BMC Medical Informatics and Decision Making

RESEARCH Open Access

Conceptual design of a generic data harmonization process for OMOP common data model

Elisa Henke[1*], Michele Zoch[1], Yuan Peng[1], Ines Reinecke[2], Martin Sedlmayr[1] and Franziska Bathelt[3]

Abstract

Background To gain insight into the real-life care of patients in the healthcare system, data from hospital information systems and insurance systems are required. Consequently, linking clinical data with claims data is necessary. To ensure their syntactic and semantic interoperability, the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) from the Observational Health Data Sciences and Informatics (OHDSI) community was chosen. However, there is no detailed guide that would allow researchers to follow a generic process for data harmonization, i.e. the transformation of local source data into the standardized OMOP CDM format. Thus, the aim of this paper is to conceptualize a generic data harmonization process for OMOP CDM.

Methods For this purpose, we conducted a literature review focusing on publications that address the harmonization of clinical or claims data in OMOP CDM. Subsequently, the process steps used and their chronological order as well as applied OHDSI tools were extracted for each included publication. The results were then compared to derive a generic sequence of the process steps.

Results From 23 publications included, a generic data harmonization process for OMOP CDM was conceptualized, consisting of nine process steps: dataset specification, data profiling, vocabulary identification, coverage analysis of vocabularies, semantic mapping, structural mapping, extract-transform-load-process, qualitative and quantitative data quality analysis. Furthermore, we identified seven OHDSI tools which supported five of the process steps.

Conclusions The generic data harmonization process can be used as a step-by-step guide to assist other researchers in harmonizing source data in OMOP CDM.

Keywords OMOP, OHDSI, Interoperability, Data harmonization, Clinical data, Claims data

*Correspondence:
Elisa Henke
elisa.henke@tu-dresden.de
[1]Institute for Medical Informatics and Biometry, Carl Gustav Carus Faculty of Medicine, Technische Universität Dresden, 01307 Dresden, Germany
[2]Data Integration Center, Center for Medical Informatics, University Hospital Carl Gustav Carus Dresden, 01307 Dresden, Germany
[3]Thiem-Research GmbH, 03048 Cottbus, Germany

Background

The use of real-world data for research is becoming increasingly important in order to gain insights into the real-life care of patients in the healthcare system and, on this basis, to gain new knowledge for the diagnosis, treatment and prevention of diseases. To promote digitization in medicine for the areas of care and research in Germany, the Medical Informatics Initiative (MII) has been funded since 2018 by the German Federal Ministry for Education and Research (BMBF) [1]. The aim of the MII is to link data from patient care by providing digital infrastructures for the integration and harmonization of health data for research purposes. However, the developments of the infrastructures are currently focused on patient data of German university hospitals. Therefore, research in the MII is limited to clinical data from patients during hospitalization. The medical care of patients in university hospitals, in contrast, affects only a small percentage, since patients are usually hospitalized only when they already have a severe disease.

Green et al. [2] pointed out that a far greater number of patients are treated outside of the hospital. In comparison to inpatient data, outpatient data provide a more comprehensive overview of patients' medical histories. A relevant data source for outpatient data is claims data from the statutory health insurance funds in Germany. By linking claims data across institutions and sectors on a person-specific basis, longitudinal analyses of treatment histories can be realized. However, due to their billing focus, claims data lack depth of content, so that information on, for example, diagnostic and laboratory data is not included. In order to integrate both detailed information on the respective inpatient stay of patients as well as the insured person-related course perspective for research with real-world data, the combination of clinical data with claims data is necessary.

To exploit the potential of linking clinical data with claims data, it is important to ensure the syntactic and semantic interoperability of both data sets. Syntactic interoperability focuses on the definition of standardized data formats and information models, while semantic interoperability aims to achieve a uniform understanding of information models and terminology content across systems [3]. Achieving syntactic and semantic interoperability requires data harmonization, i.e. transforming local source data into a standardized format [4]. For the unified representation of heterogeneous data sets, so-called standardized common data models (CDMs) are developed. In the last years, the Observational Medical Outcomes Partnership (OMOP) CDM of the Observational Health Data Sciences and Informatics (OHDSI) gained significant relevance for research with real-world data [5–9].

The main challenge for researchers is the harmonization of national and institution-specific terminologies, formats and structures into the standardized format of OMOP CDM. For this purpose, OHDSI provides tools and introduces four major steps that should help to harmonize source data in OMOP CDM: Design the Extract-Transform-Load (ETL), Create the Code Mappings, Implement the ETL, Quality Control [10] (pp. 75–94). However, our own experience in harmonizing clinical data of the MII given in the Fast Healthcare Interoperability Resources (FHIR) format to OMOP CDM [11] has shown that these steps are not detailed enough. The literature demonstrates that many researchers are concerned with harmonization of source data in OMOP CDM [6]. Nevertheless, there is no detailed guidance that would allow researchers to follow a generic process when transforming source data to OMOP CDM, which is independent of type of source data used. A generic process is necessary to ensure the reusability of methods and tools as well as the reproducibility and comparability of results.

Prior to the practical harmonization of German claims data in OMOP CDM, we first investigated how such a generic process would look like in theory. Thus, the aim of this paper is to conceptualize a generic data harmonization process for OMOP CDM that is applicable for clinical data and claims data. In this context, we focus on the following three research questions:

1. Which process steps need to be performed when harmonizing clinical data or claims data in OMOP CDM?
2. What OHDSI tools were used by researchers to support the harmonization of clinical data or claims data in OMOP CDM?
3. What sequence of identified process steps should be followed?

Methods

Literature review

Paper identification

To obtain a clearer perspective of the state of the art of methodological processes for data harmonization in OMOP CDM, we conducted a literature review on August 3, 2023. Our literature search included publications published in English between 2018 and 2023, focusing on the harmonization of clinical data or claims data in OMOP CDM. Table 1 provides an overview of the search terms used in the literature databases PubMed and Web of Science.

The resulting publications were imported into the reference management software Zotero [12]. Afterwards, duplicates were removed using Zotero's built-in duplicate detection feature.

Table 1 Search terms used for the literature search in PubMed and Web of Science

Database	Search String
PubMed	((OMOP[Title/Abstract]) OR (OHDSI[Title/Abstract])) AND ((claims data[Title/Abstract]) OR (clinical data[Title/Abstract]))
Web of Science	((TI=(OMOP) OR TI=(OHDSI)) AND (TI=(claims data) OR TI=(clinical data))) OR ((AB=(OMOP) OR AB=(OHDSI)) AND (AB=(claims data) OR AB=(clinical data)))

Table 2 Definition of exclusion criteria following to Reinecke et al. [6]

Criterion	Description of criterion
no "OMOP" or "OHDSI"	Publication does not mention "OMOP" or "OHDSI"
	Publication uses "OMOP" or "OHDSI" with other meanings
mentioned	Publication only mentions "OMOP" or "OHDSI"
evaluated	Publication focuses on the evaluation of OMOP
vocabulary	Publication focuses on vocabularies and their mapping in OMOP or use of OMOP vocabularies
extension	Publication focuses on an extension of OMOP or OHDSI tools
usage	Publication focuses on the use of OMOP, e.g. for studies, data quality analyses, development of tools or frameworks (e.g. patient level prediction)
no full text	Publication is not available as full text
foreign language	Publication is written in other languages than English
wrong type of source data	Publication focuses on types of source data other than clinical data or claims data

Paper exclusion

The process of paper exclusion consisted of a Title-Abstract-Screening and Full-Text-Screening performed by three reviewers (EH, FB, MZ). For this purpose, nine exclusion criteria (Table 2) were defined to categorize the excluded publications.

Next, all three reviewers performed the Title-Abstract-Screening for 20% of the publications found. Conflicts were then discussed and resolved. Afterwards, we utilized the kappa statistic to test the interrater reliability [13]. For this purpose, we used the Fleiss method [13] in the KappaM function of the R library *DescTools* [14]. For the analysis, an error probability of 5% was set. Depending on the result of the kappa value, we chose one of the two defined options for the further procedure of paper exclusion:

- Option 1: For a kappa value greater than 0.6 (substantial to almost perfect agreement (interpretation according to Fleiss [13])), the Title-Abstract-Screening for the remaining 80% of the publications found and afterwards, the Full-Text-Screening for the included publications should be divided as follows: (1) reviewer 1 (EH) should screen all publications; (2) reviewer 2 (FB) all included publications and (3) reviewer 3 (MZ) all excluded publications.
- Option 2: If the kappa value is less than or equal to 0.6 (poor to moderate agreement (interpretation according to Fleiss [13])), the remaining 80% of the publications and the full texts had to be screened by all three reviewers.

After both, Title-Abstract-Screening and Full-Text-Screening, conflicts between the reviewers were discussed and resolved.

Data extraction

After the paper exclusion process, we focused on data extraction from the included publications. The data extraction was performed by reviewer 1 (EH) and subsequently verified by reviewer 2 (FB). The data extraction process consisted of three iterations. The first iteration focused on extracting the process steps and OHDSI tools from the publications used to harmonize the source data in OMOP CDM. In addition, the specification of the type of source data (clinical data and/or claims data) used for data harmonization and the countries from which the data was originated were documented for each publication.

In a second iteration, for each publication, we checked which of the extracted process steps and OHDSI tools were applied during the data harmonization in OMOP CDM. For this purpose, we created a matrix. The columns of the matrix represented the extracted process steps and OHDSI tools, while the rows represented the included publications. Within the matrix, we used crosses to indicate when a process step or OHDSI tool was mentioned in the corresponding publication.

In the final third iteration, the focus was on identifying the chronological order of the applied process steps per publication. For this purpose, we replaced the crosses of the process steps in the matrix with ascending numbers. Afterwards, the distribution of the given numberings per process step was calculated and the most frequent number was highlighted. This approach was performed for each of the publications that (a) used clinical data as

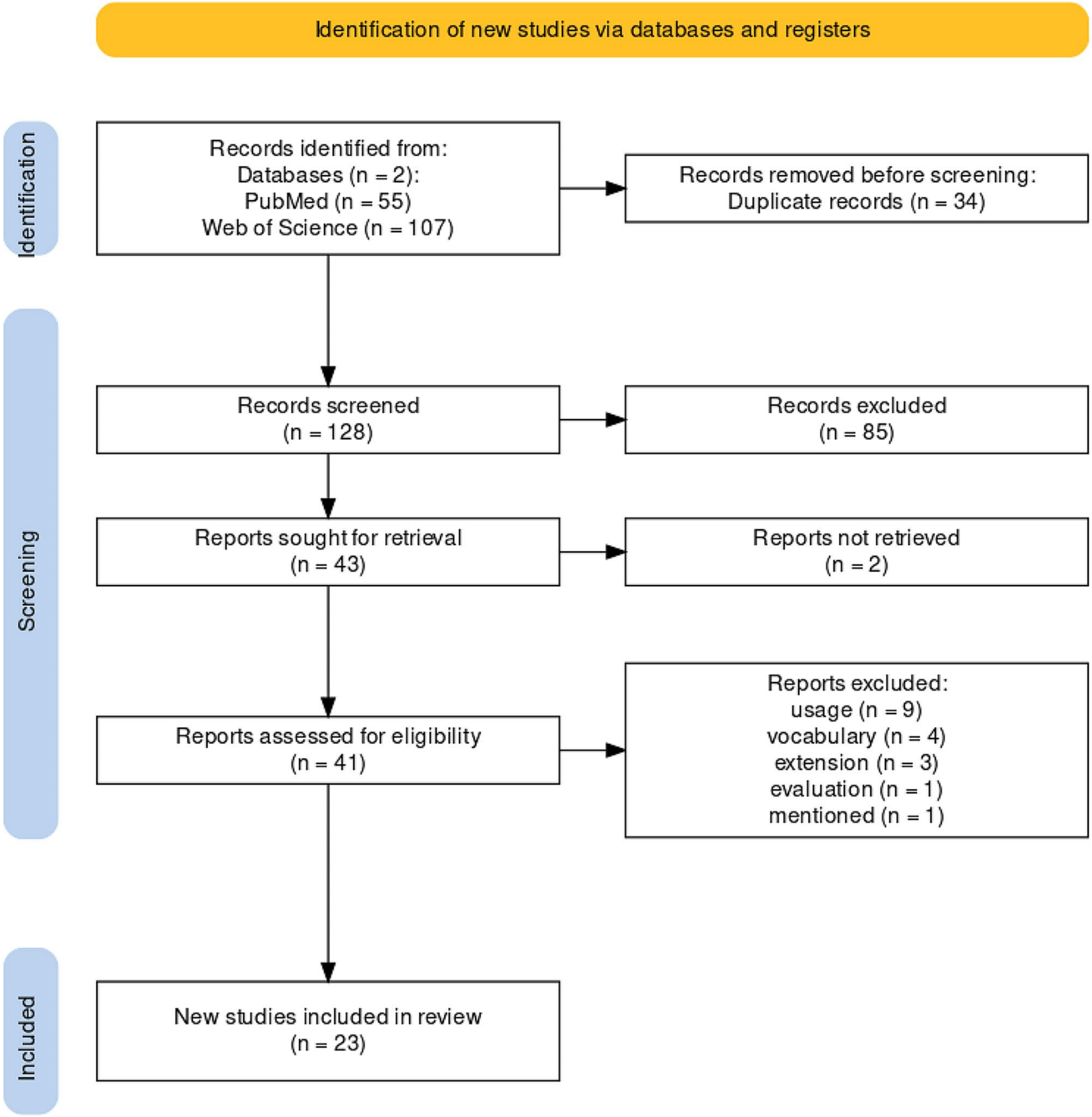

Fig. 1 PRISMA flow diagram according to [15]

source data or (b) used claims data as source data and (c) for all publications regardless of the type of source data. Publications that used both, clinical data and claims data, were categorized into all three groups.

Derivation of a generic sequence of process steps

In order to be able to apply the process steps to the harmonization of source data, it was necessary to establish a chronological classification of the process steps in an overall process. For this purpose, we compared the most frequent number(s) per process step. Our comparison started with the process step with the lowest number. Thereafter, we compared it to the number of the next process step. If the subsequent process step had a higher numbering (e.g. *1* vs. *2*), both process steps remained in their position. In the case of two identical numberings (e.g. *1*), the higher sum of the percentage of the most frequent numbering and the percentages of the preceding numberings decided the position in the comparison and all subsequent numberings were increased by the value *1*. This method of comparing predecessor and successor numbers was used until a unique numbering and thus positioning in the overall process could be defined. The derivation of the sequence of process steps was done separately for groups a)-c) defined in Sect. Data extraction. Finally, the results of the three groups were compared with regard to their agreement to derive a generic sequence for the data harmonization process.

Results

Flow diagram of the literature review

Based on the search terms, 162 publications (PubMed: 55, Web of Science: 107) were found. After removing duplicates in Zotero (34 publications), 128 publications remained for Title-Abstract-Screening. During the Title-Abstract-Screening, all three reviewers initially reviewed 20% (26/128) of the publications. Then we calculated the kappa value to check reviewer agreement. The kappa statistic resulted in a value of $\kappa=0.764$ (substantial agreement). Since this value was greater than 0.6, we chose Option 1 (see Sect. Paper exclusion) for the further paper exclusion process. After Title-Abstract-Screening, 85 publications were excluded according to the definition of the exclusion criteria in Table 2. In the subsequent screening of the full texts of 43 publications, we further excluded 20 publications. Finally, 23 publications were included for data extraction.

Figure 1 summarizes the process of the literature search, the subsequent screening of publications for inclusion and the remaining publications for the data extraction as PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) flow diagram [16]. A detailed overview of the results of the literature search, including the data extraction matrix can be found in the Additional file 1.

Process steps and OHDSI tools

The 23 reviewed publications allowed us to extract a methodological process for data harmonization in OMOP CDM [17–39]. Clinical data were used in 18 of 23 publications (78%) [17, 18, 20–29, 32–34, 36, 37, 39], claims data were used in 3 of 23 publications (13%) [30, 31, 35], and both types of data sources were used in 2 of 23 publications (9%) [19, 38] as source data for data harmonization in OMOP CDM. Clinical data originated from Belgium, Brazil, China, Denmark, Estonia, France, Germany, Italy, the Netherlands, Portugal, Serbia, Singapore, South Korea, Spain, Turkey, the United Kingdom and the United States. Claims data were used from Austria, France, and the United States (see Additional file 1).

The methodological process extracted from the publications consists of the nine process steps. Furthermore, we identified seven OHDSI tools which were used in the literature to support the harmonization of clinical data or claims data in OMOP CDM. In the following, the nine process steps are explained (in alphabetical order) and the seven OHDSI tools are assigned to them:

- In order to assess the extent to which the vocabularies found in the source data can already be mapped in OMOP CDM, a **coverage analysis of the vocabularies** of the source data is performed. The analysis helps to identify weaknesses (e.g. missing vocabularies) that would limit a full harmonization of the source data. Through this process step, Rinner et al. found that the Anatomical Therapeutic Chemical Classification and the International Statistical Classification of Diseases and Related Health Problems Tenth Revision existed in OMOP CDM, while the Austrian pharmaceutical registration number and a catalogue of medical services were missing [30].
- To get an overview of the source data including their structure, formats and unique values, a **data profiling** is performed. For this purpose, OHDSI provides the tool WhiteRabbit [40] to analyze the source data.
- **Dataset specification** refers to the definition of the scope of the source data for a specific use case. This is usually done by expert teams with clinical expertise. As a result, transformation to OMOP CDM is only performed on source data that is relevant to answering a specific research question.
- The technical transformation of the source data into OMOP CDM is realized through the implementation of **ETL-processes**. ETL-processes enable the reading of source data (Extract), the practical implementation of semantic and structural mapping (Transform), and the final writing of OMOP-compliant source data to the target database (Load).
- The **qualitative data quality analysis** examines, in particular, the plausibility, conformity and completeness of the source data in OMOP CDM (according to Kahn et al. [41]). With the Automated Characterization of Health Information at Large-scale Longitudinal Evidence Systems (Achilles) [42] and the Data Quality Dashboard (DQD) [43], two OHDSI tools exist that perform data quality checks on the transformed source data in OMOP CDM. The use of both OHDSI tools was described by Papez et al., who further investigated failed checks during multiple iterations and thus increased the qualitative data quality of their transformed data [23].
- The **quantitative data quality analysis** checks whether the number of data in the source matches the number of records in OMOP CDM. For this purpose, the OHDSI tool Atlas [44] can be used to define cohorts based on OMOP CDM. The number

of cohorts can then be compared with cohorts based on source data. Another example is provided by Yu et al., who compared record counts per variable, over time and null values for all OMOP CDM Table [20]. Haberson et al. focused on calculating and comparing descriptive statistical indicators (e.g. median value for age or number of hospitalizations per person) for the source and transformed data [31].

- **Semantic mapping** refers to the mapping of local vocabularies to the standardized vocabulary of OMOP CDM [10] (pp. 55–74). This step is necessary to be able to uniquely identify source values by concepts in OMOP CDM and to transfer source values to standard concepts to enable research in an international context. The standardized vocabulary of OMOP CDM is provided by the OHDSI vocabulary repository Athena [45]. Furthermore, the OHDSI tool Usagi [46] supports researchers in semantic mapping of source values to OMOP CDM concepts. The publication of Ji et al. describes the semantic mapping of the Korean Standard Classification of Diseases 7 to Systematized Nomenclature of Human and Veterinary Medicine Clinical Terms (SNOMED-CT) or the Korean Drug to RxNorm or RxNorm extension [27].
- The focus of the **structural mapping** is the conversion of the format of the source data into the standardized data model of OMOP CDM [10] (pp. 31–54). The structural mapping can be done by using the OHDSI tool RabbitInAHat [40].
- **Vocabulary identification** focuses on providing a comprehensive compilation of the vocabularies found in the source data, including their scope of application. For example, Papez et al. identified three vocabularies in UK primary care data (SNOMED-CT, Clinical Terms Version 3, Dictionary of Medicines and Devices), two vocabularies in hospital care data and mortality data (International Statistical Classification of Diseases and Related Health Problems Tenth Revision and Ninth Revision), one vocabulary for cancer registry data (International Classification of Diseases for Oncology, Third Edition) and two vocabularies in procedure data (OPCS Classification of Interventions and Procedures Version 3 and Version 4) [23].

Frequency

The frequency of the process step occurrence in the literature is shown in Fig. 2 in descending order. All publications included the implementation of ETL processes, semantic mapping and structural mapping. Furthermore, understanding the source data and mapping it to OMOP CDM was also important during data harmonization. This applies in particular to vocabulary identification (70%: 16/23), dataset specification (65%: 15/23), data profiling (39%: 9/23) and coverage analysis of vocabularies (35%: 8/23). About half of the publications dealt with qualitative (65%: 15/23) and quantitative data quality analyses (48%: 11/23) after successful transformation of the source data into OMOP CDM.

Furthermore, we also determined the quantitative occurrence of the seven identified OHDSI tools in the 23 included publications. The results showed that Athena and Achilles were used most frequently with 43% (10/23), followed by WhiteRabbit, RabbitInAHat and Usagi with 26% (6/23) each. The Data Quality Dashboard and Atlas were the least used in the literature with only 22% (5/23) each.

Chronological order

In order to define a generic sequence for the extracted process steps for data harmonization in OMOP CDM, the chronological order of the process steps was focused during the third iteration of data extraction. The chronological order of the process steps per publication can be found in the Additional file 1.

To calculate the percentage distribution of the given numberings per process step, we assigned the included publications to the three groups a)-c). This resulted in the following number of publications per group: group a): 20 publications, group b): five publications and group c): 23 publications. Figures 3, 4 and 5 represent the percentage distribution as well as the indication of the most frequent numbering(s) for the groups a)-c).

The three diagrams show that the numbers *1* to *9* can occur several times as the most frequent numbering per group (e.g. number *1* in Fig. 3). Therefore, it was not yet possible to obtain a clear positioning of the identified process steps in an overall data harmonization process for each group.

Generic data harmonization process

By comparing the most frequent number(s) per process step according to our approach in Sect. Derivation of a generic sequence of process steps, we were able to determine the chronological order of the data harmonization process for the groups a)-c). Table 3 summarizes the results separated by the type of source data. The results show, that there are similarities (highlighted using bold text) and deviations of the chronological order of the process steps between the three groups. All three groups match completely for process numbers *1* (dataset specification) and *7* (ETL-process). For the remaining process numbers, there are deviations between the three groups. This applies to the exchange of structural mapping and semantic mapping or quantitative and qualitative data quality analyses for clinical data, as well

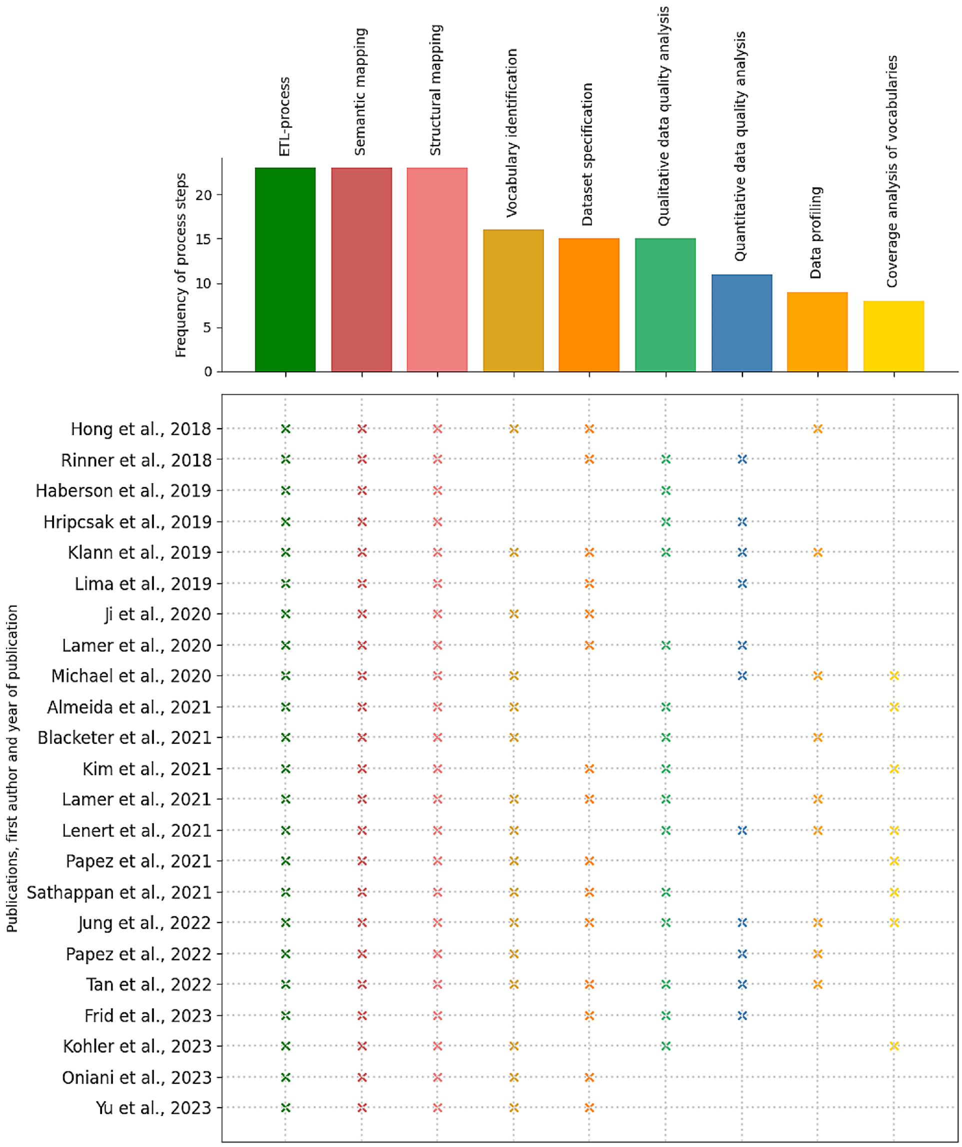

Fig. 2 Frequency distribution of the extracted process steps and their assignment to the included publications

as for the vocabulary identification, coverage analysis of vocabularies and data profiling for claims data. Nevertheless, two process steps per process number always match. This suggests that some process steps can also be interchanged with each other and the goal of harmonizing source data in OMOP has nevertheless been achieved by other researchers.

Finally, to conceptualize a generic process for data harmonization in OMOP CDM, the sequence of process steps that showed the most agreement was determined. Figure 6 represents the resulting chronological order for the generic data harmonization process for OMOP CDM.

The generic data harmonization process consists of nine process steps in the following order:

1. Dataset specification.
2. Data profiling.
3. Vocabulary identification.
4. Coverage analysis of vocabularies.
5. Semantic mapping.
6. Structural mapping.
7. ETL-process.
8. Qualitative data quality analysis.
9. Quantitative data quality analysis.

Discussion

The results show that we have achieved our goal of conceptualizing a generic data harmonization process for OMOP CDM. By conducting a literature review, we were able to answer our first research question. The literature

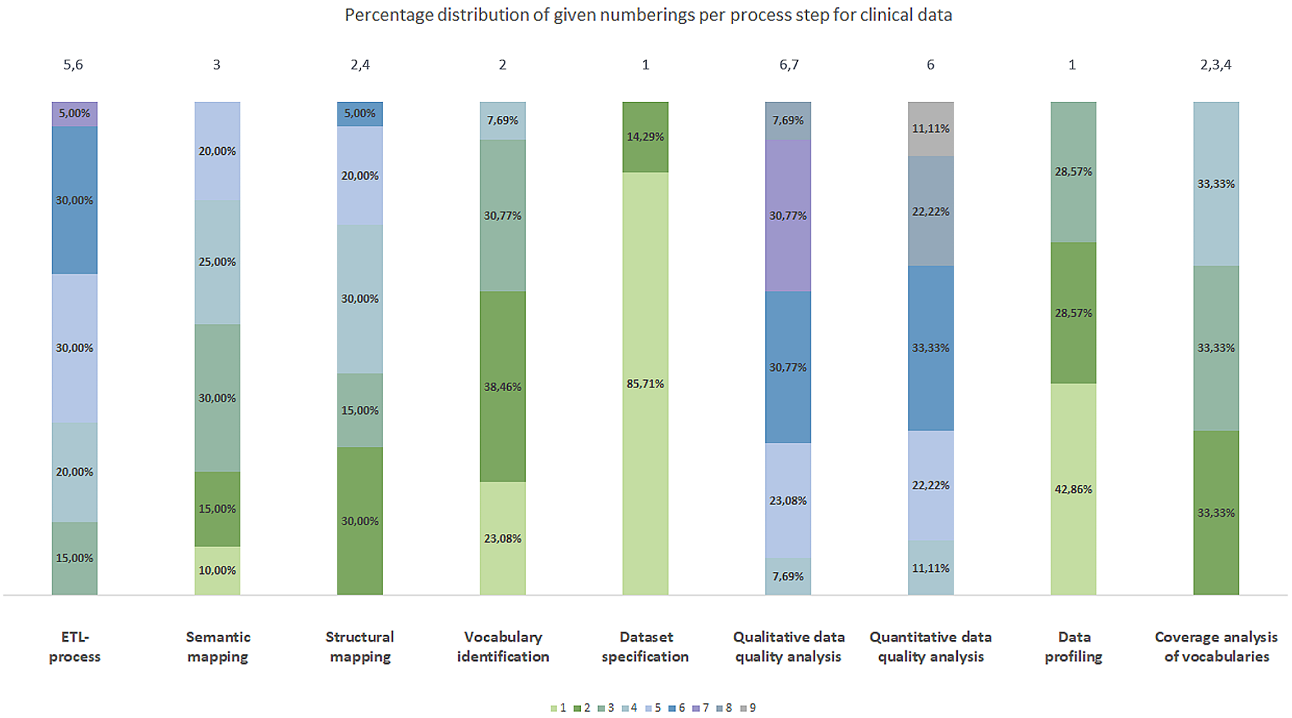

Fig. 3 Percentage distribution of given numberings per process step for clinical data (group a))

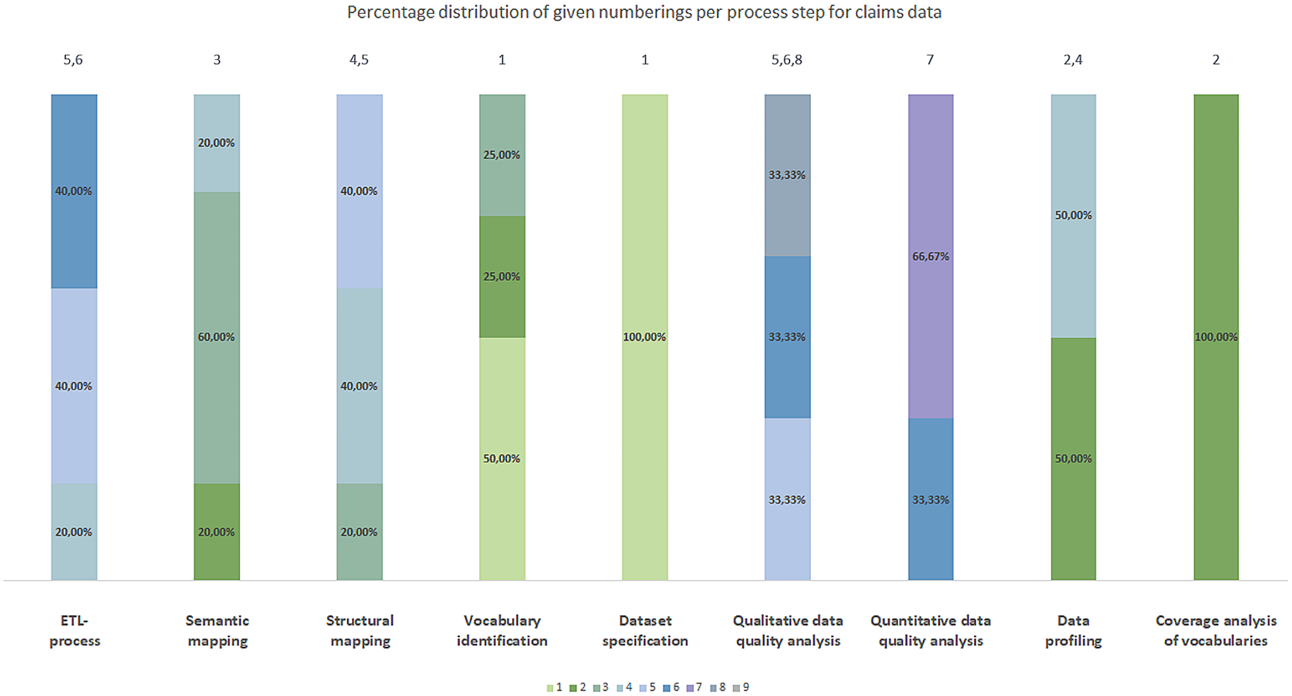

Fig. 4 Percentage distribution of given numberings per process step for claims data (group b))

review revealed nine process steps that were used by other researchers to harmonize clinical data and/or claims data in OMOP CDM. However, the results per publication show that not all steps have to be relevant. For example, Hong et al. [34], Hripcsak et al. [19], Lamer et al. [18], Lenert et al. [36] and Michael et al. [33] used only five from nine process steps. These five process steps also varied, as can be seen in the work of Hripcsak et al. [19] (dataset specification, vocabulary identification, semantic mapping, structural mapping, ETL-process)

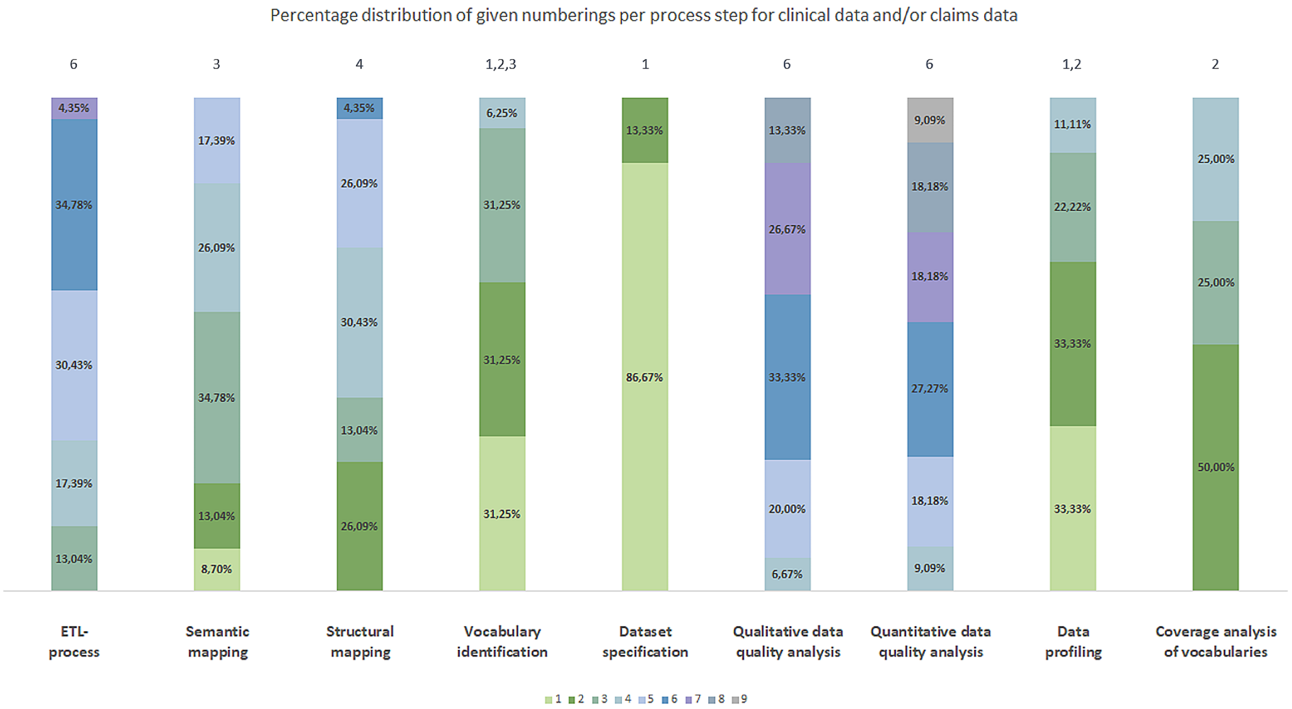

Fig. 5 Percentage distribution of given numberings per process step for clinical data and/or claims data (group c))

and Lenert et al. [36] (semantic mapping, structural mapping, ETL-process, quantitative and qualitative data quality analysis). Only the three process steps of semantic mapping, structural mapping and ETL-process are mentioned in all of the publications and thus form the mandatory part of the data harmonization process. Consequently, it is recommended to check individually for each data harmonization project whether process steps of the generic data harmonization process can be skipped if applicable.

Regarding our second research question, we identified seven OHDSI tools from the literature which were used to support the harmonization of clinical data or claims data in OMOP CDM. Overall, the results of the literature revealed that the OHDSI tools were not widely applied by researchers. This demonstrates that the use of OHDSI tools seems to be not mandatory.

Finally, our third research question was also addressed by looking at the chronological order of the identified process steps. A comparison of the results of the sequences of the three groups resulted in three different chronological orders of the nine process steps. This demonstrates that process steps can be interchanged if necessary while still achieving harmonization of source data in OMOP CDM (e.g. qualitative and quantitative data quality analysis). However, the interchange of structural mapping and semantic mapping for clinical data was surprising. Our experience during the harmonization of clinical data from FHIR format to OMOP CDM showed that in the majority of cases, it is not possible to do the structural mapping before the semantic mapping. The semantic mapping of source codes of free texts to concepts in the standardized vocabulary of OMOP CDM involves the assignment to a domain (e.g. Condition or Procedure). Domains specify the tables and fields in which source data should be transformed. Consequently, without this information a structural mapping was not possible. Nevertheless, there are some cases were an interchange of both process steps is plausible (e.g. exclusive use of demographic patient data).Despite the three different sequences, we were still able to derive a generic sequence of the nine process steps. Furthermore, it is important to mention that the generic data harmonization process should be considered as an iterative process. The last two steps of the qualitative and quantitative data quality analysis evaluate the correctness of the previous steps. If errors are identified, it may be necessary to perform the process again. In this context, it has to be checked individually whether all nine process steps have to be performed again or whether certain process steps do not contribute to the solution of the identified errors and can be skipped.

The present work is limited since only the harmonization of clinical data and claims data in OMOP CDM was focused. The literature review showed that there are also other types of data sources which are harmonized in OMOP CDM (e.g. registry data [47, 48]). In the future, it is therefore necessary to check whether the addition of other types of source data has an impact on the amount and sequence of the process steps. Furthermore, there

Table 3 Chronological order of process steps separated by type of source data

Process number/ Type of source data	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***
Clinical data	**Dataset specification**	**Data profiling**	**Vocabulary identification**	**Coverage analysis of vocabularies**	Structural mapping	Semantic mapping	**ETL-process**	Quantitative data quality analyses	Qualitative data quality analyses
Claims data	**Dataset specification**	Vocabulary identification	Coverage analysis of vocabularies	Data profiling	**Semantic mapping**	**Structural mapping**	**ETL-process**	**Qualitative data quality analyses**	**Quantitative data quality analyses**
Clinical data and/or claims data	**Dataset specification**	**Data profiling**	**Vocabulary identification**	**Coverage analysis of vocabularies**	**Semantic mapping**	**Structural mapping**	**ETL-process**	**Qualitative data quality analyses**	**Quantitative data quality analyses**

are limitations in the interpretation of the results from group b) (claims data). The reason for this is that only a small number of the included publications used claims data (5/23 publications), which limits a comparison with clinical data (20/23 publications) and a subsequent generalization of the results.

A second limitation relates to the CDM used. Our work was focused on the OMOP CDM. However, there are many other CDMs, such as Sentinel [49], Informatics for Integrating Biology & Bedside (i2b2) [50] or the National Patient-Centered Clinical Research Network (PCORnet) [51] that harmonize source data in a standardized format. We believe that our work is also relevant and useful for researchers using other CDMs as a target format. The extent to which the 9-step process for OMOP CDM can be applied to other CDMs would need to be evaluated as future work.

Notwithstanding the limitations listed above, our generic data harmonization process provides a major benefit. Compared to OHDSI's recommended 4-step process, our conceptualized 9-step process is more detailed. An analysis showed that our 9-step process can be assigned to the 4-step process of OHDSI as follows:

- OHDSI 1: Design the ETL.

  1: Dataset specification
  2: Data profiling
  3: Vocabulary identification
  6: Structural mapping

- OHDSI 2: Create the Code Mappings.

  4: Coverage analysis of vocabularies
  5: Semantic mapping

- OHDSI 3: Implement the ETL.

  7: ETL-process

- OHDSI 4: Quality Control.

  8: Qualitative data quality analysis
  9: Quantitative data quality analysis

The assignment demonstrated that all nine process steps extracted from the literature can be related to the OHDSI steps. However, we identified a difference in the chronological order of the nine assigned process steps. According to the sequence of the OHDSI steps, the structural mapping again would appear before the coverage analysis

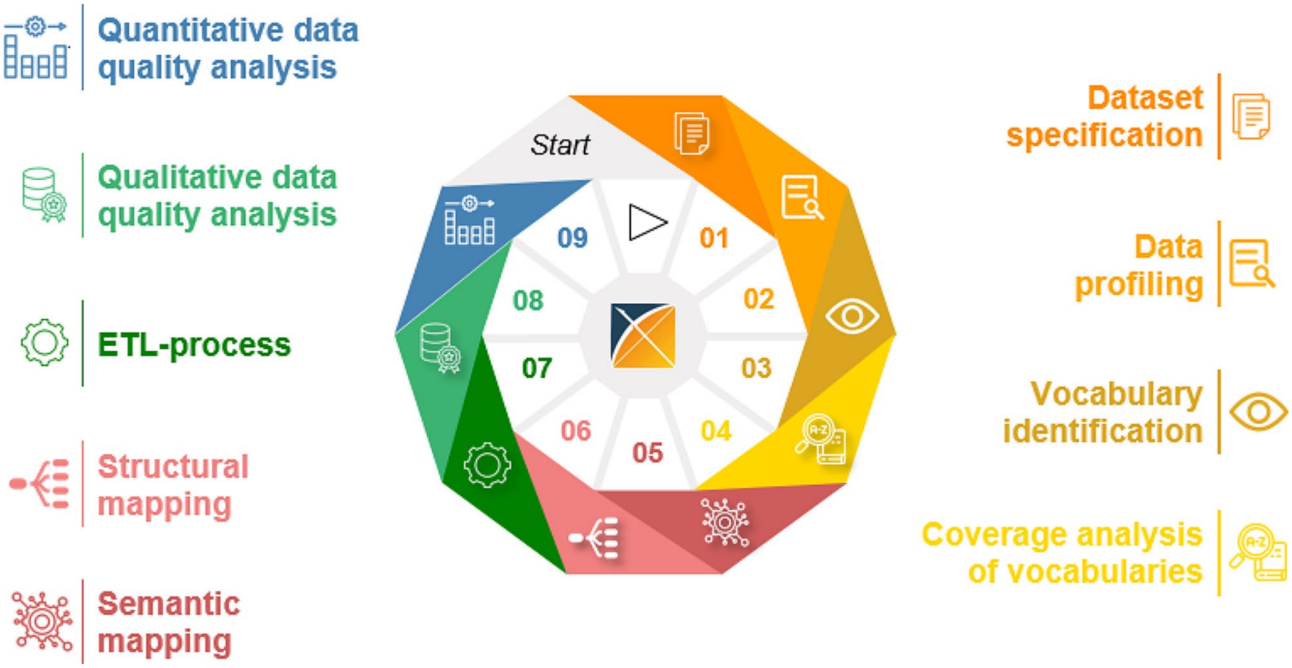
Quantitative data quality analysis
Qualitative data quality analysis
ETL-process
Structural mapping
Semantic mapping
Start
01
02
03
04
05
06
07
08
09
Dataset specification
Data profiling
Vocabulary identification
Coverage analysis of vocabularies

Fig. 6 Generic data harmonization process for OMOP CDM; icons: Flaticon.com

of vocabularies and the semantic mapping. The reason for this inconsistent order from OHDSI is that their process description [10] is focused on the use of the OHDSI tools (e.g. joint use of WhiteRabbit, RabbitInAHat) for data harmonization in OMOP CDM. In contrast, our process can be carried out independently of the OHDSI tools and used as a guide for other researchers to follow.

Conclusions

Based on a literature review, necessary process steps for harmonizing clinical data or claims data in OMOP CDM were identified and placed in a chronological order. From these findings, a generic data harmonization process was derived. This process can be used as a step-by-step guide to assist other researchers in harmonizing source data in OMOP CDM. As future work, the applicability of the generic data harmonization process and OHDSI tools to German claims data will be investigated in practice. In this context, an evaluation will show whether further additional process steps need to be considered and to what extent the derived sequence is feasible in practice. Additionally, we plan to expand the guide in future to include best practices for the practical implementation and overcoming of challenges per process step and recommendations for using the OHDSI tools in this context.

Abbreviations

Achilles	Automated Characterization of Health Information at Large-scale Longitudinal Evidence Systems
BMBF	German Federal Ministry for Education and Research
CDM	Common Data Model
DQD	Data Quality Dashboard
ETL	Extract-Transform-Load
FHIR	Fast Healthcare Interoperability Resources
i2b2	Informatics for Integrating Biology & Bedside
MII	Medical Informatics Initiative
OHDSI	Observational Health Data Sciences and Informatics
OMOP	Observational Medical Outcomes Partnership
PCORnet	National Patient-Centered Clinical Research Network
PRISMA	Preferred Reporting Items for Systematic reviews and Meta-Analyses
SNOMED-CT	Systematized Nomenclature of Human and Veterinary Medicine Clinical Terms

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-024-02458-7.

A detailed overview of the results of the literature search, including the data extraction matrix can be found in the Additional file 1.

Acknowledgements

Not applicable.

Author contributions

Conceptualization, E.H.; Data curation, E.H. and I.R.; Formal analysis, E.H., I.R. and F.B.; Investigation, E.H.; Methodology, E.H., M.Z. and F.B.; Project administration, E.H.; Resources, M.S.; Supervision, M.S.; Validation, M.Z. and F.B.; Visualization, E.H., Y.P. and I.R.; Writing– original draft, E.H.; Writing– review & editing, M.Z., Y.P., I.R., M.S. and F.B.

Funding

Open access funding enabled and organized by Projekt DEAL. No funding was received for conducting this study.
Open Access funding enabled and organized by Projekt DEAL.

Data availability

No datasets were generated or analysed during the current study.

Declarations

Competing interests
The authors declare no competing interests.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Received: 23 November 2023 / Accepted: 9 February 2024
Published online: 26 February 2024

References

1. Semler SC, Wissing F, Heyder R. German Medical Informatics Initiative - A National Approach To Integrating Health Data from Patient Care and Medical Research. Methods Inf Med. 2018;57(Suppl 1):e50–6.
2. Green LA, Fryer GE, Yawn BP, Lanier D, Dovey SM. The Ecology of Medical Care Revisited. N Engl J Med. 2001;344(26):2021–5.
3. Thun S, Dewenter H. Syntaktische und semantische Interoperabilität. In: Müller-Mielitz S, Lux T, editors. E-Health-Ökonomie [Internet]. Wiesbaden: Springer Fachmedien; 2017 [cited 2023 Mar 14]. p. 669–82. https://doi.org/10.1007/978-3-658-10788-8_34.
4. Kumar G, Basri S, Imam AA, Khowaja SA, Capretz LF, Balogun AO. Data harmonization for heterogeneous datasets: a systematic literature review. Appl Sci. 2021;11(17):8275.
5. Garza M, Del Fiol G, Tenenbaum J, Walden A, Zozus MN. Evaluating common data models for use with a longitudinal community registry. J Biomed Inf. 2016;64:333–41.
6. Reinecke I, Zoch M, Reich C, Sedlmayr M, Bathelt F. The usage of OHDSI OMOP - A Scoping Review. Stud Health Technol Inf. 2021;283:95–103.
7. EHDEN. European Health Data & Evidence Network [Internet]. 2022 [cited 2022 May 20]. Available from: https://www.ehden.eu/.
8. European Medical Agency. Data Analysis and Real World Interrogation Network (DARWIN EU) [Internet]. 2023 [cited 2022 May 20]. Available from: https://www.ema.europa.eu/en/about-us/how-we-work/big-data/data-analysis-real-world-interrogation-network-darwin-eu.
9. Observational Health Data Sciences and Informatics. HL7 International and OHDSI Announce Collaboration to Provide Single Common Data Model for Sharing Information in Clinical Care and Observational Research [Internet]. 2021 [cited 2022 May 6]. Available from: https://www.ohdsi.org/ohdsi-hl7-collaboration/.
10. Observational Health Data Sciences and Informatics. The Book of OHDSI. In: The Book of OHDSI [Internet]. 2021 [cited 2022 Apr 19]. Available from: https://ohdsi.github.io/TheBookOfOhdsi/.
11. 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. Int J Med Informatics. 2023;169:104925.
12. Digital Scholar. Zotero [Internet]. Zotero. 2023 [cited 2023 Mar 31]. Available from: https://www.zotero.org/.
13. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull. 1971;76(5):378–82.
14. Signorell A, Aho K, Alfons A, Anderegg N, Aragon T, Arachchige C et al. DescTools: Tools for Descriptive Statistics [Internet]. 2023 [cited 2023 Oct 25]. Available from: https://cran.r-project.org/web/packages/DescTools/index.html.
15. Haddaway NR, Page MJ, Pritchard CC, McGuinness LA. PRISMA2020: an R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and open synthesis. Campbell Syst Reviews. 2022;18(2):e1230.
16. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ. 2009;339:b2700.
17. Klann JG, Joss MAH, Embree K, Murphy SN. Data model harmonization for the all of Us Research Program: transforming i2b2 data into the OMOP common data model. PLoS ONE. 2019;14(2).
18. Lamer A, Abou-Arab O, Bourgeois A, Parrot A, Popoff B, Beuscart JB, et al. Development and Usability Study. J Med Internet Res. 2021;23(10):e29259. Transforming Anesthesia Data Into the Observational Medical Outcomes Partnership Common Data Model:.
19. Hripcsak G, Shang N, Peissig PL, Rasmussen LV, Liu C, Benoit B et al. Facilitating phenotype transfer using a common data model. J Biomed Inform. 2019;96.
20. Yu Y, Jiang G, Brandt E, Forsyth T, Dhruva SS, Zhang S et al. Integrating real-world data to assess cardiac ablation device outcomes in a multicenter study using the OMOP common data model for regulatory decisions: implementation and evaluation. JAMIA OPEN. 2023;6(1).
21. Papez V, Moinat M, Payralbe S, Asselbergs FW, Lumbers RT, Hemingway H et al. Transforming and evaluating electronic health record disease phenotyping algorithms using the OMOP common data model: a case study in heart failure. JAMIA OPEN. 2021;4(3).
22. Tan HX, Teo DCH, Lee D, Kim C, Neo JW, Sung C, et al. Healthc Inf Res. 2022;28(2):112–22. Applying the OMOP Common Data Model to Facilitate Benefit-Risk Assessments of Medicinal Products Using Real-World Data from Singapore and South Korea.
23. Papez V, Moinat M, Voss EA, Bazakou S, Van Winzum A, Peviani A, et al. Transforming and evaluating the UK Biobank to the OMOP Common Data Model for COVID-19 research and beyond. Volume 30. JOURNAL OF THE AMERICAN MEDICAL INFORMATICS ASSOCIATION; 2022. pp. 103–11. 1.
24. Jung H, Yoo S, Kim S, Heo E, Kim B, Lee HY, et al. Patient-level fall risk prediction using the Observational Medical Outcomes Partnership's Common Data Model. Volume 10. Pilot Feasibility Study. JMIR MEDICAL INFORMATICS; 2022. 3.
25. Almeida JR, Silva JF, Matos S, Oliveira JL. A two-stage workflow to extract and harmonize drug mentions from clinical notes into observational databases. J Biomed Inform. 2021;120.
26. Lima DM, Rodrigues-Jr JF, Traina AJM, Pires FA, Gutierrez MA. Transforming two decades of ePR Data to OMOP CDM for Clinical Research. In: OhnoMachado L, Seroussi B, editors. MEDINFO 2019: HEALTH AND WELLBEING E-NETWORKS FOR ALL. NIEUWE HEMWEG 6B, 1013 BG AMSTERDAM. Studies in Health Technology and Informatics. Volume 264. NETHERLANDS: IOS; 2019. pp. 233–7.
27. Ji H, Kim S, Yi S, Hwang H, Kim JW, Yoo S. Converting clinical document architecture documents to the common data model for incorporating health information exchange data in observational health studies: CDA to CDM. J Biomed Inform. 2020;107.
28. Kim JW, Kim S, Ryu B, Song W, Lee HY, Yoo S. Transforming electronic health record polysomnographic data into the Observational Medical Outcome Partnership's Common Data Model: a pilot feasibility study. Sci Rep. 2021;11(1).
29. Blacketer C, Voss EA, DeFalco F, Hughes N, Schuemie MJ, Moinat M, et al. Using the Data Quality Dashboard to improve the EHDEN Network. Appl SCIENCES-BASEL. 2021;11:24.
30. Rinner C, Gezgin D, Wendl C, Gall W. A Clinical Data Warehouse based on OMOP and i2b2 for Austrian Health Claims Data. In: Schreier G, Hayn D, editors. HEALTH INFORMATICS MEETS EHEALTH: BIOMEDICAL MEETS EHEALTH - FROM SENSORS TO DECISIONS. NIEUWE HEMWEG 6B, 1013 BG AMSTERDAM. Studies in Health Technology and Informatics. Volume 248. NETHERLANDS: IOS; 2018. pp. 94–9.
31. Haberson A, Rinner C, Schöberl A, Gall W. J Med Syst. 2019;43(10):314. Feasibility of Mapping Austrian Health Claims Data to the OMOP Common Data Model.
32. Sathappan SMK, Jeon YS, Dang TK, Lim SC, Shao YM, Tai ES, et al. Transformation of Electronic Health Records and Questionnaire Data to OMOP CDM: a feasibility study using SG_T2DM dataset. Appl Clin Inf. 2021;12(4):757–67.
33. Michael CL, Sholle ET, Wulff RT, Roboz GJ, Campion TR. Mapping Local Biospecimen Records to the OMOP Common Data Model. AMIA Jt Summits Transl Sci Proc. 2020;2020:422–9.
34. Hong N, Zhang N, Wu H, Lu S, Yu Y, Hou L, et al. Preliminary exploration of survival analysis using the OHDSI common data model: a case study of intrahepatic cholangiocarcinoma. Volume 18. BMC MEDICAL INFORMATICS AND DECISION MAKING; 2018. 5.
35. Lamer A, Depas N, Doutreligne M, Parrot A, Verloop D, Defebvre MM, et al. Transforming French Electronic Health Records into the Observational Medical Outcome Partnership's Common Data Model: a feasibility study. Appl Clin Inf. 2020;11(1):13–22.
36. Lenert LA, Ilatovskiy AV, Agnew J, Rudisill P, Jacobs J, Weatherston D, et al. Automated production of research data marts from a canonical fast

healthcare interoperability resource data repository: applications to COVID-19 research. J Am Med Inf Assoc. 2021;28(8):1605–11.

37. Kohler S, Boscá D, Kärcher F, Haarbrandt B, Prinz M, Marschollek M, et al. Eos and OMOCL: towards a seamless integration of openEHR records into the OMOP Common Data Model. J Biomed Inf. 2023;144:104437.
38. Oniani D, Parmanto B, Saptono A, Bove A, Freburger J, Visweswaran S, et al. ReDWINE: a clinical datamart with text analytical capabilities to facilitate rehabilitation research. Int J Med Inf. 2023;177:105144.
39. Frid S, Pastor Duran X, Bracons Cucó G, Pedrera-Jiménez M, Serrano-Balazote P, Muñoz Carrero A, et al. An ontology-based Approach for consolidating patient data standardized with European Norm/International Organization for standardization 13606 (EN/ISO 13606) into Joint Observational Medical Outcomes Partnership (OMOP) repositories: description of a methodology. JMIR Med Inf. 2023;11:e44547.
40. Observational Health Data Sciences and Informatics. White Rabbit [Internet]. 2022 [cited 2022 Nov 11]. Available from: http://ohdsi.github.io/WhiteRabbit/WhiteRabbit.html.
41. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. EGEMS (Wash DC). 2016;4(1):1244.
42. Observational Health Data Sciences and Informatics. Achilles [Internet]. Observational Health Data Sciences and Informatics; 2022 [cited 2023 Apr 12]. Available from: https://github.com/OHDSI/Achilles.
43. Blacketer C, Defalco FJ, Ryan PB, Rijnbeek PR. Increasing trust in real-world evidence through evaluation of observational data quality. J Am Med Inform Assoc. 2021;28(10):2251–7.
44. Observational Health Data Sciences and Informatics. Atlas [Internet]. Observational Health Data Sciences and Informatics; 2023 [cited 2023 Apr 12]. Available from: https://github.com/OHDSI/Atlas.
45. Observational Health Data Sciences and Informatics. Athena [Internet]. Athena– OHDSI Vocabularies Repository. 2023 [cited 2022 Nov 25]. Available from: https://athena.ohdsi.org/.
46. Schuemie M. Usagi [Internet]. Usagi. 2021. Available from: http://ohdsi.github.io/Usagi/.
47. Fischer P, Stoehr MR, Gall H, Michel-Backofen A, Majeed RW. Data Integration into OMOP CDM for Heterogeneous Clinical Data collections via HL7 FHIR bundles and XSLT. In: PapeHaugaard L, Lovis C, Madsen I, Weber P, Nielsen P, Scott P, editors. DIGITAL PERSONALIZED HEALTH AND MEDICINE. NIEUWE HEMWEG 6B, 1013 BG AMSTERDAM. Studies in Health Technology and Informatics. Volume 270. NETHERLANDS: IOS; 2020. pp. 138–42.
48. Biedermann P, Ong R, Davydov A, Orlova A, Solovyev P, Sun H et al. Standardizing registry data to the OMOP Common Data Model: experience from three pulmonary hypertension databases. BMC Med Res Methodol. 2021;21(1).
49. Sentinel Initiative. Sentinel Common Data Model [Internet]. 2024 [cited 2024 Jan 11]. Available from: https://sentinelinitiative.org/methods-data-tools/sentinel-common-data-model.
50. i2b2 TranSMART Foundation. i2b2: Informatics for Integrating Biology & the Bedside [Internet]. 2024 [cited 2024 Jan 11]. Available from: https://www.i2b2.org/.
51. The National Patient-Centered Clinical Research Network. The National Patient-Centered Clinical Research Network - Data [Internet]. The National Patient-Centered Clinical Research Network. 2024 [cited 2024 Jan 11]. Available from: https://pcornet.org/data/.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.