Conversion of Bulgarian Observational Data to OMOP Common Data Model: Initial Results

Petko Kovachev [1], Evgeniy Krastev [1], Dimitar Tcharaktchiev [2], Emanuil Markov [3] and Ivan Evg. Ivanov [3]

[1] *Sofia University "St. Kliment Ohridski", Faculty of Mathematics and Informatics, James Bourchier blvd., No. 5, Sofia, 1164, Bulgaria*
[2] *Medical University of Sofia, University Hospital of Endocrinology, Zdrave street No. 2, Sofia, 1431, Bulgaria*
[3] *Technical University of Sofia, Faculti of Automatics, Kliment Ohridsky blvd. No. 8, Sofia, Bulgaria*

Abstract

Common data models (CDMs) offer a standardized approach for data persistence and exchange. This is especially useful when nowadays-clinical data is distributed among heterogeneous sharing systems. Besides, OHDSI provides software tools in support on each stage of the ETL and ensure quality control. Therefore, data presented in CDM possesses all the features of a reliable source for a broad range of statistical analyses.

This paper presents initial results of a research work done with the objective to transfer outpatient records from the Bulgarian Diabetes register into the OMOP CDM. One of the major challenges has been the extraction of clinical data from native language text as well as the use of international OMOP concepts to annotate data recorded in a Bulgarian context. The mapping of national encoding for drug codes was one of the serious obstacles to conceptual mapping that requires adaptation of such codes to corresponding drug codes in the International Classification of Diseases 9th Revision.

Keywords

eHealth, observational health data, common data model, ETL, data harmonization, electronic health records

Information Systems & Grid Technologies: Fifteenth International Conference ISGT'2022, May 27–28, 2022, Sofia, Bulgaria
EMAIL: az@petko.info (P. Kovachev); eck@fmi.uni-sofia.bg (E. Krastev); dimitardt@gmail.com (D. Tcharaktchiev); emospy@gmail.com (E. Markov); ivan.evgeniev@gmail.com(I. Ivanov)
ORCID: 0000-0001-7509-4636 (P. Kovachev); 0000-0001-8740-5497 (E. Krastev);0000-0001-5765-840X (D. Tcharaktchiev); 0000-0002-8332-5884 (E. Markov); 0000-0002-0307-1600 (I. Ivanov)

1. Introduction

Digital health technologies produce huge amounts of data related to patient health collected as part the execution of routine healthcare services under real-world conditions. Data collected from such sources is collectively known as observational data (OD) OD is generated from a number of sources such as electronic health. OD is a valuable source for clinical evidence, which can be used to evaluate the safety and effectiveness of medical products in treatment of socially significant diseases like diabetes or cancer. Moreover, results from analysis of OD provide evidence in support for clinical decision-making [1]. Therefore, many research groups around the world attempt to integrate OD into a common data model that can serve as a reliable source for analyses of healthcare data [2] [3].

The Observational Health Data Sciences and Informatics [4] [5] (or OHDSI, pronounced "Odyssey") program is a multi-stakeholder, interdisciplinary collaborative initiative that aims to bring out the value of health data through large-scale analytics. OHDSI objective is to generate accurate, reproducible, and well-calibrated evidence and promote better health decisions and better care.

The Observational Medical Outcomes Partnership (OMOP) [6]Common Data Model (CDM) [7] is an open community data standard, designed to standardize the structure and content of observational data and to enable efficient analyses that can produce reliable evidence. It is a unified database model that allows integrating various OD sources including EHRs based on the standard.

The European Health Data and Evidence Network project (EHDEN) [8] under the Innovative Medicines Initiative (IMI) drives the adoption of the OMOP-CDM in Europe in close collaboration with OHDSI.

The OHDSI community and the software tools it is using follow the FAIR Data Guiding Principles [9] [10]:

- Findability – Any healthcare database that is mapped to OMOP and used for analytics should persist for future reference and reproducibility. Therefore, data are described with rich metadata, where metadata early and explicitly include a globally unique and persistent identifier of the data it describes.
- Accessibility – Accessibility of OMOP mapped data through an open protocol is typically achieved through the SQL interface. The protocol must provide a procedure for authentication and authorization.
- Interoperability – data use a formal, accessible, shared, and broadly applicable language for knowledge representation. Additionally, data must be accompanied with vocabularies that follow FAIR principles with qualified references to other data.
- Reusability – Metadata and data should be well described so that they can be replicated and/or combined in different settings. Moreover, data must satisfy domain-relevant community standards.

The OMOP Common Data Model (CDM) [7] is an open community data standard, designed to standardize the structure and content of observational data and to enable efficient analyses that can produce reliable evidence. A central component of the OMOP CDM are the OHDSI standardized vocabularies (Figure 1).

The OMOP Common Data Model allows systematic analysis of disparate observational databases. The concept behind this approach is to transform data contained within those databases into a common format (data model) as well as a common representation (terminologies, vocabularies, coding schemes), and then perform systematic analyses using a library of standard analytic routines that have been written based on the common format.

Routine health databases, based on routine electronic health records (EHRs) differ in purpose, content and design. Common Data Models (CDM) can enable standardized analysis of disparate data sources simultaneously.

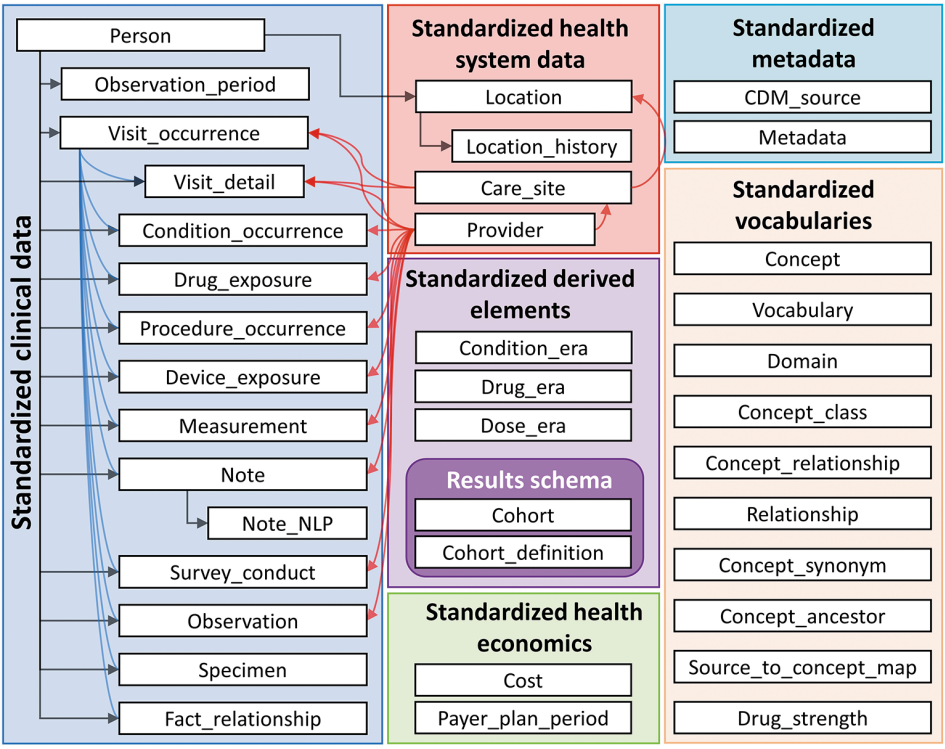

Figure 1: Overview of all tables in the CDM version 6.0. Not all relationships between tables are shown

The CDM contains standardized tables grouped in 16 Clinical Event tables, 10 Vocabulary tables, 2 metadata tables, 4 health system data tables, 2 health economics data tables, 3 standardized derived elements, and 2 Results schema tables.

- The development of the CDM follows the following design elements:
- Suitability for purpose – CDM data is organized in a way that is suitable for analysis
- Data protection – Personalized data, such as names, birthdays, and living address and so on is anonymized format.
- Design of domains – The domains are modeled in a person-centric relational data model, where for each record refers to a person and the date when the OD is captured.
- Rationale for domains – Domains are identified and separately defined in an Entity-relationship diagram, where each domain has specific attributes that are not otherwise applicable. All other data can be preserved as an observation in an entity-attribute-value structure.
- Standardized Vocabularies – CDM relies on the Standardized Vocabularies such as SNOMED containing all necessary and appropriate corresponding standard healthcare concepts.
- Reuse of existing vocabularies- definitions of codes drugs, diseases from national, industry standardization or vocabulary definitions are mapped to international coding systems or reused.
- Maintaining source codes – The original source code is persisted together with their corresponding codes in Standardized Vocabularies, so that the model loses no information from the OD.
- Technology neutrality – The CDM does.t depend on specific technology. It can be implemented on any relational database, such as MS SQL Server, Oracle etc.
- Scalability – The CDM is optimized for data processing and computational analysis to accommodate data sources that vary in size, including databases with up to hundreds of millions of persons and billions of clinical observations.
- Backwards compatibility – All changes from previous CDMs are clearly delineated. Older versions of the CDM can be easily created from this CDMv5, and no information is lost that was present previously

There are implicit and explicit conventions that adopted in the CDM:

- General Conventions of the Model – The CDM is considered a “person-centric” model, meaning that all clinical Event tables are linked to the PERSON table.
- General Conventions of Schemas – Most of the schemas are considered as read only, writable tables are only COHORT and COHORT_DEFINITION in “Results” schema.

- General Conventions of Data Tables – The CDM is platform independent. Data types are defined generically using ANSI SQL data types (VARCHAR, INTEGER, FLOAT, DATE, DATETIME, and CLOB). Precision is provided only for VARCHAR.
- General Conventions of Domains – Events of different nature are organized into Domains. These Events are stored in tables and fields, which are Domainspecific, and represented by Standard Concepts that are also Domainspecific as defined in the Standardized Vocabularies.

2. Methods

2.1. Environment preparation

Preparing the CDM database environment by installing a PostgreSQL DBMS, Java JDK and Docker compose in a Linux workstation. The database setup includes:

a) Creating the required database users;

b) Creating the OMOP CDM tables with the Common Data Model/PostgreSQL DDL scripts; and

c) Importing the standard OMOP vocabularies from athena.ohdsi.org.

d) All the OHDSI sources are available at github.org/OHDSI. We installed the White Rabbit, Rabbit-In-a-Hat, Usagi, [11] Achilles [12] and Broadsea repositories required for the OHDSI web applications, configured the addresses and JDBC URLs and started their respective Docker containers.

2.2. ETL – Extract Transform Load

In order to get from the native/raw data to the OMOP Common Data Model (CDM) we have to create an extract, transform, and load (ETL) process. This process should restructure the data to the CDM, and add mappings to the Standardized Vocabularies with these steps (Figure 2).

1. Design the ETL – To initiate an ETL process on a database we need to understand source data, including the tables, fields, and content. The White Rabbit software from OHDSI to perform a scan of the source data. The scan generates a report used as a reference when designing the ETL. With the White Rabbit scan in hand, we have a clear picture of the source data. We also know the full specification of the CDM. Rabbit-In-a-Hat [13] is used in next step is to perform mapping of source fields to the target in CDM database. Rabbit-In-a-Hat is designed to read and display a White Rabbit scan document and generates documentation for the ETL process but it does not generate code to create an ETL.

2. Create the Code Mappings- With Usagi form OHDSI tools we perform manual process of creating a code mappings with standard source codes to Vocabulary concepts.
3. Implement the ETL – Once the design and code mappings are completed, the ETL process is implemented with ETL-CDM Builder. As result, we have CDM relevant database populated with the data from the source database.
4. Quality Control – For the extract, transform, load process, quality control is iterative. The typical pattern is to write logic > implement logic > test logic > fix.

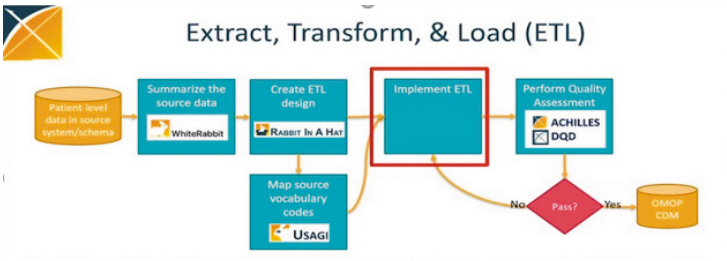

Figure 2: CDM ETL processes and tools /write logic

The result of the ETL process is a CDM compliant database/schema ready to be used for analyses.

2.3. Study execution

The most convenient and precise approach to perform observational study against CDM database is to use ATLAS [14]- free, publicly available, web based tool developed by the OHDSI that facilitates the design and execution of analyses on standardized, patient level, observational data in the CDM format. ATLAS is deployed as a web application in combination with the OHDSI WebAPI and is hosted on Apache Tomcat and could be deployed and started as Docker container or cloud service.

The screenshot of ATLAS in Figure 3 shows the various functionalities provided by ATLAS:

- Data Sources – provides the capability review descriptive, standardized reporting for each of the configured data sources.
- Vocabulary Search – provides the ability to search and explore the OMOP standardized vocabulary.
- Concept Sets provides the ability to create collections of logical expressions that can be used to identify a set of concepts to be used throughout your standardized analyses.

- Cohort Definitions – ability to construct a set of persons who satisfy one or more criteria for a duration of time.
- Characterizations – an analytic capability that allows you to look at one or more cohorts and to summarize characteristics about those patient populations.
- Cohort Pathways Cohort pathways is an analytic tool that allows you to look at the sequence of clinical events that occur within one or more populations.
- Incidence Rates – a tool that allows you to estimate the incidence of outcomes within target populations of interest.
- Profiles – tool that allows exploring of an individual patients longitudinal observational data to summarize what is going on within a given individual.
- Population Level Estimation- a capability that allows defining a population study for level effect estimation using a comparative cohort design whereby comparisons between one or more target and comparator cohorts can be explored for a series of outcomes.
- Patient Level Prediction – allows to apply machine-learning algorithms to conduct prediction analyses at patient level whereby you can predict an outcome within any given target exposures.
- Jobs – used to explore the state of processes that are running through the WebAPI.
- Configuration – to review the configured data sources that have been in the source configuration section.

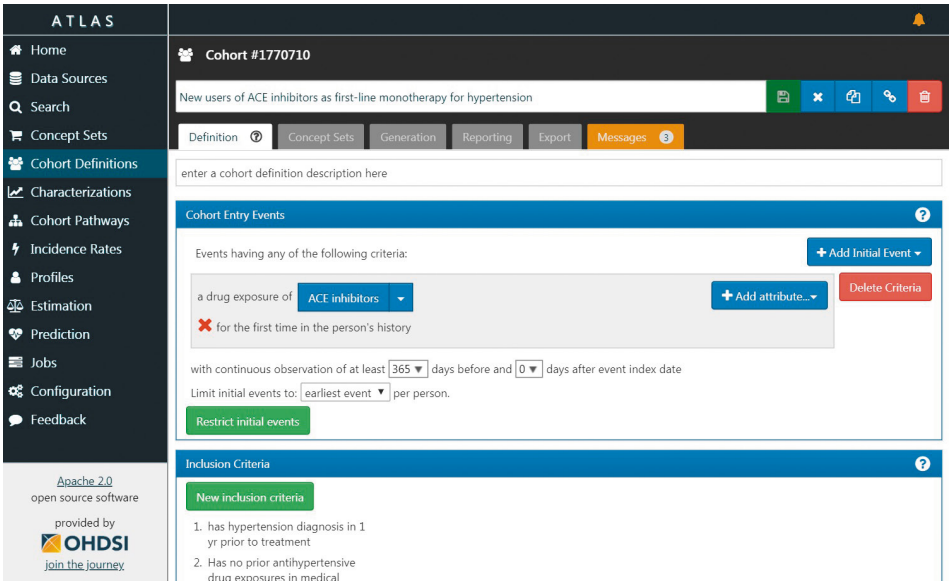

Figure 3: ATLAS user interface

3. Results

In Bulgaria outpatient records are produced by the General Practitioners (GPs) and the Specialists from Ambulatory Care for every contact with the patient. Outpatient records from patients with diabetes are maintained by the Bulgarian Diabetes Register. These records represent a true example of OD that can be produce valuable evidence for improving the treatment and management the healthcare services for such patients. The outpatient records are semi-structured files with predefined XML schema. The source XML documents included more than 1 600 000 pseudonymized outpatient records. The most important indicators in the records like Age, Gender, Location, Diagnoses are stored in explicit tags. The Case history is presented as free text in the Anamnesis. Additionally, these records include in native text information about the Patient status described the patient state, symptoms, syndromes, patients' height and weight, body mass index (BMI), blood pressure and other clinical concepts. The values of clinical tests and lab data are enumerated also as free text in a separate section of the XML document. A special section is dedicated to the prescribed treatment.

This paper presents first results from a research work whose objective is to convert this OD into OMOP CDM version 5.3.0. Here we will shortly describe the first stage of the ETL process (Design the ETL) that will be used for mapping of the source fields to the target in CDM database.

Our first task has been to parse data from the available XML documents and store it in a relational database, where the relational model matches the XML schema of the source XML documents. Each one of the XML documents contains a set of outpatient records of patients with diabetes. Therefore, the relational database is referred to as Diabetes2018. Both the source database Diabetes2018 and the target CDM database are MS SQL Server databases. The Entity Relationship Diagram of the source database and a description of the tables in the source database are shown correspondingly in Table 1 and Figure 4.

Table 1
Description of tables in the source database

Source Table	English Name	Description
AmbLists	AmbLists	Outpatient records generated during a visits to GP or Specialist of Ambulatory Care, coded in ICD9.
Diagnoses	Diagnoses	Diagnoses set in Outpatient record
Doctors	Doctors	Doctors data
HospitalizationLists	Hospitalization Lists	List with hospitalization directions related to an outpatient record
HospitalNapravlenie	HospitalNapravlenie	Direction for hospitalization

MDdirections	MDdirections	Direction for medical examination by Specialist in Ambulatory care
Patients	Patients	Patient description data
PrimaryVisit	PrimaryVisit	Initial visit to a Specialist in Ambulatory Care
Procedures	Procedures	Procedures assigned in Outpatient record
Profilact	Profilact	Describes a visit for Disease prevention
RecipeBooks	RecipeBooks	Recipe Books of Patient
Recipies	Recipies	Recipes for reimbursable medicinal products. These are stored in the patient's Recipe Book
Reimbursables	Reimbursables	Reimbursable medicinal products
SecondaryVisit	SecondaryVisit	Secondary visit to a Specialist in Ambulatory Care
SIMPConsults	SIMPConsults	Specialized Medical care
VSDSIMPConsults	VSDSIMPConsults	Highly Specialized Medical care
BloodPressure	BloodPressure	Blood pressure values measured in **mmHg** extracted from natural language text description of patient status and examination data
BloodSugar	BloodSugar	Blood sugar profile first value measured in mmol/l extracted from natural text description of patient status and examination data
BMIdata	BMIdata	Body Mass Index data extracted from natural language text description of patient status and examination data. Table includes also Height and Weight measurements in **sm** and **kg**, when Height and Weight data are found in the text
HbAc1Data	HbAc1Data	Contains values of HbAc1 extracted from natural language text description of patient status and examination data measured in **mmol/mol**
DiabetTimeData	DiabetTimeData	Contains the number of years before the illness diabetes has been established for the first time, extracted from natural language text description of patient status and examination data
TrigData	TrigData	Contains values of triglycerides extracted from natural language text description of patient status and examination data and measured in **mmol/L**

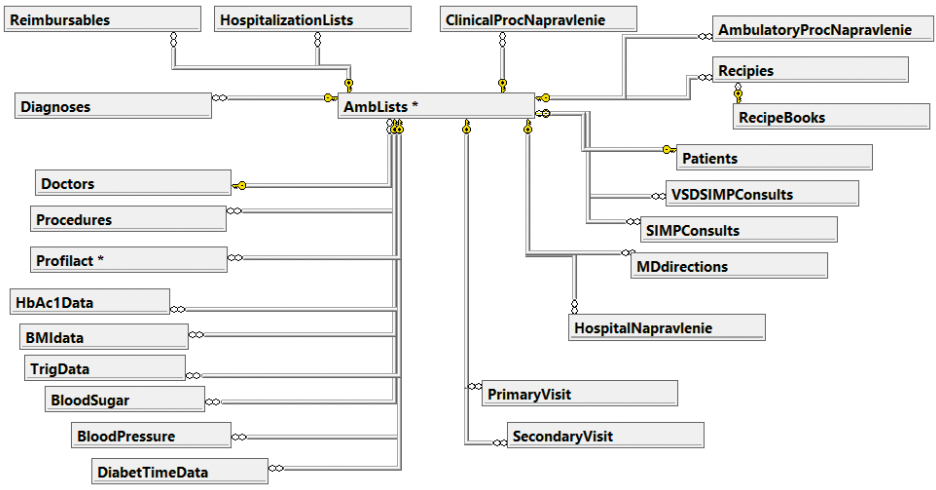

Figure 4: Source data from DIAB2018 to CDMV5.3.0 database mapping

Thus, clinical data for more than 502 000 patients with diabetes have been loaded in Diabetes2018. In order to execute this task, we had to write programs in Java and Python scripts that parse the XML documents and extract clinical data as blood pressure, glucose, height and weight body mass index and many other measurements recorded in the source XML documents in natural language text. Next, we used the Rabbit-In-A-Hat tool for the mapping definition of data (Figure 5).

Figure 5: Source data from DIAB2018 to CDMV5.3.0 database mapping

The mapping rules identified in Figure 5 allow proceeding with the extract, transform, and load stages of the ETL process using SQL queries implemented in Microsoft SQL Server 2019. For completeness, with the help of White Rabbit we generated a data dictionary for all the tables and fields that have been profiled. This data dictionary includes the English translation of the local name of fields and their description (Table 1). All the tasks at this stage have been accompanied by continuous and tedious quality control so that the CDM database has all the attributes of a reliable evidence, namely, repeatable, reproducible, replicable, generalizable, robust and calibrated [15].

4. Conclusion

Common data models (CDMs) offer a standardized approach for data persistence and exchange. This is especially useful when nowadays-clinical data is distributed among heterogeneous sharing systems. Besides, OHDSI provides software tools in support on each stage of the ETL and ensure quality control. Therefore, data CDM possesses all the features of a reliable source for a broad range of statistical analyses.

This paper presents initial results of a research work done with the objective to transfer outpatient records from the Bulgarian Diabetes register into the OMOP CDM. One of the major challenges has been the extraction of clinical data from native text as well as the use of international OMOP concepts to annotate data recorded in a Bulgarian context. The mapping of national encoding for drug codes was one of the serious obstacles to conceptual mapping that requires adaptation of such codes to corresponding drug codes in the International Classification of Diseases 9th Revision.

5. Acknowledgements

The presentation of this paper is supported by the National Scientific Program “Electronic Healthcare in Bulgaria” (eHealth).

6. References

[1] Y. Jeon, Y. Choi, E. Kim, S. Oh and H. Lee, “Common data model-based real-world data for practical clinical practice guidelines,” Transl Clin Pharmacol., vol. 28, no. 2, pp. 67–72, 2020.

[2] A. Lamer, N. Depas, M. Doutreligne, A. Parrot, D. Verloop, M. Defebvre, G. Ficheur, E. Chazard and J. Beuscart, “Transforming French Electronic Health Records into the Observational Medical Outcome Partnership’s Common Data Model: A Feasibility Study,” Appl Clin Inform., vol. 11, no. 1, pp. 13–22, 2020.

[3] B. Ryu, E. Yoon, S. Kim, S. Lee, H. Baek, S. Yi, H. Na, J. Kim, R. Baek, H. Hwang and S. Yoo, “Transformation of Pathology Reports Into the Common Data Model With Oncology Module: Use Case for Colon Cancer.,” J Med Internet Res., vol. 22, no. 12:e18526, 2020.

[4] OHDSI, “Observational Health Data Sciences and Informatics – OHDSI,” www.ohdsi.org, 2022. [Online]. Available: https://www.ohdsi.org/web/wiki/doku.php?id=welcome. [Accessed 10 April 2022].

[5] Observational Health Data Sciences and Informatics, The Book of OHDSI, https://ohdsi.github.io/TheBookOfOhdsi, 2021.

[6] OHDSI CDM Working Group, “Welcome to OMOP,” OHDSI, 2022. [Online]. Available: https://ohdsi.org/omop. [Accessed 2 April 2022].
[7] Observational Health Data Sciences and Informatics, “OMOP Common Data Model,” ohdsi.org, 2022. [Online]. Available: https://ohdsi.github.io/CommonDataModel. [Accessed 10 April 2022].
[8] EHDEN, “The European Health Data & Evidence Network,” EHDEN, 2022. [Online]. Available: https://www.ehden.eu. [Accessed 10 April 2022].
[9] OHDSI, “FAIR Principles,” 2016. [Online]. Available: https://www.go-fair.org/fair-principles. [Accessed 10 April 2022].
[10] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg et al., “The FAIR Guiding Principles for scientific data management and stewardship,” Scientific Data, vol. 3:160018, no. 1, 2016.
[11] Observational Health Data Sciences and Informatics, “Software Tools,” ohdsi.org, 2022. [Online]. Available: https://www.ohdsi.org/software-tools. [Accessed 10 April 2022].
[12] Observational Health Data Sciences and Informatics, “ACHILLES for data characterization,” ohdsi.org, 2022. [Online]. Available: https://www.ohdsi.org/analytic-tools/achilles-for-data-characterization. [Accessed 10 April 2022].
[13] Observational Health Data Sciences and Informatics, “Rabbit-In-a-Hat,” 2022, 15 February 2022. [Online]. Available: http://ohdsi.github.io/WhiteRabbit/RabbitInAHat.html. [Accessed 10 April 2022].
[14] OHDSI, ”ATLAS,” 2022. [Online]. Available: https://github.com/OHDSI/Atlas/wiki. [Accessed 10 April 2022].
[15] C. Blacketer, M. Kallfelz and P. Rijnbeek, “WP5 – Data Workflow Implementation & Service Deployment. D5.2 Report on Quality Assurance and Control Procedures,” EHDEN, 2020.