Towards a Bavarian Oncology Real World Data Research Platform

Jasmin ZIEGLER[1a,b], Julian GRUENDNER[c], Lorenz ROSENAU[d], Marcel ERPENBECK[a], Hans-Ulrich Prokosch[c], and Noemi DEPPENWIESE[a]

[a] *Medical Center for Information and Communication Technology, Universitätsklinikum Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany*

[b] *Bavarian Cancer Research Center (BZKF)*

[c] *Chair of Medical Informatics, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany*

[d] *IT Center for Clincal Research, University of Lübeck, Lübeck, Germany*

Abstract. Introduction In the last decade numerous real-world data networks have been established in order to leverage the value of data from electronic health records for medical research. In Germany, a nation-wide network based on electronic health record data from all German university hospitals has been established within the Medical Informatics Initiative (MII) and recently opened for researchers´ access through the German Portal for Medical Research Data (FDPG). In Bavaria, the six university hospitals have joined forces within the Bavarian Cancer Research Center (BZKF). The oncology departments aim at establishing a federated observational research network based on the framework of the MII/FDPG and extending it with a clear focus on oncological clinical data, imaging data and molecular high throughput analysis data. **Methods** We describe necessary adaptions and extensions of existing MII components with the goal of establishing a Bavarian oncology real world data research platform with its first use case of performing federated feasibility queries on clinical oncology data. **Results** We share insights from developing a feasibility platform prototype and presenting it to end users. Our main discovery was that oncological data is characterized by a higher degree of interdependence and complexity compared to the MII core dataset that is already integrated into the FDPG. **Discussion** The significance of our work lies in the requirements we formulated for extending already existing MII components to match oncology specific data and to meet oncology researchers needs while simultaneously transferring back our results and experiences into further developments within the MII.

Keywords. federated feasibility queries, oncology, ADT/GEKID, FHIR

1. Introduction

In the last decade numerous real world data networks (e.g. SPHN [1], PCORnet [2], OHDSI [3]) have been established in order to leverage the value of data from electronic health records for medical research. Some of those networks cover a broad spectrum of

[1] Corresponding author: Jasmin Ziegler, Medical Center for Information and Communication Technology, Universitätsklinikum Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg, jasmin.ziegler@uk-erlangen.de

clinical data, while others have focused on disease specific data (e.g. the National COVID Cohort Collaborative (N3C) [4], the 4CE consortium [5], the CODEX platform [6]) or clinical specialty registries, such as the German Emergency Department Data Registry AKTIN [7]. Many scientific results have illustrated that large-scale international observational research is feasible and that such collaborative approaches can overcome many of the logistic and methodological challenges associated with observational study designs (e.g. [8]). In Germany, a nation-wide network based on electronic health record data from all German university hospitals has been established within the Medical Informatics Initiative (MII) [9] and recently opened for researchers´ access through the German Portal for Medical Research Data (FDPG) [10]. Most of the software code of the MII/FDPG infrastructure [11] is published as open source, to support wide stream cooperative development and participation of researchers from many areas.

In Bavaria, the six university hospitals have joined forces within the Bavarian Cancer Research Center (BZKF) to combine outstanding oncological research with high-performance cutting-edge medicine. Besides performing numerous multicenter clinical trials involving all six BZKF partners, the oncology departments have also decided to establish a federated observational research network based on an Oncology Real World Data Research Platform. To avoid reinventing the wheel, BZKF researchers have decided to build their federated research infrastructures on the framework (technological components, regulatory groundwork and organizational structures) of the MII/FDPG, reuse as many components as possible and extend it with a clear focus on oncological clinical data, imaging data and molecular high throughput analysis data. Thus, several independent IT infrastructure subprojects have been initiated, each associated with a different data type focus (e.g. genomic analysis data from molecular tumor boards, radiology imaging data and clinical data), but all following the general strategy to add subsets of an oncology patient journey to research data repositories located within the local university hospital IT infrastructures. Finally, the concept of privacy preserving federated analysis and federated learning approaches [12] was adopted as the main strategic foundation by the BZKF partners to support network-wide analysis. This strategic focus is to be further specified and implemented within the MII data integration center (DIC) infrastructures.

The objective of this paper is to describe necessary adaptions and extensions of existing MII components with the goal of establishing a Bavarian oncology real world data research platform with its first use case of performing federated feasibility queries. We share insights from developing a feasibility platform prototype and evaluating it with end users.

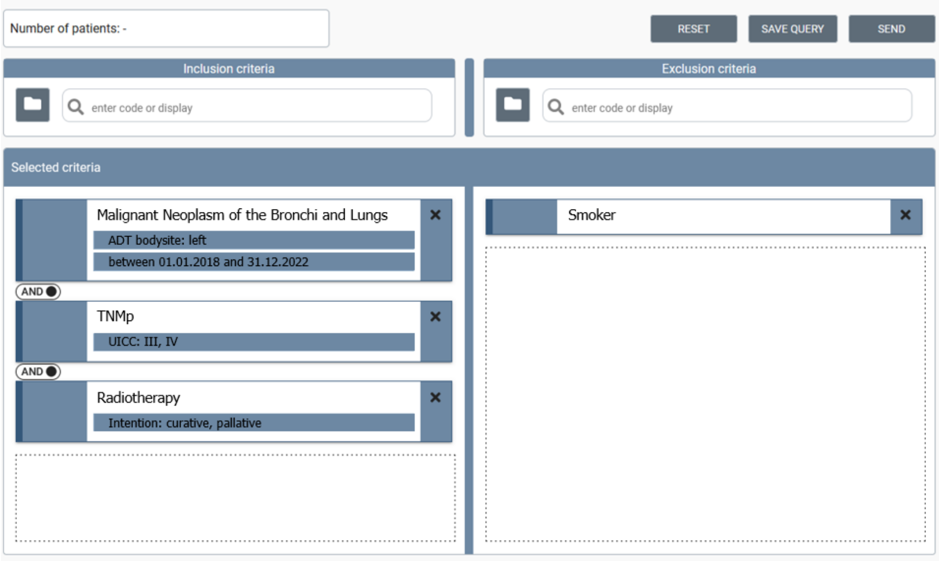
2. Methods

In a first step we analyzed the dedicated tumor documentation systems at the six Bavarian university hospitals and their data export capabilities. We compared the variations of systems in use, access to tumor data from within the DIC and the format of data present in the data bases. Secondly, we further evaluated the existing FDPG components and functionalities for their potential to serve as a basis for a specialized oncology data feasibility portal within the BZKF. We assessed compatibility of the FDPG components with all six sites and with oncological data. Important criteria lie within the technical infrastructure inside of each DIC which should be able to embed an input data server populated with oncology specific data in HL7 FHIR® format. Finally, the stepwise defined high-level BZKF IT strategy framed all work focusing on clinical oncology data (within this subproject) to serve as a joined link between all IT and data gathering subprojects within the BZKF. As already pursued within earlier projects for establishing harmonized multicenter IT networks [13–15], we adapted the strategy of data-driven development accompanied by stakeholder workshops while aiming at a detailed requirements analysis built on experiences from oncology researchers within the BZKF network. The stakeholder workshop focused on discussing the most important types of queries and how we can accurately represent them, bringing together a diverse group of oncology researchers.

3. Results

3.1. Feasibility Platform Prototype

We developed a first prototype of the BZKF feasibility platform (see figure 1). The prototype is based on the existing MII FDPG platform which is constructed from generic and thus extensible components to enable secure and privacy-preserving distributed feasibility queries for FHIR® data [11]. To safeguard privacy, the query result is presented as an aggregated number of patients and obfuscated at each site before being sent to the central platform and displayed to the end user. The user can select desired criteria via a user interface (UI), an intermediate structured query is assembled and distributed to all participating sites for execution within each hospital, before the user receives the results. A search ontology service is required for displaying selectable hierarchical concepts and query composition, while the middleware ensures secure transportation of the query. Finally, an execution service translates the structured query to a FHIR search query or a Clinical Quality Language (CQL) query that can be processed by a FHIR server, which stores all the relevant data.



Number of patients: -

Malignant Neoplasm of the Bronchi and Lungs
ADT bodysite: left
between 01.01.2018 and 31.12.2022
AND
TNMp
UICC: III, IV
AND
Radiotherapy
Intention: curative, palliative

Smoker

Figure 1. Prototype of the Feasibility Platform with Oncological Search Criteria.

For the first version of our prototype, we extended the FDPG platform by the following oncological criteria: histology (ICD-O-3 and grading), TNM staging, distant metastases and procedures like surgery, system therapy or radiotherapy with their intention and residual status. Tumor documentation systems are established at all six BZKF sites which transfer data to the state cancer registries in an ADT/GEKID XML schema [16]. This standard format was specified by the Working Group of German Tumor Centers (ADT) and the Society of Epidemiological Cancer Registries in Germany (GEKID) as a legal obligation for clinical cancer registration and provides a suitable representation of relevant oncological concepts. From those we selected key criteria to be displayed in the UI and identified attributes to each key element that help to refine a query. To display oncological concepts in the user interface, we extended the FDPG search ontology generator [17]. The tool simplifies FHIR profiles to the relevant criteria and their associated hierarchies and displays them in a tree-like format for ease of use. The search ontology generator requires FHIR structure definitions of oncological concepts. The harmonized MII core data set (CDS) represented in HL7 FHIR is presently limited to basic clinical data (e.g. patient demographics and encounter data, diagnosis, procedures, laboratory analysis and medication), but a dedicated oncology extension module is currently in development [18]. Since the German Cancer Consortium (DKTK) FHIR representation [19] of the ADT/GEKID schema was the starting point for the definition efforts of the MII oncology extension module, we applied this schema as a temporary basis for the search ontology generation to display oncological concepts in the BZKF feasibility tool prototype.

3.2. Stakeholder Involvement for Feasibility Tool Prototype Evaluation

In order to learn about the most important concepts to be used for oncological feasibility queries, we conducted a workshop involving stakeholders from various groups, including directors and employees of comprehensive cancer centers, a Department of Radiation Oncology director, and the Head of the Working Group on Clinical-Scientific Digitalization. We presented the prototype and discussed relevant data attributes for oncological feasibility queries. After the workshop, we had asked oncology researchers to provide us with typical example cohort queries. We sent out worksheets with all available criteria and asked for a written feedback with exemplary queries. The occurrence count of each concept is specified in brackets hereinafter. From a total of nine feedback responses, the ICD10 diagnosis code (9 times) was the most frequently mentioned category closely followed by radiotherapy/chemotherapy with or without intention and the tumor entity (6 times each).

Table 1. Original Example Queries from Oncological Researchers.

Example Queries
what is the number of
• all patients with anal carcinomas with date of surgery in 2021 or 2022
• all patients with NSCLC (Non-Small Cell Lung Cancer) diagnosed in 2022
• all patients with mamma carcinoma, diagnosed between 2018 and 2023, HER2-positive
• all patients with glioblastoma, histology WHO stage IV, chemotherapy protocol Procarbazin or Lomustin, diagnosed in 2022
• all patients with malignant neoplasm of the bronchi and lungs, UICC stage III or IV and primary radiotherapy with curative or palliative intention

- all non-smoking patients with NSCLC, T1-T3, N0, with curative radiotherapy alone without consideration of subsequent palliative radiotherapy, from 2020 to 2021
- all patients staged T3 N2 before surgery, N3 or R1-R2 past surgery, who need postoperative radio or chemotherapy
- all female patients < 30 years of age, C50, ICD-O-3 8500/3, T1-T3 who have received radiotherapy
- all patients with C43 and age <50 at diagnosis, BRAF mutation, UICC II - IV, >T2, neoadjuvant chemotherapy

In those queries tumor entities classify cancer types by body parts/organs (e.g. anal carcinomas = C21.1 anal canal carcinoma + C44.5 carcinoma of the anal margin) [20] or histological features (e.g. basal cell carcinomas include ICD-O-3 morphology codes 8090-8110) [21]. The following criteria were mentioned frequently: date of diagnosis and TNM (4 times each), UICC/WHO stage (3 times). The data elements ICD-O-3, date of surgery, residual status after surgery, smoking status, sex, recurrent tumor, distant metastasis location, WHO stage and ADT extension module Mamma were mentioned once each.

We then categorized the requested concepts into three distinct categories:

- Represented Criteria: search criteria already represented in feasibility queries with the current FDPG feasibility tool release (ICD10 diagnosis code, date of diagnosis)
- Modifiable Criteria: Search criteria requiring modifications to the current FDPG feasibility tool release to meet more complex user requirements *(see 3.3)*
- Unqueryable Elements: Elements that cannot be queried because this information is not directly part of the ADT specification; nevertheless, such information can be retrieved by more complex analysis later *(see 4.)*

3.3. Defining Adaptions to Enhance Platform Utility for Oncology Data

Workshop attendees requested a search feature that incorporates the use of tumor entities instead of listing ICD10 or ICD-O-3 codes, respectively long OR-Clauses to cover all possible classification system codes which match one tumor entity term. As a result, users do not have to select all concepts individually. Moreover, we observed a discrepancy in the interdependence of criteria between the oncological data elements and the existing MII CDS criteria. It is crucial to have precise references from any type of therapy (surgery, radiotherapy, system therapy) to the original diagnosis in oncological search criteria. Otherwise, a patient with multiple tumors may be mistakenly counted as part of a cohort even if the therapy was performed for the other tumor. This “vagueness” issue can be addressed by adjusting the platform backend to ensure that the reference from the therapy to the corresponding diagnosis is always included in the query. Furthermore, a modification of the representation of interdependent concepts in the graphical user interface is necessary. Certain criteria cannot be queried currently due to the absence of FHIR search parameters. While the FHIR specification offers a variety of search parameters for many data fields of every resource type [22], some fields do not have corresponding parameters or need to be queried in combination with multiple other fields to avoid vagueness. Thus, the definition of custom search parameters is required. In the DKTK oncology FHIR profiles, we identified 14 relevant data elements without corresponding custom search parameters although they greatly enhance utility and were repeatedly emphasized as crucial criteria in the feedback of experts. To enable querying of these criteria, we generated custom search parameters for the DKTK profiles.

4. Discussion

Our article examines the requirements for an oncological real-world data platform to conduct federated feasibility queries on clinical oncology data across six German hospitals using HL7 FHIR conform data. We presented a prototype of the platform, analyzed user feedback and example queries collected from an end-user workshop that identified several needs for adaption to the current platform setup. Our key finding is that oncological data is more complex and interdependent than the basic modules of the MII CDS provided by the current FDPG release. Based on our initial experiences, we identified the need for adjustments to the backend, user interface, and search ontology generation of the feasibility platform, which is currently integrated by the FDPG platform developer team. To display and query complex interdependent oncological concepts, the structured query and the search ontology generator require major adaptions. The ability to query interdependent concepts with multiple attributes is not only a benefit for oncological concepts, but can also be helpful for future CDS extension modules and the entire MII. At present, one exemplary query cannot be expressed because the DKTK oncology FHIR profiles do not encompass the organ-specific supplementary modules (e.g. module mamma, module prostate) that ADT/GEKID recently introduced. However, since the MII extension module oncology has already integrated these modules into its information model, criteria associated with these modules will be queryable in the future once the transition to the MII extension module oncology is completed. Those examples nicely illustrate the advantage of an open source development strategy, since it allowed us to quickly build our BZKF work on previous FDPG developments, and now transfer our results and experiences back into further developments and projects within the MII.

Our approach to conducting federated feasibility queries in oncology requires harmonized data in FHIR format across all participating sites. A major future challenge is to create an extract-transform-load (ETL) job that can be delivered to all BZKF sites regardless of varying DIC infrastructures, software setups and tumor documentation systems. Furthermore, the workshop with medical professionals provided only a glimpse of example queries and is not fully comprehensive yet. We intend to conduct further workshops where we iteratively present development steps of the BZKF RWD platform and integrate stakeholder feedback immediately. Further, we aim at extending the search capabilities in close collaboration with the research experts and their assessment of relevant search parameters for feasibility queries in oncology.

To further improve the effectiveness of feasibility queries, there are two potential options to consider. The first is expanding the criteria available for querying by extending the MII oncology extension module with further data elements which are currently not part of the ADT/GEKID dataset. Tumor documentation systems today contain a wealth of information that goes beyond what is currently included in the ADT/GEKID dataset. However, extracting this additional data may present challenges due to heterogeneity of tumor documentation systems and IT infrastructures across different sites. Furthermore, adding data elements can cause a challenge in terms of data mapping and harmonization which might require adaptions to the existing FHIR profiling. Therefore, standardization and harmonization work, as planned to be pursued in the MII PM4Onco project, is very important and needs to then be reflected also in the take-over of such extensions in the local documentation systems. In the future, additional CDS extension modules will be developed to also include many more relevant oncology data elements (e.g. collected in molecular tumor boards or from patient reported outcome data).

Finally, it is important to differentiate between data elements that improve feasibility queries and are essential during the time of the feasibility inquiry, and information that can be obtained through distributed analysis on exported complete datasets. While some information may not be essential at the outset, it may still be accessible later in the project (e.g. examining an entire exported dataset in detail can provide insights into whether a particular therapy qualifies as a "primary" or "sole" treatment option). Even though some elements could not be queried at the time, they may still become available during the subsequent analysis phase.

Declarations

Conflict of Interest: The authors declare that they have no conflict of interest.

Contributions of the authors: JZ developed prototype and wrote manuscript; HUP, ME, JG, JZ: conducted workshops; JG, LR: support with FDPG components; HUP: conceived and supervised the project, provided input on manuscript. All authors have approved the manuscript as submitted and accept responsibility for the scientific integrity of the work.

The present work was performed in (partial) fulfillment of the requirements for obtaining the degree “Dr. rer. biol. hum.” from the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) (JZ, ND).

Acknowledgement: The present work was funded by the Bavarian Cancer Research Center (BZKF). We would like to thank all BZKF researchers for their valuable input.

References

[1] Lawrence AK, Selter L, Frey U. SPHN - The Swiss Personalized Health Network Initiative. Stud Health Technol Inform. 2020 Jun.;270:1156–1160.

[2] Fleurence RL, Curtis LH, Califf RM, et al. Launching PCORnet, a national patient-centered clinical research network. J Am Med Inform Assoc JAMIA. 2014;21:578–582.

[3] Hripcsak G, Duke JD, Shah NH, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform. 2015;216:574–578.

[4] Haendel MA, Chute CG, Bennett TD, et al. The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment. J Am Med Inform Assoc JAMIA. 2021 Mar.;28:427–443.

[5] Brat GA, Weber GM, Gehlenborg N, et al. International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium. NPJ Digit Med. 2020;3:109.

[6] Prokosch H-U, Bahls T, Bialke M, et al. The COVID-19 Data Exchange Platform of the German University Medicine. Stud Health Technol Inform. 2022 May;294:674–678.

[7] Brammen D, Greiner F, Kulla M, et al. Das AKTIN-Notaufnahmeregister – kontinuierlich aktuelle Daten aus der Akutmedizin : Ergebnisse des Registeraufbaus und erste Datenauswertungen aus 15 Notaufnahmen unter besonderer Berücksichtigung der Vorgaben des Gemeinsamen Bundesausschusses zur Ersteinschätzung [AKTIN - The German Emergency Department Data Registry - real-time data from emergency medicine : Implementation and first results from 15 emergency departments with focus on Federal Joint Committee’s guidelines on acuity assessment]. Med Klin Intensivmed Notfallmedizin. 2022 Feb.;117:24–33.

[8] Hripcsak G, Ryan PB, Duke JD, et al. Characterizing treatment pathways at scale using the OHDSI network. Proc Natl Acad Sci U S A. 2016 Jul.;113:7329–7336.

[9] Semler SC, Wissing F, Heyder R. German Medical Informatics Initiative. Methods Inf Med. 2018 Jul.;57:e50–e56.

[10] Prokosch H, Gebhardt M, Gruendner J. Towards a national portal for medical research data (FDPG): Vision, Status, and Lessons Learned. Stud Health Technol Inform. 2023 in press;

[11] Gruendner J, Deppenwiese N, Folz M, et al. 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 May;10:e36709.

[12] Wirth FN, Meurers T, Johns M, et al. Privacy-preserving data sharing infrastructures for medical research: systematization and comparison. BMC Med Inform Decis Mak. 2021 Aug.;21:242.

[13] Schüttler C, Buschhüter N, Döllinger C, et al. Anforderungen an eine standortübergreifende Biobanken-IT-Infrastruktur : Erhebung des Stakeholderinputs zum Aufbau eines Biobankennetzwerks der German Biobank Alliance (GBA) [Requirements for a cross-location biobank IT infrastructure : Survey of stakeholder input on the establishment of a biobank network of the German Biobank Alliance (GBA)]. Pathol. 2018 Jul.;39:289–296.

[14] Sedlmayr B, Sedlmayr M, Kroll B, et al. Improving COVID-19 Research of University Hospitals in Germany: Formative Usability Evaluation of the CODEX Feasibility Portal. Appl Clin Inform. 2022 Mar.;13:400–409.

[15] Schüttler C, Zerlik M, Gründner J, et al. Empowering researchers to query medical data and biospecimens by ensuring appropriate usability: Evaluation study of the ABIDE_MI feasibility tool (Preprint). JMIR Hum Factors. 2022 Oct.;

[16] Altmann U, Katz FR, Dudeck J. A reference model for clinical tumour documentation. Stud Health Technol Inform. 2006;124:139–144.

[17] Rosenau L, Majeed RW, Ingenerf J, et al. Generation of a Fast Healthcare Interoperability Resources (FHIR)-based Ontology for Federated Feasibility Queries in the Context of COVID-19: Feasibility Study. JMIR Med Inform. 2022 Apr.;10:e35789.

[18] MII Kerndatensatz Modul Onkologie [Internet]. Medizininformatik-Initiative; 2023 Feb. [cited 2023 Mar 12]. Available from: https://github.com/medizininformatik-initiative/kerndatensatzmodul-onkologie.

[19] DKTK Oncology [Internet]. Dedktkoncology 120 - Simpl. [cited 2023 Mar 29]. Available from: https://simplifier.net/packages/de.dktk.oncology/1.2.0.

[20] Leichman LP, Cummings BJ. Anal carcinoma. Curr Probl Cancer. 1990;14:117–159.

[21] DIMDI - ICD-O-3 Erste Revision [Internet]. [cited 2023 Mar 29]. Available from: https://www.dimdi.de/static/de/klassifikationen/icd/icd-o-3/icdo3rev1html/zusatz-10-kodierrichtlinien.htm#multiple.

[22] Search - FHIR v4.3.0 [Internet]. [cited 2023 Mar 15]. Available from: https://hl7.org/fhir/search.html.