Collaboration across Disciplines for the Health of People, Animals and Ecosystems
L. Stoicu-Tivadar et al. (Eds.)

doi:10.3233/SHTI241086

Suitability of the OMOP Common Data Model for Mapping Datasets of Medical Research Studies Using the Example of a Multicenter Registry

Milla KURTZ[a,1], Alfred WINTER[a] and Matthias LÖBE[a]
[a] *Institute for Medical Informatics, Statistics and Epidemiology (IMISE), University of Leipzig, Leipzig, Germany*
ORCiD ID: Milla Kurtz https://orcid.org/0009-0008-2696-5659,
Alfred Winter 0000-0003-0179-954X,
Matthias Löbe 0000-0002-2344-0426

Abstract. Common Data Models (CDM) are developed to solve integration problems that arise in the secondary use of health data. The OMOP CDM is such a model that is mainly used for healthcare data, so this paper examines whether it is also suitable for mapping research data. An exemplary research dataset is mapped to the model and the model is tested for suitability. For this purpose, an ETL process is first designed with the OHDSI tools and finally implemented with Talend Open Studio for Data Integration. The data quality is checked, and the mapping and the model, together with the tools, are evaluated. Overall, all but three data fields from the source dataset could be mapped to the OMOP model, so that the model's suitability for research data can be confirmed.

Keywords. Common Data Model, Data Accuracy, OMOP, Registry

1. Introduction

Common Data Models (CDM) are developed to solve integration problems that arise in the secondary use of health data. The OMOP (Observational Medical Outcome Partnership) CDM is one such model that serves as a mechanism for standardizing health data structure, content and semantics [1].

OHDSI (Observational Health Data Science and Informatics) develops and maintains the model. The feasibility of mapping German healthcare data to the OMOP CDM has already been demonstrated in previous work and was significantly advanced by OHDSI working groups [2,3,4,5,6]. So far, however, there is less experience with mapping research data to the model. The following work is, therefore, intended to show whether an exemplary mapping of a disease registry to the OMOP CDM is possible. The mapped registry data could ultimately be analysed quickly and efficiently internationally and across locations with the additional use of healthcare data.

[1] Corresponding Author: Milla Kurtz, Institute for Medical Informatics, Statistics and Epidemiology at Universität Leipzig, Germany; E-mail: mk18zufa@studserv.uni-leipzig.de.

2. Methods

Figure 1 gives a visual overview of the methods used in this work.

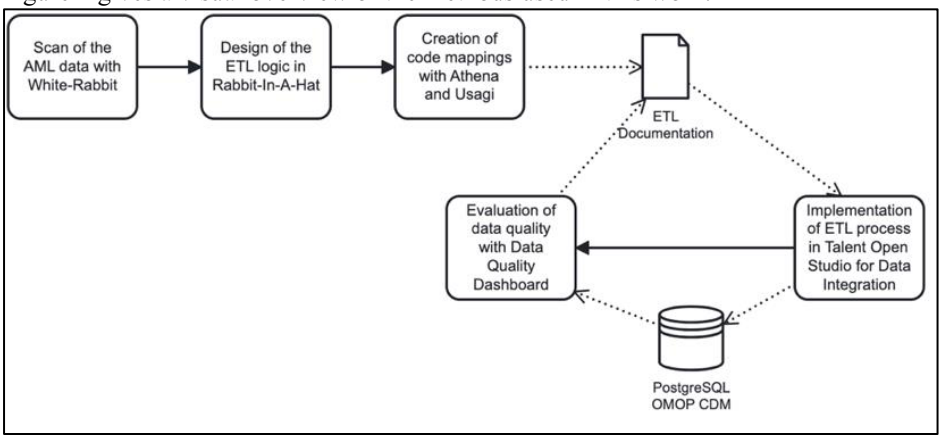

Figure 1. Schematic illustrations of used methods

2.1. Data set

An AML (acute myeloid leukemia) registry serves as the source data set for the study. The registry study consists of six events documented using eight electronic Case Report Forms (eCRF). The documentation report includes study inclusion, diagnosis, treatment cycles, recurrence, follow-ups, and treatment discontinuation. The events contained the following reports: Inclusion report, pre-therapy report, comorbidity report, biomaterial report, therapy report, recurrence report, follow-up report, and discontinuation report. Cycles can occur between therapy and recurrence; the events are read longitudinally. Therefore, the example data set is sufficiently complicated to show whether the OMOP CDM can generally be used to map research data. Accordingly, mappings had to be created for the eight different eCRFs of the registry study to the OMOP CDM.

2.2. Design of the ETL process

An extraction, transformation and loading (ETL) process was designed to map the eCRFs to the OMOP CDM. The OHDSI tool White-Rabbit was used to scan the source data and generate a report, which was read into Rabbit-In-A-Hat to create a logic for the mapping [7]. In the end, Rabbit-In-A-Hat generates documentation for the ETL process, but there is intentionally no ETL code.

2.3. Content mapping

The author then created the concept code mappings. The terminology repository Athena [8] was used to search for appropriate concepts, and the OHDSI-tool Usagi [9] was used for classifications to simplify content mapping.

2.4. Implementation of the ETL process

The ETL process was finally implemented in Talend Open Studio for Data Integration, and the data was loaded into a PostgreSQL database.

2.5. Evaluation of the mapping

The quality of the mapped data was checked with OHDSI's Data Quality Dashboard [10], and the mapping was evaluated quantitatively.

3. Results

3.1. Results of the structural mapping

It was possible to fill 14 OMOP tables and 51.9% (92 from 177) of the fields in the model. All mandatory fields were filled, except for 'Race' and 'Ethnicity', as the source data contained no information on these fields. Both were mapped to 0 ('no matching concept'). The OBSERVATION table was used for non-clinical observations, and 'event-supplementary' fields were used for some source fields. Three data fields could not be mapped for data protection reasons: First and last name and the clinic's internal ID.

3.2. Results of the content mapping

61,6% of the values could be assigned to a unique concept, for 3% of the values' uphill mapping' was performed, and 11,5% of the values had to be mapped to 0 ('no matching concept'). 3,9% of the values could not be mapped either since the non-existence of an event cannot be stored in OMOP. As the manual mapping for free texts, such as comorbidities, was too time-consuming, these were also mapped with 0. In some cases, fields were character-limited, meaning that character strings had to be shortened.

3.3. Qualitative results

DATA QUALITY ASSESSMENT

AML-REGISTER

DataQualityDashboard Version: 2.4.0
Results generated at 2023-08-16 12:05:27 in 2 mins

	Verification				Validation				Total			
	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass
Plausibility	2167	4	2171	100%	287	0	287	100%	2454	4	2458	100%
Conformance	700	28	728	96%	106	0	106	100%	806	28	834	97%
Completeness	370	26	396	93%	17	0	17	100%	387	26	413	94%
Total	3237	58	3295	98%	410	0	410	100%	3647	58	3705	**98%**

2909 out of 3647 passed checks are Not Applicable, due to empty tables or fields.
40 out of 58 failed checks are SQL errors.
Corrected pass percentage for NA and Errors: 98% (738/756).

Figure 2. Results of the data quality checks from the Data Quality Dashboard

An overall data quality metric, defined by the Data Quality Dashboard, of 98% was achieved, as shown in Figure 2. It also presented 'fails' in the mapping, caused mainly by the author's lack of medical knowledge. For plausibility, for example, a 'fail' occurred because the concept of blast measurement does not match the unit. This error was caused by the author's lack of medical knowledge but can be corrected.

4. Discussion

4.1. General experience

It was shown that it is possible to map most of the data from the AML registry to the OMOP CDM. The tools provided by OHDSI supported the mapping process and checked the quality of the generated CDM logic.

The mapping of the source dataset was possible to a limited extent. All data fields could be mapped except for three fields containing personal identifying information. Since not all research datasets contain so little personal data, a loss of information must always be expected when mapping research data, as the OMOP CDM for personal data is limited [6]. The OBSERVATION table was 'misused' for non-clinical observations like mobility from the G8-Screening, but OHDSI members suggested this and thus seems appropriate as a solution [11]. The mapping was incomplete because OMOP lacks vocabularies, and some concepts, such as karyotype, could not be mapped to standard concepts for some genetic alterations. The vocabulary should be expanded to include terminology for such less common concepts. Nevertheless, most of the information could be mapped, and the next step would be to analyze the data using the OHDSI tools.

4.2. Problems

One problem largely responsible for the loss of information is the inability to store events in the OMOP CDM that did not occur. Especially research data that collect information through questionnaires often contain such information, so this problem has been recognized before [12]. This problem can be prevented if the OMOP CDM is expanded to include the possibility of storing non-occurring events.

The content mapping of free text fields, such as the comorbidities of the comorbidity report, was not carried out as this was not feasible for resource reasons. Using free text fields sparingly when collecting medical data and storing expected values for diagnoses, medications, and other information with a classification code makes sense. Thus, standards for data collection can facilitate the reuse of data elements, such as in the OMOP CDM, interoperability, and the conduct of studies [13].

The character limitation of some fields in the OMOP CDM also led to a loss of information, which meant, for example, that the comorbidities had to be shortened.

The final illustrations could not be checked for correctness due to the author's lack of medical knowledge. Therefore, there is the risk of imprecision in the mapping and wrong interpretation. Medical experts and terminology specialists should work together for productively used mapping projects. The OHDSI tools supported the ETL design process, but OHDSI decided not to develop an ETL tool, so the mapping process requires technical knowledge.

4.3. Further Steps

It would be advisable to query the data from the AML register and other sources in aggregated form via the OMOP CDM in the future in cooperation with clinicians to provide evidence of the usability of the mapped register data. In addition, further steps should be taken to check whether the format of the OMOP CDM is suitable for typical analyses of registry data and whether the OHDSI tool palette can be used.

Acknowledgements

The work was part of the bachelor thesis of MK. Research was supported by DFG grant NFDI4Health (442326535) and DFG grant WI 1605/10-2.

References

[1] OHDSI. The Book of OHDSI: Observational Health Data Sciences and Informatics. 2021.
[2] OHDSI Europe: OHDSI Europe [Internet]. (n.d.) [updated (n.d.); cited 2023 Aug 08]. Available from: https://www.ohdsi-europe.org/index.php/info
[3] OHDSI Germany: OHDSI Germany [Internet]. (n.d.) [updated (n.d.); cited 2023 Aug 23]. Available from: https://www.uniklinikum-dresden.de/de/das-klinikum/universitaetscentren/zentrum-fuer-medizinische-informatik/zentrum/professur-fuer-medizinische-informatik-1/ohdsi-germany
[4] Maier C, Lang L, Storf H, Vormstein P, Bieber R, Bernarding J, Hermann T, Haverkamp C: Towards implementation of OMOP in a German university hospital consortium. Appied Clinical Informatics. 2018 Jan;0(1):54-61.
[5] Lang LJ. Mapping eines deutschen, klinischen Datensatzes nach OMOP Common Data Model [Dissertation]. [Erlangen-Nürnberg]: Friedrich-Alexander-Universität Erlangen-Nürnberg; 2020, 144 p.
[6] Peng Y, Henke E, Reinecke I, Zoch M, Sedlmayr M, Bathelt F: An ETL-process design for data harmonization to participate in international research with German real-world data based on FHIR and OMOP CDM. International Journal of Medical Informatics. 2023 Jan; 169:104925.
[7] OHDSI: OHDSI White Rabbit [Internet]. (n.d.) [updated (n.d.); cited 2023 Aug 23]. Available from: http://ohdsi.github.io/WhiteRabbit/
[8] Athena: Athena [Internet]. (n.d.) [updated (n.d.); cited 2023 August 23]. Available from: https://athena.ohdsi.org/search-terms/start
[9] OHDSI Usagi: OHDSI Usagi [Internet]. (n.d.) [updated (n.d.); cited 2023 August 2023], Available from: http://ohdsi.github.io/Usagi/
[10] OHDSI: OHDSI Data Quality Dashboard [Internet]. (n.d.) [updated (n.d.); cited 2023 Aug 23]. Available from: https://github.com/OHDSI/DataQualityDashboard
[11] Blacketer M, Voss E, Ryan P: Applying the OMOP Common Data Model to Survey Data [Internet]. 2015 [cited 2023 Aug 23]. 1 p. Available from: https://www.ohdsi.org/web/wiki/lib/exe/fetch.php?media=resources:using_the_omop_cdm_with_survey_and_registry_data_v6.0.pdf
[12] Biedermann P, Ong R, Davydov A, Orlova A, Solovyev O, Sun H, Wetherill G, Brand M: Standardizing registry data to the OMOP Common Data Model: Experience from three pulmonary hypertension databases. BMC Medical Research Methodology. 2021 Nov 2;21(1):238.
[13] Richesson R, Nadkarni P: Data Standards for Clinical Research Data Collection Forms: Current status and challenges. Journal of the American Medical Informatics Association. 2011 May-Jun; 18(3): 341-346.