	8140/6-C34.3	8480/6-C76.3	9040/3-C38.1*
8010/9-C54.1	8070/6-C17.9	8091/3-C44.1*	8140/6-C63.2	8500/3-C44.9	9040/3-C76.5
8010/9-C67.9	8070/6-C32.9	8091/3-C44.2*	8160/3-C25.0*	8500/3-C76.1	9061/3-C77.2
8032/3-C44.9*	8070/6-C34.3	8094/3-C44.2*	8200/3-C63.2	8500/3-C76.4	9180/3-C76.5
8032/3-C51.9*	8070/6-C38.4	8095/3-C44.9*	8211/3-C15.5*	8584/3-C38.3	9370/3-C76.3
8041/3-C38.4*	8070/6-C44.2	8097/3-C44.1*	8246/3-C11.9*	8742/3-C76.0	9687/3-C76.2*
8041/3-C76.3*	8070/6-C44.5	8097/3-C44.2*	8246/3-C49.5	8743/3-C49.1	9690/3-C76.2*
8070/3-C38.1*	8070/6-C44.6	8120/3-C76.3*	8246/3-C71.9	8772/3-C49.0	9930/3-C62.9*
8070/3-C38.3*	8070/6-C44.7	8140/3-C38.3*	8246/3-C76.2*	8800/3-C06.9	

** indicates codes that are already planned for inclusion in the next ICD-O-3 vocabulary update.*