Bridging Data Silos in Oncology with Modular Software for Federated Analysis on FHIR: A Multisite Implementation Study

Jasmin Ziegler, Marcel Erpenbeck, Timo Fuchs, Anna Saibold, Paul-Christian Volkmer, Günter Schmidt, Johanna Eicher, Peter Pallaoro, Renata De Souza Falguera, Fabio Aubele, Marlien Hagedorn, Ekaterina Vansovich, Johannes Raffler, Stephan Ringshandl, Alexander Kerscher, Julia Maurer, Brigitte Kühnel, Gerhard Schenkirsch, Marvin Kampf, Lorenz A. Kapsner, Hadieh Ghanbarian, Helmut Spengler, Iñaki Soto-Rey, Fady Albashiti, Dirk Hellwig, Maximilian Ertl, Georg Fette, Detlef Kraska, Martin Boeker, Hans-Ulrich Prokosch, Christian Gulden

Submitted to: Journal of Medical Internet Research
on: August 22, 2024

Table of Contents

Original Manuscript **5**
Supplementary Files **27**
Figures **28**
Figure 1 ..... **29**
Figure 2 ..... **30**
Figure 3 ..... **31**
Figure 4 ..... **32**
Figure 5 ..... **33**

Bridging Data Silos in Oncology with Modular Software for Federated Analysis on FHIR: A Multisite Implementation Study

Jasmin Ziegler[1, 2]; Marcel Erpenbeck[1]; Timo Fuchs[2, 3, 4]; Anna Saibold[2, 5]; Paul-Christian Volkmer[2, 6]; Günter Schmidt[2, 7]; Johanna Eicher[8, 9]; Peter Pallaoro[2, 8, 9]; Renata De Souza Falguera[9, 10]; Fabio Aubele[11]; Marlien Hagedorn[11]; Ekaterina Vansovich[2, 12]; Johannes Raffler[2, 12] Dr rer nat; Stephan Ringshandl[13] Dr rer nat; Alexander Kerscher[2, 6] Dr med; Julia Maurer[2, 14] Dr med; Brigitte Kühnel[2, 15]; Gerhard Schenkirsch[2, 16] Dr med; Marvin Kampf[1]; Lorenz A. Kapsner[17, 18] Dr med; Hadieh Ghanbarian[17]; Helmut Spengler[2, 8] Dr rer nat; Iñaki Soto-Rey[2, 12] Dr rer med; Fady Albashiti[2, 11] Dr; Dirk Hellwig[2, 3, 4] Prof Dr; Maximilian Ertl[7]; Georg Fette[7]; Detlef Kraska[1] Dr; Martin Boeker[2, 9] Prof Dr; Hans-Ulrich Prokosch[1, 2, 17] Prof Dr; Christian Gulden[2, 17] Dr

[1]Medical Center for Information and Communication Technology, Universitätsklinikum Erlangen Erlangen DE
[2]Bavarian Cancer Research Center (BZKF) Erlangen DE
[3]Department of Nuclear Medicine, University Hospital Regensburg Regensburg DE
[4]Medical Data Integration Center (MEDIZUKR), University Hospital Regensburg Regensburg DE
[5]Department of Information Technology, University Hospital Regensburg Regensburg DE
[6]Comprehensive Cancer Center Mainfranken, University Hospital Würzburg 97080 Würzburg DE
[7]University Hospital Würzburg 97080 Würzburg DE
[8]Data Integration Center, Klinikum rechts der Isar, School of Medicine and Health, Technical University of Munich Munich DE
[9]Institute for Artificial Intelligence and Informatics in Medicine (AIIM), Chair of Medical Informatics, Klinikum rechts der Isar, School of Medicine and Health, Technical University of Munich Munich DE
[10]Section of Precision Psychiatry, Clinic for Psychiatry and Psychotherapy, LMU Munich Munich DE
[11]Medical Data Integration Center (MeDICLMU), LMU University Hospital, LMU Munich Munich DE
[12]Digital Medicine, University Hospital of Augsburg Augsburg DE
[13]Department of Medicine, Data Integration Center (DIC), Philipps-University Marburg Marburg DE
[14]University Cancer Center Regensburg, University Hospital Regensburg Regensburg DE
[15]Comprehensive Cancer Center Munich, Klinikum rechts der Isar – Technical University of Munich Munich DE
[16]Comprehensive Cancer Center Augsburg, University Hospital of Augsburg Augsburg DE
[17]Friedrich-Alexander-Universität Erlangen-Nürnberg, Institute for Medical Informatics, Biometrics and Epidemiology, Medical Informatics Erlangen DE
[18]Institute of Radiology, Uniklinikum Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) Erlangen DE

Corresponding Author:
Jasmin Ziegler
Medical Center for Information and Communication Technology, Universitätsklinikum Erlangen
Krankenhausstr. 12
Erlangen
DE

Abstract

Background: Real-world data (RWD) from sources like administrative claims, electronic health records, and cancer registries offer insights into patient populations beyond the tightly regulated environment of randomized controlled trials. To leverage this and to advance cancer research, six university hospitals in Bavaria have established a joint research IT infrastructure.

Objective: This article aims to outline the design, implementation, and deployment of a modular data transformation pipeline that transforms oncological RWD into HL7 (Health Level 7) FHIR (Fast Healthcare Interoperability Resources) format and then into a tabular format in preparation for a federated analysis (FA) across the six BZKF university hospitals.

Methods: To harness RWD effectively, we designed a pipeline to convert the oncological basic dataset (oBDS) into HL7 FHIR format and prepare it for federated analysis. The pipeline handles diverse IT infrastructures and systems while maintaining privacy by keeping data decentralized for analysis. To assess the functionality and validity of our implementation, we defined a cohort to address two specific medical research questions. We evaluated our findings by comparing the results of the FA with reports from the Bavarian Cancer Registry and the original data from local tumor documentation systems.

Results: We conducted a federated analysis of 17,885 cancer cases from 2021/2022. Breast cancer was the most common diagnosis at three sites, prostate cancer ranked in the top two at four sites, and malignant melanoma was notably prevalent. Gender-specific trends showed larynx and esophagus cancers were more common in males, while breast and thyroid cancers were more frequent in females. Discrepancies between the Bavarian Cancer Registry and our data, such as higher rates of malignant melanoma (5 % vs. 11 %) and lower representation of colorectal cancers (13 % vs. 7 %) likely result from differences in the time periods analyzed (2019 vs. 2021/2022) and the scope of data sources used. The Bavarian Cancer Registry reports approximately three times more cancer cases than the six university hospitals alone.

Conclusions: The modular pipeline successfully transformed oncological RWD across six hospitals, and the federated approach preserved privacy while enabling comprehensive analysis. Future work will add support for recent oBDS versions, automate data quality checks, and integrate additional clinical data. Our findings highlight the potential of federated health data networks and lay the groundwork for future research that can leverage high-quality RWD, aiming to contribute valuable knowledge to the field of cancer research.

(JMIR Preprints 22/08/2024:65681)
DOI: https://doi.org/10.2196/preprints.65681

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Bridging Data Silos in Oncology with Modular Software for Federated Analysis on FHIR: A Multisite Implementation Study

Jasmin Ziegler (0009-0005-5362-5228)[1,2*], Marcel Erpenbeck (0009-0007-5468-6510)[1], Timo Fuchs (0009-0002-3896-6786)[2,3,4], Anna Saibold (0009-0004-2088-1025)[2,5], Paul-Christian Volkmer (0009-0007-0967-9696)[2,6], Günter Schmidt[2,7], Johanna Eicher (0000-0003-4871-0282)[8,18], Peter Pallaoro[2,8,18], Renata De Souza Falguera (0000-0002-0475-4892)[8,9], Fabio Aubele (0009-0006-9970-7058)[10], Marlien Hagedorn (0009-0002-6998-5429)[10], Ekaterina Vansovich [2,11], Johannes Raffler (0000-0003-2495-4020)[2,11], Stephan Ringshandl (0000-0002-0544-4298)[12], Alexander Kerscher (0000-0001-7742-570X)[2,6], Julia Maurer (0009-0006-5340-2793)[2,13], Brigitte Kühnel[2,14], Gerhard Schenkirsch (0000-0003-0510-6069)[2,15], Marvin Kampf (0000-0002-9108-0469)[1], Lorenz A. Kapsner (0000-0003-1866-860X)[16,17], Hadieh Ghanbarian (0009-0009-2903-3504)[16], Helmut Spengler[2,18], Iñaki Soto-Rey (0000-0003-3061-5818)[2,11], Fady Albashiti (0000-0002-0671-152X)[2,10], Dirk Hellwig (0000-0002-3056-0143)[2,3,4], Maximilian Ertl (0000-0002-1290-9444)[7], Georg Fette (0000-0002-0369-3805)[7], Detlef Kraska[1], Martin Boeker (0000-0003-2972-2042)[2,8], Hans-Ulrich Prokosch (0000-0001-6200-753X)[1,2,16], Christian Gulden (0000-0003-1261-3691)[2,16]

[1] *Medical Center for Information and Communication Technology, Universitätsklinikum Erlangen, Erlangen, Germany*

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

[3] *Department of Nuclear Medicine, University Hospital Regensburg, Regensburg, Germany*

[4] *Medical Data Integration Center (MEDIZUKR), University Hospital Regensburg, Regensburg, Germany*

[5] *Department of Information Technology, University Hospital Regensburg, Regensburg, Germany*

[6] *Comprehensive Cancer Center Mainfranken, University Hospital Würzburg, 97080 Würzburg, Germany*

[7] *University Hospital Würzburg, 97080 Würzburg, Germany*

[8] *Institute for Artificial Intelligence and Informatics in Medicine (AIIM), Chair of Medical Informatics, Klinikum rechts der Isar, School of Medicine and Health, Technical University of Munich, Munich, Germany*

[9] *Section of Precision Psychiatry, Clinic for Psychiatry and Psychotherapy, LMU Munich, Munich, Germany*

[10] *Medical Data Integration Center (MeDIC[LMU]), LMU University Hospital, LMU Munich, Munich,*

Germany

[11] *Digital Medicine, University Hospital of Augsburg, Augsburg, Germany*

[12] *Department of Medicine, Data Integration Center (DIC), Philipps-University Marburg, Marburg, Germany*

[13] *University Cancer Center Regensburg, University Hospital Regensburg, Regensburg, Germany*

[14] *Comprehensive Cancer Center Munich, Klinikum rechts der Isar – Technical University of Munich, Munich, Germany*

[15] *Comprehensive Cancer Center Augsburg, University Hospital of Augsburg, Augsburg, Germany*

[16] *Friedrich-Alexander-Universität Erlangen-Nürnberg, Institute for Medical Informatics, Biometrics and Epidemiology, Medical Informatics, Erlangen, Germany*

[17] *Institute of Radiology, Uniklinikum Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany*

[18] *Data Integration Center, Klinikum rechts der Isar, School of Medicine and Health, Technical University of Munich, Munich, Germany*

** Corresponding Author*

Abstract

Background: Real-world data (RWD) from sources like administrative claims, electronic health records, and cancer registries offer insights into patient populations beyond the tightly regulated environment of randomized controlled trials. To leverage this and to advance cancer research, six university hospitals in Bavaria have established a joint research IT infrastructure.

Objective: This article aims to outline the design, implementation, and deployment of a modular data transformation pipeline that transforms oncological RWD into HL7 (Health Level 7) FHIR (Fast Healthcare Interoperability Resources) format and then into a tabular format in preparation for a federated analysis (FA) across the six BZKF university hospitals.

Methods: To harness RWD effectively, we designed a pipeline to convert the oncological basic dataset (oBDS) into HL7 FHIR format and prepare it for federated analysis. The pipeline handles diverse IT infrastructures and systems while maintaining privacy by keeping data decentralized for analysis. To assess the functionality and validity of our implementation, we defined a cohort to address two specific medical research questions. We evaluated our findings by comparing the results of the FA with reports from the Bavarian Cancer Registry and the original data from local tumor documentation systems.

Results: We conducted a federated analysis of 17,885 cancer cases from 2021/2022. Breast cancer was the most common diagnosis at three sites, prostate cancer ranked in the top two at four sites, and malignant melanoma was notably prevalent. Gender-specific trends showed larynx and esophagus cancers were more common in males, while breast and thyroid cancers were more frequent in females. Discrepancies between the Bavarian Cancer Registry and our data, such

as higher rates of malignant melanoma (5 % vs. 11 %) and lower representation of colorectal cancers (13 % vs. 7 %) likely result from differences in the time periods analyzed (2019 vs. 2021/2022) and the scope of data sources used. The Bavarian Cancer Registry reports approximately three times more cancer cases than the six university hospitals alone. **Conclusion:** The modular pipeline successfully transformed oncological RWD across six hospitals, and the federated approach preserved privacy while enabling comprehensive analysis. Future work will add support for recent oBDS versions, automate data quality checks, and integrate additional clinical data. Our findings highlight the potential of federated health data networks and lay the groundwork for future research that can leverage high-quality RWD, aiming to contribute valuable knowledge to the field of cancer research.

keywords: real-world data; real-world evidence; oncology; electronic health records; federated analysis; HL7® FHIR®; cancer registries; interoperability; observational research network

Introduction

Real-world data (RWD), including information from various sources such as administrative claims data, electronic health records (EHRs), and cancer registries, offers a broad perspective on real-world patient populations, beyond the tightly regulated environment and specific conditions of randomized controlled trials (RCT) [1–3]. RWD enables the generation of real-world evidence (RWE) concerning patient care by providing a comprehensive understanding of how interventions perform in real-life clinical settings and in diverse and unselected patient populations. This includes individuals often beyond the scope of RCTs, such as patients with frailty, comorbidities or pregnant women, regardless of their social, cultural, or educational background [4–9].

The Medical Informatics Initiative (MII) has established a large-scale data sharing network across Germany based on electronic health record data from university hospitals, using the Health Level 7 (HL7®) Fast Healthcare Interoperability Resources (FHIR®) standard for data integration [10]. Hospitals harmonize heterogeneous clinical data in local data integration centers (DIC) nationwide, and a central portal has been established to access this data [11]. However, oncological data have not yet been integrated into the MII network. In Bavaria, the six university hospitals have united to form the Bavarian Cancer Research Center (BZKF) to provide comprehensive access to the latest methods of early detection, prevention, diagnosis, and treatment of cancer and build networked structures for cutting-edge research with a broad impact for all patients in Bavaria.

In this context, their oncology departments together with the six university hospitals` DIC have established a federated observational research network, building on the groundwork laid by the MII..Analyzing data from multiple hospitals enhances statistical validity by increasing the sample size, which enables rare event analysis in more diverse patient populations. However, the challenge

of data protection in multi-site scenarios underscores the need for implementing federated and privacy-preserving methods in data analysis [12–14].

This article aims to outline the design, implementation, and deployment of a modular data transformation pipeline that transforms oncological RWD into HL7 FHIR format and then into a tabular format in preparation for a federated analysis (FA) across the six BZKF university hospitals.

Methods

In previous work, we detailed the necessary adaptations and extensions of existing MII components with the goal to enable federated feasibility queries on clinical oncology data [15], setting the groundwork for the BZKF Oncology Real World Data Platform (oRWDP). Our current goal is to extend the oRWDP and implement a data transformation pipeline with an initial use case of performing a federated analysis with a particular focus on data quality and comparability between the sites. As a source of RWD, we use output from the six hospitals` tumor documentation systems. Four of the hospitals use the same commercial system (ONKOSTAR™) [16], whereas two hospitals apply CREDOS, a tumor documentation system closely integrated into their EHR system, which was developed by one of the German Comprehensive Cancer Centers [17]. Because of the German law on National Cancer Registry Data (Bundeskrebsregisterdatengesetz), both systems have to be able to export data in the oncological basic dataset (oBDS) format, a standardized dataset definition employed nation-wide for the collection of cancer data in cancer registries [18–20]. Since data pseudonymization is an important step in our pipeline, the respective pseudonymization tools already applied within the hospitals' DIC (twice entici [21] and four times gPAS [22–24]) had to be generically integrated into the pipeline. Further, the pipeline endpoint was set as the DataSHIELD Opal database, as we chose to use the privacy-preserving DataSHIELD framework [25] as our FA environment.

In designing our system architecture, because of the six sites` heterogeneous software mix and our aim to keep our approach scalable for future deployments in further hospitals with even other systems to be included, we established the following key objectives in accordance with related work in the field of FA in health care [12,13,26–29]:

1. **Modular Adaptability**: Create a flexible architecture to address diverse site requirements.
2. **Multi Institutional FA**: Data remains onsite, only aggregated results are shared.
3. **Security and Privacy:** Secure and non-disclosive analysis of pseudonymized patient data.

4. **Interoperability**: Enhance standard conformity by utilizing HL7 FHIR, improve data management and stewardship.
5. **Open Source**: Use open-source software for (cost) efficiency, longevity, community collaboration, and transparency.

To test the functionality and validity of our implementation, we defined a cohort to address specific medical research questions. We planned to include all patients who were diagnosed with cancer in 2022 and reported to the cancer registry as our data foundation for the following research questions:

Q1: What is the distribution of tumor entities across the six university hospitals for cases diagnosed in 2022?

Q2: What is the distribution of the administrative gender among the cases of tumor entities diagnosed in 2022?

To evaluate our pipeline, we compared the FA results with reports from the Bavarian Cancer Registry and with the original data from the local tumor documentation systems.

Results

Architecture

The complete pipeline architecture comprises five major modules and four transformation steps (figure 1) and described in more detail in the subsequent sections.

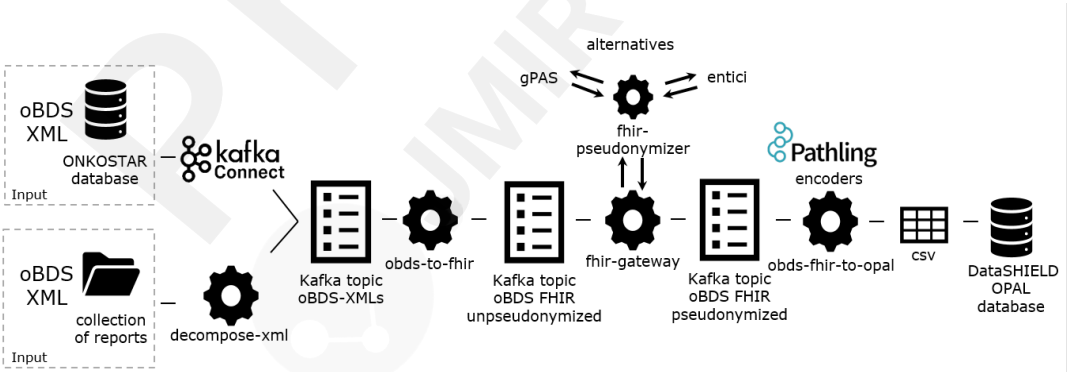


Figure 1 Architecture of the pipeline for transforming oBDS data into a final analysis format. This figure illustrates the key components, including two input interfaces, generic integration with pseudonymization services, and support for two output formats, enabling the conversion of oBDS data from XML to HL7 FHIR and to a tabular format suitable for FA.

Input interfaces: ONKOSTAR database connector and decompose-xml folder import

We incorporated two input interfaces: one that connects directly to the ONKOSTAR database, and another that functions through a folder import mechanism for locations without ONKOSTAR or access to its database. All tumor documentation systems offer an export of reported oBDS collections encoded in XML, the official format in which they are transmitted to state mandated cancer registries. The oBDS collections differ structurally from the oBDS single reports stored in the ONKOSTAR database. Therefore, we provide a preprocessing service that reads in oBDS collections from a folder and decomposes them to match the format of single reports. As a second input interface, we provide an Apache Kafka Connect [30] connector to directly read in oBDS single reports from the ONKOSTAR database. In both import scenarios, an Apache Kafka producer [30] publishes the (decomposed) single report XML to a topic for use by subsequent services.

Mapping oncology RWD to FHIR: obds-to-fhir

We developed an ETL process that transforms oBDS XML-data to HL7 FHIR resources [31]. This component reads single oBDS XML-reports from an Apache Kafka topic, maps them to FHIR resources of the oncology FHIR model developed by Lambarki et al. [32] and publishes the results to another Apache Kafka topic.

Pseudonymization: FHIR Gateway and FHIR Pseudonymizer

To de-identify the resources generated by the obds-to-fhir job, we deploy two services: the FHIR Gateway [33] and the FHIR Pseudonymizer [34]. The former reads resources from a given Kafka topic and sends them to the latter for pseudonymization based on configurable de-identification rules. For pseudonym generation, four sites use the pseudonymization service gPAS and two sites use the entici software, which we integrated with the FHIR Pseudonymizer. The FHIR Gateway publishes the resulting pseudonymized FHIR resources to a new output topic.

Transformation to tabular data: obds-fhir-to-opal

In the previous step, pseudonymized FHIR resources have been generated which can be used as the endpoint for feasibility queries as illustrated in our previous work [15]. The DataSHIELD FA framework however with its OPAL database requires a tabular data format [25]. Therefore we use the Pathling library FHIR encoders [35] in the obds-fhir-to-opal service to transform the nested FHIR

resources into structured, tabular data. The library builds upon Apache Spark to convert FHIR bundles into Spark datasets. Following successful transformation of FHIR resources to dataframes, we use SQL and Spark functions for joining and grouping of relevant data elements tailored to the research queries. The result is a CSV file.

Upload to OPAL and Federated Analysis with DataSHIELD

In the final step, we upload the CSV file resulting from the obds-fhir-to-opal service to the local OPAL servers. Figure 2 shows the FA network where the OPAL servers form the local analysis endpoints within each of the six BZKF sites. A central DataSHIELD client manages FA processes by distributing the analysis script across the network sites. These scripts are then locally executed, accessing the oBDS data stored in the local OPAL servers and returning aggregated results to the central DataSHIELD client, thus ensuring the confidentiality of private data by design.

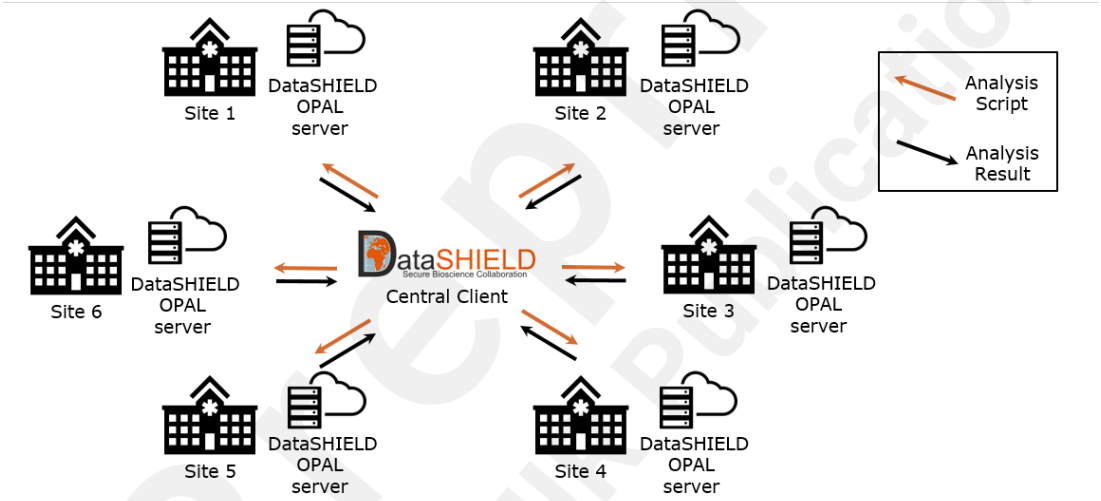


Figure 2 The FA network illustrating the OPAL servers as the local analysis endpoints at each of the six BZKF sites.

Software distribution to all locations

We distribute the previously described software components from a public GitHub-repository and the GitHub Container Registry [36]. Apart from a full setup, we provide multiple Docker Compose files which allow for a modular deployment of each individual component enabling an easily adaptable setup at all sites and a generic integration with the different software systems already available at the sites (e.g. ONKOSTAR, CREDOS, gPAS, entici). Additionally, we supply Helm charts which allow for deployment and orchestration of all containerized applications in a Kubernetes cluster [37,38]. As several sites deploy the software on servers without Internet connectivity, we provide an air-gapped installer which includes all container images compressed into an archive file for convenient download.

Federated Analysis of Oncology data

To address the two research questions, we utilized the data elements ICD-10 diagnosis code, date of diagnosis, and gender. We aggregated all diagnoses from 2022 for site 1-5. For site 6, only data from 2021 was available and therefore used.

The total volume of cancer data analyzed for the one-year time span across all six BZKF sites comprised 17,885 patients, including 7,969 women, 9,913 men, and 3 individuals of other or unknown genders. The ten most frequent cancer diagnoses included prostate cancer (14%), breast cancer (11%), malignant melanoma of the skin (11%), cancer of the trachea, bronchus, and lungs (8%), cancer of the lip, oral cavity, and pharynx (7%), and cancer of the colon and rectum (7%). Non-Hodgkin Lymphoma, cervical cancer, pancreatic cancer, and thyroid cancer each accounted for 4%.

In the latest report for the year 2019, the Bavarian Cancer Registry reported a total of 61,031 cancer diagnoses with the following ten most frequent entities: breast cancer (18 %), prostate cancer (13 %), cancer of the colon and rectum (13 %), cancer of the trachea, bronchus and lungs (9 %), malignant melanoma of the skin (5 %), bladder cancer (5 %), and cervical cancer (5 %). Pancreas cancer, Non-Hodgkin lymphoma and stomach cancer each represented 3% of the cases.

Figure 3 illustrates the distribution of cancer incidences among various cancer types diagnosed in 2022 (site 1-5) and 2021 (site 6) within the BZKF (research question Q1). Both breast cancer (C50, D05) and prostate cancer (C61) rank among the top two in five sites, with breast cancer being the most prevalent in three of these sites and prostate cancer being the most prevalent in two. Malignant melanoma of skin (C43) also shows a significant representation, particularly in Site 5 with 24.4%.

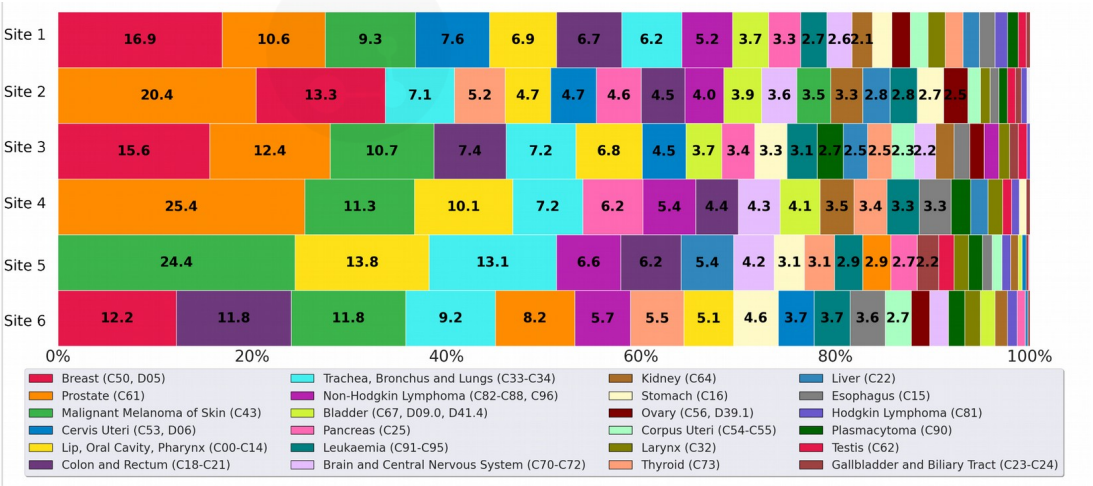


Figure 3 Distribution of tumor entities at each hospital for cases diagnosed in 2022 (site 1-5) and 2021 (site 6). Results of the FA across six locations in relative numbers per site (research question Q1).

Site 4 reported no instances of breast cancer or uterine cancers, as its gynecology department does neither use ONKOSTAR nor CREDOS and therefore has not yet been integrated into our pipeline. Furthermore, Site 5 exhibited notably fewer cases of gynecological cancers (breast, cervix, uterus), as the university professorships for gynecology and obstetrics are based at affiliated hospitals separate from the university hospital, and thus, this data was not fully accessible for our analysis. This site reported the lowest relative number of prostate cancer cases, likely because the Department of Urology is also based at a partner hospital. As a result, the urological cancer data from site 5 is probably incomplete in our dataset. Site 6 showed a lower prevalence of prostate cancer and a relatively higher prevalence of colon and rectal cancers compared to the other sites. This site did not report any cases of testicular cancer.

Figure 4 presents an overview of how different cancer diagnoses are distributed among female and male patients in 2022 (site 1-5) and 2021 (site 6). It depicts the aggregated frequencies of cancer diagnoses for each entity group across all six locations and highlights relative distribution for female and male patients (pertaining to research question Q2). Apart from cancer affecting sex-specific organs, such as cancers of the prostate and uterus, there are notable differences in the frequency of other cancer diagnoses between sexes.

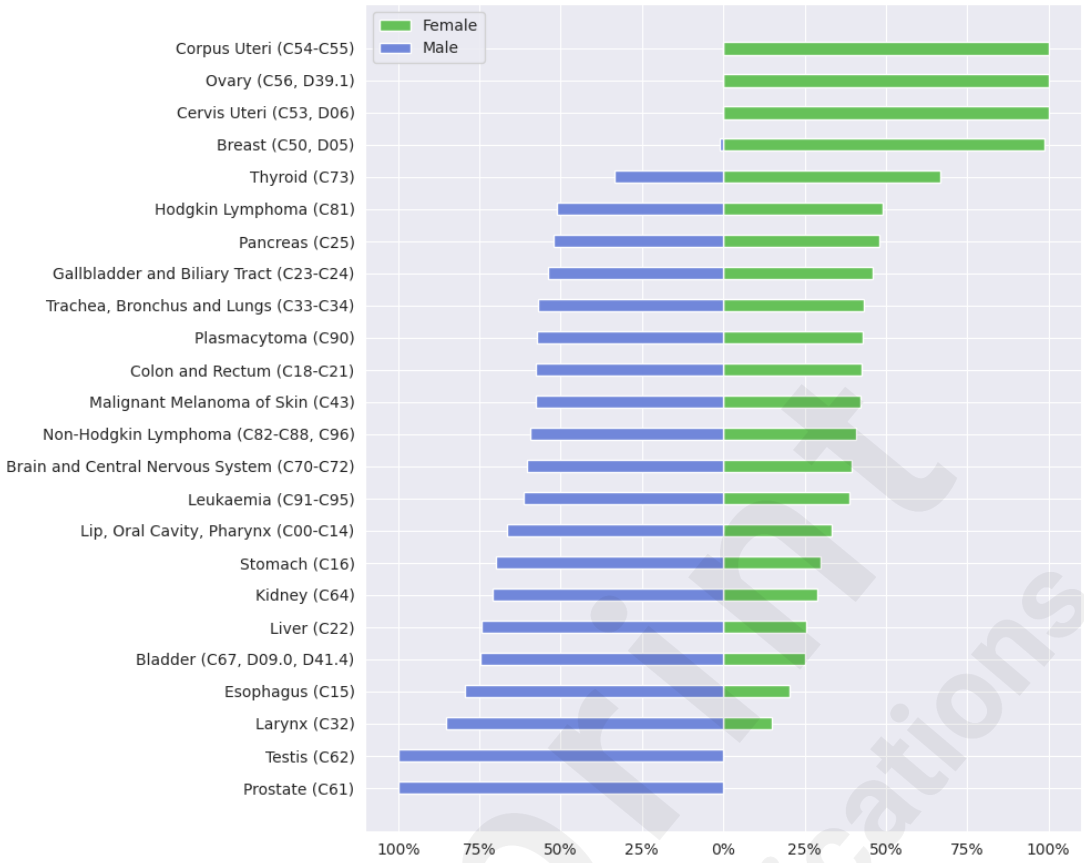


Figure 4 Distribution of administrative gender among the tumor entities for cases diagnosed in 2022 (site 1-5) and 2021 (site 6). Mean results of the FA across six locations in relative numbers (research question Q2). Other genders are omitted from the visualization due to the presence of only three cases.

Cancer types such as larynx, esophagus, bladder, liver, kidney or stomach see higher frequency rates in males compared to females, a trend also observed with cancers of the lip and oral cavity as well as leukemia, which are predominantly diagnosed in males. In contrast, breast and thyroid cancer frequency is significantly higher in females. These findings are consistent with the results of reviewed literature, which explored sex differences in cancer incidence [39–43]. We compared our results with the 2019 Bavarian Cancer Registry report and identified the distribution of the five most frequently diagnosed conditions among female and male patients, focusing on the gender distribution of each specific condition and excluding sex-specific organs (table 1).

Table 1 Distribution of the top five most frequently diagnosed conditions among female and male patients, excluding sex-specific organs (uterus, ovary, prostate, testis). Data from BZKF (2021/2022) compared to the Bavarian Cancer Registry report (2019). The table highlights the gender-specific distribution of each condition.

	Female BZKF (%)	Female Registry (%)	Male BZKF (%)	Male Registry (%)
Breast	99	99	1	1
Thyroid	67	69	33	31

Hodgkin Lymphoma	49	45	51	55
Pancreas	48	48	52	52
Gallbladder and Biliary Tract	46	44	54	56
Larynx	15	16	85	84
Esophagus	21	19	79	81
Bladder	25	23	75	77
Liver	26	29	74	71
Kidney	29	30	71	70

Comparison of Federated Analysis Results with Tumor Documentation Systems

To evaluate validity, we compared the total number of diagnoses in the original data from the local tumor documentation systems to the total number of diagnoses after being processed by the presented pipeline and aggregated through the FA framework DataSHIELD. Figure 5 provides details on the calculation of the entity-wise deviation.

$\hat{y}$: predicted value for entity i (federated analysis result)
y: gold standard for entity i (evaluation with tumor documentation system)

For each entity i, calculate entity-wise deviation and return mean value over all 24 entity-wise deviations (Mean Absolute Percentage Error):

$$\overline{Deviation - entitywise} = \frac{1}{n}\sum_{i=1}^{n}\frac{|\hat{y}_i - y_i|}{y_i} \times 100 \quad with \quad n = 24$$

Figure 5 Calculation of the entity-wise deviation (Mean Absolute Percentage Error).

This evaluation, conducted by tumor documentation specialists querying the tumor documentation systems or utilizing a custom built tool that automates the majority of the process [44], unveiled entity-wise mean deviations detailed in Table 2.

Table 2 FA evaluation - entity-wise deviations (mean and median) of the FA and the original data [%].

	Site 1	Site 2	Site 3	Site 4	Site 5	Site 6
deviation entity-wise (mean)	1.6	1.9	1.4	3.7	11.6	1.22

For sites 1, 2, 3, and 6, the mean of entity-wise deviations remains under 2 %, contrasting with site 4 and 5, which exhibited a mean of entity-wise deviation of 3.7 % and 11.6 %, respectively.

Discussion

Previous research has highlighted the importance of utilizing RWD for generating RWE on patient

care in diverse, unselected populations [1–9,11]. Geldorf et al. have argued for the development of federated RWD infrastructures on a common data model, capable of bringing the centrally-conducted big data analysis to the de-centrally kept biomedical data [45]. Following this paradigm the BZKF multi-institutional research network offers the foundation to leverage insights into oncology RWD from not only one, but six sites. However, our process of creating a harmonized foundation of care-related RWD from tumor documentation systems across the BZKF university hospitals with heterogeneous IT infrastructures also illustrated challenges arising in such real world environments.

We have outlined the successful development and deployment of a modular pipeline for extracting, harmonizing and transforming oBDS data across the six BZKF university hospitals. Unlike traditional statewide cancer registries, which centralize data collection and analysis, our approach uses federated analysis, keeping data decentralized and preserving privacy by design.

We demonstrated the functionality of our pipeline through a federated analysis using the DataSHIELD framework to address two research questions. Our analysis shows breast cancer (C50, D05) as the most common at three sites and prostate cancer (C61) among the top two at four sites. Additionally, cancers of the larynx, esophagus, bladder, and liver are more frequent in males, while breast and thyroid cancers are more common in females (excluding sex-specific cancers).

Our findings generally align with expected incidence rates or can be attributed to local specializations in treatment and data availability [39–43]. However, our data shows a slightly higher frequency of malignant melanoma of the skin (11 %) compared to the Bavarian Cancer Registry (5 %) [39]. Conversely, colorectal cancers are underrepresented in our data (7 %) compared to the registry (13 %) [39]. These discrepancies might be partially attributed to the broader data sources utilized by the Bavarian Cancer Registry, which include additional clinics, outpatient facilities, and other reporting institutions and ultimately reporting approximately three times more cancer diagnoses than the BZKF, as well as the different time periods analyzed (2019 for the registry versus 2021/2022 for our study).

In light of the comparison of our reported figures and the data recorded in the tumor documentation systems, several factors may account for the discrepancies observed. One significant issue is the occurrence of retrospectively documented cases. These arise in cases of a large time delay between diagnosis and documentation or if cases initially diagnosed externally are later incorporated. The latter situation arises when patients, who were diagnosed elsewhere in the relevant time period but are now receiving treatment at one of the six facilities, have their therapy documented now with a diagnosis date in the past within the relevant time period.

Another contributing factor is the timing mismatch between data extraction and the data quality

evaluation. Data extraction was performed in January, while the evaluation occurred in May. This delay may have led to an increase in cases recorded in the tumor documentation systems due to the reasons outlined above. Additionally, re-extraction of data was not feasible for Site 5, which showed the highest entity-wise deviation of 11.6%. This site had transitioned to oBDS version 2.2.3 in February, while our pipeline only supports up to oBDS version 2.2.2, preventing us from processing the updated data from this site.

Moreover, we found a discrepancy between the data elements defined in the oBDS standard and those available in the oBDS XML-files reported to the cancer registry. We had intended to investigate the UICC (Union for International Cancer Control) stage of cancer diagnoses, but this data was largely missing in the oBDS reports from most sites. Since we currently only process oBDS data from XML-reports sent to the cancer registry, our dataset could be enhanced by extracting additional data elements from other database tables within the tumor documentation systems, leveraging even more of the documented data. However, this is not feasible with the current decompose-xml folder import interface, which is limited to reading oBDS XML-reports. If expanding the dataset in this manner is a future goal, we would need to establish direct access to the tumor documentation databases at all locations to retrieve additional data beyond the oBDS XML-reports. The planned transition of two locations from CREDOS to ONKOSTAR would help streamline this process by eliminating the use of two different systems.

Lessons Learned

Healthcare research IT infrastructure requires tailored solutions and adherence to established processes and security standards. Heterogeneous IT systems across sites introduce multiple challenges. Certain locations require air-gapped installations, isolated from unsecured networks to protect sensitive data, complicating development, deployment, and maintenance. Additionally, the DataSHIELD framework imposes strict restrictions on analytics to ensure data privacy. To address these issues, we iteratively adapted the obds-fhir-to-opal module, implementing various groupings and mappings directly into the dataset, which was crucial for effective analysis within the framework’s constraints. Significant challenges such as data incompleteness, the employment of various documentation systems, and the heterogeneity of documentation practices across different hospitals or sub-clinics per site persist. Similar to Maier et al. we found that it is an essential requirement to have precise information about the conditions under which documentation was conducted and in what time frame after the original event documentation is pursued [46]. We also learned that data from some sites should not be integrated into future analysis of dedicated cancer

entities (e.g. breast and prostate cancer) since their provided dataset is not representative because of local organizational structures or the documentation in a particular clinic still being pursued with a tumor documentation system not yet integrated into our pipeline. Thus, our insights add further perspectives to the barriers to RWD analysis mentioned by Saesen et al. (methodological and operational challenges) [47], illustrating that the knowledge about the documentation practice, context and potential incompleteness of the real word data integrated into a RWD network is essential to avoid misinterpretation of analysis results.

Future Work

We plan to support all oBDS versions in our ETL job. To date, data quality and completeness checks have predominantly depended on human intervention. Ru et al. highlighted the absence of interoperable data quality standards and observed significant variability in the quality of two real-world datasets following data quality assessment [48]. The inclusion of data from six sites introduces even more variability and further underscores the significance of addressing data quality and completeness. Alongside addressing future queries, we will develop a unified evaluation strategy that incorporates automated data quality and plausibility checks into the pipeline, aligning with the standards of the State Cancer Registry [49] for completeness, validity and plausibility such as ensuring date variables follow a logical sequence (e.g., birth date <= diagnosis date) or verifying valid combinations of histology, tumor localization, and TNM staging. For all data quality measures and for accurate interpretation of results, thorough communication with domain professionals is essential.

Berger et al. emphasize the critical need to integrate various often siloed RWD sources to produce high-quality RWE in oncology [9]. Addressing this gap involves incorporating RWD such as laboratory findings, pathology reports, radiology reports and molecular genetic data from molecular tumor boards.

The Munich Online Comprehensive Cancer Analysis (MOCCA) platform is a tool that allows analysis and visualization of oncology-related data from the tumor documentation system CREDOS by utilizing all documented data, not just the oBDS [50]. However, the deployment of MOCCA is limited to sites using the CREDOS tumor documentation system due to its strong dependence on its internal data model, making it incompatible with sites using the ONKOSTAR or other tumor documentation systems. Converting data to the FHIR data model enhances interoperability across systems and sites and facilitates the integration of these - in the past - siloed data sources at the DIC. Thus, the next steps for the BZKF sites will involve integrating the various sources of oncology data, with the oBDS datasets leveraged within our project and the MII core dataset data, already available

within the DIC. Inspired by the findings of Swinckels et al. [51], who showed in their scoping review of 20 studies that machine learning (ML) and deep learning (DL) applied to longitudinal EHR data can greatly enhance early disease detection and prevention across various conditions, we plan to integrate various data sources and analyze longitudinal data from the past decade. This approach will enable us to develop new machine learning models for detecting or predicting oncological diseases. Expanding the number of hospitals involved is also essential to increase sample size and diversify patient populations. Therefore we will contribute our open source pipeline as well as our experiences and insights into ongoing work within oncology-related Germany-wide projects, such as e.g. the expansion of the national portal for medical research data [11] and the MII project PM4Onco [52].

Conclusion

Our modular approach demonstrates the feasibility of converting oncological RWD into HL7 FHIR and tabular data and querying it in a federated way across six sites. These findings motivate us to build on this work and integrate the full set of oBDS data from the six university hospitals to leverage the value of more than 200.000 oncological cases from the last decade in the future, growing by about 20,000 new cases annually. The dataset can be leveraged for cohort searches, hypothesis generation, study planning, and the development of new AI-models. In their 2021 systematic review of research applications of FA, Hunger et al. emphasized that additional efforts are necessary to promote awareness about the significant potential of FA in leveraging readily available RWE to address key research questions in cancer [53]. Our study contributes to achieving this goal, and we will continue to explore the benefits of FA for RWD in our future research. Through a focus on iterative processes aimed at integrating further clinical data and improving data quality, we aim to generate valuable RWE from previously untapped sources of care-related information, ultimately aiming to make significant contributions to cancer research.

Declarations

Acknowledgements

We would like to thank the Comprehensive Cancer Centers of the six Bavarian university hospitals for their support in providing the original tumor documentation data and for verifying our findings. We also extend our gratitude to all BZKF researchers for their valuable input.

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).

Funding

The present work was funded by the Bavarian Cancer Research Center (BZKF).
This publication was partially funded by the German Federal Ministry of Education and Research (BMBF) Network of University Medicine 2.0: “NUM 2.0", Grant No. 01KX2121, Project: NUM-DIZ.

Conflicts of Interest

The authors declare that they have no conflict of interest.

Ethics approval

This retrospective study was approved by the relevant ethics committees and permission for data usage was obtained from the use and access committees across all sites.
University Hospital Erlangen: application number 23–160_1-Br, approved on 21.09.23
Klinikum rechts der Isar of Technical University of Munich: approval University Hospital Erlangen is sufficient
University Hospital Würzburg: approval University Hospital Erlangen is sufficient
University Hospital LMU Munich, application number 23-0559, approved on 14.11.23
University Hospital of Augsburg: ethics committee of the University Hospital LMU Munich, application number 23-0583, approved on 28.11.23
University Hospital Regensburg: application number 23-3587-104, approved on 5.12.23

Data Availability

The original and processed data are not publicly available due to privacy restrictions.

Code Availability

The software can be retrieved from the GitHub-Repository [36].

Author Contributions

Conceptualizing the overall research idea and framework: JZ, ME, HUP, AK, CG
Designing the methodology used in the research.: JZ, ME, HUP, AK, CG
Collection and curation of the data required for the research.: JZ, ME, TF, AS, PCV, GS, JE, PP, RDSF, FAu, MH, EV, JR, GeS, HS, ISR, FA, DK, CG

Developing and conducting the federated analysis and evaluation of results: JZ, ME, TF, AS, PCV, GS, JE, PP, RDSF, FAu, MH, EV, JR, SR, AK, JM, BK, GeS, HS, ISR, FA, DH, MErt, GF, CG

Developing software for the modular pipeline: JZ, CG, ME, PCV, JE, PP, FAu, MK, LAK

Writing the manuscript - original draft: JZ

Writing the manuscript - review & editing: JZ, TF, AS, PCV, GS, JE, PP, RDSF, JR, SR, AK, JM, GeS, LAK, HS, ISR, FA, DH, MB, HUP, CG

Visualization of results: JZ, HG

Supervision and guidance: HUP, CG

All authors have read and approved the final version of the manuscript.

Abbreviations

BZKF - Bavarian Cancer Research Center

CREDOS - Cancer Retrieval Evaluation and Documentation System

DIC - Data Integration Center

EHR - Electronic Health Records

FA - Federated Analysis

FHIR - Fast Healthcare Interoperability Resources

HL7 - Health Level 7

ICD-10 - International Classification of Diseases, 10th Revision

MII - Medical Informatics Initiative

PM4Onco - Personalized Medicine for Oncology

oBDS - Oncological Basic Data Set

oRWDP - Oncology Real World Data Platform

RWE - Real-World Evidence

RWD - Real-World Data

RCT - Randomized Controlled Trials

XML - Extensible Markup Language

References

1. Penberthy LT, Rivera DR, Lund JL, Bruno MA, Meyer A-M. An overview of real-world data sources for oncology and considerations for research. CA Cancer J Clin 2021;(72):287–300. doi: 10.3322/caac.21714
2. Liu F, Panagiotakos D. Real-world data: a brief review of the methods, applications, challenges and opportunities. BMC Med Res Methodol 2022;(22):287. doi: 10.1186/s12874-022-01768-6

3. Julian GS, Shau W-Y, Chou H-W, Setia S. Bridging Real-World Data Gaps: Connecting Dots Across 10 Asian Countries. JMIR Med Inform 2024;(12):e58548. doi: 10.2196/58548
4. Saesen R, Hemelrijck MV, Bogaerts J, Booth CM, Cornelissen JJ, Dekker A, Eisenhauer EA, Freitas A, Gronchi A, Hernán MA, Hulstaert F, Ost P, Szturz P, Verkooijen HM, Weller M, Wilson R, Lacombe D, Graaf WT van der. Defining the role of real-world data in cancer clinical research: The position of the European Organisation for Research and Treatment of Cancer. Eur J Cancer Elsevier; 2023;(186):52–61. PMID:37030077
5. Bastarache L, Brown JS, Cimino JJ, Dorr DA, Embi PJ, Payne PRO, Wilcox AB, Weiner MG. Developing real-world evidence from real-world data: Transforming raw data into analytical datasets. Learn Health Syst 2022;(6):e10293. doi: 10.1002/lrh2.10293
6. Kyriazakos S. Editorial: The Role of Real World Evidence (RWE) for Digital Health. Front Comput Sci 2022;(4). doi: https://doi.org/10.3389/fcomp.2022.862712
7. Cowie MR, Blomster JI, Curtis LH, Duclaux S, Ford I, Fritz F, Goldman S, Janmohamed S, Kreuzer J, Leenay M, Michel A, Ong S, Pell JP, Southworth MR, Stough WG, Thoenes M, Zannad F, Zalewski A. Electronic health records to facilitate clinical research. Clin Res Cardiol 2017;(106):1–9. doi: 10.1007/s00392-016-1025-6
8. Mahon P, Hall G, Dekker A, Vehreschild J, Tonon G. Harnessing oncology real-world data with AI. Nat Cancer Nature Publishing Group; 2023;(4):1627–1629. doi: 10.1038/s43018-023-00689-7
9. Berger ML, Ganz PA, Zou KH, Greenfield S. When Will Real-World Data Fulfill Its Promise to Provide Timely Insights in Oncology? JCO Clin Cancer Inform 2024;(8):e2400039. PMID:38950323
10. HL7 FHIR v4.0.1. Available from: http://hl7.org/fhir/R4/index.html [accessed Aug 19, 2024]
11. 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):307–311. PMID:37203668
12. Casaletto J, Bernier A, McDougall R, Cline MS. Federated Analysis for Privacy-Preserving Data Sharing: A Technical and Legal Primer. Annu Rev Genomics Hum Genet 2023;(24):347–368. doi: 10.1146/annurev-genom-110122-084756
13. Hallock H, Marshall SE, ’t Hoen PAC, Nygård JF, Hoorne B, Fox C, Alagaratnam S. Federated Networks for Distributed Analysis of Health Data. Front Public Health 2021;(9):712569. PMID:34660512
14. 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. PMID:35038764
15. Ziegler J, Gruendner J, Rosenau L, Erpenbeck M, Prokosch H-U, Deppenwiese N. Towards a Bavarian Oncology Real World Data Research Platform. Stud Health Technol Inform 2023;(307):78–85. PMID:37697840
16. ONKOSTAR Tumordokumentation, IT-Choice Software AG. Available from: https://www.onkostar.de/tumordokumentation/ [accessed Apr 21, 2024]
17. CREDOS (Tumordokumentation) Software Universitätsklinikum Ulm. Available from: https://www.uniklinik-ulm.de/comprehensive-cancer-center-ulm-cccu/klinisches-krebsregister/software-eigenentwicklungen/credos-tumordokumentation.html [accessed Apr 21, 2024]
18. Einheitlicher onkologischer Basisdatensatz. Available from: https://basisdatensatz.de/basisdatensatz [accessed Apr 19, 2024]
19. Bundesministerium für Gesundheit. Bekanntmachung – Aktualisierter einheitlicher onkologischer Basisdatensatz der Arbeitsgemeinschaft Deutscher Tumorzentren e. V. (ADT) und der Gesellschaft der epidemiologischen Krebsregister in Deutschland e. V. (GEKID). Amtliche Veröff – Bundesanz. 2021. Available from: https://www.bundesanzeiger.de/pub/de/amtliche-veroeffentlichung?1 [accessed Apr 19, 2024]

20. Bayerische Staatskanzlei. Bayerisches Krebsregistergesetz (BayKRegG). 2017. Available from: https://www.gesetze-bayern.de/Content/Document/BayKRegG/true [accessed Apr 19, 2024]
21. entici. 2024. Available from: https://gitlab.com/mri-tum/aiim/entici [accessed May 6, 2024]
22. Geidel L, Bahls T, Hoffmann W. A generic pseudonymization tool as a module of Central Data Management for medical research data (Ein generisches Pseudonymisierungswerkzeug als Modul des Zentralen Datenmanagements medizinischer Forschungsdaten). Abstr 8th Annu Conf Ger Soc Epidemiol DGEpi EV 1st Int LIFE Symp Abstr 8 Jahrestag Dtsch Ges Für Epidemiol 1 Int LIFE Symp Leipzig; 2013. p. pp 245-246.
23. Bialke M, Bahls T, Havemann C, Piegsa J, Weitmann K, Wegner T, Hoffmann W. MOSAIC – A Modular Approach to Data Management in Epidemiological Studies. Methods Inf Med Georg Thieme Verlag KG; 2015;(54):364–371. doi: 10.3414/ME14-01-0133
24. Bialke M, Penndorf P, Wegner T, Bahls T, Havemann C, Piegsa J, Hoffmann W. A workflow-driven approach to integrate generic software modules in a Trusted Third Party. J Transl Med 2015;(13):176. doi: 10.1186/s12967-015-0545-6
25. 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):1929–1944. PMID:25261970
26. Deist TM, Dankers FJWM, Ojha P, Scott Marshall M, Janssen T, Faivre-Finn C, Masciocchi C, Valentini V, Wang J, Chen J, Zhang Z, Spezi E, Button M, Jan Nuyttens J, Vernhout R, van Soest J, Jochems A, Monshouwer R, Bussink J, Price G, Lambin P, Dekker A. Distributed learning on 20 000+ lung cancer patients - The Personal Health Train. Radiother Oncol J Eur Soc Ther Radiol Oncol 2020;(144):189–200. PMID:31911366
27. Rootes-Murdy K, Gazula H, Verner E, Kelly R, DeRamus T, Plis S, Sarwate A, Turner J, Calhoun V. Federated Analysis of Neuroimaging Data: A Review of the Field. Neuroinformatics 2022;(20):377–390. doi: 10.1007/s12021-021-09550-7
28. Geleijnse G, Chiang RC-J, Sieswerda M, Schuurman M, Lee KC, van Soest J, Dekker A, Lee W-C, Verbeek XAAM. Prognostic factors analysis for oral cavity cancer survival in the Netherlands and Taiwan using a privacy-preserving federated infrastructure. Sci Rep 2020; (10):20526. PMID:33239719
29. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, ’t Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data Nature Publishing Group; 2016;(3):160018. doi: 10.1038/sdata.2016.18
30. Apache Kafka. Apache Kafka. Available from: https://kafka.apache.org/documentation/ [accessed May 6, 2024]
31. bzkf/obds-to-fhir. 2024. Available from: https://github.com/bzkf/obds-to-fhir [accessed Aug 16, 2024]
32. Lambarki M, Kern J, Croft D, Engels C, Deppenwiese N, Kerscher A, Kiel A, Palm S, Lablans M. Oncology on FHIR: A Data Model for Distributed Cancer Research. Stud Health Technol Inform 2021;(278):203–210. PMID:34042895

33. MIRACUM FHIR Gateway. Available from: https://github.com/miracum/fhir-gateway [accessed Jun 19, 2024]
34. Gulden C, Stöcker S. MIRACUM FHIR Pseudonymizer. Available from: https://github.com/miracum/fhir-pseudonymizer [accessed Jun 19, 2024]
35. Grimes J, Szul P, Metke-Jimenez A, Lawley M, Loi K. Pathling: analytics on FHIR. J Biomed Semant 2022;(13):23. doi: 10.1186/s13326-022-00277-1
36. Bayerisches Zentrum für Krebsforschung. onco-analytics-on-fhir. 2024. Available from: https://github.com/bzkf/onco-analytics-on-fhir [accessed Jun 25, 2024]
37. Helm - The package manager for Kubernetes. Available from: https://helm.sh/ [accessed May 6, 2024]
38. kubernetes: Production-Grade Container Orchestration. Available from: https://kubernetes.io/ [accessed May 6, 2024]
39. Bayerisches Landesamt für Gesundheit und Lebensmittelsicherheit. Jahresberichte des Bayerischen Krebsregisters. Ausgabe 2023. Krebs in Bayern in den Jahren 2015 bis 2019. Available from: https://www.lgl.bayern.de/gesundheit/krebsregister/auswertung_forschung/jahresberichte/index.htm [accessed Mar 25, 2024]
40. Robert Koch Institut, Gesellschaft der epidemiologischen Krebsregister in Deutschland e.V. Zentrum für Krebsregisterdaten - Krebs in Deutschland für 2019/2020, 14. Ausgabe. Available from: https://www.krebsdaten.de/Krebs/DE/Content/Publikationen/Krebs_in_Deutschland/krebs_in_deutschland_node.html [accessed Jun 19, 2024]
41. Kim H-I, Lim H, Moon A. Sex Differences in Cancer: Epidemiology, Genetics and Therapy. Biomol Ther 2018;(26):335–342. PMID:29949843
42. Jackson SS, Marks MA, Katki HA, Cook MB, Hyun N, Freedman ND, Kahle LL, Castle PE, Graubard BI, Chaturvedi AK. Sex disparities in the incidence of 21 cancer types: Quantification of the contribution of risk factors. Cancer 2022;(128):3531–3540. PMID:35934938
43. Harvey BJ, Harvey HM. Sex Differences in Colon Cancer: Genomic and Nongenomic Signalling of Oestrogen. Genes 2023;(14):2225. PMID:38137047
44. Comprehensive Cancer Center Mainfranken. CCC-MF/bzkf-rwdp-check. 2024. Available from: https://github.com/CCC-MF/bzkf-rwdp-check [accessed Apr 22, 2024]
45. Geldof T, Huys I, Van Dyck W. Real-World Evidence Gathering in Oncology: The Need for a Biomedical Big Data Insight-Providing Federated Network. Front Med 2019;(6):43. PMID:30906740
46. Maier D, Vehreschild JJ, Uhl B, Meyer S, Berger-Thürmel K, Boerries M, Braren R, Grünwald V, Hadaschik B, Palm S, Singer S, Stuschke M, Juárez D, Delpy P, Lambarki M, Hummel M, Engels C, Andreas S, Gökbuget N, Ihrig K, Burock S, Keune D, Eggert A, Keilholz U, Schulz H, Büttner D, Löck S, Krause M, Esins M, Ressing F, Schuler M, Brandts C, Brucker DP, Husmann G, Oellerich T, Metzger P, Voigt F, Illert AL, Theobald M, Kindler T, Sudhof U, Reckmann A, Schwinghammer F, Nasseh D, Weichert W, von Bergwelt-Baildon M, Bitzer M, Malek N, Öner Ö, Schulze-Osthoff K, Bartels S, Haier J, Ammann R, Schmidt AF, Guenther B, Janning M, Kasper B, Loges S, Stilgenbauer S, Kuhn P, Tausch E, Runow S, Kerscher A, Neumann M, Breu M, Lablans M, Serve H. Profile of the multicenter cohort of the German Cancer Consortium's Clinical Communication Platform. Eur J Epidemiol 2023;(38):573–586. doi: 10.1007/s10654-023-00990-w
47. Saesen R, Lacombe D, Huys I. Real-world data in oncology: a questionnaire-based analysis of the academic research landscape examining the policies and experiences of the cancer cooperative groups. ESMO Open 2023;(8):100878. PMID:36822113
48. Ru B, Sillah A, Desai K, Chandwani S, Yao L, Kothari S. Real-World Data Quality Framework for Oncology Time to Treatment Discontinuation Use Case: Implementation and Evaluation Study. JMIR Med Inform 2024;(12):e47744. doi: 10.2196/47744
49. Bayerisches Landesamt für Gesundheit und Lebensmittelsicherheit. Manual der

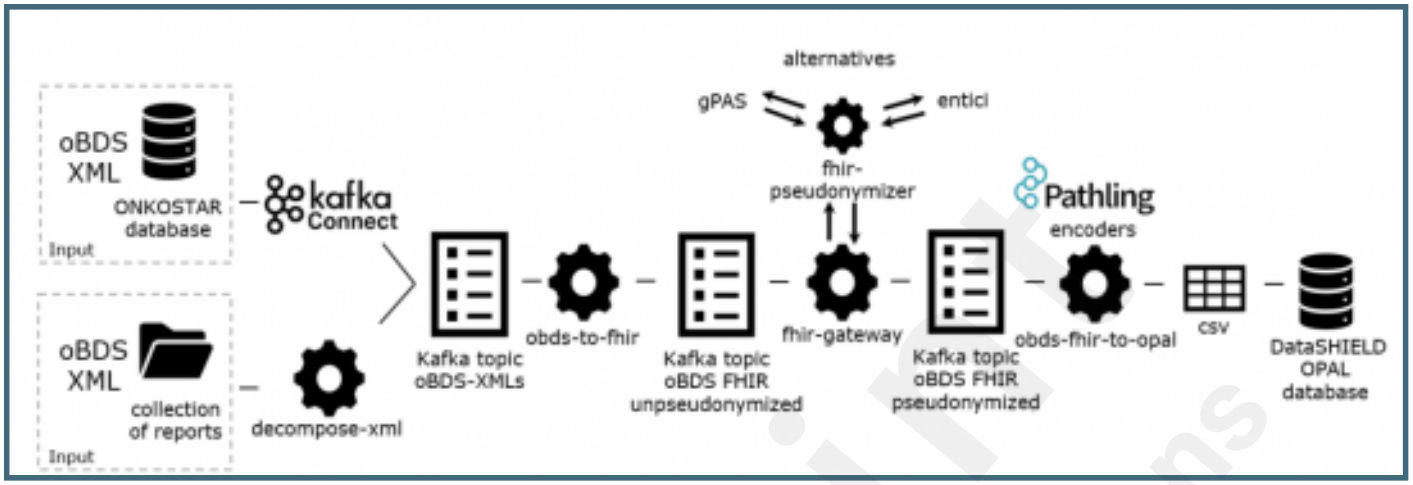
Krebsregistrierung (2018) (GEKID). 2024. Available from: https://www.lgl.bayern.de/downloads/gesundheit/krebsregister/ [accessed Apr 24, 2024]

50. Nasseh D, Schneiderbauer S, Lange M, Schweizer D, Heinemann V, Belka C, Cadenovic R, Buysse L, Erickson N, Mueller M, Kortuem K, Niyazi M, Marschner S, Fey T. Optimizing the Analytical Value of Oncology-Related Data Based on an In-Memory Analysis Layer: Development and Assessment of the Munich Online Comprehensive Cancer Analysis Platform. J Med Internet Res 2020;(22):e16533. doi: 10.2196/16533
51. Swinckels L, Bennis FC, Ziesemer KA, Scheerman JFM, Bijwaard H, Keijzer A de, Bruers JJ. The Use of Deep Learning and Machine Learning on Longitudinal Electronic Health Records for the Early Detection and Prevention of Diseases: Scoping Review. J Med Internet Res 2024;(26):e48320. doi: 10.2196/48320
52. Metzger P, Boerries M. [The collaborative project “Personalized medicine for oncology” (PM4Onco) as part of the Medical Informatics Initiative (MII)]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2024;(67):668–675. PMID:38739266
53. Hunger M, Bardenheuer K, Passey A, Schade R, Sharma R, Hague C. The Value of Federated Data Networks in Oncology: What Research Questions Do They Answer? Outcomes From a Systematic Literature Review. Value Health 2022;(25):855–868. doi: 10.1016/j.jval.2021.11.1357

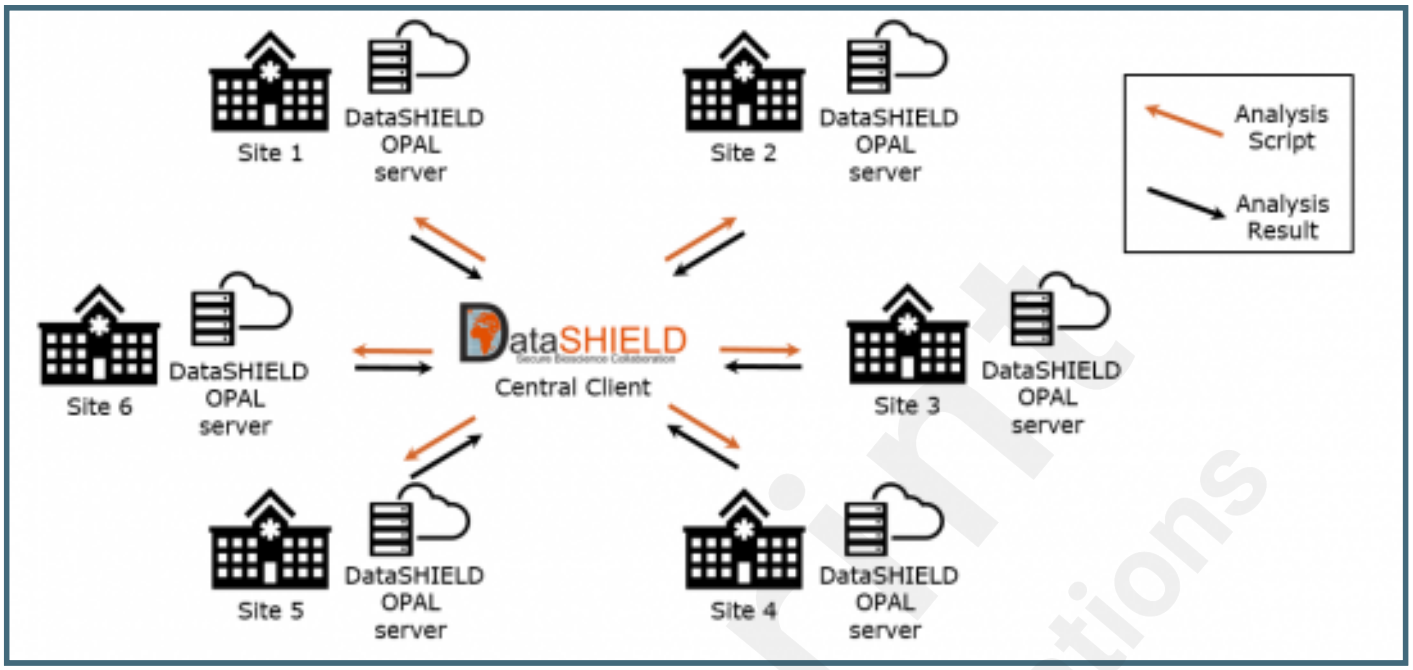
Supplementary Files

Figures

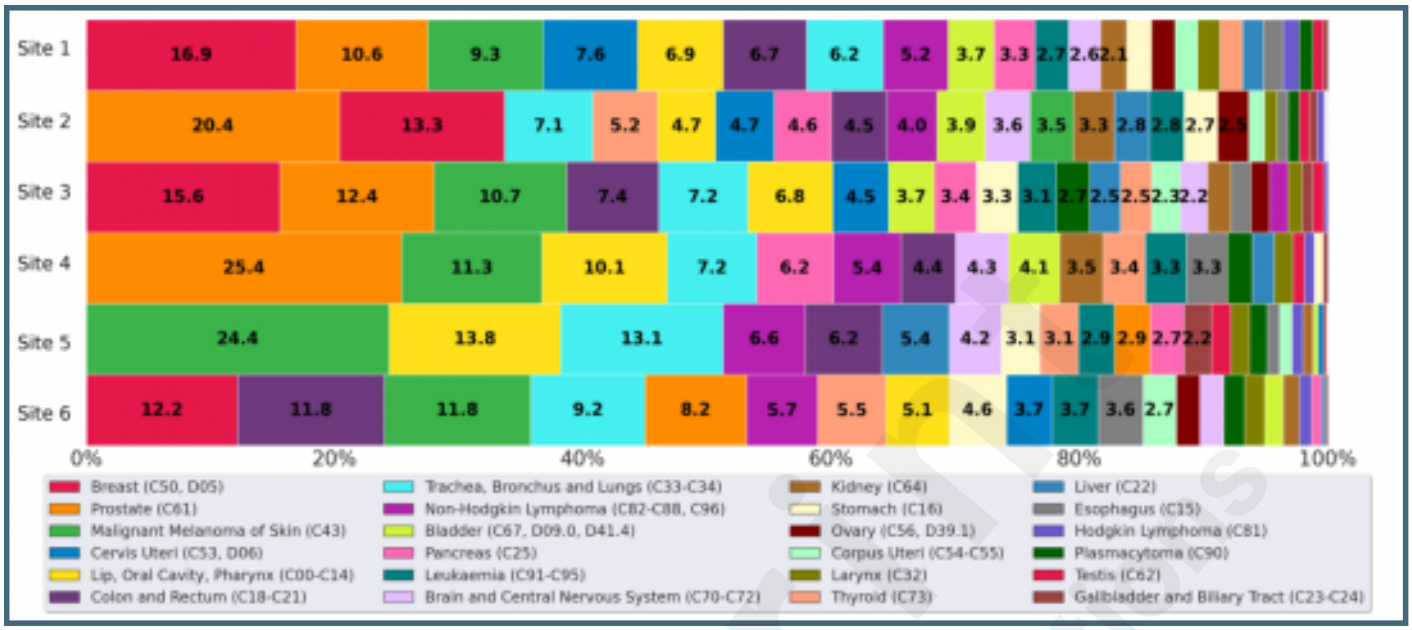
Architecture of the pipeline for transforming oBDS data into a final analysis format. This figure illustrates the key components, including two input interfaces, generic integration with pseudonymization services, and support for two output formats, enabling the conversion of oBDS data from XML to HL7 FHIR and to a tabular format suitable for FA.



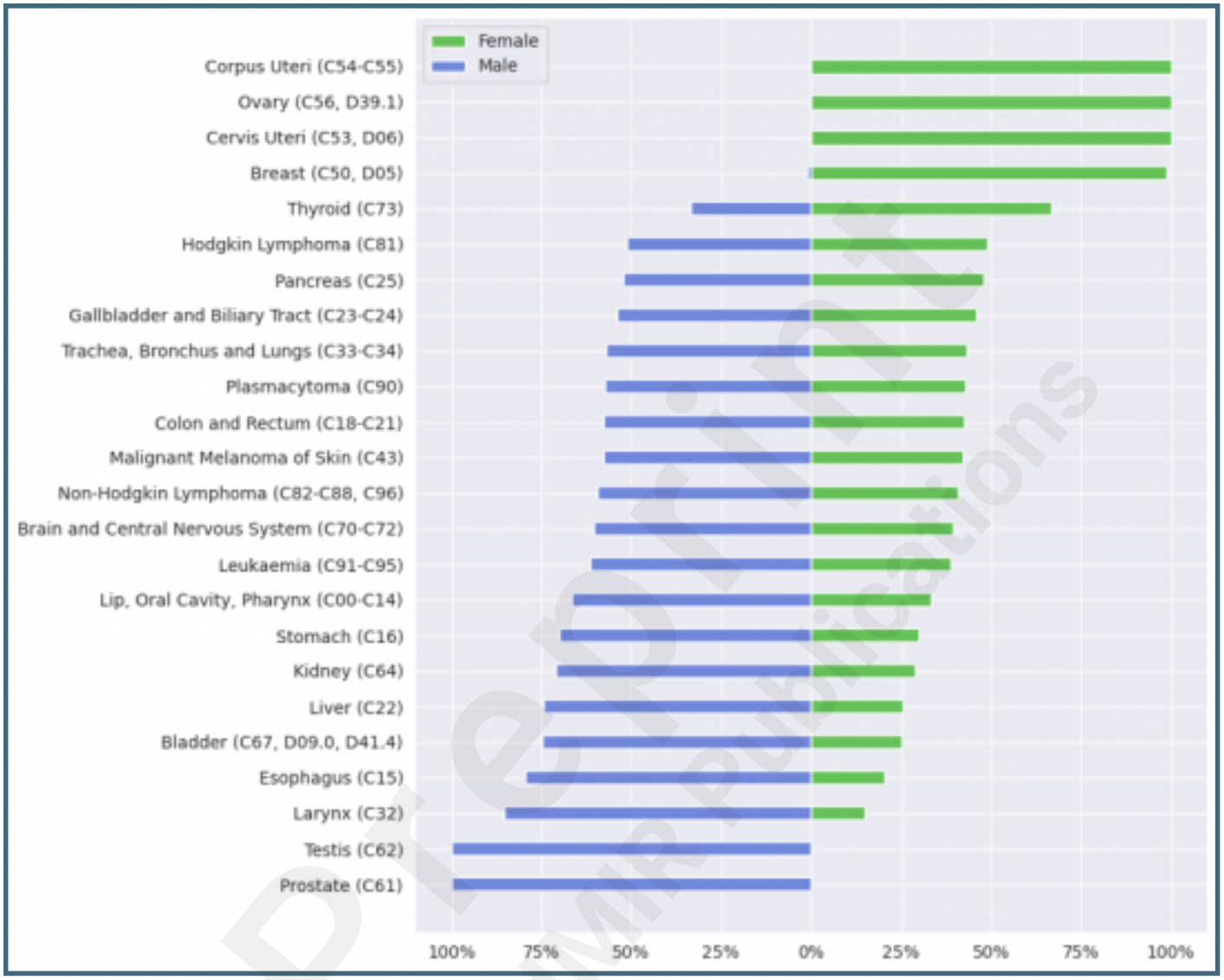
The FA network illustrating the OPAL servers as the local analysis endpoints at each of the six BZKF sites.



Distribution of tumor entities at each hospital for cases diagnosed in 2022 (site 1-5) and 2021 (site 6). Results of the FA across six locations in relative numbers per site (research question Q1).



Distribution of administrative gender among the tumor entities for cases diagnosed in 2022 (site 1-5) and 2021 (site 6). Mean results of the FA across six locations in relative numbers (research question Q2). Other genders are omitted from the visualization due to the presence of only three cases.



Calculation of the entity-wise deviation (Mean Absolute Percentage Error).

$\hat{y}$: predicted value for entity i (federated analysis result)
y: gold standard for entity i (evaluation with tumor documentation system)

For each entity i, calculate entity-wise deviation and return mean value over all 24 entity-wise deviations (Mean Absolute Percentage Error):

$$\overline{Deviation - entitywise} = \frac{1}{n}\sum_{i=1}^{n}\frac{|\hat{y}_i - y_i|}{y_i} \times 100 \quad with \quad n = 24$$