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

doi:10.3233/SHTI230245

Fitness for Use of Anatomical Therapeutic Chemical Classification for Real World Data Research

Ines REINECKE[a,1], Elisa HENKE[a], Yuan PENG[a], Martin SEDLMAYR[a] and Franziska BATHELT[a]
[a] *Carl Gustav Carus Faculty of Medicine, Center for Medical Informatics, Institute for Medical Informatics and Biometry, Technische Universität Dresden, Dresden, Germany*

Abstract. *Introduction:* Real-world data (RWD) is gaining importance in research. For instance, the European Medicines Agency (EMA) is currently in the process of establishing a cross-national research network that utilizes RWD for research. However, data harmonization across countries must be carefully considered to avoid misclassification and bias. *Objectives:* This paper aims to investigate the extent to which a correct assignment of RxNorm ingredients is possible for medication orders that include only ATC codes. *Methods:* In this study, we analyzed 1,506,059 medication orders from the University Hospital Dresden (UKD) and merged them with the ATC vocabulary in the Observational Medical Outcomes Partnership (OMOP) including relevant relationship mappings to RxNorm. *Results:* We identified 70.25% of all medication orders were single ingredients with direct mapping to RxNorm. However, we also identified a significant complexity in mappings for the other medication orders that was visualized in an interactive scatterplot. *Discussion:* The majority of medication orders under observation (70.25%) are single ingredients and can be standardized to RxNorm, combination drugs pose a challenge due to the different approaches of ingredient assignments in ATC and RxNorm. The provided visualization can help research teams gain a better understanding of problematic data and further investigate identified issues.

Keywords. OHDSI, OMOP, RxNorm, ATC, interoperability, medication, RWD

1. Introduction

Real-world data (RWD) is health data that is routinely collected from multiple data sources during treatment and offers new research opportunities complementary to randomized controlled trials. Digitization is leading to the availability of electronic health records and are the catalyst for establishing retrospective research with RWD. Huge efforts are currently undertaken in Europe to benefit from RWD in research by establishing large, multi-national networks such as demonstrated by the European Health Data Evidence Network (EHDEN) [1]. Additionally, the European Medicines Agency (EMA) is currently in the process of establishing the Data Analysis and Real World Interrogation Network (DARWIN) to access and analyze healthcare data from across the

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

European Union with the goal of running large RWD studies to support regulatory decision-making [2]. The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM), developed by the Observational Health Data Sciences and Informatics (OHDSI) community, is used by EHDEN and DARWIN and is becoming increasingly important for retrospective research with RWD [3]. A major challenge in harmonizing RWD from different sources and countries into a common data model is the transition of non-standardized data and terminologies into a standardized format and terminologies. However, these activities are necessary to ensure semantic interoperability within cross-country research networks [4,5]. The translation activity is subject of bias due to the risk of misclassification of clinical facts such as drug exposures because of multiple mapping options [6]. Medication data are very heterogeneous and need to be harmonized for the use in the OMOP CDM. It is necessary to use the mandatory standard terminology RxNorm. RxNorm is a standardized nomenclature for all clinical drugs in the United States provided by the U.S. National Library of Medicine [7]. Unfortunately, hospital information systems often document drug exposure as free text or with ingredient information only as Anatomical Therapeutic Chemical Classification (ATC) codes [8]. A major difference between the active ingredients in ATC and RxNorm is that a single ATC code can combine more than one active ingredient in one code, whereas in RxNorm each active ingredient is represented by a separate code [9]. Therefore, this paper reports on the evaluation of the fitness for use of ATC for research in the OMOP CDM by checking to what extent the correct active ingredient in RxNorm can be assigned unambiguously. We present a visualization solution that helps interdisciplinary teams to gain better understanding of their medication data in OMOP CDM, when source data only contains ATC codes and mapping to a standard terminology is required.

2. Methods

This work is based on the medication order data from the University Hospital Carl Gustav Carus Dresden (UKD) from 2016 to 2020. It consists of 1,768,153 medication orders that have been already analyzed and improved in a previous work as described by Reinecke et al. [8]. For the present work, a subset of 85.18% (1,506,059) medication orders that has an ATC Level 5 code assigned, is used as input data. The most recent ATC vocabulary version from September, 07 2021 (version flag RxNorm 20210907) from the OHDSI ATHENA vocabulary service [10] containing 6,497 valid concepts has been imported into an OMOP database version 5.3.1. The required ATC to RxNorm vocabulary relationship information has been exported from the *concept* and *concept_relationship* tables in OMOP CDM for further exploration utilizing SQL statements. This export has been limited to the relationships of interest ("ATC -RxNorm pr up", "ATC -RxNorm pr lat", "ATC -RxNorm sec lat", "ATC -RxNorm sec up") since those represent the relationships for active ingredients available as ATC codes and their related RxNorm ingredients in accordance to the official OHDSI ATC vocabulary documentation [11]. In a first step, the relationship information has been enriched to provide a single row for each ATC Level 5 code with at least one of the above-mentioned relationship types containing the number of relationships types each calculated as numerical value in a separate column. Additionally, the total number of existing relationships between ATC Level 5 code and RxNorm has been calculated and added into an additional column. Second, the medication orders have been transformed according to the *drug_exposure* table in the OMOP CDM. Third, medication data in the

OMOP format has been merged with the transformed relationship information outcome from the first step. The three above described steps allow the determination of all ATC codes with a relationship to one or more RxNorm codes and show the challenges due to multi mappings during the transformation of medication orders. Since numerical results are rather confusing due to the high number of medical orders containing combination drugs, we developed an interactive visualization in addition to the table that contains the results. The visualization is designed interactively with search and hover capabilities to support identification of medication orders containing ATC codes that need special attention due to relationship complexity and high occurrence. Data analyzes and visualization has been implemented in Python Version 3.9.1 utilizing the following libraries: Pandas, Matplotlib and Bokeh. The complete source code and the interactive visualization can be accessed and downloaded on Zenodo [12].

3. Results

We identified 4811 ATC codes of the vocabulary (71.38%, 4811/6479) that have at least one mapping relationship to any of the 4 investigated types. The number of relationship mappings per ATC code can be large as exemplified in Table 1 for the 5 most frequent mapping relationship type combinations from a total of 363 different mapping combinations. There are 3655 (56.41%, 3655/6479) ATC codes with a single ingredient mapping to a RxNorm ingredient. ATC codes that contain combination ingredients can still have an exact mapping of ingredients to RxNorm, e.g.: J05AR13= "lamivudine, abacavir and dolutegravir; systemic" with relationships to the exact 3 RxNorm ingredients "abacavir", "dolutegravir" and "lamivudine". But there are more generic ATC codes with the primary ingredient mapped to RxNorm and additional ingredients unknown, e.g. ATC code C09BA05= "ramipril and diuretics" with a mapping to the RxNorm ingredient "ramipril" and 30 options for the diuretics. Thus, a specific mapping without the drug product information is not possible.

Table 1. Most frequent mapping combinations for ATC to RxNorm ingredient relationships

Number of ATC codes (exist in RWD)	ATC RxNorm pr lat	ATC RxNorm pr up	ATC RxNorm sec lat	ATC RxNorm sec up	# of Medication orders (%)
3655 (567)	1	0	0	0	1,057,986 (70.25)
143 (10)	1	0	1	2	8,859 (0.59)
95 (17)	2	0	0	0	28,717 (1.91)
55 (0)	1	0	2	3	0 (0.00)
34 (4)	1	0	1	3	2,717 (0.18)

The merge of the medication orders with the existing mapping relationships lead to a total of 567 ATC codes in our data with exactly one relationship to RxNorm. This identifies 70.25% (1,057,986/1,506,059) of all medication orders as single ingredients that can be mapped directly to a single RxNorm ingredient. A visualization with a search option, a result table and a scatterplot was developed (Figure 1). The search field offers filtering capabilities for ingredient names or ATC codes. The result table shows details on the ATC codes including its frequency in the medication order data, total number of mapping relationships to RxNorm and the ingredient name. The scatterplot represents each ATC code occurred in the medication orders as a single dot. The frequency of each ATC code in the medication order data is shown on the x-axis and the level of mapping

complexity to RxNorm is represented on the y-axis. For example, ATC code B05BB01 was prescribed 74,314 (frequency) times and has a total of 349 different mapping relationship types to RxNorm. This interactive visualization is used together with the pharmacist experts at the UKD to investigate the data further.

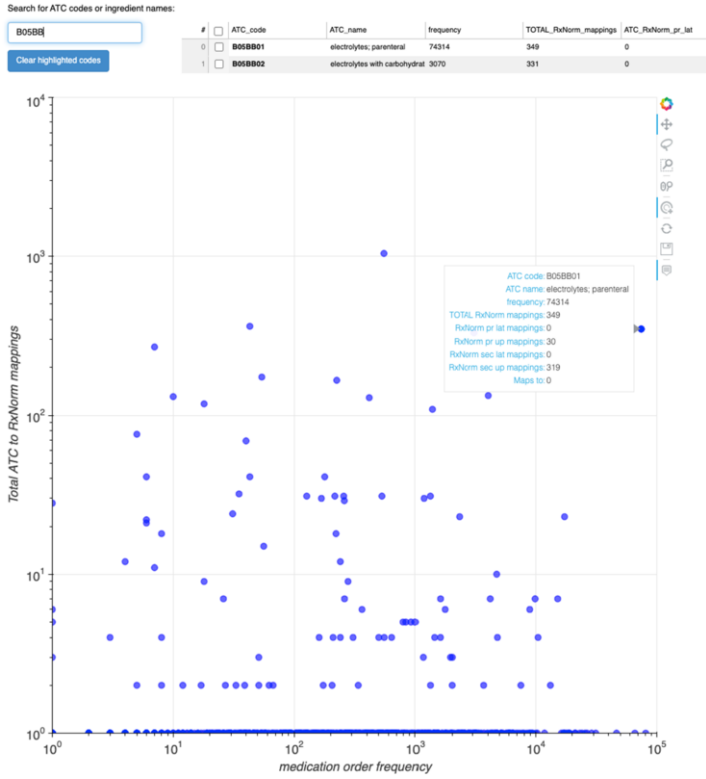

Figure 1. Medication orders by frequency and number of mapping relationships

4. Discussion and Conclusion

In this paper we were able to apply RxNorm ingredients to the majority of the medication orders at the UKD and thus enable research on OMOP CDM for single ingredient drugs. However, ATC is challenging when a single ATC code represents ingredient combinations. For some ATC codes a large number of mapping relationships to the RxNorm were identified (Figure 1) that cannot be assigned without further knowledge about the medication orders to the appropriate standard terminology required by the OMOP CDM. The interactive visualization created in this work was applied to only one site. The developed visualization can be used by other research teams, because our implementation uses the OMOP CDM as an input format. A prerequisite is the presence of the vocabulary in ATHENA and the mappings to a standardized terminology. In a next step we are going to analyze study protocols of studies executed by the OHDSI community teams to check how often ATC codes with combination ingredients have been used for cohort definitions and to answer research questions in the past years.

Declaration

Conflict of Interest: The authors declare that there is no conflict of interest.
Funding: This work is part of the MIRACUM project funded by the German Ministry of Education and Research (FKZ 01ZZ1801A/L).

References

[1] EHDEN, EHDEN - European Health Data & Evidence Network, (n.d.). https://www.ehden.eu/ (accessed January 2, 2023).
[2] European Medicines Agency (EMA), Data Analysis and Real World Interrogation Network (DARWIN EU), (n.d.). https://www.ema.europa.eu/en/about-us/how-we-work/big-data/data-analysis-real-world-interrogation-network-darwin-eu (accessed January 2, 2023).
[3] Reinecke I, Zoch M, Reich C, Sedlmayr M, Bathelt F. The Usage of OHDSI OMOP–A Scoping Review. German Medical Data Sciences 2021: Digital Medicine: Recognize–Understand–Heal. 2021:95-103. doi:10.3233/SHTI210546.
[4] 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. doi:10.1016/j.ijmedinf.2022.104925.
[5] Puttmann D, De Keizer N, Cornet R, Van Der Zwan E, Bakhshi-Raiez F. FAIRifying a Quality Registry Using OMOP CDM: Challenges and Solutions. InChallenges of Trustable AI and Added-Value on Health 2022 (pp. 367-371). IOS Press. doi:10.3233/SHTI220476.
[6] EMA, A common data model in Europe? – Why? Which? How?, (2018). https://www.ema.europa.eu/en/documents/report/common-data-model-europe-why-which-how-workshop-report_en.pdf (accessed January 20, 2023).
[7] National Institutes of Health (NIH), National Library of Medicine - RxNorm, (n.d.). https://www.nlm.nih.gov/research/umls/rxnorm/index.html (accessed January 2, 2023).
[8] Reinecke I, Siebel J, Fuhrmann S, Fischer A, Sedlmayr M, Weidner J, Bathelt F. Assessment and Improvement of Drug Data Structuredness From Electronic Health Records: Algorithm Development and Validation. JMIR Medical Informatics. 2023 Jan 25;11(1):e40312. doi:10.2196/40312.
[9] Bodenreider O, Rodriguez LM. Analyzing US prescription lists with RxNorm and the ATC/DDD Index. InAMIA Annual Symposium Proceedings 2014 (Vol. 2014, p. 297). American Medical Informatics Association.
[10] ATHENA standardized vocabularies – OHDSI, (n.d.). https://www.ohdsi.org/analytic-tools/athena-standardized-vocabularies/ (accessed January 17, 2022).
[11] OHDSI, Vocabulary ATC description, (n.d.). https://www.ohdsi.org/web/wiki/doku.php?id=documentation:vocabulary:atc.
[12] Reinecke I. Source code of the visualization, (2023). https://zenodo.org/record/7521226 (accessed January 10, 2023).