Transfer of Clinical Drug Data to a Research Infrastructure on OMOP – A FAIR Concept

Ines REINECKE[a,1], Michéle ZOCH[a], Markus WILHELM[a], 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*

Abstract. Generating evidence based on real-world data is gaining importance in research not least since the COVID-19 pandemic. The Common Data Model of Observational Medical Outcomes Partnership (OMOP) is a research infrastructure that implements FAIR principles. Although the transfer of German claim data to OMOP is already implemented, drug data is an open issue. This paper provides a concept to prepare electronic health record (EHR) drug data for the transfer to OMOP based on requirements analysis and descriptive statistics for profiling EHR data developed by an interdisciplinary team and also covers data quality issues. The concept not only ensures FAIR principles for research, but provides the foundation for German drug data to OMOP transfer.

Keywords. EHR, data quality, drug administration, OHDSI, OMOP, FAIR

1. Introduction

The pandemic of coronavirus disease 2019 (COVID-19) has shown the need of standardized and reproducible research data, especially regarding drug administration, as observational studies are important to gain evidence, learn on real-word data and improve the COVID-19 patient treatment and their effects in the future [1]. However, those studies highly depend on the level of data quality, interoperability and reproducibility, even more if they are proceeded in a multi-centric environment [2].

The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) developed by the Observational Health Data Sciences and Informatics (OHDSI) is one option to foster reliability of retrospective, observational studies on real-world data [3] and compared to other CDMs, e.g. i2b2, PROCNet it best supports those studies [4]. OMOP comes with standardized vocabularies and terminologies, methods for data analyses and data quality checks while ensuring FAIR principles [5,6]. However, the main prerequisite to take advantage of OMOP and corresponding tools is the storage of patient electronic healthcare record (EHR) data in an OMOP conformed way. Although this is already tackled (for a nationally consolidated core data set) by the MIRACUM project [7] of the German Medical Informatics Initiative, drug data is still an open issue. Mainly because of the fact that drug administration is often documented in a non-

---
[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

standardized way and using unstructured data. Therefore this paper aims to provide a concept on data preparation for EHR drug data which is documented during in-patient visits at a German university hospital to be used in the OMOP for research in order to increase FAIR data principles for observational research on real-world data.

2. Methods

To move from heterogenous and proprietary EHR data to OMOP that aligns to the FAIR principles [5], we developed a target oriented concept based on medical expertise and an EHR as well as an OMOP analysis (Figure 1). Working with EHR data in research requires a deep understanding of the original data (e.g data origin, data completeness, data correctness, data structure) and the given target environment for research [2]. Thus we built a multidisciplinary team of data and computer scientists as well as pharmacists.
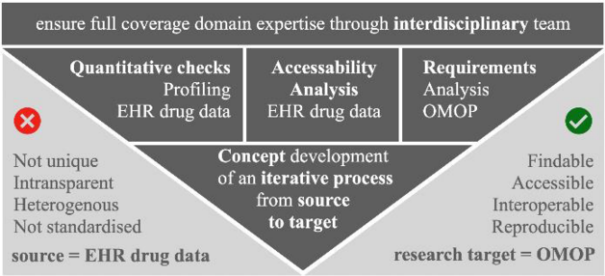

Figure 1. Methodology to derive a concept

A **requirements analysis** was done to determine the data elements and terminologies necessary to work with drugs in OMOP and to run OHDSI network studies. The analysis was done by [1] identifying relevant OMOP tables and data elements and [2] reviewing existing OHDSI network studies. Task (1) was done based on OMOP version 5.3.1 since this is the latest version supported by the current available OHDSI software stack. Task (2) was done for 29 OHDSI network studies [8] identified by Reinecke [9] in a Scoping Review. We checked the study protocols for those publications and determined if drug data was relevant and whether the drug dose information was required to answer the research question. The **accessibility** of EHR drug administration data in the corresponding IT system was analyzed in general and in comparison to the identified requirements. Access to drug documentation for intensive care was restricted and therefore excluded, additionally this study is limited to drug products registered by the German Federal Institute for Drugs and Medical Devices (BfArM) in a drug catalogue. A data profiling by **quantitative checks** of the drug prescription data element *dose unit* was done for all in-patient cases (approx. 55000) in the year 2020, for example we determined the quantitative ratio between free-text usage and drug catalogue reference in the drug administration data element *drug name*. Additionally a quantitative analysis of the *dose unit* values of the drug prescription data was done. Based on the results of the above methods a **concept of an iterative process** to convert EHR drug data into the OMOP was developed, which includes repetitive discussions with pharmacists (see Figure 3).

3. Results

The requirements analysis was limited to the OMOP *drug_exposure* table*.* The OMOP *drug_strength* table does not contain clinical data but drug concept information with dose and unit of drug ingredients, components and procucts that supports the standardization for drug utilization analysis. Table 1 shows a minimum list of required data elements in the OMOP table *drug_exposure*. The OMOP table column *quantity* was included, although it is not required by the CDM conventions, but we identified this information as required to conduct studies that need the drug dose information to answer the research question.

Table 1. Minimum list of required data elements in the OMOP table drug_exposure

drug_exposure	description	EHR availability
drug_exposure_id	unique key in the table	yes
person_id	reference to the patient identifier in the person table	yes
drug_concept_id	standard concept of domain drug	only drug names (catalogue/free-text)
drug_exposure_start_date	determines the start date of a drug exposure	yes
drug_exposure_end_date	determines the end date of a drug exposure	yes
drug_type_concept_id	specifies the type of a drug concept (e.g. EHR medication list, EHR prescription)	yes
quantity	based on dose form it refers to - The amount of tablets for clinical drugs with a fixed dose form - The amount of ingredient for divisible, liquid dose forms like injections	yes

The review of the studies identified 23 of 29 studies requiring drug data. Most of them used drug data based on RxNorm ingredient level, with no drug dose information. Only 2 studies were taking dose information into account for research. The analysis of the EHR drug data determined the availability of the identified and required data elements in the EHR system as shown in Table 1. The drug_concept_id is not available in the EHR system. Rather the drug name exists in the EHR system either as drug catalogue entry identifier or free-text information. The BfArM drug catalogue includes ingredient with ATC codes of active ingredients for each catalogue entry. Figure 2 visualizes the distribution of drug catalogue entries and free-text information. 59.50% of the drug prescription have a reference to the catalogue with ATC code and dose information. The other 40.50% have free-text only. Figure 3 shows the iterative process to prepare the EHR drug data to facilitate research based on OMOP. The data clean-up has to be done for free-text drug data by an appropriate domain expert. First computer scientists develop algorithm to extract ATC codes, dose and unit information from the free-text. Second the results get evaluated by pharmacists. The data clean-up is not needed for EHR data comprising drug catalogue entries. The concept mapping consists of step 3 and 4, where first the ATC concept gets mapped to the RxNorm ingredient concepts and second the RxNorm ingredient concept and the dose information has to be mapped to the RxNorm drug component. In step 5 the prepared drug data is moved to the OMOP database by an ETL job. Finally a data quality assessment using the OHDSI data quality dashboard [10] will be done. The results of the data quality assessment will be used to improve the previous steps if needed in the next iteration.
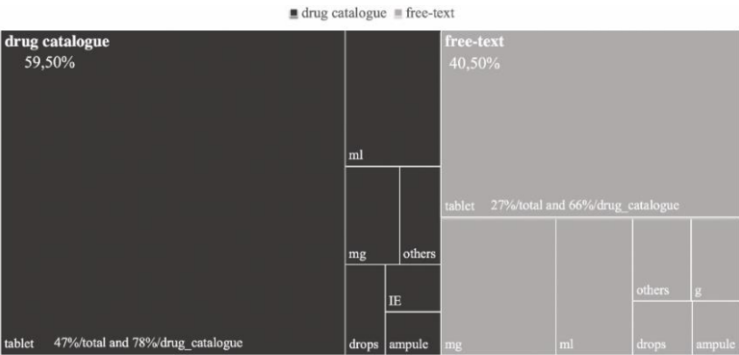
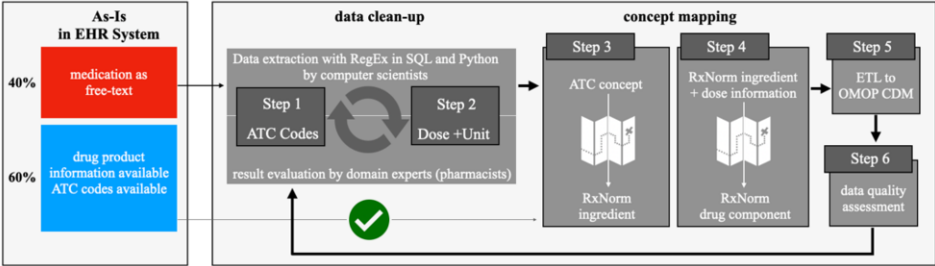**Figure 2.** Availability of drug data in the EHR system

Figure 3. Concept drug data transfer from EHR system to OMOP, ensuring data quality (Kahn et. al.)

4. Discussion

The concept for transforming real-world drug data to OMOP works towards the standardization and interoperability of data as well as the reproducibility of studies while ensuring data quality. In particular the extraction of ATC codes, dose and unit information from free-text is promising regarding the step towards applying FAIR principles for in-patient care data. With the continuous result evaluation by domain experts we ensure correctness of the original data and its meaning. The steps from ATC concept to RxNorm drug components further increase the standardization and thus the semantic interoperability. The subsequent implementation of the ETL processes enables research based on OMOP. Although ETL processes on German EHR data already exist [11], our approach extends the limitation on claim data to support drug data and thus is crucial to participate in observational research on OMOP in the future. It is essential for the introduction of the OHDSI Data Quality Dashboard to check for completeness, conformity and plausibility [12] of data in OMOP. Our concept is limited to EHR drug data prescription during a hospital stay. Integrating Intensive care unit (ICU) data needs further investigation since drugs get often applied continuously with a changing dose rates over time that rises new challenges in terms of dose calculations and conversion to OMOP. Drug history data is often part of free-text medical history records and requires implementation of NLP algorithm. In a first proof of concept the data clean-up has been applied to a small data set for a very specific clinical research question. In a next step, the concept will be applied systematically to all available EHR drug data for evaluation and quality assessment and to provide feedback on overall EHR drug data quality back to the patient care teams.

5. Conclusions

This paper provides a concept that closes the gap between EHR drug data and the requirements given by the common data model OMOP that focuses on improving the FAIRness of real-world data for research. It builds the foundation on converting German drug data to international standardized research environments and is an important step to enable German research groups to participate in studies in the OHDSI community.

Declarations

Author contributions: IR: conception of work, data analysis, concept development, results generation; All authors contributed ideas to the study and manuscript editing/revising. All authors approved the submitted manuscript and take responsibility for its scientific integrity. This study was performed by IR to (partially) fulfill the requirements for obtaining the academic degree "Dr. rer. medic." from the Technische Universität Dresden.

References

[1] EMA sets up infrastructure for real-world monitoring of treatments and vaccines [Internet]. [cited 2021 Jul 30]. Available from: https://www.ema.europa.eu/en/news/covid-19-ema-sets-infrastructure-real-world-monitoring-treatments-vaccines

[2] Kohane IS, Aronow BJ, Avillach P, Beaulieu-Jones BK, Bellazzi R, Bradford RL, et al. What Every Reader Should Know About Studies Using Electronic Health Record Data but May Be Afraid to Ask. J Med Internet Res. 2021 Mar 2;23(3):e22219.

[3] Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform. 2015;216:574–8.

[4] G Garza M, Del Fiol G, Tenenbaum J, Walden A, Zozus MN. Evaluating common data models for use with a longitudinal community registry. J Biomed Inform. 2016 Dec;64:333-341.

[5] Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Dec;3(1):160018.

[6] OHDSI. The book of OHDSI, Chapter 3.7 FAIR Guiding Principles [Internet]. [cited 2021 Jul 30. Available from: https://ohdsi.github.io/TheBookOfOhdsi/OpenScience.html#ohdsi-and-the-fair-guiding-principles

[7] Prokosch H-U, Acker T, Bernarding J, Binder H, Boeker M, et al. MIRACUM: Medical Informatics in Research and Care in University Medicine. Methods Inf Med. 2018 Jul;57(S 01):e82–91.

[8] Reinecke I. literature list of OHDSI studies [Internet]. Zenodo; 2021 [cited 2021 Jul 29]. Available from: https://zenodo.org/record/5145048

[9] Reinecke I, Zoch M, Reich C, Sedlmayr M, Bathelt F. The usage of OHDSI OMOP – A Scoping Review. Stud Health Technol Inform. Forthcoming 2021

[10] Blacketer C, Defalco FJ, Ryan PB, Rijnbeek PR. Increasing Trust in Real-World Evidence Through Evaluation of Observational Data Quality [Internet]. Health Informatics; 2021 Mar [cited 2021 Jul 29]. Available from: http://medrxiv.org/lookup/doi/10.1101/2021.03.25.21254341

[11] Maier C, Lang L, Storf H, Vormstein P, Bieber R, Bernarding J, et al. Towards Implementation of OMOP in a German University Hospital Consortium. Appl Clin Inform. 2018 Jan;9(1):54–61.

[12] Kahn MG, Callahan TJ, Barnard J, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. eGEMs. 2016 Sep 11;4(1):18.