German Medical Data Sciences 2024
R. Röhrig et al. (Eds.)

doi:10.3233/SHTI240838

The Concept of a Versatile Computing Tool Chain for Utilizing the Core Data Set of the Medical Informatics Initiative in the INTERPOLAR Project

Sebastian STÄUBERT[a,1], Alexander STRÜBING[a], Florian SCHMIDT[a], Maryam YAHIAOUI-DOKTOR[a], Matthias REUSCHE[a], Frank MEINEKE[a], Daniel NEUMANN[a] and Markus LOEFFLER[a]
[a] *Institute for Medical Informatics, Statistics and Epidemiology (IMISE), University of Leipzig, Leipzig, Germany*
ORCiD ID: Sebastian Stäubert https://orcid.org/0000-0002-7221-7415

Abstract. Introduction: To support research projects that require medical data from multiple sites is one of the goals of the German Medical Informatics Initiative (MII). The data integration centers (DIC) at university medical centers in Germany provide patient data via FHIR® in compliance with the MII core data set (CDS). Requirements for data protection and other legal bases for processing prefer decentralized processing of the relevant data in the DICs and the subsequent exchange of aggregated results for cross-site evaluation. **Methods:** Requirements from clinical experts were obtained in the context of the MII use case INTERPOLAR. A software architecture was then developed, modeled using 3LGM², finally implemented and published in a github repository. **Results:** With the CDS tool chain, we have created software components for decentralized processing on the basis of the MII CDS. The CDS tool chain requires access to a local FHIR endpoint and then transfers the data to an SQL database. This is accessed by the DataProcessor component, which performs calculations with the help of rules (input repo) and writes the results back to the database. The CDS tool chain also has a frontend module (REDCap), which is used to display the output data and calculated results, and allows verification, evaluation, comments and other responses. This feedback is also persisted in the database and is available for further use, analysis or data sharing in the future. **Discussion:** Other solutions are conceivable. Our solution utilizes the advantages of an SQL database. This enables flexible and direct processing of the stored data using established analysis methods. Due to the modularization, adjustments can be made so that it can be used in other projects. We are planning further developments to support pseudonymization and data sharing. Initial experience is being gathered. An evaluation is pending and planned.

Keywords. distributed system, distributed processing, Datasets as Topic, core data set, medical informatics, medical informatics initiative, HL7 FHIR, Data Exchange, Medical Data Science

[1] Corresponding Author Sebastian Stäubert, Institut für Medizinische Informatik, Statistik und Epidemiologie, Universität Leipzig, Härtelstraße 16-18, 04107 Leipzig, Germany;
E-mail: sebastian.staeubert@imise.uni-leipzig.de

1. Introduction

The German Medical Informatics Initiative (MII) aims to promote digitalization in the healthcare sector [1, 2]. One of these aims is to make data collected in patient care available for research projects. To this end, data integration centers (DIC) were set up at university medical centers in Germany starting in 2018 [3]. The DICs enable access to the patient data available at distributed locations. One of their tasks is to map the patient data available in the heterogeneous primary systems to the core data set (CDS) [4] of the MII [5] and making it available via a FHIR® endpoint in the MII infrastructure [6–8]. This is a necessary prerequisite for carrying out distributed research projects. We assume the regulatory constraints like federal states hospital laws, GDPR [9], MII Broad Consent [10], etc. are fulfilled as they are not focus of this paper. In the following, we also assume that the decentralized processing of patient data is helpful given the background of the legal bases and that decentralized processing of selected patient data is possible within the context of cross-site research projects.

Patient data recorded during treatment in hospital and processed by the DIC into the MII CDS may not be sufficient on its own, which means (research) data (e.g. scores, cut-of points for specific values, etc.) calculated by algorithms on the basis of this data should be verified and evaluated by clinical researchers. This means that it may be necessary to record additional characteristics or to record feedback on calculated data, e.g. using electronic data capture (EDC) applications such as REDCap [11], OpenClinica, etc.

If the data processed as part of the research project is available on site, it must be distributed in a technically suitable, legally permitted, data protection-compliant manner, e.g. in the form of anonymous aggregates or using data sharing procedures such as DataSHIELD [12], Personal Health Train [13], etc., which is out of scope of this article and will be addressed in the course of the project.

The aim of this work is to introduce an IT system that receives MII CDS-compliant patient data provided by the DIC in a FHIR endpoint and processes it in a modular fashion. In addition, the calculated results can be presented and supplemented in the context of patient data from the health care provider.

2. Methods

Our work is conducted within the framework of the multi-center MII use case project INTERPOLAR [14, 15]. We therefore had the opportunity to obtain expert input from physicians, pharmacists, pharmacologists and clinical researchers by means of surveys, questionnaires, focus group meetings and interviews.

Potential solutions were developed within the IT team, evaluated on the basis of defined criteria using a decision matrix and then the final decision for the technical implementation was made.

We modeled the IT-architecture of the technical solution using the 3-level graph-based metamodel (3LGM²) [16] and the associated 3LGM² tool [17, 18]. This allowed us to model the functions to be supported and the required information objects (domain layer), the application systems and their interfaces with each other (logical tool layer) and the physical or virtual data processing systems including communication links (physical tool layer). In addition, the relationships between the elements of the different layers could be modeled.

Software development, which is not finalized, is based on an agile approach with weekly SPRINTs. The code is managed in github and is public available in the INTERPOLAR repository [19]. The repository contains documented toml files that allow the CDS tool chain components to be configured. The modularization also makes it possible to replace components with your own. Mockups and prototypes are used as proofs of concept and for feedback from users, for bug fixing, refinement and further development.

3. Results

We have created a software architecture called the “Core Data Set tool chain” in short “CDS tool chain”, which takes into account the requirements mentioned in the introduction and present it below. The graphic of the logical tool layer from the model provides an overview, see Figure 1.

The model file is also in the repository and can be viewed even without the 3LGM² tool by means of representations in the repository wiki [19].

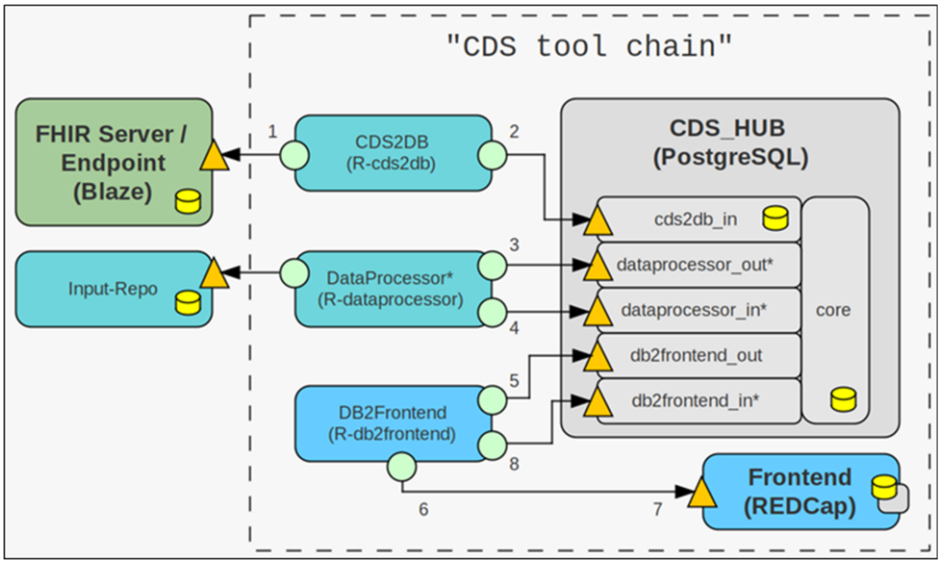

Figure 1. "Core Data Set tool chain” architecture. Component diagram modeled in 3LGM² (logical tool layer): application systems (rounded rectangles) are connected (arrows) via providing (triangles) and invoking (circles) interfaces. Storage of information objects in a certain application system is indicated with yellow barrels. The numbers on the connecting arrows represent a typical processing sequence. The entries in brackets are assigned software products or artifacts.

When applying our CDS tool chain, we assume that FHIR resources can be queried via an FHIR endpoint in accordance with the specification of the core data set (CDS) of the MII (1). This endpoint is configured in CDS2DB, together with filter criteria (FHIR Search or list of patient IDs) to select the relevant cases and patients. CDS2DB retrieves them, transforms the structured FHIR resources into flat tables based on a specification (table description) using the R [20] package fhircrackr [21] and writes the result to the CDS_HUB database (scheme cds2db_in) (2). This allows access to the data using SQL and reduces the effort required for further processing. Apart from the transfer from the

CDS FHIR representation to flat tables, no changes are made to the content of the data up to this step. This is important to know if only the presentation without further processing is desired or permitted.

Such processing can optionally be carried out by the DataProcessor. This component can be used to validate, harmonize or enrich the content of the data, as well as for any other processing (3). The DataProcessor component operates in combination with parameter sets from a separate Input-Repo. These can be assignments to classifications, such as LOINC mappings or lists of characteristic combinations provided using tables, or rules for calculating scores or checking the data for defined combinations of existing characteristics. Further calculation modules will be added in the future, but can principally also be custom made. The results produced are written back to the database (scheme dataprocessor_in) (4).

The data filtered for a project and the values computed from it can be presented in a suitable application system (frontend). We chose REDCap as the frontend for our current development because it has a well-documented API and suitable libraries in R [22]. In addition, there is solid experience with it in the IT team as well as in the MII community and even if a project requires data from patients (Patient Reported Outcome Messages, PROM), this can be supported in the future. However, other EDC applications are also conceivable, including user interfaces developed in-house, which are connected via a modified DB2Frontend component or directly via SQL.

The DB2Frontend component is currently optimized for the connection of REDCap. This reads provided data from the database (scheme db2frontend_out) (5) and uses the aforementioned REDCap API to fill in forms or generate dashboards previously created for the project (6). To get an impression, some forms are already part of the documentation in the repository wiki. Human actors can view and supplement the forms with feedback (validation, comments, additions, etc.) and send them back via the frontend's form logic. DB2Frontend retrieves sent form data (7) and plays it back into the database (scheme db2frontend_in) for further use or analysis (8).

The CDS tool chain is available in the github repository. Using yml and toml files, it can be configured and deployed container-based via Docker (compose) for easy installation. Further development of the components will continue as part of the INTERPOLAR project and the documentation (github pages) will be expanded accordingly. Feature requests and bug reports are welcome via github issues.

4. Discussion and Outlook

We are aware that there are other possible solutions for the requirements described above. SMART on FHIR [23], FHIRBase [24], Phenomen [25], REDCap CDIS [26] and others like future SQLonFHIR [27] are conceivable. We deliberately opted for a solution with an SQL database as the data hub and R as programming language in accordance with our decision matrix. It enables the flexible and direct processing of the data stored there using established analysis methods without further transformations. We also wanted to use tried-and-tested techniques that were familiar to the existing development team so that we could get started as quickly as possible and without further training. We call it a "hub" because the database is structured in such a way that it provides an input and output schema for each module, which act as interfaces.

For the extract, transform and load (ETL) part (step 1 and 2), we decided to use fhircracker rather than other solutions like FhirExtinguisher [28] because it is written in

the R programming language and integrates well with our other software modules. The use of fhircrackr during ETL, that is at the start of processing, has the advantage that the data is then available in tabular form in the CDS_HUB and can be processed directly by analysis scripts. Otherwise, fhircrackr would have to be called before each processing.

Our approach has limitations, e.g. the transformation of FHIR into flat tables with fhircrackr can lead to deeply nested information not being taken into account. This can be managed with knowledge of the existing data and the data relevant to the project defined in the CDS2DB table description [29]. Another point is that in the architecture model of the CDS tool chain (Figure 1) we assign software products or artifacts (in brackets) to the application systems (rounded rectangles). From the perspective of the model, these are interchangeable. In practice, however, this can entail additional effort. For example, the software product Blaze, which implements the application system 'FHIR Server', could be replaced by another FHIR Server with little effort. Others may require more effort, which we may take into account in future development.

Additional components may be required, e.g. if the data processed from the FHIR endpoint is pseudonymized but must be displayed un-pseudonymized in the frontend so that clinical staff can assign the patients. In this case, interaction with a trustee can be considered and implemented in DB2Frontend.

If the resulting data is to be shared across location and project boundaries, a re-transformation to FHIR can be reasonable. To this end, we are planning to expand the CDS tool chain to include components that perform pseudonymization with the connection of a trusted third party and enable further processing steps such as anonymization, aggregation or data sharing using DataSHIELD. The first releases of the CDS tool chain are currently being tested by the DICs participating in the INTERPOLAR project. A subsequent evaluation and presentation of the results is planned.

In principle, our approach can be used by other projects that require data from a FHIR endpoint and process it in tabular form after appropriate adaptation. It is intended that (project-specific) processing rules are made available to the DataProcessor in the Input-Repo. In order to make the processed data available to the frontend and to write back feedback, corresponding forms must be designed in the frontend and DB2Frontend must be adapted accordingly.

Declaration

Conflict of Interest: There is no conflict of interest.

Contributions of the authors: Sebastian Stäubert: Conceptualization, Methodology, Software, Investigation, Writing – Original Draft Alexander Strübing: Conceptualization, Methodology, Software Florian Schmidt: Conceptualization, Software, Writing – Review & Editing Maryam Yahiaoui-Doktor: Conceptualization, Software, Investigation, Writing – Review & Editing Matthias Reusche: Conceptualization, Software, Writing – Review & Editing Frank Meineke: Conceptualization Daniel Neumann: Conceptualization, Methodology, Writing – Review & Editing, Project administration Markus Löffler: Writing – Review & Editing, Supervision, Project administration, Funding acquisition

Acknowledgements

This work was supported by BMBF grants INTERPOLAR (01ZZ2320A), SMITH (01ZZ1803A) and DFG grant WI 1605/9-2. FHIR® is the registered trademark of HL7 and is used with the permission of HL7. Thanks to André Medek (University Hospital Bonn) and his help in developing standard-compliant R packages and DB connectivity in R.

References

[1] Gehring S, Eulenfeld R. German Medical Informatics Initiative: Unlocking Data for Research and Health Care. Methods Inf Med. 2018;57:e46-e49. doi:10.3414/ME18-13-0001.

[2] Semler SC, Wissing F, Heyder R. German Medical Informatics Initiative. Methods Inf Med. 2018;57:e50-e56. doi:10.3414/ME18-03-0003.

[3] Albashiti F, Thasler R, Wendt T, Bathelt F, Reinecke I, Schreiweis B. Die Datenintegrationszentren – Von der Konzeption in der Medizininformatik-Initiative zur lokalen Umsetzung in einem Netzwerk Universitätsmedizin. [Data integration centers-from a concept in the Medical Informatics Initiative to its local implementation in the Network of University Medicine]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2024. doi:10.1007/s00103-024-03879-5.

[4] Networked Medical Research e.V. The Medical Informatics Initiative's core data set. https://www.medizininformatik-initiative.de/en/medical-informatics-initiatives-core-data-set. Accessed 25 Jun 2024.

[5] Ammon D, Kurscheidt M, Buckow K, Kirsten T, Löbe M, Meineke F, Prasser F, Saß J, Sax U, Stäubert S, Thun S, Wettstein R, Wiedekopf JP, Wodke JA, Boeker M, Ganslandt T. Arbeitsgruppe Interoperabilität: Kerndatensatz und Informationssysteme für Integration und Austausch von Daten in der Medizininformatik-Initiative. [Interoperability Working Group: Core Data Set and Information Systems for Data Integration and Data Exchange in the Medical Informatics Initiative]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz. 2024. [accepted].

[6] Hund H, Wettstein R, Heidt CM, Fegeler C. Executing Distributed Healthcare and Research Processes - The HiGHmed Data Sharing Framework. Stud Health Technol Inform. 2021;278:126–33. doi:10.3233/SHTI210060.

[7] Prokosch H-U, Gebhardt M, Gruendner J, Kleinert P, Buckow K, Rosenau L, Semler SC. Towards a National Portal for Medical Research Data (FDPG): Vision, Status, and Lessons Learned. Stud Health Technol Inform. 2023;302:307–11. doi:10.3233/SHTI230124.

[8] Gruendner J, Deppenwiese N, Folz M, Köhler T, Kroll B, Prokosch H-U, Rosenau L, Rühle M, Scheidl M-A, Schüttler C, Sedlmayr B, Twrdik A, Kiel A, Majeed RW. The Architecture of a Feasibility Query Portal for Distributed COVID-19 Fast Healthcare Interoperability Resources (FHIR) Patient Data Repositories: Design and Implementation Study. JMIR Med Inform. 2022;10:e36709. doi:10.2196/36709.

[9] European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council: of 27 Apri} 2016 on the protection of natural persons with regard to the processing of personal data and on

the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) 27.04.2016.
[10] Zenker S, Strech D, Jahns R, Müller G, Prasser F, Schickhardt C, Schmidt G, Semler SC, Winkler E, Drepper J. National standardisierter Broad Consent in der Praxis: erste Erfahrungen, aktuelle Entwicklungen und kritische Betrachtungen. [Nationally standardized broad consent in practice: initial experiences, current developments, and critical assessment]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2024. doi:10.1007/s00103-024-03878-6.
[11] Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. 2009;42:377–81. doi:10.1016/j.jbi.2008.08.010.
[12] Gaye A, Marcon Y, Isaeva J, LaFlamme P, Turner A, Jones EM, Minion J, Boyd AW, Newby CJ, Nuotio M-L, Wilson R, Butters O, Murtagh B, Demir I, Doiron D, Giepmans L, Wallace SE, Budin-Ljøsne I, Oliver Schmidt C, Boffetta P, Boniol M, Bota M, Carter KW, deKlerk N, Dibben C, Francis RW, Hiekkalinna T, Hveem K, Kvaløy K, Millar S, Perry IJ, Peters A, Phillips CM, Popham F, Raab G, Reischl E, Sheehan N, Waldenberger M, Perola M, van den Heuvel E, Macleod J, Knoppers BM, Stolk RP, Fortier I, Harris JR, Woffenbuttel BHR, Murtagh MJ, Ferretti V, Burton PR. DataSHIELD: taking the analysis to the data, not the data to the analysis. Int J Epidemiol. 2014;43:1929–44. doi:10.1093/ije/dyu188.
[13] Welten S, Mou Y, Neumann L, Jaberansary M, Yediel Ucer Y, Kirsten T, Decker S, Beyan O. A Privacy-Preserving Distributed Analytics Platform for Health Care Data. Methods Inf Med. 2022;61:e1-e11. doi:10.1055/s-0041-1740564.
[14] Networked Medical Research e.V. INTERPOLAR - Reducing medication related problems and drug interactions. https://www.medizininformatik-initiative.de/en/interpolar-reducing-medication-related-problems-and-drug-interactions. Accessed 25 Jun 2024.
[15] Loeffler M, Maas R, Neumann D, Scherag A. INTERPOLAR – prospektive, interventionelle Studien im Rahmen der Medizininformatik-Initiative zur Verbesserung der Arzneimitteltherapiesicherheit in der Krankenversorgung. [INTERPOLAR-prospective, interventional studies as part of the Medical Informatics Initiative to improve medication therapy safety in healthcare]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz. 2024;67:676–84. doi:10.1007/s00103-024-03890-w.
[16] Winter A, Ammenwerth E, Haux R, Marschollek M, Steiner B, Jahn F. Technological Perspective: Architecture, Integration, and Standards. In: Winter A, Ammenwerth E, Haux R, Marschollek M, Steiner B, Jahn F, editors. Health Information Systems. Cham: Springer International Publishing; 2023. p. 51–152. doi:10.1007/978-3-031-12310-8_3.
[17] Institut für Medizinische Informatik, Statistik und Epidemiologie (IMISE). Download of the 3LGM² Tool. https://3lgm2.de/en/Downloads/3LGM2_Tool/. Accessed 25 Jun 2024.
[18] Winter A. The 3LGM2-tool to support information management in health care. GMS Med Inform Biom Epidemiol. 2011;7:Doc6. doi:10.3205/mibe000120.
[19] INTERPOLAR Project. CDS Tool Chain Repository. https://github.com/medizininformatik-initiative/INTERPOLAR. Accessed 25 Jun 2024.

[20] R Core Team. R: A Language and Environment for Statistical Computing. 2018. https://www.R-project.org/. Accessed 30 Apr 2024.
[21] Palm J, Meineke FA, Przybilla J, Peschel T. "fhircrackr": An R Package Unlocking Fast Healthcare Interoperability Resources for Statistical Analysis. Appl Clin Inform. 2023;14:54–64. doi:10.1055/s-0042-1760436.
[22] Nutter B, Garbett S, Obregon S, Obadia T, Lehr M, High B, et al. redcapAPI: Interface to 'REDCap'. https://rdrr.io/cran/redcapAPI/. Accessed 25 Jun 2024.
[23] Mandel JC, Kreda DA, Mandl KD, Kohane IS, Ramoni RB. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. J Am Med Inform Assoc. 2016;23:899–908. doi:10.1093/jamia/ocv189.
[24] Health Samurai. fhirbase - Your persistence layer for FHIR data. https://github.com/fhirbase/fhirbase.
[25] Uciteli A, Beger C, Kirsten T, Meineke FA, Herre H. Ontological representation, classification and data-driven computing of phenotypes. J Biomed Semantics. 2020;11:15. doi:10.1186/s13326-020-00230-0.
[26] Cheng AC, Duda SN, Taylor R, Delacqua F, Lewis AA, Bosler T, Johnson KB, Harris PA. REDCap on FHIR: Clinical Data Interoperability Services. J Biomed Inform. 2021;121:103871. doi:10.1016/j.jbi.2021.103871.
[27] Health Level Seven International (HL7). SQL on FHIR. https://build.fhir.org/ig/FHIR/sql-on-fhir-v2/. Accessed 25 Jun 2024.
[28] Oehm J, Storck, Michael, Fechner M, Brix TJ, Yildirim K, Dugas, Martin. FhirExtinguisher: A FHIR Resource Flattening Tool Using FHIRPath. Stud Health Technol Inform. 2021;281:1112–3. doi:10.3233/SHTI210369.
[29] INTERPOLAR Project. CDS Tool Chain Repository: Table Description. https://github.com/medizininformatik-initiative/INTERPOLAR/tree/v0.1.0-gmds/R-cds2db/cds2db/inst/extdata. Accessed 2 Jul 2024.