Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230053

German Claims Data for Real-World Research: Content Coverage Evaluation in OMOP CDM

Elisa HENKE[a,1], Michéle ZOCH[a], Ines REINECKE[a], Melissa SPODEN[b], Thomas RUHNKE[b], Christian GÜNSTER[b], Martin SEDLMAYR[a], and Franziska BATHELT[a]
[a] *Institute for Medical Informatics and Biometry, Carl Gustav Carus Faculty of Medicine, Technische Universität Dresden, Dresden, Germany*
[b] *AOK Research Institute, WIdO, Berlin, Germany*

Abstract. Research on real-world data is becoming increasingly important. The current restriction to clinical data in Germany limits the view of the patient. To gain comprehensive insights, claims data can be added to the existing knowledge. However, standardized transfer of German claims data into OMOP CDM is currently not possible. In this paper, we conducted an evaluation regarding the coverage of source vocabularies and data elements of German claims data in OMOP CDM. We point out the need to extend vocabularies and mappings to support research on German claims data.

Keywords. OMOP CDM, OHDSI, interoperability, claims data

1. Introduction

Research based on real-world data is becoming increasingly important to gain new insights for personalized diagnoses and treatments. In this context, the German Federal Ministry of Education and Research (BMBF) has extensively funded the Medical Informatics Initiative [1] to provide digital infrastructures for the integration and harmonization of health data. This is currently limited to university hospitals and their patient data. However, this limits the view of patients to the time of hospitalization. This only relates to a small percentage of patients. The far greater number of patients and treatments take place outside the hospital [2]. To get a comprehensive view of the patient's journey, clinical data must be combined with outpatient data. An important data source for outpatient data are health insurance claims data. Although claims data and clinical data differ in structure and periodicity, they can be linked using the patients' unique health insurance number. To provide a complete picture of a patient, the two different datasets have to be merged into a common data model, such as the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM)

[1] Corresponding Author: Elisa Henke, Institute for Medical Informatics and Biometry, Carl Gustav Carus Faculty of Medicine, Technische Universität Dresden, Dresden, Germany; E-mail: Elisa.Henke@tu-dresden.de.

[3,4]. OMOP CDM comes with internationally standardized terminologies and a wide range of tools that enable statistical analyzes and machine learning. While many ways exist to import clinical data into OMOP CDM [5,6], there is no way yet for German claims data. However, the conformance is necessary for using the available analysis tools and the international comparability. To gain initial insights into the current content coverage of German claims data in OMOP CDM, we conducted an evaluation of the extent to which source vocabularies and data elements can be mapped to OMOP CDM.

2. Methods

To ensure successful harmonization of German claims data in OMOP CDM, we focused the two central components of OMOP CDM: the OHDSI standardized vocabularies and the data tables. Transforming German claims data to OMOP CDM requires the standardization of national vocabularies and the mapping of source data elements into OMOP CDM data tables. To examine the current content coverage of German claims data vocabularies and data elements in OMOP CDM, we focused on the following preliminary considerations (Figure 1). We 1) first performed a data profiling to get an overview of German claims data, especially their structure, content and references between data elements. Then 2) we compared German claims data and clinical data to check similarities that allow reuse of existing mappings. Finally, 3) a feasibility analysis of mapping German claims data to OMOP CDM was done to identify any current obstacles that prevent the mapping.

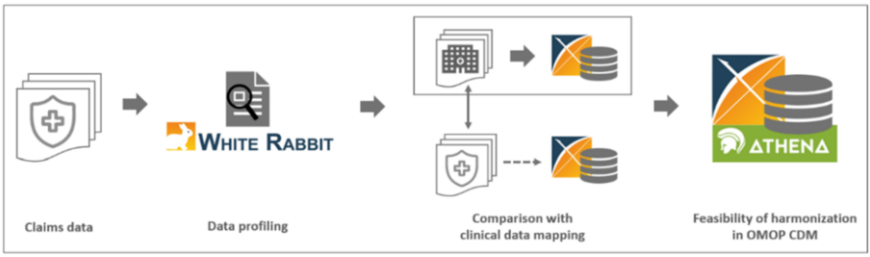

Figure 1. Concept on content coverage evaluation of German claims data in OMOP CDM (own illustration)

2.1. Data profiling

In Germany, healthcare providers and institutes are legally bound to transfer claims data to health insurance companies for billing purposes using a national standardized format. To gain a detailed insight into German claims data, we used synthetic claims data from the German local health care funds (Allgemeine Ortskrankenkassen, AOK), comprising data from 10.000 patients. This data set included demographic data, inpatient and outpatient hospital visits, inpatient rehabilitation visits, contract-medical care, outpatient drug prescriptions, therapeutic services, and care data. For the purpose of data profiling we used the open-source OHDSI tool WhiteRabbit[2] [7]. WhiteRabbit analyzes the provided source data and automatically generates a scan report including detailed information about tables, their data elements and data types as well as frequency

[2] https://github.com/OHDSI/WhiteRabbit

distributions of source values. The scan report can be used as a starting point for further comparison to German clinical data and feasibility analysis.

2.2. Comparison with clinical data mapping

As stated by Henke et al. [8] there is an intersection between German claims data and clinical data with respect to demographic, visit, procedure and diagnosis data from inpatient hospital care. Overlapping data elements are then focused on to assess mapping similarities between German claims data and clinical data. For the mapping of clinical data, we used the documentation of the Extract-Transform-Load (ETL) processes implemented by Peng et al. [5] and Zoch et al. [6] that focuses on mapping clinical data from German university hospitals to OMOP CDM [5] and mapping German claims data to OMOP CDM restricted to inpatient billing data [6], respectively.

2.3. Feasibility of harmonizing claims data in OMOP CDM

The final step of our concept took the results from the data profiling and the comparison with clinical data mapping into account to analyze the feasibility of harmonizing German claims data in OMOP CDM v5.3.1. First, we identified all vocabularies used in German claims data. Next, we evaluated the presence of required vocabularies in the OHDSI vocabulary web application ATHENA [9] to map the source data. Moreover, we examined the ETL processes for existing source code mappings to Standard OMOP Concepts through the source_to_concept_map table in OMOP CDM (interim mappings). To measure the current vocabulary coverage of German claims data in OMOP CDM, we categorized the source vocabularies into “available in ATHENA”, “available through interim mapping” and “not available”. After considering the vocabularies, we evaluated the feasibility of mapping source data elements to the standardized OMOP CDM fields. As part of this, the influence of any missing vocabularies on the mapping of data elements was included. For the measurement of the mapping coverage, we distinguished between the categories “possible”, “missing vocabularies”, “missing OMOP CDM fields” and “missing vocabularies and OMOP CDM fields”.

3. Results

3.1. Vocabulary

Based on the categories assigned to the source vocabularies, we calculated the vocabulary coverage in OMOP CDM in percentage (see supplementary file[3] for details). Figure 2 shows that only 15% of the source vocabularies are available in OMOP CDM but 55% are not. Missing vocabularies mainly concern vocabularies for therapeutic services and their indications, billing vocabularies such as the German Uniform Assessment Standard (EBM[4]) or services provided by psychiatric institutional outpatient

[3] https://caruscloud.uniklinikum-dresden.de/index.php/s/cs4HCB7LkJyjaFK

[4] Einheitlicher Bewertungsmaßstab

clinics (PIA[5]). Nevertheless, the remaining 30% of the source vocabularies can be mapped to standard OMOP CDM vocabularies through interim mappings.

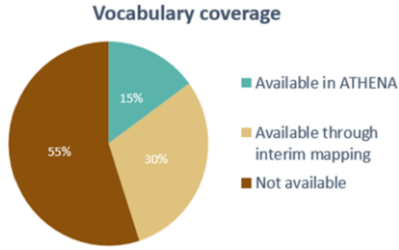

Figure 2. Vocabulary coverage of German claims data in OMOP CDM

3.2. Mapping

In order to gain knowledge about the current feasibility of mapping German claims data to OMOP CDM, we calculated a mapping coverage based on the flags mentioned in Section 2.3. As shown in Figure 3, 87% of the data elements can already be mapped to OMOP CDM. Mapping is currently not possible for 13% of the data elements. Reasons for this are missing vocabularies (8%), missing OMOP CDM fields (3%) or both, missing vocabularies and OMOP CDM fields (2%). In this context, we identified that the cost table in OMOP CDM is solely constructed for US specific hospital charges leaving no possibilities to map German charge data.

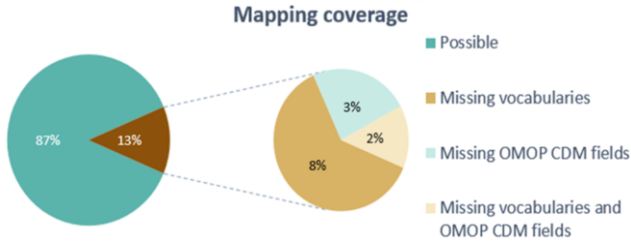

Figure 3. Mapping coverage of German claims data in OMOP CDM

4. Discussion and Conclusions

The results show a high mapping coverage of German claims data in OMOP CDM. However, we identified some obstacles regarding missing vocabularies and OMOP CDM fields, resulting in a mapping coverage lower than 90%. When reviewing the literature, our results are comparable to findings in other countries [10–14]. During our next steps, we address the current obstacles to prevent data loss during the transformation of German claims data to OMOP CDM. First, we focus on the preparation of missing vocabularies and their integration in the standardized vocabulary of OMOP CDM. Afterwards, we develop an approach to handle missing OMOP CDM fields and implement an ETL process to transform German claims data to OMOP CDM including

[5] Psychiatrische Institutsambulanzen

a qualitative control mechanism for the prepared vocabularies and the semantic mapping of source data elements to OMOP CDM. Consequently, we are building the basis for making German claims data available for international research as well as for the linkage with clinical data in OMOP CDM.

Declaration

Conflict of Interest: The authors declare that there is no conflict of interest.

Author contributions: All authors contributed substantial ideas and participated in editing and revising of the manuscript. All authors approved the manuscript in the submitted version and take responsibility for the scientific integrity of the work.

Funding: The study was funded by the German Innovations Fund of the Federal Joint Committee in Germany (G-BA) (grant number: 01VSF20013).

References

[1] Semler SC, Wissing F, Heyder R, German Medical Informatics Initiative. Methods Inf. Med. 2018;57:e50-6.
[2] Green LA, Fryer Jr GE, Yawn BP, Lanier D, Dovey SM. The ecology of medical care revisited. New England Journal of Medicine. 2001 Jun 28;344(26):2021-5. doi:10.1056/NEJM200106283442611.
[3] Garza M, Del Fiol G, Tenenbaum J, Walden A, Zozus MN. Evaluating common data models for use with a longitudinal community registry. Journal of biomedical informatics. 2016 Dec 1;64:333-41. doi:10.1016/j.jbi.2016.10.016.
[4] Data Standardization – OHDSI, (n.d.). https://www.ohdsi.org/data-standardization/ (accessed November 7, 2022).
[5] 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 1;169:104925.
[6] M. Zoch, E. Henke, I. Reinecke, Y. Peng, R. Gebler, M. Gruhl, and M. Sedlmayr, Extract, Transform and Load German Claim Data to OMOP CDM – Design and Implications, in: German Medical Science GMS Publishing House, 2022: p. DocAbstr. 153. doi:10.3205/22gmds057.
[7] White Rabbit, (n.d.). http://ohdsi.github.io/WhiteRabbit/WhiteRabbit.html (accessed November 11, 2022).
[8] E. Henke, M. Zoch, F. Bathelt, M. Spoden, T. Ruhnke, C. Günster, and M. Sedlmayr, Identifikation von Gemeinsamkeiten klinischer Daten und GKV-Routinedaten für deren Verknüpfung in OMOP CDM zur stationären Qualitätsmessung, in: German Medical Science GMS Publishing House, 2022: p. DocAbstr. 85. doi:10.3205/22gmds007.
[9] Athena, (n.d.). https://athena.ohdsi.org/search-terms/start (accessed November 25, 2022).
[10] Haberson A, Rinner C, Schöberl A, Gall W. Feasibility of mapping austrian health claims data to the OMOP common data model. Journal of Medical Systems. 2019 Oct;43:1-5. doi:10.1007/s10916-019-1436-9.
[11] Kim H, Choi J, Jang I, Quach J, Ohno-Machado L. Feasibility of representing data from published nursing research using the OMOP common data model. InAMIA Annual Symposium Proceedings 2016 (Vol. 2016, p. 715). American Medical Informatics Association..
[12] Cho S, Sin M, Tsapepas D, Dale LA, Husain SA, Mohan S, Natarajan K. Content coverage evaluation of the OMOP vocabulary on the transplant domain focusing on concepts relevant for kidney transplant outcomes analysis. Applied Clinical Informatics. 2020 Aug;11(04):650-8. 650–658. doi:10.1055/s-0040-1716528.
[13] Lamer A, Depas N, Doutreligne M, Parrot A, Verloop D, Defebvre MM, Ficheur G, Chazard E, Beuscart JB. Transforming French electronic health records into the Observational Medical Outcome Partnership's common data model: a feasibility study. Applied Clinical Informatics. 2020 Jan;11(01):013-22.. doi:10.1055/s-0039-3402754.
[14] Sathappan SM, Jeon YS, Dang TK, Lim SC, Shao YM, Tai ES, Feng M. Transformation of electronic health records and questionnaire data to OMOP CDM: a feasibility study using SG_T2DM Dataset. Applied Clinical Informatics. 2021 Aug;12(04):757-67. doi:10.1055/s-0041-1732301.