German Medical Data Sciences 2021: Digital Medicine: Recognize - Understand - Heal
R. Röhrig et al. (Eds.)

doi:10.3233/SHTI210545

Towards the Representation of Genomic Data in HL7 FHIR and OMOP CDM

Yuan PENG[a,1], Azadeh NASSIRIAN[a], Najia AHMADI[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. High throughput sequencing technologies have facilitated an outburst in biological knowledge over the past decades and thus enables improvements in personalized medicine. In order to support (international) medical research with the combination of genomic and clinical patient data, a standardization and harmonization of these data sources is highly desirable. To support this increasing importance of genomic data, we have created semantic mapping from raw genomic data to both FHIR (Fast Healthcare Interoperability Resources) and OMOP (Observational Medical Outcomes Partnership) CDM (Common Data Model) and analyzed the data coverage of both models. For this, we calculated the mapping score for different data categories and the relative data coverage in both FHIR and OMOP CDM. Our results show, that the patients genomic data can be mapped to OMOP CDM directly from VCF (Variant Call Format) file with a coverage of slightly over 50%. However, using FHIR as intermediate representation does not lead to further information loss as the already stored data in FHIR can be further transformed into OMOP CDM format with almost 100% success. Our findings are in favor of extending OMOP CDM with patient genomic data using ETL to enable the researchers to apply different analysis methods including machine learning algorithms on genomic data.

Keywords. Genomic data, FHIR, OMOP CDM, VCF

1. Introduction

Digitalization plays an increasingly important role in modern Healthcare. The digitally stored data however needs to be interoperable between different databases or systems, which is not yet realized for most of the medical data [1]. To address this issue, the German government has started the Medical Informatics Initiative (MII) to enable collaboration on data among all German University Hospitals and to promote as well as further enable the secondary use of data for research [2]. In this context, the National Steering Committee of the MII has specified profiles based on the R4 version of FHIR (Fast Healthcare Interoperability Resources)[2] as a national communication standard. FHIR itself is an international standard for the exchange of electronic health records, which was introduced by HL7 in 2014 and is since increasingly used in medical information technology [3]. In order to store and analyze data for research questions, MIRACUM (Medical Informatics in Research and Care in University Medicine), as one

[1] Corresponding Author: Yuan Peng, Institute for Medical Informatics and Biometry at Carl GustavCarus Faculty of Medicine at Technische Universität Dresden, Fetscherstraße 74, 01407 Dresden, Germany; E-mail: yuan.peng@tu-dresden.de.

[2] https://www.hl7.org/fhir/

of four consortia of the MII, uses OMOP (Observational Medical Outcomes Partnership)[3] CDM (Common Data Model) as research repository. OMOP CDM, which has been developed by OHDSI (Observational Health Data Sciences and Informatics)[4], is a relational data model that comprises standardized vocabularies to harmonize data from different sources. Due to its strict structural specification, including the need for using standard vocabularies a true standardization is supported. In contrast to that, the communication standard FHIR pro-vides a set of modular components called "Resources" to define various medical concepts and by that is less restrictive in terms of its structure as so-called extensions and the use of non-standard vocabularies are allowed. However, both FHIR and OMOP CDM can be used to store patient and clinical data.

Besides patients' metadata, genomic data is also of increasing importance to clinical care and secondary analysis, especially in the personalized treatment of tumor disease [4, 5]. In the past few years NGS (Next Generation Sequencing) is widely used in cancer researches [6]. The result of NGS from a cancer research can be reported using a tab separated data format called, VCFs (Variant Call Format), which is one of the most commonly used file format for presenting the result of a sequencing process [7, 8]. Each VCF file contains information of all variants of a person. The information of a variant is presented in each row, via chromosome number, location of the variant on the chromosome and a unique identifier for the variant. The identifier is usually an rs id from dbSNP (Single Nucleotide Polymorphism Database)[5], but the id from other databases could also be used here. It would be highly desirable to harmonize and standardize genomic patient data taken from VCF files with patient data taken from clinical systems as research base using OMOP CDM.

Originally OMOP CDM was not supporting genetic data, until 2019, when G-CDM (Genome Common Data Model), an extension of OMOP CDM to store NGS analysis data, was published [9]. Genomic_test, Target_gene, Variant_occurrence, and Variant_annotation are the four tables introduced by G-CDM, to store NGS analysis data from CSV (comma separated values) files. Based on the previous work of G-CDM, the OHDSI genetic work group has successfully built a new vocabulary using a nomenclature called HGNC (HUGO Gene Nomenclature Committee)[6] and variants information extracted from different gene-bank [10], to enable genetic data storage in OMOP using MEASUREMENT table. MEASUREMENT table is a part of standard CDM used to store the records of measurements for a certain patient e.g., laboratory tests, vital signs, quantitative findings from pathology reports, etc.

To be as independent as possible from the source systems, the benefits of FHIR as communication standard and the extensive activities should be used. Since the first standard Genetics profile was developed in 2014, the HL7 Clinical Genomics Work Group[7] has been trying to integrate genomic data into clinical care and support interoperable ex-change of genomic data using a resource named MolecularSequence. This resource describes an atomic sequence which was designed for storing the alignment sequencing test result and multiple variations[8], such as VCF files. Besides MolecularSequence profile, there are many different extended profiles that could also be

[3] https://www.ohdsi.org/data-standardization/

[4] https://www.ohdsi.org/

[5] https://www.ncbi.nlm.nih.gov/snp/

[6] https://www.genenames.org/

[7] https://www.hl7.org/fhir/genomics.html

[8] https://www.hl7.org/fhir/molecularsequence.html

used for storing VCF data. For example, the Observation-genetics profile is extended from Observation resource and is used to interpret variants from the sequence resource [11]. A tool, namely VCF2FHIR, is implemented based on the profiles, which are defined in General Genomic Reporting from Genomics Reporting Implementation Guide for converting VCF data into HL7 FHIR R4 format[12]. This tool is also part of the SMART Cancer Navigator, which needs to be connected to an EHR system [13].

In this work, we analyze the potential data coverage of VCF data in OMOP CDM with and without using FHIR as intermediate layer. Therefore, we distinguish between

- VCF data→OMOP CDM, using OMOP G-CDM
- VCF data→OMOP CDM, using standard MEASUREMENT table
- VCF data→FHIR→OMOP CDM, using OMOP G-CDM
- VCF data→FHIR→OMOP CDM, using standard MEASUREMENT table

2. Method

In order to determine the data coverage in OMOP CDM, we have used data from PPGL (Pheochromocytomas and paragangliomas) pipeline [14] for creating semantic mappings. The data from PPGL pipeline contains patient information, blood analysis, tumor analysis and VCF data from each patient. Since some types of cancer could also be inherited, the disease histories of family members are also included. Additionally, the blood test and tumor laboratory result, including amino acid and nucleotide alteration along with the specific gene names, are recorded.

2.1. Semantic mappings

Based on the source data, we have designed two mapping tables for OMOP CDM. One uses G-CDM for storing genomic data (Figure 1) and the other one uses the standard MEASUREMENT table (Figure 2). As for FHIR profiles, we have decided to use MolecularSeqeunce resource, Observation-Genetics and FamilyMemberHistory-Genetic profiles to store tumor analysis data based on our test data (Figure 3, Figure 4).We have then transformed the data stored in FHIR profiles to both versions of OMOP CDM (G-CDM and MEASURE-MENT tables).

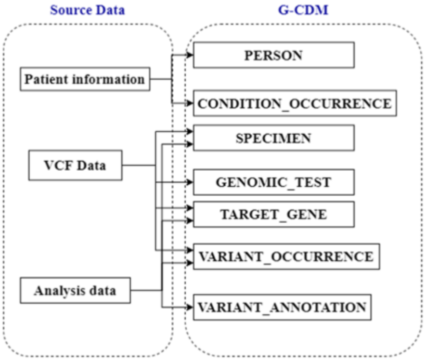

Figure 1. Data Mapping Concept for OMOP using G-CDM Table.

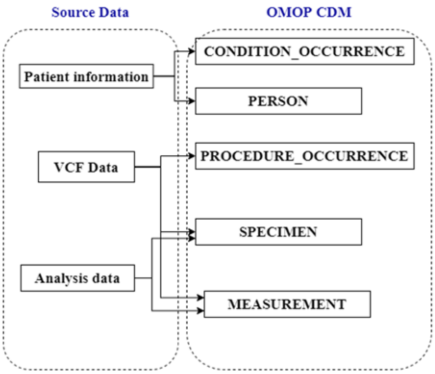

Figure 2. Data Mapping Concept for OMOP using MEASUREMENT Table.

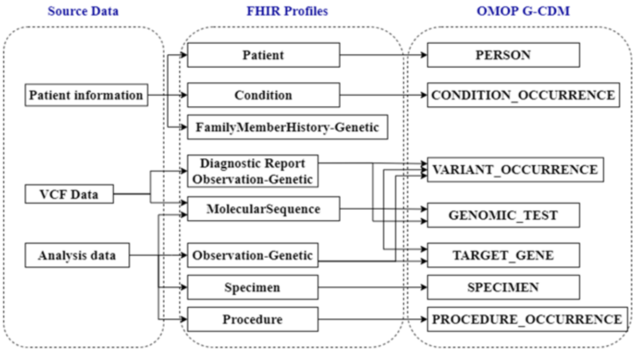

Figure 3. Data Mapping Concept for FHIR to OMOP using G-CDM.

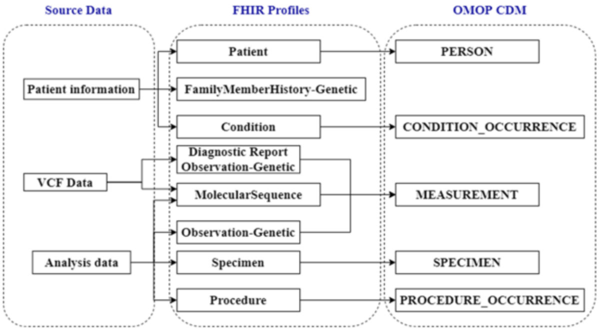

Figure 4. Data Mapping Concept for FHIR to OMOP using MEASUREMENT.

2.2. Evaluation of mappings

To evaluate the data coverage in FHIR and OMOP CDM, we designed a score system, which is inspired from the field of bioinformatics [15]. If a source data element can be stored in the target system, it will be given a score one, otherwise a score zero. This score system was implemented for 9 different combinations of transforming three source files into FHIR and into two different versions of OMOP CDM (Table 1). According to the scores for each data element, we calculate the percentage of data coverage for each category. The higher the scores of the system is, the higher is the data coverage rate in the target system. Additionally, we looked into each not-mappable data element, and its importance for the evaluation of cancer studies.

Table 1. Comparison criteria

Target system	Data type
FHIR	Patient information, VCF, Analysis data
OMOP CDM (1st and 2nd version)	Patient information, VCF, Analysis data

3. Results

3.1. Semantic mapping tables

Each target standard system has three different mapping tables namely, patient information data mapping table, VCF mapping table and analysis data mapping table. Each mapping table contains information from source data and target system, as shown in the example table (Table 2).

Table 2. An example of semantic mapping table

Column Name	Data	FHIR	OMOP
Patient ID	1	Patient.id	PERSON.person_source_value
Gender	f	Patient.gender	PERSON.gender_source_value PERSON.gender_concept_id
Diagnosis	Pheochromocytoma	Condition.code.coding.display Condition.code.coding.code	CONDITIONOCCURRENCE.conditionsourcevalue CONDITIONOCCURRENCE.conditionconceptid CONDITIONOCCURRENCE.conditionsourceconceptid

3.2. Mappability

Based on the semantic mapping table and the score system, we have calculated the mapping scores for the 9 criteria mentioned before. The results are presented in both: Relation between number of mapped elements (blue) and not mappable/missing elements (orange) as well as the relative coverage of mapped elements in percentage (Figure 5). In our analysis, FHIR receives relatively higher score than OMOP CDM for storing VCF data since the MolecularSequence, Observation-genetics and other extensions for genetics in FHIR are designed for storing those kind of data. Moreover, the G-CDM is designed for storing both VCF data and cancer analysis data, therefore the data coverage in G-CDM is also higher than it using MEASUREMENT table in both categories. The only not-mappable data element in OMOP CDM is the id of the variant from a public gene database, how-ever this information is not necessarily recorded in the VCF file, if the variant not exists in any gene database. As of patient information data and cancer analysis data, some data elements from the source data of PPGL pipeline are quite particular for this study only, since these two file formats are simple tab separated files. Therefore this can also lead to data loss in both FHIR and OMOP[9]. The missing information from nucleotide change in FHIR could be retrieved from the concept name of the HGNC vocabulary.

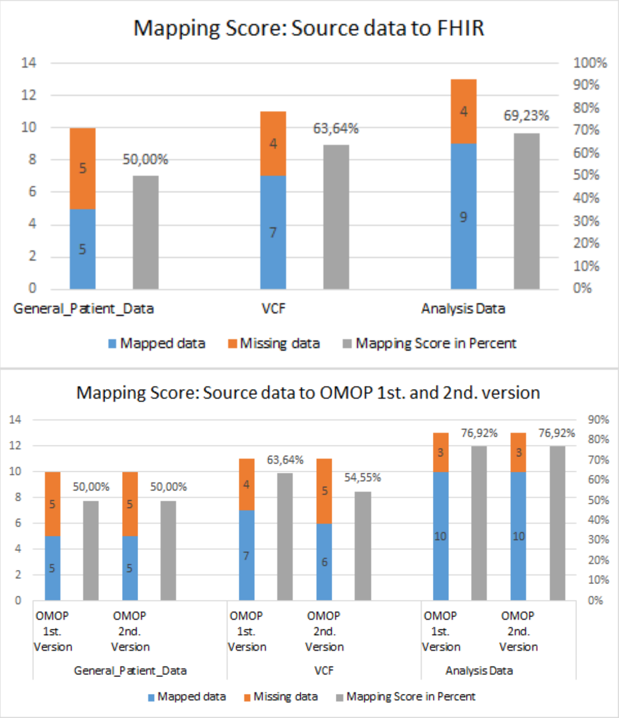

Figure 5. Mapping result from source data to target system.

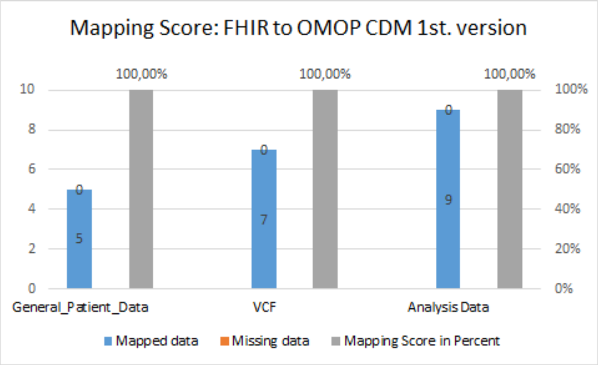

Figure 6. Mapping result from FHIR to 1st version of OMOP (G-CDM).

[9] Supplemental File:https://caruscloud.uniklinikum-dresden.de/index.php/s/88grQzSYRm3TkaZ

However, using FHIR as an intermediate representation between the source data and OMOP CDM, this leads into really promising results (Figure 6, Figure 7). As stated previously, the only missing data is the variant id from VCF file, which, via combination of CHROM, POS, REF and ALT from VCF, the exact variant can also be found in the gene database. All data that already stored in FHIR can be transformed into OMOP CDM without further data loss.

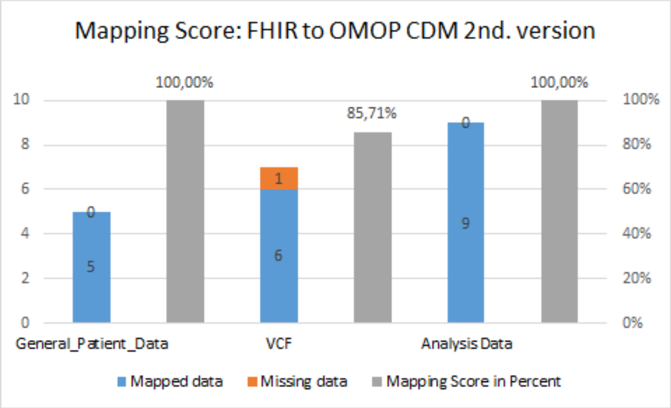

Figure 7. Mapping result from FHIR to 2nd. Version of OMOP (MEASUREMENT).

4. Discussion

FHIR and OMOP are the two most used standards in the field of Medical informatics. As the need of integrating genomic data into EHR is increasing, it becomes more important to transform the genomic data, especially the cancer analysis data, into these two standards. There are a few limitations when using MEASUREMENT table for storing cancer analysis data. The new vocabulary HGNC contains only the variants of certain gene databases, such as CIVic, ClinVar, and NCIt[10]. And databases such as dbSNP and COSMIC[11] are not included in the new HGNC vocabulary. This can cause data loss in researches that are using excluded databases. But, this problem can be solved using a self-generated vocabulary of desired databases. Although it is already possible to store genomic data using FHIR standards and apply some precision medicine modules on the data using SMART on FHIR[12] [5], the mainly supported FHIR version is DSTU2, which is not compatible with the FHIR version R4 that we are using. Therefore we suggest OMOP CDM as the standard to store the genomic data, use this data for further analysis and thus ensure our active role in medical studies as part of the OHDSI community in the future.

Findings of our research indicate that while FHIR and OMOP CDM as two known standards for storing medical data have the potential to solve many data sharing problems, each has limitation when it comes to genomic data. FHIR can be used as an intermediate data point between raw genomic data and OMOP CDM, for its high compatibility with

[10] https://athena.ohdsi.org/search-terms/terms?vocabulary=Vocabulary&page=1&pageSize=15&query=

[11] https://cancer.sanger.ac.uk/cosmic

[12] https://apps.smarthealthit.org/apps/category/genomics

our test data. The data can be stored into OMOP using ETL (Extract, Transform and Load) process, which is our next step of this research project. Storing genomic data in OMOP CDM enables analysis studies using machine learning methods that can be used in early prediction and diagnosis and improvement of personalized cancer care [4, 16].

Declarations

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

Authors Contribution: YP, AN, NA: conception of the work. All authors contributed substantial ideas 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.

Acknowledgement: Many thanks to Ms. Doreen William for preparing the cancer study data.

References

[1] M. Lehne, J. Sass, A. Essenwanger, J. Schepers, and S. Thun, Why digital medicine depends on interoperability, *Npj Digital Medicine*. **2** (2019) 1–5. doi:10.1038/s41746-019-0158-1.

[2] S.C. Semler, F. Wissing, and R. Heyder, German Medical Informatics Initiative, *Methods Inf Med*. **57** (2018) e50–e56. doi:10.3414/ME18-03-0003.

[3] 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.

[4] H.K. Brittain, R. Scott, and E. Thomas, The rise of the genome and personalised medicine, *Clin Med*. **17** (2017) 545–551. doi:10.7861/clinmedicine.17-6-545.

[5] J.L. Warner, S.K. Jain, and M.A. Levy, Integrating cancer genomic data into electronic health records, *Genome Med*. **8** (2016). doi:10.1186/s13073-016-0371-3.

[6] C. Meldrum, M.A. Doyle, and R.W. Tothill, Next-Generation Sequencing for Cancer Diagnostics: a Practical Perspective, *Clin Biochem Rev*. **32** (2011) 177–195.

[9] S.J. Shin, S.C. You, Y.R. Park, J. Roh, J.-H. Kim, S. Haam, C.G. Reich, C. Blacketer, D.-S. Son, S. Oh, and R.W. Park, Genomic Common Data Model for Seamless Interoperation of Biomedical Data in Clinical Practice: Retrospective Study, *Journal of Medical Internet Research*. **21** (2019) e13249. doi:10.2196/13249.

[10] R. Belenkaya, M.J. Gurley, A. Golozar, D. Dymshyts, R.T. Miller, A.E. Williams, S. Ratwani, A. Siapos, V. Korsik, J. Warner, W.S. Campbell, D. Rivera, T. Banokina, E. Modina, S. Bethusamy, H.M. Stewart, M. Patel, R. Chen, T. Falconer, R.W. Park, S.C. You, H. Jeon, S.J. Shin, and C. Reich, Extending the OMOP Common Data Model and Standardized Vocabularies to Support Observational Cancer Research, *JCO Clinical Cancer Informatics*. (2021) 12–20. doi:10.1200/CCI.20.00079.

[11] B. Ryu, S.-Y. Shin, R.-M. Baek, J.-W. Kim, E. Heo, I. Kang, J.S. Yang, and S. Yoo, Clinical Genomic Sequencing Reports in Electronic Health Record Systems Based on International Standards: Implementation Study, *J Med Internet Res*. **22** (2020). doi:10.2196/15040.

[12] R.H. Dolin, S.R. Gothi, A. Boxwala, B.S.E. Heale, A. Husami, J. Jones, H. Khangar, S. Londhe, F. Naeymi-Rad, S. Rao, B. Rapchak, J. Shalaby, V. Suraj, N. Xie, S. Chamala, and G. Alterovitz, vcf2fhir: a utility to convert VCF files into HL7 FHIR format for genomics-EHR integration, *BMC Bioinformatics*. **22** (2021) 104. doi:10.1186/s12859-021-04039-1.

[13] J.L. Warner, I. Prasad, M. Bennett, M. Arniella, A. Beeghly-Fadiel, K.D. Mandl, and G. Alterovitz, SMART Cancer Navigator: A Framework for Implementing ASCO Workshop Recommendations to Enable Precision Cancer Medicine, *JCO Precis Oncol*. **2** (2018). doi:10.1200/PO.17.00292.

[14] L. Gieldon, D. William, K. Hackmann, W. Jahn, A. Jahn, J. Wagner, A. Rump, N. Bechmann, S. Nölting, T. Knösel, V. Gudziol, G. Constantinescu, J. Masjkur, F. Beuschlein, H.J. Timmers, L. Canu,

K. Pacak, M. Robledo, D. Aust, E. Schröck, G. Eisenhofer, S. Richter, and B. Klink, Optimizing Genetic Workup in Pheochromocytoma and Paraganglioma by Integrating Diagnostic and Research Approaches, *Cancers*. **11** (2019) 809. doi:10.3390/cancers11060809.

[15] D.A. Hendrix, Chapter 3: Sequence Alignments, in: Applied Bioinformatics, n.d.

[16] M. Khalilia, M. Choi, A. Henderson, S. Iyengar, M. Braunstein, and J. Sun, Clinical Predictive Modeling Development and Deployment through FHIR Web Services, *AMIA Annu Symp Proc*. **2015** (2015) 717–726.