dHealth 2024
D. Hayn et al. (Eds.)

doi:10.3233/SHTI240007

Mapping the Bulgarian Diabetes Register to OMOP CDM: Application Results

Evgeniy KRASTEV[a], Emanuil MARKOV[b], Simeon ABANOS[a], Ralitsa KRASTEVA[c] and Dimitar TCHARAKTCHIEV[d,1]

[a] *Sofia University St. Kliment Ohridski, Bulgaria (EFMI institutional member)*
[b] *Technical University, Sofia, Bulgaria*
[c] *Specialized Hospital for Active Treatment of Children's Diseases, Sofia, Bulgaria*
[d] *Medical University, Sofia, Bulgaria (EFMI institutional member)*

Abstract. *Background*: The Bulgaria Diabetes Register (BDR) contains more than 380 millions of pseudonymized outpatient records with proprietary data structures and format. *Objectives:* This paper presents the application results and experience acquired during the process of mapping such observational health data to OMOP CDM with the objective of publishing it in the European Health Data and Evidence Network (EHDEN) Portal. *Methods:* The data mapping follows the activities of the well-structured Extract-Transform-Load process. Unlike other publications, we focus on the need for preprocessing the data structures of raw data, cleaning data and procedures for assuring quality of data. *Results:* This paper provides quantitative and statistical measures for the records in the CDM database as published in the EHDEN Portal. *Conclusion:* The mapping of data from the BDR to OMOP CDM provides the EHDEN community with opportunities for including these data in large-scale project for evidence generation by applying standard analytical tools.

Keywords. Medical Informatics Applications, Diabetes Mellitus, Registries, OMOP CDM, Electronic Health Records

1. Introduction

Diabetes mellitus is a socially significant illness that rapidly increases its prevalence over the world. National diabetes registries (NDR) like the Bulgarian Diabetes Register (BDR) provide opportunities to analyze huge amounts of clinical data and develop new approaches for timely prevention of this disease by focusing on risk factors with age-specific quantitative effects on diabetes [1] [2].

The BDR manages pseudonymized electronic health records of the patients with diabetes (Type 1 and Type 2) since 2013, where the latest dataset is from 2018 [3] [4]. Over 380 million outpatient records were collected during this time period by the National Health Insurance Fund (NHIF) from all General Practitioners and Health Professionals in specialized healthcare nationwide for every visit of a patient suffering from diabetes. The data structure of the outpatient records follows a proprietary XML schema introduced by the NHIF for its specific data processing purposes. Similarly to

[1] Corresponding Author: Dimitar Tcharaktchiev, Department of Medical Informatics, Medical University-Sofia, University Hospital of Endocrinology, 2 Zdrave street, Sofia 1431, Bulgaria; E-mail: dimitardt@gmail.com, ORCID: https://orcid.org/0000-0001-5765-840X

other NDR, this proprietary structure of observational health data strongly constrains the exchange and integration of data as well as the interoperability of the tools for data analysis and evidence generation [5].

In this paper we describe the implementation of the Extract, Transform and Load (ETL) process for mapping the dataset with outpatient records in the BDR from year 2018 to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) in the Observational Health Data Sciences and Informatics (OHDSI) community [6] [7]. The OMOP CDM offers a feasible solution to interoperability challenges in health data processing by transforming data into a common format and standardized vocabulary [8] [9]. It is the CDM employed to build the European Health Data and Evidence Network (EHDEN), a growing network of 187 Data Partners across 29 countries, with access to more than 850 million anonymized health records [10]. We focus on overcoming some of the major difficulties in the implementation of the ETL process that emerge at the stage of preprocessing the source data, vocabulary mapping and overall quality assurance of data workflow [11]. Although each step of the ETL process is supported by OHDSI open-source software tools [12], the nature of these problems makes the ETL process a challenging experience. It is required to overcome a lot of technical difficulties, knowledge management problems and quality concerns in order to make the OMOP CDM part of EHDEN [13] [14]. Thus, the OMOP CDM of the BDR with data for more than 500K distinct patients was published in the EHDEN portal [15] [16]. Section 2 describes the methods to execute the ETL process, while Section 3 presents the OMOP CDM database in terms of quantitative measures and results from data quality tests. Section 4 summarizes the findings in this paper, compares the results with existing literature sources and outlines future research plans.

2. Methods

The source dataset comprises all the outpatient records of 501,065 patients of all ages with diabetes (42,452 Type 1; 458,613 Type 2) collected in 2018 by the NHIF. These records were provided for research purposes to the National eHealth Scientific Program after applying pseudonymization procedures [4].

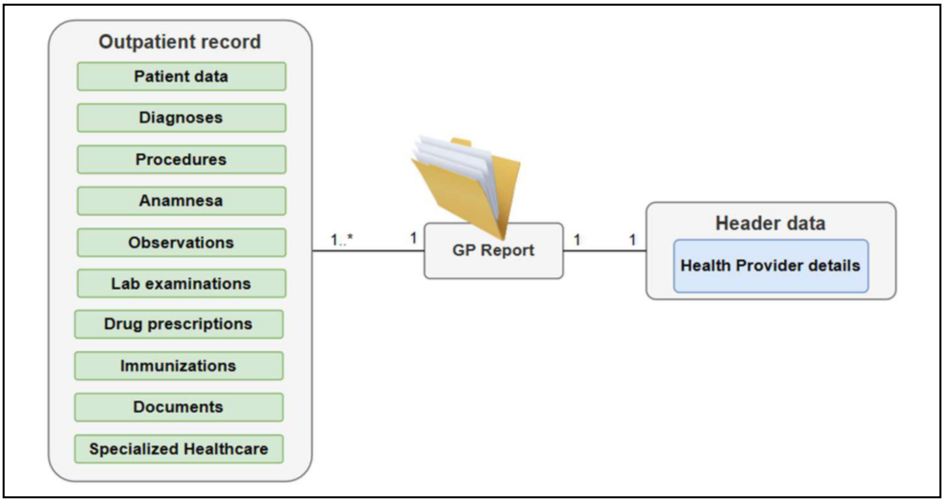


Figure 1. Patient-centric structure of the source dataset with outpatient records.

Mapping these nationally representative observational health data to OMOP CDM allows to extend the scope of shared data at international level and generate evidence using standardized analytics tools. Besides, the ETL process of mapping is enhanced by the patient-centric structure of both the source dataset (Figure 1) and the OMOP CDM. It is noteworthy, that in our case we were dealing with a large number (6,887,876) of original clinical documents in XML format whose XML structure appeared to be damaged by the pseudonymization procedure, for instance, by inserting the '&' character or various control symbols between opening and closing XML tags. Therefore, a stage of converting the pseudonymized records into valid XML documents and transforming XML data into relational database tables preceded the well-structured ETL process.

Java applications were developed for this purpose and it facilitated significantly the execution of the Extraction and Transform stages in the ETL process. For example, these applications created a relational model of the source database that retains the patient-centric structure of the outpatient records and resembles the semantics of the tables in the OMOP CDM (v. 5.3.0). On the other side, the extraction of observational health data by processing original clinical documents is a problem that requires non-traditional approach for finding a solution. It is common to discover such data recorded in raw text in sections Anamneses, Observations or Lab examinations of outpatient records. In our case, this text was provided in unstructured native (Cyrillic) language, where one and the same clinical concept appeared with different abbreviations in huge number of records and besides, sometimes with wrongly specified measurement values or units. Python scripts using regular expressions for extraction of observational health data (blood pressure, blood sugar etc.) proved to be more effective than traditional Natural Language Processing (NLP) tools in resolving extraction problems with such complexities [17].

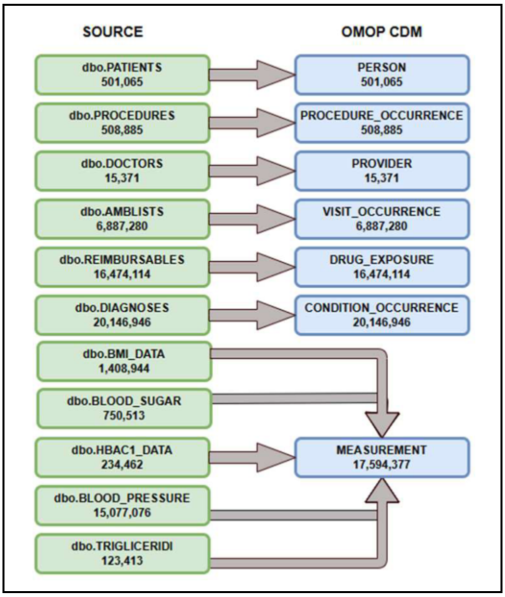


Figure 2. Data table mapping and number of rows per table after the ETL process.

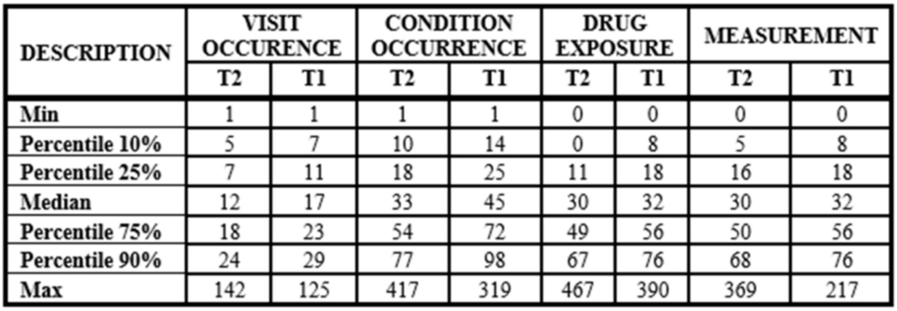
Dealing with raw data required to complement the extraction of observational health data with procedures for cleaning data and ensuring EHDEN data quality criteria (conformance, completeness and plausibility) are satisfied [11] [13] [18]. For example, blood pressure measurements might appear with a different separator ('-', ',', '/') or without any separator between the systolic and diastolic values located inside the unstructured text in elements Anamneses or Observations of the XML documents. Moreover, the systolic and diastolic values might appear exchanged. It is just an example, that illustrates the complexity of the procedures for measurement extraction.

At the end of the preprocessing stage, the source database was loaded with all the data from the original dataset and the ETL process followed the development process suggested by the OHDSI community [7]. It entails the usage of specific OHDSI tools like White Rabbit, Rabbit-in-a-Hat, Achilles, Data Quality Dashboard at each step of the process [12]. For example, the White Rabbit report stated that there were no missing values in the source database. The standardization of the medical terminologies in the source usually requires significant efforts in this process. In this use case it helped that the NHIF applied the ATC classifier for drug encoding. Similarly, the source encoding for diagnoses and procedures followed ICD10 and ICD9Proc, respectively. Data from the national specialty classifier is mapped manually to the Medicare Specialty dictionary. It significantly facilitated the translations of these common source terminologies to standard concepts using the OMOP Vocabularies [19]. Once the structural and semantic standardization design was finalized, a complete set of instructions were produced to map the source database tables to the OMOP CDM tables (Figure 2).

3. Results

The preprocessing of the original dataset described in the previous section contributed to the successful completion of the ETL process. For better efficiency, both the source database and the OMOP CDM database were hosted on the same instance of Microsoft SQL Server. It enabled direct data extraction and transformation via SQL scripts making the procedure to load the OMOP CDM database stepwise and fully reproducible.

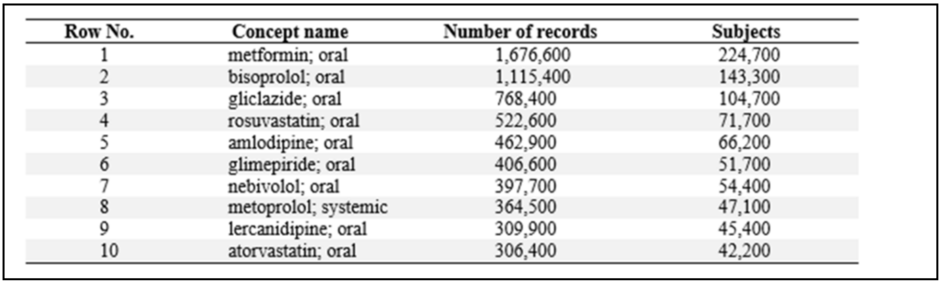
Table 1. Distribution of records per person with diabetes Type 1 (T1) and Type 2 (T2) in major data domains.



DESCRIPTION	VISIT OCCURENCE		CONDITION OCCURRENCE		DRUG EXPOSURE		MEASUREMENT	
	T2	T1	T2	T1	T2	T1	T2	T1
Min	1	1	1	1	0	0	0	0
Percentile 10%	5	7	10	14	0	8	5	8
Percentile 25%	7	11	18	25	11	18	16	18
Median	12	17	33	45	30	32	30	32
Percentile 75%	18	23	54	72	49	56	50	56
Percentile 90%	24	29	77	98	67	76	68	76
Max	142	125	417	319	467	390	369	217

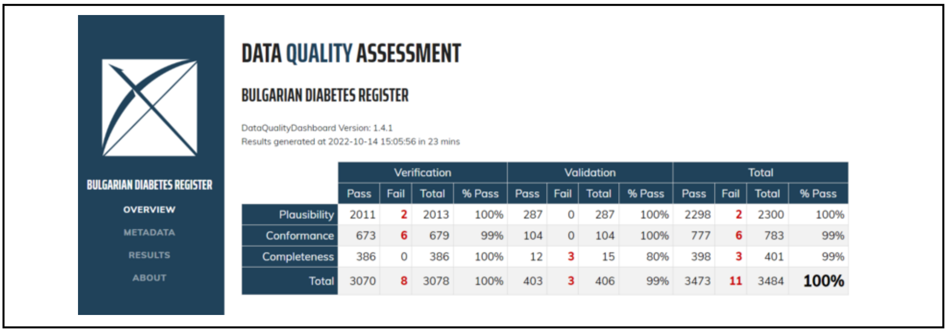
This way, all the clinical documents (100%) from the source dataset for 501,065 patients with diabetes were mapped to the OMOP CDM. Accordingly, data from the source dataset were loaded in tables of a Microsoft SQL Server database with relational model matching the OMOP CDM (Figure 2). Table 1 and Table 2 provide insight about the burden caused by the diabetes illness on the healthcare system during 2018. Although detailed analysis of the numerical data in these tables is outside the scope of this paper, they are good example for the potential benefits of using the OMOP CDM in research.

Table 2. Top 10 of the mapped drugs.



Row No.	Concept name	Number of records	Subjects
1	metformin; oral	1,676,600	224,700
2	bisoprolol; oral	1,115,400	143,300
3	gliclazide; oral	768,400	104,700
4	rosuvastatin; oral	522,600	71,700
5	amlodipine; oral	462,900	66,200
6	glimepiride; oral	406,600	51,700
7	nebivolol; oral	397,700	54,400
8	metoprolol; systemic	364,500	47,100
9	lercanidipine; oral	309,900	45,400
10	atorvastatin; oral	306,400	42,200

At the end of the ETL process we ran the Achilles and the Data Quality Dashboard tools provided by the OHDSI community for data quality assessment. Figure 3 shows the results of the quality checks executed by the Data Quality Dashboard. These results prove (1) conformance with OMOP CDM standards and format; (2) plausibility i.e. data values are valid with respect to integrity rules for domains, concept classes and vocabulary IDs and (3) completeness, meaning that data values are present.



DATA QUALITY ASSESSMENT

BULGARIAN DIABETES REGISTER

DataQualityDashboard Version: 1.4.1
Results generated at 2022-10-14 15:05:56 in 23 mins

	Verification				Validation				Total			
	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass
Plausibility	2011	2	2013	100%	287	0	287	100%	2298	2	2300	100%
Conformance	673	6	679	99%	104	0	104	100%	777	6	783	99%
Completeness	386	0	386	100%	12	3	15	80%	398	3	401	99%
Total	3070	8	3078	100%	403	3	406	99%	3473	11	3484	**100%**

Figure 3. The Total summary result of executing the DataQualityDashboard v1.4.1 tool is 100% pass.

The quality assessment procedures confirm the successful completion of the ETL process for mapping the BDR in OMOP CDM. It is among the 153 OMOP CDM from 28 countries that are currently published in the EHDEN Portal [16].

4. Discussion

Mapping of the BDR to the OMOP CDM resolves interoperability issues across similar data sources by standardizing management of observational health data in terms of common structure (data model) and terminology (vocabulary). Unlike other papers [5] [9], this study addresses the usage of data quality assessment procedures supporting high-quality in evidence generation. Preprocessing of the source dataset and the extraction of measurements of observational health data were the major challenges to overcome. The mapping of a nationally representative sample of outpatient records of patients with diabetes to OMOP CDM allows us to join the EHDEN community and apply standard analytical tools in the large-scale EHDEN MegaStudy project entitled “Studying Drug shortages in Europe: A Multinational, Multidatabase Network Study”.

Acknowledgement

This research is supported by Grant 80-10-8/11.04.2023 of Sofia University, Bulgaria.

References

[1] Bak JCG, Serné EH, Kramer MHH, Nieuwdorp M, Verheugt CL. National diabetes registries: do they make a difference? Acta diabetologica. 2021; 58(3): 267–278.
[2] Boytcheva S, Angelova G, Angelov Z, Tcharaktchiev D. Data Mining and Analytics for Exploring Bulgarian Diabetic Register. In Data Analytics and Management in Data Intensive Domains.: Springer International Publishing; 2018.
[3] University Specialized Hospital for Active Treatment in Endocrinology. Diabetes Register. [Online].; 2024 [cited 2024 January 11. Available from: https://usbale.org/bg/registar-zaharen-diabet/.
[4] National Scientific Program eHealth. National Scientific Program "Electronic Healthcare in Bulgaria" (e-health). [Online].; 2024 [cited 2024 January 10. Available from: https://ehealth.fmi.uni-sofia.bg/.
[5] Korntheuer RL, Katsch F, Duftschmid G. Transforming Documents of the Austrian Nationwide EHR System into the OMOP CDM. In Pfeifer B, et. al.. Volume 301: dHealth 2023, Studies in health technology and informatics (dHealth 2023).: IOS Press; 2023. p. 54–59.
[6] Observational Health Data Sciences and Informatics. OMOP Common Data Model. [Online].; 2024 [cited 2024 January 10. Available from: https://ohdsi.github.io/CommonDataModel/.
[7] Observational Health Data Sciences and Informatics. The Book of OHDSI: https://ohdsi.github.io/TheBookOfOhdsi/; 2021.
[8] Kent S, Burn E, Dawoud D, Jonsson P, Østby JT, Hughes N, et al. Common Problems, Common Data Model Solutions: Evidence Generation for Health Technology Assessment. Pharmacoeconomics. 2021; 39(3): 275-285.
[9] Reinecke I, Zoch M, Reich C, Sedlmayr M, Bathelt F. The Usage of OHDSI OMOP - A Scoping Review. In German Medical Data Sciences 2021: Digital Medicine: Recognize – Understand – Heal, Studies in health technology and informatics.: IOS Press; 2021. p. 95–103.
[10] EHDEN. A federated network of Data Partners. [Online].; 2022 [cited 2024. Available from: https://www.ehden.eu/datapartners/.
[11] Blacketer C, Voss EA, DeFalco F, Hughes N, Schuemie MJ, Moinat M, et al. Using the Data Quality Dashboard to Improve the EHDEN Network. Applied Sciences. 2021; 11(24): 11920.
[12] Observational Health Data Sciences and Informatics. Software Tools. [Online].; 2024 [cited 2024 January 10. Available from: https://www.ohdsi.org/software-tools/.
[13] Voss EA, Blacketer C, van Sandijk S, Moinat M, Kallfelz M, van Speybroeck M, et al. European Health Data & Evidence Network-learnings from building out a standardized international health data network. Journal of the American Medical Informatics Association : JAMIA. 2024; 31(1): 209–219.
[14] 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 medical research methodology. 2021; 21(1): 238.
[15] Krastev E, Markov E. EHDEN Community Calls. [Online].; 2023 [cited 2024 January 12. Available from: https://www.ehden.eu/bulgarian-diabetes-register-mapping-onto-omop-cdm-cc10/.
[16] Krastev E, Tcharaktchiev D, Abanos S. Application of OMOP Common Data Model for Data Integration: The Bulgarian Diabetes Register. In Volume 309: Telehealth Ecosystems in Practice, Studies in Health Technology and Informatics. Torino, Italy: IOS Press; 2023. p. 141 - 142.
[17] Wiesmüller F, Hayn D, Kreiner K, Pfeifer B, Pölzl G, Kastner P, et al. Natural Language Processing for Free-Text Classification in Telehealth Services: Differences Between Diabetes and Heart Failure Applications. In Volume 279: Navigating Healthcare Through Challenging Times, Studies in Health Technology and Informatics (dHealth 2021).: IOS Press; 2021. p. 157 - 164.
[18] Brownlee J. Data Preparation for Machine Learning. Data Cleaning, Feature Selection, and Data Transforms in Python: Machine Learning Mastery; 2020.
[19] Reich C, Ostropolets A, Ryan P, Rijnbeek P, Schuemie M, Davydov A, et al. OHDSI Standardized Vocabularies-a large-scale centralized reference ontology for international data harmonization. Journal of the American Medical Informatics Association : JAMIA. 2024; 31(3): ocad247.