German Medical Data Sciences: Bringing Data to Life
R. Röhrig et al. (Eds.)

doi:10.3233/SHTI210055

openEHR Mapper – A Tool to Fuse Clinical and Genomic Data Using the openEHR Standard

Niklas REIMER [a,1], Hannes ULRICH [b], Hauke BUSCH [c], Ann-Kristin KOCK-SCHOPPENHAUER [b] and Josef INGENERF [a,b]
[a] *Institute of Medical Informatics, University of Lübeck, Germany*
[b] *IT Center for Clinical Research (ITCR-L), University of Lübeck, Germany*
[c] *Institute for Experimental Dermatology, University of Lübeck, Germany*

Abstract. Precision medicine is an emerging and important field for health care. Molecular tumor boards use a combination of clinical and molecular data, such as somatic tumor mutations to decide on personalized therapies for patients who have run out of standard treatment options. Personalized treatment decisions require clinical data from the hospital information system and mutation data to be accessible in a structured way. Here we introduce an open data platform to meet these requirements. We use the openEHR standard to create an expert-curated data model that is stored in a vendor-neutral format. Clinical and molecular patient data is integrated into cBioPortal, a warehousing solution for cancer genomic studies that is extended for use in clinical routine for molecular tumor boards. For data integration, we developed *openEHR Mapper*, a tool that allows to (i) process input data, (ii) communicate with the openEHR repository, and (iii) export the data to cBioPortal. We benchmarked the mapper performance using XML and JSON as serialization format and added caching capabilities as well as multi-threading to the *openEHR Mapper*.

Keywords. openEHR, tumor board, HiGHmed, cBioPortal, Health Information Interoperability

1. Introduction

Precision medicine is envisioned and the future of patient care in the healthcare sector, especially for personalized cancer treatment using genomic data. Due to reduced sequencing costs and turn-around time for analysis, panel, exome, and genome sequencing have found their way into treatment routine. Such data is usually discussed in so-called molecular tumor boards that have developed from specialized entity-specific tumor boards. Case discussions involve multiple parties and require integration of clinical and patient data with diagnostic results and mutation information that are needed to be extracted from the hospital information system (*HIS*) and other relevant clinical and research systems [1]. However, patient and diagnostic data is currently stored in proprietary formats maintained by various vendors, often preventing direct database access. Instead, information must be accessed using proprietary interfaces or those based

[1] Corresponding Author: Niklas Reimer, Institute of Medical Informatics, Ratzeburger Allee 160, 23562 Lübeck, Germany; E-mail: niklas.reimer@student.uni-luebeck.de.

on the older underspecified HL7 V2 standard [3]. In order to open the vendor-locked information, the openEHR [2] community develops flexible and reliable health platforms that integrate interfaces between research and routine care as well. The aim is to establish an open platform that provides direct access to the data. This is also the intention of the HiGHmed consortium [4] which is part of the Medical Informatics Initiative Germany [5]. HiGHmed uses openEHR as their primary platform for cross-institutional data analysis on data of participating locations. A viable extension to the HiGHmed infrastructure and as a cross-consortium application we therefore determined the possibility to fuse HiGHmed's openEHR platform with the tumor board application based on the data model of the MIRACUM consortium [6]. This work on openEHR Mapper shall demonstrate that interoperability within the Medical Informatics Initiative is possible across the participating consortia, instead of establishing multiple distinct networks of interoperability.

To unlock the potential of an open platform, various patient and diagnostic information must be integrated, e.g., tumor mutation data must be reconciled with the patients' current condition and treatment history. Theoretically, this can be done in three steps: first, the patient's data is extracted from the HIS and stored together with the annotated tumor genome in the openEHR-based platform [1]. Next, the tool supporting the molecular tumor board accesses and visualizes the data from the openEHR platform. Lastly, therapy recommendations of the molecular tumor board are transferred from the supporting tool back into the openEHR repository and eventually into the local HIS. Here, we focus on the first step of the proposed workflow.

2. Material and Methods

2.1. openEHR

openEHR is an open standard for Electronic Health Records (*EHR*s) based on the ISO 13606 standard. It defines a reference model and the Archetype Definition Language enabling domain experts to model basic concepts, i.e., clinical parameters like blood pressure in archetypes. Archetypes can be combined in blocks to build a template that represents a more complex concept like a report. This architecture allows using the same archetype for multiple purposes as well as sharing archetypes and templates on national or international platforms like the *Clinical Knowledge Managers* [7].

For storing clinical information in an openEHR repository, the data is assembled as *Composition*s and uploaded via a RESTful interface. Data inside the repository can be accessed using the Archetype Query Language (*AQL*), a special language for openEHR with a syntax similar to query languages like SQL. To ensure data quality and interoperability, openEHR offers terminology binding for local terminologies that are provided by the archetypes themselves as well as bindings to external terminologies like LOINC or SNOMED CT.

2.2. Clinical data and whole-exome sequencing (WES)

Clinical data contains demographic information like sex, age, ICD-O codes, gradings, histologic results, and information about surgeries. It is provided in a structured message format based on HL7 V2. In the particular case of the authors' local environment, the

clinical data for cancer patients is based on the ADT/GEKID standard [8]. The technology of next-generation sequencing has revolutionized life sciences for the last 15 years. The exome is the sequence within the human DNA that codes for proteins, yet makes up only about 1.1 % of the three billion bases of the human genome [9]. It can be sequences as so-called Whole Exome Sequencing (WES) at much less effort than the whole genome. Still, it contains most of the pathogenetic mutations relevant for tumor progression making it a reliable method for the diagnostic of cancer. Sequencing data is stored in *FASTQ* files that contain an identifier, the sequence of nucleobases, and a quality value for each base. WES results in two files (reads), having a size of about 4 gigabytes each. In the field of oncology, WES is used to analyze differences between a sample from healthy tissue and tumor tissue. The WES data sets used to evaluate our work were obtained from Open-Access Data provided by the Texas Cancer Research Biobank (*TCRB*) [10]. Bioinformatic analysis of WES generally consists of aligning the reads from the FASTQ files to a reference genome, subsequent variant calling to determine the somatic, i.e. the acquired mutations of the tumor and annotation of the variants in terms of their effect on protein structure and function. In November 2019 the MIRACUM-Pipe was released as a part of the Medical Informatics Initiative Germany [11]. It provides a fully automatic pipeline that integrates various analysis tools for alignment, variant calling, and annotation of genomic raw data. While the results of each step can be used for further processing, the pipeline also generates a human-readable report.

2.3. cBioPortal

For warehousing of cancer genomic studies, the Memorial Sloan Kettering Cancer Center maintains the cBioPortal project [12][13]. It provides the functionality to compare different cancer studies and visualizes clinical and mutation information as well as image data. In addition to displaying the data, cBioPortal is also capable of integrating online information from databases like OncoKB to provide additional information about the impact of mutations. cBioPortal has the capability of data exchange and import. Data must be provided in a format called *cancer study* **Error! Reference source not found.**. It contains an identifier and the tumor entity in a metadata file, the annotated mutations, e.g., from the MIRACUM-Pipe described in 2.12.2, and the clinical data from the openEHR repository exported to tab-separated files.

2.4. Evaluation Methods

The evaluation of the mapping results covers the following aspects:

1. Which attributes of the clinical data could be mapped?
2. Which attributes of the genetic data could be mapped?
3. How is the difference in performance, depending on the output format?
4. How well does the mapping scale using multiple threads?

3. Results

We designed and implemented a tool called *openEHR Mapper* to support the integration of clinical data into the cBioPortal data warehouse and visualization tool. The integration has three parts as shown in Figure 1.

Our implementation of the *openEHR Mapper* is based on Java and uses the *Archie* library [15] as an implementation of the openEHR reference model on the client-side. For the server-side, we chose EHRbase [16], a server that implements the openEHR standard, but can also be applied to other openEHR platforms that are compliant to the most recent version 1.0.4 of the openEHR reference model [17] and openEHR REST API 1.0.1.

The first step is mapping the input schemes of the clinical and the genetic data processed by the MIRACUM-Pipe to the openEHR templates. The second part fuse and serialize the previously generated model objects [17] and templates into a composition. The data is then submitted to the openEHR repository and is available to other (third-party) applications. In a last step, we transform the processed data from the two source systems into a tab-separated file format that is required by cBioPortal.

Step 1 Since openEHR provides a strict terminology, each attribute bound to a specific value set is being mapped using lookup tables that store the corresponding values. Attributes that are not linked to the terminology like the attributes from ADT/GEKID can be assigned directly. Table 1 gives an overview of the clinical attributes we focused on. The annotated genetic data is provided by the MICRACUM-Pipe and is separately provided in two different formats: a *VCF* file describing all somatic mutations and an *extended MAF* file containing the annotated mutations.
In order to import the genetic information, the openEHR template *Molecular Pathology Report,* provided by the HiGHmed consortium [18][19], is used. The template processes most of the VCF file attributes directly and maintains additional information about the sample and the way it was processed. The *MAF* file can be directly processed by cBioPortal. Table 2 lists the attributes, we mapped from the VCF file to the openEHR template.

Step 2 The openEHR Mapper combines the mapped data and a template file into a composition. The template file contains the structure, data types, and terminologies of the contained attributes. An implemented generator parses the template dynamically and builds objects containing the previously mapped attributes. Afterward, the information model objects are serialized and committed to the openEHR repository. The attributes of the clinical data were mapped, as shown in Table 1. As the openEHR specification [2] offers XML and JSON for serializing compositions, both formats were implemented. For the evaluation purpose, EHR compositions of artificially oversized patients' were generated containing the somatic mutations from TCRB [10] in both formats and compared the processing time. The results are shown in Figure 2. To ensure that each attribute will only be processed once, the mapped data is stored in queues. The generator implements Java Runnables for the openEHR classes *OBSERVATION, ADMIN_ENTRY*, and *CLUSTER* to have a thread-safe data structure that enables the use of multi-threading. This improves the processing performance, and an integrated caching mechanism prevents excessive memory access as shown in Figure 3.

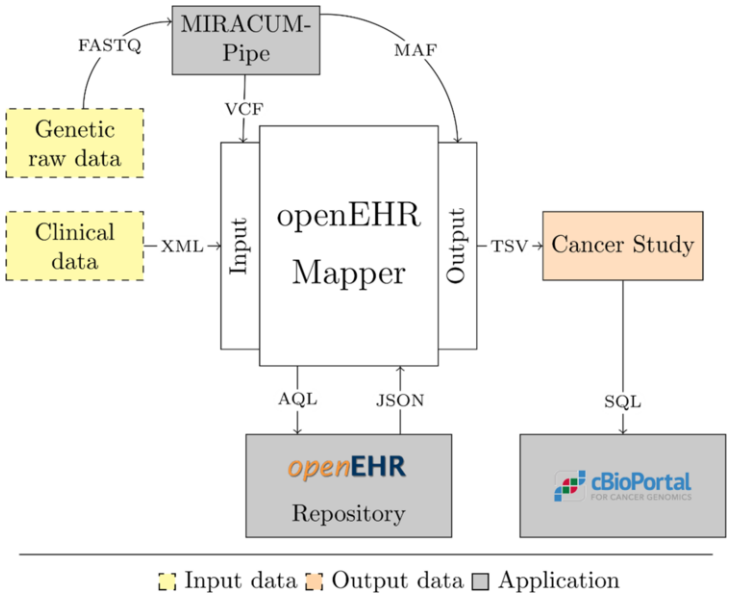

Figure 1. The Architecture of openEHR Mapper – Mapping clinical and preprocessed genetic data to the openEHR information model at the input side, querying data using AQL for generating compositions, and managing EHRs in the core component and exporting a *cancer study* for cBioPortal at the output side.

Step 3 Clinical and mutation data is fetched and stored in a specific folder structure required by cBioPortal to build a cancer study. The clinical data is retrieved using AQL queries on the openEHR repository. Queries can be defined individually and granularly down to the level of single attributes. The annotated mutation data is retrieved from the MIRACUM-Pipe. After the study is assembled, it is stored in the cBioPortal database by executing SQL inserts from a Python script.

4. Discussion

We have demonstrated the use of openEHR as an open platform for storing and managing EHR data is the context of personalized oncology in clinical routine. We have successfully stored genetic data alongside with the corresponding clinical information in an openEHR repository. A central repository is placed as a mediator between source and the target system cBioPortal. Adhering to the harmonized data model and archetypes allows replacing either the input or the output system. The performance scales with processing thread count and allows the integration of multiple sites into a single target system. Unfortunately, our use case implementation currently lacks the representation of annotations for the genetic data. This would require changes in the archetypes describing genetic variants. Also, binding the mutation data to biomedical ontologies could add more value by enabling the capability to query mutations by the affection of cell processes. The mapping was done exemplary using the clinical tumor documentation software on our local site. For now, the mapped attributes of the clinical data provide the majority of important information for the use case but currently lack information about prior patient history and treatment. While the genetic variants contain essential attributes

like the chromosome, position, reference base, and alternative alleles, additional attributes were mapped to document further information about the filtering and sample genotypes. Especially the genotype fields are currently limited in the corresponding openEHR archetypes. It should be considered to add more parameters or slots which would allow custom fields if needed.

Table 1. Attributes that are being mapped from clinical data to openEHR archetypes

Attribute	Mapping
Birth / Death	✓ / ✓
Diagnosis insurance	×
Diagnosis (ICD-10)	×
Gender	✓
Histology (ICD-O)	✓
Medications	×
Procedures	×
TNM status	✓
UICC classification	✓

Table 2. Attributes that are being mapped from VCF files to openEHR archetypes

Attribute	Mapping
Chromosome	✓
Position	✓
ID	O
Reference Base	✓
Alternative alleles	✓
Quality score	O
Filter	✓
INFO fields	1/3
Genotype fields	3/7

✓ Attribute successfully mapped
× Attribute not successfully mapped
O Attribute could be mapped but the field was empty in our source data

The evaluation of the different data formats shows that the JSON implementation is at least 21% faster than the XML one. The JSON format supports key-value pairs and arrays while XML only uses key-value pairs. This means that XML generates many more objects and is less efficient for this use case.

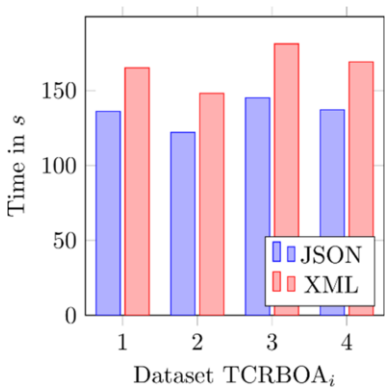

Figure 2. Comparing the amount of time that is needed to generate an EHR composition when using XML or respectively JSON as target data format

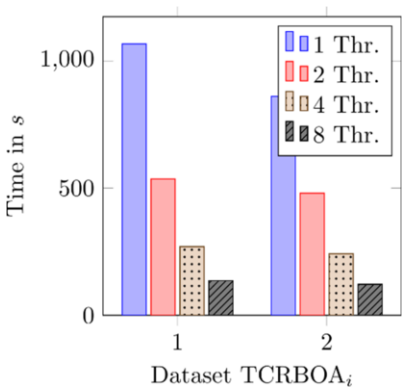

Figure 3. Overview of how the generation of clinical EHR compositions (JSON) scales with an increasing number of threads (*Thr.*)

5. Conclusion

In this paper, we have shown the possibility of integrating clinical and genomic data into EHRs stored in repositories based on the openEHR standard. The developed tools enable standardized EHR data for further use in the molecular tumor boards to assist personalized treatment decisions. As the input data is already structured, the integration

process could be redesigned to run fully automated, minimizing labor, cost, and time. During the development, it was discovered that the generation of EHR compositions could be heavily improved by using JSON instead of XML, caching Xpath queries, and the use of multiple threads. The openEHR Mapper provides a workflow to store clinical and genomic data using the openEHR data model built by domain experts, establishing easy access for use in tumor boards. The combination of openEHR standard, AQL, and modern web technologies has a high potential for further application, e.g., the structured data analysis of clinical information using AI technologies.

Acknowledgement

The project is funded by the German Federal Ministry of Education and Research (BMBF, grand id: 01ZZ1802Z).
The authors acknowledge support through the HiGHmed and MIRACUM consortia as part of the Medical Informatics Initiative Germany.

References

[1] Maranhao PA, Bacelar-Silva G, Goncalves-Ferreira D, Vieira-Marques P, Cruz-Correia R. Challenges in Design and Creation of Genetic openEHR-Archetype. Stud Health Technol Inform. 2018;247:835–839.
[2] openEHR Foundation. openEHR - Working Baseline. https://specifications.openehr.org/ (accessed January 7, 2019).
[3] Benson T, Grahame G. Principles of Health Interoperability - SNOMED CT, HL7 and FHIR. Springer, 2016.
[4] Haarbrandt B, Schreiweis B, Rey S, Sax U, Scheithauer S, Rienhoff O, et al. HiGHmed – An Open Platform Approach to Enhance Care and Research across Institutional Boundaries. Methods of Information in Medicine. 2018 jul;57(S 01):e66–e81.
[5] Semler S, Wissing F, Heyder R. German Medical Informatics Initiative. Methods of Information in Medicine. 2018 jul;57(S 01):e50–e56.
[6] Prokosch HU, Acker T, Bernarding J, Binder H, Boeker M, Boerries M, et al. MIRACUM: Medical Informatics in Research and Care in University Medicine. Methods of Information in Medicine. 2018jul;57(S 01):e82–e91.
[7] The HiGHmed Consortium. HiGHmed Clinical Knowledge Manager. https://ckm.highmed.org/ckm/ (accessed January 20, 2020).
[8] Arbeitsgemeinschaft Deutscher Tumorzentren e.V. (ADT), Gesellschaft der epidemiologischen Krebsregister in Deutschland e.V. (GEKID). Einheitlicher onkologischer Basisdatensatz ADT/GEKID. https://www.gekid.de/adt-gekid-basisdatensatz (accessed January 20, 2020).
[9] Choi M, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proceedings of the National Academy of Sciences of the United States of America. 2009 Nov;106(45):19096–19101.
[10] Becnel LB, et al. An open access pilot freely sharing cancer genomic data from participants in Texas. Scientific Data. 2016 Feb;3:160010.
[11] Metzger P, et al.. MIRACUM-Pipe. https://github.com/AG-Boerries/MIRACUM-Pipe (accessed January 7, 2020).
[12] Gao J, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal. 2013 Apr;6(269):pl1.
[13] Cerami E, et al. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. Cancer Discovery. 2012;2(5):401–404.
[14] Memorial Sloan Kettering Cancer Center (MSKCC). File Formats - cBioPortal. https://docs.cbioportal.org/5.1-data-loading/data-loading/file-formats (accessed March 16, 2020).
[15] Nedap Healthcare. Archie: openEHR Library. https://github.com/openEHR/archie (accessed July 17, 2020).
[16] Vitasystems, Hannover Medical School. EHRbase. https://ehrbase.org (accessed January 7, 2020).

[17] openEHR Foundation. openEHR - EHR Information Model. https://specifications.openehr.org/releases/RM/latest/ehr.html (accessed March 16, 2020).
[18] Mascia C, Uva P, Leo S, Zanetti G. OpenEHR modeling for genomics in clinical practice. Int J Med Inform. 2018 12;120:147–156.
[19] Tomczak A.. Molecular Pathology Report. https://ckm.openehr.org/ckm/templates/1013.26.249 (accessed March 11, 2020).