doi:10.3233/SHTI210546

The Usage of OHDSI OMOP – A Scoping Review

Ines REINECKE[a,1] Michéle ZOCH[a], Christian REICH[b,c], Martin SEDLMAYR[a] and Franziska BATHELT[a]

[a] *Institute for Medical Informatics and Biometry at Carl Gustav Carus Faculty of Medicine at Technische Universität Dresden, Germany*

[b] *IQVIA, Cambridge, MA, USA*

[c] *Observational Health Data Sciences and Informatics (OHDSI), New York, NY, USA*

Abstract. OHDSI, a fast growing open-science research community seeks to enable researchers from around the globe to conduct network studies based on standardized data and vocabularies. There is no comprehensive review of publications about OHDSI's standard: the OMOP Common Data Model and its usage available. In this work we aim to close this gap and provide a summary of existing publications including the analysis of its meta information such as the choice of journals, journal types, countries, as well as an analysis by topics based on a title and abstract screening. Since 2016, the number of publications has been constantly growing and the relevance of the OMOP CDM is increasing in terms of multi-country studies based on observational patient data.

Keywords. OHDSI, OMOP, scoping review

1. Introduction

The Observational Medical Outcomes Partnership (OMOP) was formed in 2008 originally as public-private partnership to establish best practices of retrospective observational research. To overcome technical challenges of executing a large number of analytical methods in a network of multiple databases to detect 10 outcomes in 10 drug classes, standardization of data formats, content (coding) and methods became a necessity[1]. The result was the development of a common data model (CDM). This enabled other research groups around the world to exchange and use observational data for research in a standardized way[2]. The OMOP project ended in 2013. Based on the promising results of the OMOP project and the foundational CDM, an open-science community named Observational Health Data Science and Informatics (OHDSI) was founded in 2014.

The overall goal of the OHDSI community is to improve health care for patients [3]. Therefore, they provide an open-source software portfolio and methods for data standardization and analysis. This includes the OMOP CDM that provides the foundation for data storage. With it comes the OMOP Standardized Vocabularies for harmonizing the data, guidelines and a set of tools for the development of ETL jobs to transfer data into the OMOP CDM and a software framework for data analytics. It enables researchers

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

around the globe to perform studies based on standardized data. Until now, the OHDSI community is already represent in more than 19 countries, with more than 200 million patient records from outside the US and with more than 2,500 collaborators [4].

Since its inception, the adoption of OHDSI has steadily increased. Examples in Europe are (i) the European Health Data and Evidence Network (EHDEN) [5], (ii) the collaboration with Health Level Seven International (HL7) [6] and (iii) the relevance in the German Medical Informatics Initiative [7]. The EHDEN consortium is working on patient record harmonization across 22 countries in Europe based on OMOP. In June 2020, the European Medicines Agency (EMA) announced a project for building a research framework for multi-center studies on COVID-19 patients that includes the collaboration with EHDEN and is also built on OMOP as data standardization foundation. In March 2021, a collaboration between HL7 and the OHDSI community was announced with the goal to create a single common data model that integrates HL7 Fast Healthcare Interoperability Resources (FHIR) [8] and OMOP [9]. The consortium "Medical Informatics in Research and Care in University Medicine" (MIRACUM) [10], part of the German Medical Informatics Initiative, focusses on open-source software to enable data integration and distributed analyses. Thus, a Clinical Trial Recruitment Support System based on OMOP is being developed in MIRACUM [11].

All standards are as relevant as their adoption. There is no comprehensive overview of the adoption rate of the OMOP CDM in the field of observational patient data research. Therefore, we aim to provide this overview of the existing literature for the last 5 years. Our scoping review includes details on the origin of the publications and a categorization based on the content to gain a detailed overview of main research areas OMOP is used for. Additionally, we discuss the current state of the usage of OMOP in Germany.

2. Methods

To ensure transparency and sufficient reporting of our scoping review we followed the "Preferred Reporting Items for Systematic reviews and Meta-Analyses" (PRISMA) statement for the process of paper identification, screening, eligibility checking and inclusion [12].

2.1. Paper identification

To search for relevant paper to get included in our review, we used Pubmed, Web of Science and IEEExplore. Table 1 provides a detailed overview on our search that was executed on 22nd of February 2021.

Table 1. Search engine with search strings

search engine	search string
Pubmed	All Fields: OHDSI or OMOP or "Observational Health Data Sciences and Informatics" or "Observational Medical Outcomes Partnership"
Web of Science	ALL FIELDS: OHDSI or OMOP or " Observational Health Data Sciences and Informatics" or "Observational Medical Outcomes Partnership"
IEEEXplore	("Full Text & Metadata":OMOP) OR ("Full Text & Metadata":OHDSI) OR ("Full Text & Metadata":Observational Medical Outcomes Partnership) OR ("Full Text & Metadata":Observational Medical Outcomes Partnership)

2.2. Paper exclusion

The results from the three sources were imported to the Zotero Citation Manager [13] into a new library. The exclusion process consisted of two steps. First, duplicate findings were removed using the built-in function of Zotero for duplicate item handling. Second, papers that are off topic, have no open access to full text, or are not in English were excluded.

The exclusion of papers based on the above aspects was done by one author (IR) based on the paper titles and abstracts as well as a check for access and language of full text in English. A list of included publications has been exported as comma separated (CSV) file for further categorization by content.

2.3. Paper categorization

The categorization of paper has been done by country, journal and thematic focus. To categorize paper by country, we determined the country of the first author. Furthermore, we checked the origin of the clinical data that has been used in the publication and marked all paper when the usage of multi country data was described in the methods section of each paper. To categorize by journal we first used the item type that provides information whether a publication is a journal paper, a conference paper, a book or another type of publication. We categorized all journal paper by journal and grouped the journals into the three categories medical journals, medical informatics journals and informatics journals. Additionally, all paper abstracts and if needed the full text were screened by three co-authors (IR, FB, MZ) for categorization by content. Based on an initial tagging during the title-abstract-screening the reviewers agreed on six major thematic dimensions (see Figure 1). The dimension "usage" is very generic and covers a large number of subjects. Figure 1 contains nine subcategories for this dimension that provide more inside on the usage of OMOP. We determined that tool development based on OMOP is an important topic in research and therefore we checked each paper for this criterion as well. From the three annotations of the co-authors, a final result was agreed upon. The paper categorization process resulted in a single CSV file to be used for the next step of the paper analysis.

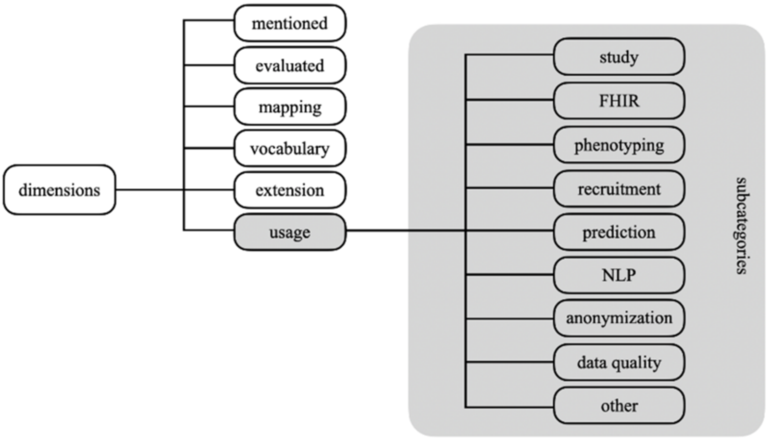

Figure 1: Paper categorization by content

3. Results

The number of publications identified by the initial search was 415 from all three data sources. After the removal of duplicates, 267 publications remained for the initial screening to check the eligibility for inclusion in the review. Of these, 94 publications were excluded because of missing full-text, no full text in English or paper are off topic. The latter ones are the result of us including authors that are involved in OHDSI research and have been picked up by the search, but may also write on unrelated subjects. Figure 2 shows the PRISMA flow chart result for the literature review. A complete list of publications included in our analysis is available here http://doi.org/10.5281/zenodo.4635599.

The literature research led to an inclusion of 173 publications, 53 of which are conference papers and 120 of which are journal publications. While the number of conference articles remains constant over the last five years, the number of journal publications is growing linearly.

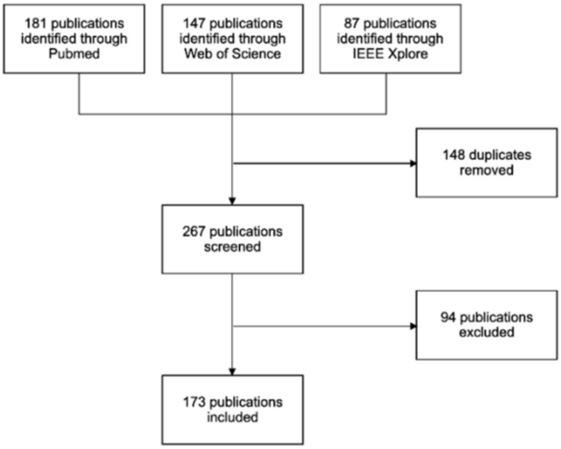
181 publications identified through Pubmed
147 publications identified through Web of Science
87 publications identified through IEEE Xplore
148 duplicates removed
267 publications screened
94 publications excluded
173 publications included

Figure 2: PRISMA flow chart diagram

Figure 3 provides an overview of the number of paper per journal type that are informatics, medical (med) and medical informatics (mi). We determined a constantly growing number of publications in mi journals. Moreover, the number of publication in medical journals increased to 14 in the year 2020.

The analysis of publications regarding the country of the authorship (Figure 5) shows a high activity in the United States (109 single + 2 joint main authorships with other countries). This is followed by South Korea (17 single main authorships), UK (7 single + 2 joint main authorships with other countries) and Germany (10 single main authorships).

We additionally marked those publications that are using data from multiple countries and counted 21 multi country publications in terms (full list of literature [14], column "country of data"). This led to a total number of 14 publications with multi country publication in the subcategory studies (Figure 4).

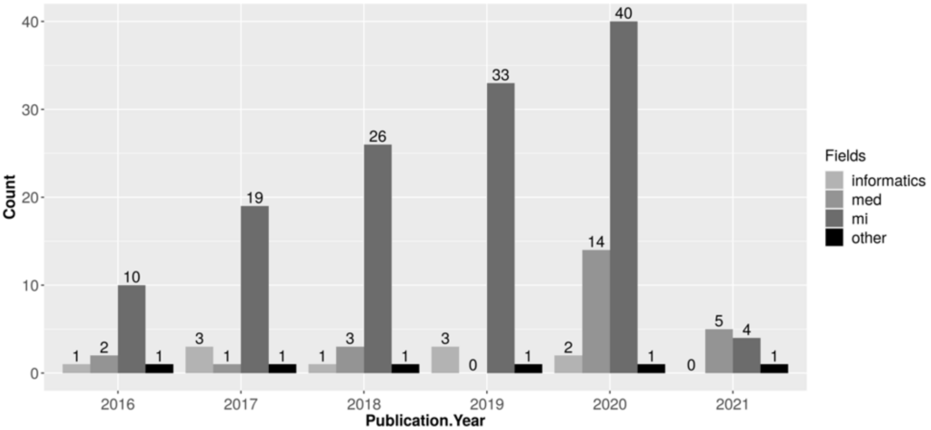

Figure 3: Annual distribution of different journal fields

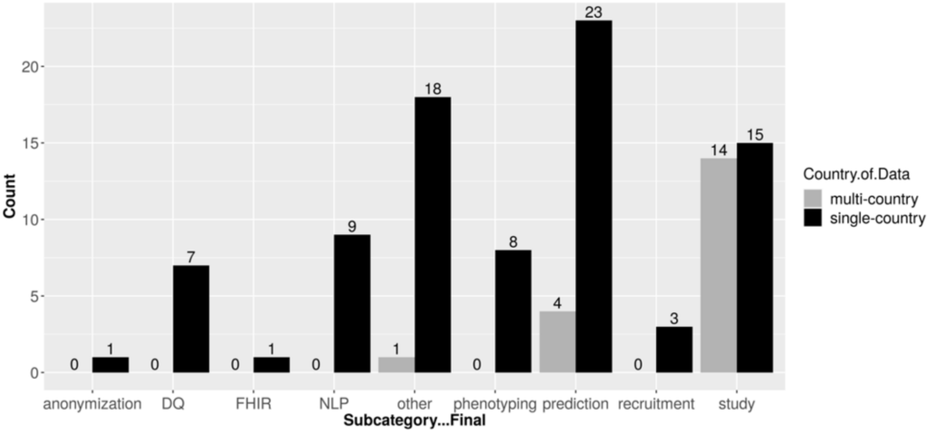

Figure 4: Distribution of Articles by Specific Usage (subcategories) and country of data

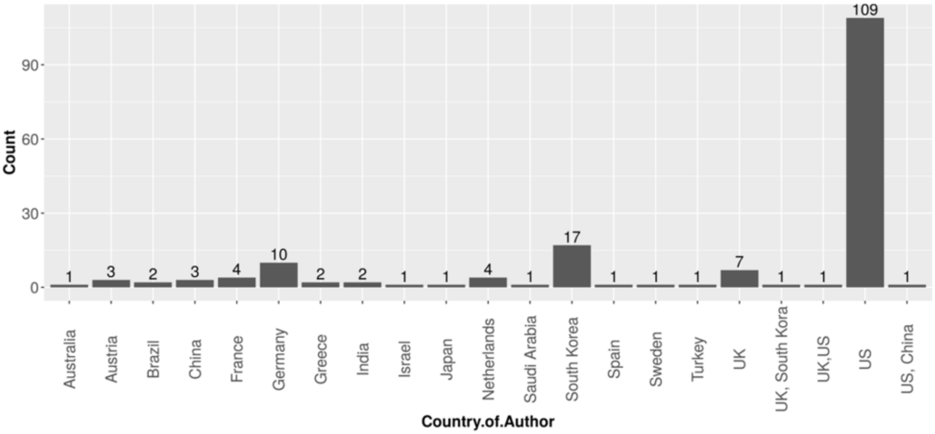

Figure 5: Distribution of articles by country of the main author

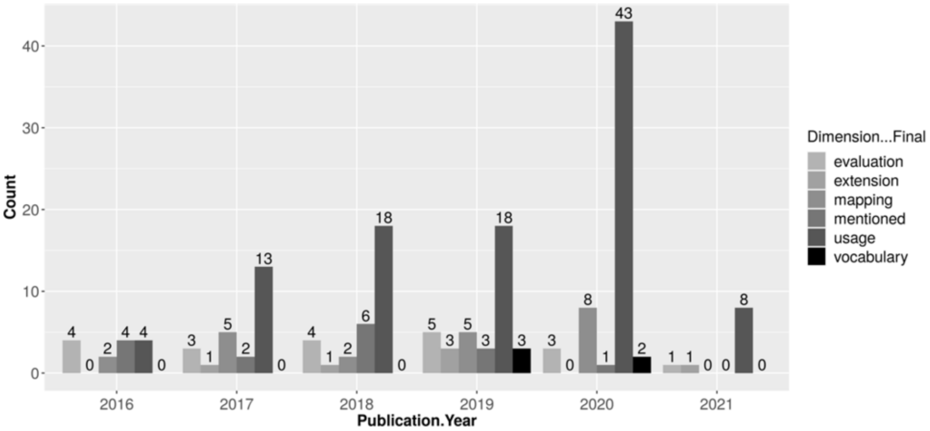

Figure 6: Annual distribution of articles by dimension

The analysis of the dimensions shows an extensive growth of the usage of OMOP in 2020 in comparison to the last years (see Figure 6). All other dimensions remain relatively steady.

The analysis of the subcategories (see Figure 7) from the main category "usage" shows that OMOP CDM is mostly used for conducting studies (29 publications) and for machine learning and prediction (25 Publications). Other important topics are phenotyping, Natural Language Processing (NLP), data quality (DQ) and patient recruitment for clinical studies. Most of the publications in the subcategory NLP are related to the extraction of inclusion criteria out of free-text to improve the overall recruitment process identifying potential participants for clinical studies. The results of the simple text and abstract free text analysis indicate a good match to the developed content dimensions and subcategories as shown in Figure 1.

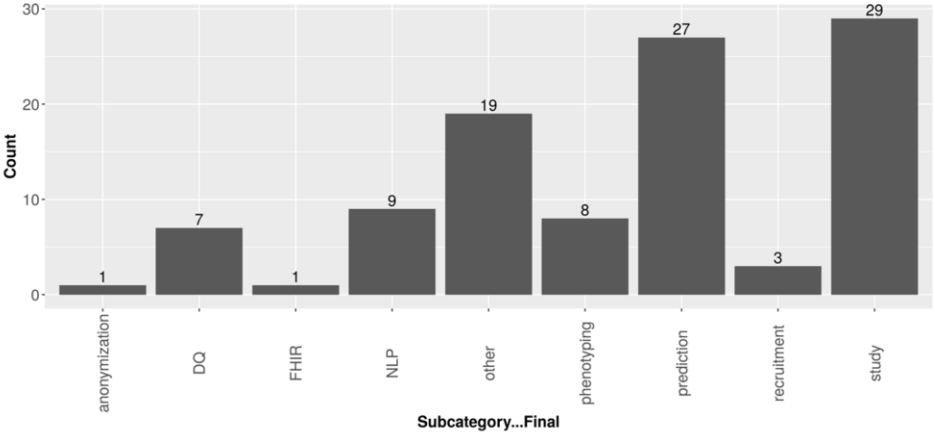

Figure 7: Distribution of articles by specific usage (subcategories)

A further analysis led to the identification of 34 publications that present newly developed tools based on OMOP CDM.

4. Discussion

Our review includes all publications listed on PubMed, Web of Science and IEEEXplore on OHDSI and OMOP. This covers the overwhelming part of the existing scientific work in this area. However, our review is limited since it does not reflect the complete range of research work done within the OHDSI community during regular community meetings or OHDSI symposia, because those publications are not indexed in the databases, we used for the paper identification.

Our work showed a steady growth of publications year by year since 2016, in particular the number of journal publications increased substantially. Compared to the years before and in contrast to the other dimension the "usage" of OMOP, the 2020 rates more than doubled compared to the previous years. Additionally we determined a high correlation between publications using patient data from multiple countries and publications that used OMOP for clinical network studies. This matches the goal of the OHDSI community to establish a worldwide research network for multi country studies on observational patient data to improve healthcare [3].

In the subcategory "mapping", we found many research publications from outside the United States (e.g., France, Austria, Germany and others), developing solutions to map local data terminologies to predefined and standardized vocabulary. The reason might be that as a precondition for joining the worldwide research network studies within the OHDSI community these countries had to do some catch up to do. Furthermore, the mapping and transfer of data to the OMOP CDM is a constantly seen topic since 2016. We identified only one publication that covers the topic of incremental data transfer to OMOP. The incremental load of data to OMOP is still an open issue and requires higher focus on the quality assurance process [15]. However, a generic concept on how to process changing data over time to the OMOP CDM is not available.

Another interesting fact is the low correlation between OHDSI OMOP and FHIR in the reviewed literature by now. FHIR is an international standard for exchanging healthcare data with increasing impact and popularity [15, 16]. Only one publication in the year 2017 was published in the context of FHIR and OMOP.

We identified 10 publications from German author teams but none of them is related to a clinical study based on observational patient data with a medical background [10, 11, 17–24]. All of them are related to mapping issues, architectural concepts or tool development based on OMOP. Though in 2020 two new data partners in Germany for the EHDEN project have been announced [26], Germany did not participate on multi country OHDSI OMOP studies based on observation in-patient data. Hence, it is crucial to empower the German research community to use OMOP productively to participate on international network studies on observational patient data.

5. Conclusions

This paper provides a review of existing publications on OHDSI OMOP over the last 5 years, showing the focus of groups while adopting the standard and putting it to use. We investigated the existing publications based on different aspects. We complemented our review by outlining latest trends and confirmed an increasing importance of OHDSI OMOP conducting network studies with observational patient data across different countries. This will help research teams to quickly explore relevant research topics around OMOP.

Declarations

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

Acknowledgement: This work is part of the project MIRACUM, funded by the German Ministry of Education and Research (FKZ 01ZZ1801A/L)

Author contributions: IR: conception of the work; IR, MZ, FB: literature screening. All authors contributed substantial ideas to the study 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.

References

[1] P.E. Stang, P.B. Ryan, J.A. Racoosin, J.M. Overhage, A.G. Hartzema, C. Reich, E. Welebob, T. Scarnecchia, and J. Woodcock, Advancing the Science for Active Surveillance: Rationale and Design for the Observational Medical Outcomes Partnership, *ANNALS OF INTERNAL MEDICINE*. **153** (2010) 600–606. doi:10.7326/0003-4819-153-9-201011020-00010.

[2] J.M. Overhage, P.B. Ryan, C.G. Reich, A.G. Hartzema, and P.E. Stang, Validation of a common data model for active safety surveillance research, *JOURNAL OF THE AMERICAN MEDICAL INFORMATICS ASSOCIATION*. **19** (2012) 54–60. doi:10.1136/amiajnl-2011-000376.

[3] G. Hripcsak, P.B. Ryan, J.D. Duke, N.H. Shah, R.W. Park, V. Huser, M.A. Suchard, M.J. Schuemie, F.J. DeFalco, A. Perotte, J.M. Banda, C.G. Reich, L.M. Schilling, M.E. Matheny, D. Meeker, N. Pratt, and D. Madigan, Characterizing treatment pathways at scale using the OHDSI network, *PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA*. **113** (2016) 7329–7336. doi:10.1073/pnas.1510502113.

[4] OHDSI, The book of OHDSI Observational Health Data Sciences and Informatics, 2019.

[5] EHDEN, EHDEN - European Health Data & Evidence Network, (n.d.). https://www.ehden.eu/.

[6] HL7, Health Level Seven International, (n.d.). http://www.hl7.org/index.cfm.

[7] S. Semler, F. Wissing, and R. Heyder, German Medical Informatics Initiative: A National Approach to Integrating Health Data from Patient Care and Medical Research, *Methods Inf Med*. **57** (2018) e50–e56. doi:10.3414/ME18-03-0003.

[8] HL7, HL7 FHIR Standard, (n.d.). http://hl7.org/fhir/.

[9] OHDSI, OHDSI HL7 Collaboration, (n.d.). https://www.ohdsi.org/ohdsi-hl7-collaboration/.

[10] H.-U. Prokosch, T. Acker, J. Bernarding, H. Binder, M. Boeker, M. Boerries, P. Daumke, T. Ganslandt, J. Hesser, G. Höning, M. Neumaier, K. Marquardt, H. Renz, H.-J. Rothkötter, C. Schade-Brittinger, P. Schmücker, J. Schüttler, M. Sedlmayr, H. Serve, K. Sohrabi, and H. Storf, MIRACUM: Medical Informatics in Research and Care in University Medicine: A Large Data Sharing Network to Enhance Translational Research and Medical Care, *Methods Inf Med*. **57** (2018) e82–e91. doi:10.3414/ME17-02-0025.

[11] I. Reinecke, C. Gulden, M. Kümmel, A. Nassirian, R. Blasini, and M. Sedlmayr, Design for a Modular Clinical Trial Recruitment Support System Based on FHIR and OMOP., in: Studies in Health Technology and Informatics, Netherlands, 2020: pp. 158–162. doi:10.3233/SHTI200142.

[12] D. Moher, A. Liberati, J. Tetzlaff, D.G. Altman, and The PRISMA Group, Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement, *PLoS Med*. **6** (2009) e1000097. doi:10.1371/journal.pmed.1000097.

[13] Zotero, Zotero, (n.d.). https://www.zotero.org/.

[14] I. Reinecke, The Use of OHDSI OMOP – A Scoping Review: list of included publications, (2021). doi:10.5281/ZENODO.4635599.

[15] K.E. Lynch, S.A. Deppen, S.L. DuVall, B. Viernes, A. Cao, D. Park, E. Hanchrow, K. Hewa, P. Greaves, and M.E. Matheny, Incrementally Transforming Electronic Medical Records into the Observational Medical Outcomes Partnership Common Data Model: A Multidimensional Quality Assurance Approach., *Appl Clin Inform*. **10** (2019) 794–803. doi:10.1055/s-0039-1697598.

[16] M. Lehne, S. Luijten, P. Vom Felde Genannt Imbusch, and S. Thun, The Use of FHIR in Digital Health - A Review of the Scientific Literature, *Stud Health Technol Inform*. **267** (2019) 52–58. doi:10.3233/SHTI190805.

[17] Information Technology Industry Council, Cloud Healthcare Pledge, (n.d.). https://www.itic.org/news-events/news-releases/tech-industry-looks-to-improve-healthcare-through-cloud-technology.

[18] M. Gruhl, I. Reinecke, and M. Sedlmayr, Specification and Distribution of Vocabularies Among Consortial Partners., in: Studies in Health Technology and Informatics, Netherlands, 2020: pp. 1393–1394. doi:10.3233/SHTI200458.

[19] P. Fischer, M.R. Stöhr, H. Gall, A. Michel-Backofen, and R.W. Majeed, Data Integration into OMOP CDM for Heterogeneous Clinical Data Collections via HL7 FHIR Bundles and XSLT., in: Studies in Health Technology and Informatics, Netherlands, 2020: pp. 138–142. doi:10.3233/SHTI200138.

[20] P. Unberath, H.U. Prokosch, J. Gründner, M. Erpenbeck, C. Maier, and J. Christoph, EHR-Independent Predictive Decision Support Architecture Based on OMOP., *Appl Clin Inform*. **11** (2020) 399–404. doi:10.1055/s-0040-1710393.

[21] J. Gruendner, T. Schwachhofer, P. Sippl, N. Wolf, M. Erpenbeck, C. Gulden, L.A. Kapsner, J. Zierk, S. Mate, M. Stürzl, R. Croner, H.-U. Prokosch, and D. Toddenroth, KETOS: Clinical decision support and machine learning as a service - A training and deployment platform based on Docker, OMOP-CDM, and FHIR Web Services., *PLoS One*. **14** (2019) e0223010. doi:10.1371/journal.pone.0223010.

[22] C. Maier, L. Lang, H. Storf, P. Vormstein, R. Bieber, J. Bernarding, T. Herrmann, C. Haverkamp, P. Horki, J. Laufer, F. Berger, G. Höning, H.W. Fritsch, J. Schüttler, T. Ganslandt, H.U. Prokosch, and M. Sedlmayr, Towards Implementation of OMOP in a German University Hospital Consortium., *Appl Clin Inform*. **9** (2018) 54–61. doi:10.1055/s-0037-1617452.

[23] H. Spengler, I. Gatz, F. Kohlmayer, K.A. Kuhn, and F. Prasser, Improving Data Quality in Medical Research: A Monitoring Architecture for Clinical and Translational Data Warehouses, in: 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), IEEE, Rochester, MN, USA, 2020: pp. 415–420. doi:10.1109/CBMS49503.2020.00085.

[24] H. Freitas da Cruz, B. Bergner, O. Konak, F. Schneider, P. Bode, C. Lempert, and M.-P. Schapranow, MORPHER - A Platform to Support Modeling of Outcome and Risk Prediction in Health Research, in: 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE), IEEE, Athens, Greece, 2019: pp. 462–469. doi:10.1109/BIBE.2019.00090.

[25] V. Tresp, J. Marc Overhage, M. Bundschus, S. Rabizadeh, P.A. Fasching, and S. Yu, Going Digital: A Survey on Digitalization and Large-Scale Data Analytics in Healthcare, *Proc. IEEE*. **104** (2016) 2180–2206. doi:10.1109/JPROC.2016.2615052.

[26] EHDEN, EHDEN Data Partner, (n.d.). https://www.ehden.eu/datapartners/.