Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240808

An Open Information Model-Based Repository for Sustainable Re-Use of Heterogeneous Pandemics Research Data

Kristina SCHEUERMANN[a,1*], Philip HUFELAND[b*], Birger HAARBRANDT[a], Sabine HANSS[b], Mareike JOSEPH[a], Severin KOHLER[c], Dagmar KREFTING[b], Jendrik RICHTER[b], Erik TUTE[a], Klaus-Hendrik WOLF[a] and Michael MARSCHOLLEK[a]

[a] *Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Germany*

[b] *Department for Medical Informatics, University Medical Center Göttingen, Germany*

[c] *Berlin Institute of Health at Charité - Universitätsmedizin Berlin, Digital Health Center, Berlin, Germany*

ORCiD ID: Kristina SCHEUERMANN https://orcid.org/0009-0002-1895-1587,
Philip HUFELAND https://orcid.org/0000-0001-8110-5039

Abstract. The COVID-19 Research Network Lower Saxony (COFONI) is a German state network of experts in Coronavirus research and development of strategies for future pandemics. One of the pillars of the COFONI technology platform is its established research data repository[2], which enables provision of pseudonymised data and cross-location data retrieval for heterogeneous datasets. The platform consistently uses open standards (openEHR) and open source components (EHRbase) for its data repository, taking into account the FAIR criteria. Available data include both clinical and socio-demographic patient information. A comprehensive AQL query builder interface and an integrated research request process enable new research approaches, rapid cohort assembly and customized data export for researchers from participating institutions. Our flexible and scalable platform approach can be regarded as a blueprint. It contributes, to pandemic preparedness by providing easily accessible cross-location research data in a fully standardised and open representation.

Keywords. health information interoperability, data base, data sharing, open source, pandemics, pandemic preparedness, COVID-19

1. Introduction

The recent pandemic has shown that establishment of comprehensive data repositories with flexible data models is essential to tackle the challenges in both, understanding the mechanisms of spread of an infectious disease, as well as the pathophysiologic mechanisms on different levels in an constantly changing environment. Therefore, static data models appear as suboptimal solutions for such repositories. Furthermore, there is

[1] Corresponding Author: Kristina Scheuermann; E-mail: Scheuermann.Kristina@mh-hannover.de.
[*] These authors contributed equally.
[2] Available at https://forschungsdb.cofoni.de/

an immanent call for sustainable data collections in the public as well as in the research community, that not only help us understand Covid-19 and its mechanisms, but also help us to prepare for and tackle the next pandemic more efficiently.

In the light of the above challenges, Germany has set up a central data repository for Covid-19 data within the Network University Medicine - Routine Data Platform NUM-RDP [1]. This platform is based on an open-source data repository (EHRbase[3]) and a corresponding data querying web portal for end users. The German state of Lower Saxony has established a Covid-19 research network (COFONI), which funds numerous research projects and studies that – amongst others – investigate cellular pathways, pathomechanisms, social determinants, and various outcome parameters for acute Covid-19 infections and their long-term impact on individual health and society.

Part of this network is a data repository, the objective of which is to enable re-use of these research data across studies and across institutions in accordance with the FAIR [2] criteria for research data management.

2. Aims

The aim of this paper is to describe the concept and practical implementation of a centralized, model-driven research data repository for heterogeneous data, enabling sustainable, cross-institutional and long-term research. We will discuss the challenges and our experiences in modeling and integrating data in such an infrastructure.

3. Methods

The COFONI research data repository builds on previous open source components and projects of NUM, as the technical basis of the platform for practical implementation and concretization of data. In COFONI, this platform was customized to meet project-specific requirements, by establishing a wide range of openEHR models within the data structure and providing the ability to filter within a large dataset. The resulting research data platform comprises of PostgreSQL clusters, Keycloak, EHRbase, the NUM backend for data processing, and the customized NUM web application - a web portal with full authentication and data protection mechanisms in place- for data filtering and application for data export. Applications for data access are reviewed and approved by a use-and-access committee and made available to researchers at the participating institutions. A visual step-by-step manual will support this access. All components are data flow processes that have been approved by the data protection officers of the participating university hospitals.

3.1. Environment

During technical implementation, the components were containerized for deployment within a Kubernetes cluster, and the individual applications were combined in a Helm Chart. Kubernetes is responsible for provision, management, and scaling of the individual modules. Other tools, such as DevOps pipelines, enable fast and agile

[3] Available at https://ehrbase.org/

development of the portal. These pipelines are used in conjunction with a development, test and production environment.

Researchers from participating institutions and funded project partners are provided with automatic rights assignment, enabling them to use the portal directly for data filtering. Additional rights are required for further processes such as data requests.

3.2. Data management and integration

Currently, two of eight projects have been completed, and the pseudonymised data of the completed studies are represented in the database. The data integration process is depicted in **Figure 1**. The core of the data integration process is the modelling of clinical data in openEHR archetypes, in close collaboration with domain experts. Information managers, called data stewards, carry out the modeling process. They categorize the requirements and domain knowledge derived from clinical experts, map existing standards and terminologies, such as openEHR, and create new archetypes according to the standard for defining clinical models of the HiGHmed [3,4] Clinical Knowledge Manager. After several reviews by domain experts from different areas, these archetypes are published and used for data integration and retrieval. Archetypes developed this way are public and can be reused by the openEHR community and beyond [5].

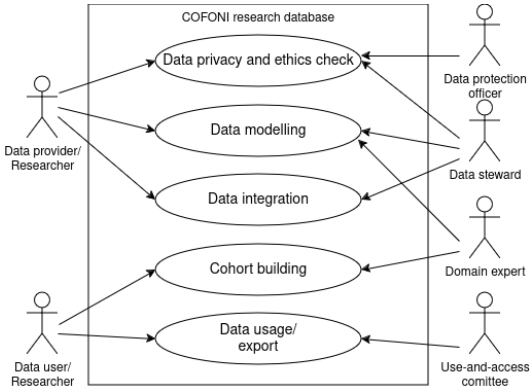

Figure 1. Use case diagram data management and data integration in COFONI.

For querying and filtering in the portal, a standardized, concept-based query graphical user interface has been implemented. A researcher can build new cohorts from the data intuitively and without knowledge of a query language **(Figure 2)**. All data stem from trials where participants signed an informed consent to share their data for scientific purposes. After a sufficient number of study data integration and data use projects, the system will be evaluated for user satisfaction, e.g., employing the SUS [6] score for the web application.

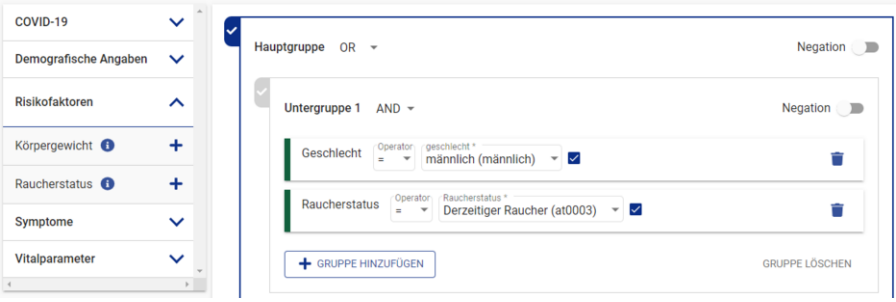

Figure 2. Web application cohort builder of the COFONI research data repository.

4. Results

Concept-based Archetype Query Language (AQL) queries applied to standardized template structures generate cross-project results from two completed projects, with over 180 different data elements from over 520 patients. The systematic structure of the individual template sections for each study allows a uniform query for data items that occur repeatedly in multiple projects, such as lab values, biomarkers, gender, age, etc. The customized portal can process a wide range of openEHR data, well beyond the German Corona Consensus (GECCO) [7] dataset definition.

For the database containing the data from the two completed studies, 14 pre-existing, fully interoperable archetypes were used, some of them repeatedly. Three new archetypes have been created and published for reuse in the openEHR community.[4]

After the integration of the first studies, additional projects with sociological content and data structures will follow. Additional archetypes need to be created for new data structures, including, e.g., sociological surveys that are innovative in terms of their outcomes and data for the CKM compared to other data repositories. Finally, the research database will implement eight of the funded COFONI projects and will be open for integration of further, also international datasets and can also ingest data from other research projects or health records.

Since openEHR follows a 100% modeling approach, including multi-language support, using detailed clinical models enables us to connect our data to international external data repositories, e.g., in OMOP CDM (Common Data Model) [8] representation, safeguarding FAIR data management.

A step-by-step guide provides visual instructions and uses real external test data from over one million patients integrated into the system. It systematically guides the user through all web application functions using test data, from data filtering to project application for a retrospective study.

5. Discussion

Our approach to implement a universally adaptable research data repository stresses the consistent use of open-source components and open standards. Among the benefits of using openEHR for research data are its flexibility and extensibility to cover all kinds of clinical concept information and assessment tests or questionnaires and its concept-based, standardized query language that does not require knowledge of the persistence layer (e.g., database schemata).

The data can be utilized in various ways. Covid-19-specific data can be recompiled and analyzed for retrospective studies with a different focus, enabling to test new hypotheses. Data-driven models may support to identify preventive measures for patients. Comparisons with other infectious diseases are also possible, as the repository – being fully compliant with the FAIR criteria - is in no way limited to Covid-19-related data. However, obtaining comparative data with similar characteristics and features from other data repositories must be done individually. Once selected and exported, standardized data formats can be easily converted to other data formats. Tools such as the OMOP Conversion Language (OMOCL) [8] and the GECCO dataset with the web tool 'openEHR-to-FHIR' [9] can be useful for this purpose.

[4] Available at https://ckm.highmed.org/ckm/projects/1246.152.40

Although the software components used were developed in accordance with FAIR criteria, currently only partners of the studies and institutions funded by COFONI may access data, in compliance with data protection regulations.

6. Conclusions

More and more applications require readily available data repositories that provide different data in different areas. Standards exist to meet the needs in these areas. With openEHR, other existing tools and a lot of preparation and effort in COFONI, it will be possible to simplify and further standardize the interoperability of different types of data in the future. The chosen modelling standard or standards-based data repository is not a one-way street. Many components/modules are interchangeable and can be converted into each other, using other helpful tools. The customized components of the research database made it possible to store, find, access and use a wide variety of fully interoperable data in openEHR.

Acknowledgment

We thank the team of the central routine data platform of the German network university medicine for their support. Supported by the COVID-19-Research Network Lower Saxony (COFONI) through funding from the Ministry of Science and Culture of Lower Saxony in Germany (14-76403-184).

References

[1] Heyder R, NUM Coordination Office, NUKLEUS Study Group, NUM-RDP Coordination, RACOON Coordination, AKTIN Coordination, et al. The German Network of University Medicine: technical and organizational approaches for research data platforms. Bundesgesundheitsbl. 2023;66:114-125. doi: 10.1007/s00103-022-03649-1. PMID: 36688978. PMCID: PMC9870206.

[2] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar 15;3:160018. doi: 10.1038/sdata.2016.18. PMID: 26978244. PMCID: PMC4792175.

[3] Wulff A, Haarbrandt B, Marschollek M. Clinical Knowledge Governance Framework for Nationwide Data Infrastructure Projects. Stud Health Technol Inform. 2018; 248:196-203. PMID: 29726437.

[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 Inf Med. 2018 Jul; 57(S 01):e66-e81. doi: 10.3414/ME18-02-0002. Epub 2018 Jul 17. PMID: 30016813.

[5] Ballout S, Biermann P, Gietzelt M, Haarbrandt B, Wulff A, Marschollek M, et al. Die offene IT-Plattform für die datengetriebene Medizin mit openEHR. 5. Workshop Antibiotikaresistenz des RKI/DGHM; 2019 Nov 14-15; Berlin, Germany. doi: 10.13140/RG.2.2.18946.84169.

[6] Brooke J. SUS: A Quick and Dirty Usability Scale. Usability Eval Ind. 1996;189-194.

[7] Sass J, Bartschke A, Lehne M, Essenwanger A, Rinaldi E, Rudolph S, et al. The German Corona Consensus Dataset (GECCO): a standardized dataset for COVID-19 research in university medicine and beyond. BMC Med Inform Decis Mak. 2020 Dec 21;20(1):341. doi: 10.1186/s12911-020-01374-w.

[8] Kohler S, Boscá D, Kärcher F, Haarbrandt B, Prinz M, Marschollek M, et al. Eos and OMOCL: Towards a seamless integration of openEHR records into the OMOP Common Data Model. J Biomed Inform. 2023 Aug;144:104437. doi: 10.1016/j.jbi.2023.104437. Epub 2023 Jul 12. PMID: 37442314.

[9] Ladas N, Franz S, Haarbrandt B, Sommer KK, Kohler S, Ballout S, et al. openEHR-to-FHIR: Converting openEHR Compositions to Fast Healthcare Interoperability Resources (FHIR) for the German Corona Consensus Dataset (GECCO). Stud Health Technol Inform. 2022 Jan 14;289:485-486. doi: 10.3233/SHTI210963. PMID: 35062196.