Advancing FAIR sharing for epidemiological, clinical and public health study data Services and Community Integration of the German National Research Data Infrastructure NFDI4Health

Iris Pigeot
Leibniz Institute for Prevention Research and Epidemiology – BIPS, Bremen, Germany; University of Bremen, Bremen, Germany

Manuela Peters
mpeters@leibniz-bips.de

Leibniz Institute for Prevention Research and Epidemiology – BIPS, Bremen, Germany
https://orcid.org/0000-0001-6012-1654

Johannes Darms
ZB MED – Information Centre for Life Sciences, Bonn, Germany; University of Bonn, Bonn, Germany

Jens Dierkes
University of Cologne, Cologne, Germany

Martin Golebiewski
Heidelberg Institute for Theoretical Studies (HITS), Heidelberg, Germany

Carsten Oliver Schmidt
University Medicine Greifswald, Greifswald, Germany

Ulrich Sax
University Medical Center Göttingen, Göttingen, Germany

Matthias B. Schulze
German Institute of Human Nutrition Potsdam-Rehbruecke (DIfE), Nuthetal, Germany; University of Potsdam, Nuthetal, Germany

Hajo Zeeb
Leibniz Institute for Prevention Research and Epidemiology – BIPS, Bremen, Germany; University of Bremen, Bremen, Germany

Roman Baum
ZB MED – Information Centre for Life Sciences, Bonn, Germany; University of Bonn, Bonn, Germany

Oya Beyan
University of Cologne, Cologne, Germany; University Hospital Cologne, Cologne, Germany

Holger Fröhlich

University of Bonn, Bonn, Germany; Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin, Germany

Julia Fürst
ZB MED – Information Centre for Life Sciences, Bonn, Germany

Marie L. Gebhardt
TMF - Technology, Methods, and Infrastructure for Networked Medical Research, Berlin, Germany

Theresa Idda
ZB MED – Information Centre for Life Sciences, Bonn, Germany; University of Bonn, Bonn, Germany

Timm Intemann
Leibniz Institute for Prevention Research and Epidemiology – BIPS, Bremen, Germany

Franziska Jannasch
German Institute of Human Nutrition Potsdam-Rehbruecke (DIfE), Nuthetal, Germany

Elisa Kasbohm
University Medicine Greifswald, Greifswald, Germany

Dennis-Kenji Kipker
Leibniz Institute for Prevention Research and Epidemiology – BIPS, Bremen, Germany; cii.research gGmbH, Frankfurt am Main, Germany

Toralf Kirsten
Leipzig University, Leipzig, Germany

Sophie Anne Inès Klopfenstein
Berlin Institute of Health at Charité - Universitätsmedizin Berlin, Berlin, Germany

Ekaterina Kutafina
University of Cologne, Cologne, Germany; University Hospital Cologne, Cologne, Germany

Johanna Liebig
Robert Koch Institute (RKI), Berlin, Germany

Birte Lindstädt
ZB MED – Information Centre for Life Sciences, Bonn, Germany

Matthias Löbe
Leipzig University, Leipzig, Germany

Frank Meineke
Leipzig University, Leipzig, Germany

Katharina Nimptsch
Max Delbrück Center for Molecular Medicine (MDC), Berlin, Germany

Ute Nöthlings
University of Bonn, Bonn, Germany

Tracy Bonsu Osei
Max Delbrück Center for Molecular Medicine (MDC), Berlin, Germany

Ines Perrar
University of Bonn, Bonn, Germany

Tobias Pischon
Max Delbrück Center for Molecular Medicine (MDC), Berlin, Germany
Fabian Prasser
Berlin Institute of Health at Charité - Universitätsmedizin Berlin, Berlin, Germany
Julia Sasse
ZB MED – Information Centre for Life Sciences, Bonn, Germany; University of Bonn, Bonn, Germany
Lina H. Schaare
U Bremen Research Alliance e.V., Bremen, Germany
Florian Schwarz
German Institute of Human Nutrition Potsdam-Rehbruecke (DIfE), Nuthetal, Germany
Sofia Maria Siampani
Max Delbrück Center for Molecular Medicine (MDC), Berlin, Germany
Sylvia Thun
Berlin Institute of Health at Charité - Universitätsmedizin Berlin, Berlin, Germany
Carina Nina Vorisek
Berlin Institute of Health at Charité - Universitätsmedizin Berlin, Berlin, Germany
Juliane Fluck
ZB MED – Information Centre for Life Sciences, Bonn, Germany; University of Bonn, Bonn, Germany

---

Research Article

Keywords: Data sharing, Federated data analysis, Health Study Hub, Metadata, Reusability, Synthetic data, FAIR, Health data literacy, Data quality

Posted Date: June 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10082861/v1

Advancing FAIR sharing for epidemiological, clinical and public health study data

Services and Community Integration of the German National Research Data Infrastructure NFDI4Health

Iris Pigeot[1,2], Manuela Peters[1], Johannes Darms[3,4], Jens Dierkes[5], Martin Golebiewski[6], Carsten Oliver Schmidt[7], Ulrich Sax[8], Matthias B. Schulze[9,10], Hajo Zeeb[1,2], Roman Baum[3,4], Oya Beyan[5,11], Holger Fröhlich[4,12], Julia Fürst[3], Marie L. Gebhardt[13], Theresa Idda[3,4], Timm Intemann[1], Franziska Jannasch[9], Elisa Kasbohm[7], Dennis-Kenji Kipker[1,14], Toralf Kirsten[15], Sophie Anne Inès Klopfenstein[16], Ekaterina Kutafina[5,11], Johanna Liebig[17], Birte Lindstädt[3], Matthias Löbe[15], Frank Meineke[15], Katharina Nimptsch[18], Ute Nöthlings[4], Tracy Bonsu Osei[18], Ines Perrar[4], Tobias Pischon[18], Fabian Prasser[16], Julia Sasse[3,4], H. Lina Schaare[19], Florian Schwarz[9], Sofia Maria Siampani[18], Sylvia Thun[16], Carina Nina Vorisek[16], Juliane Fluck[3,4]

[1] Leibniz Institute for Prevention Research and Epidemiology – BIPS, Bremen, Germany

[2] University of Bremen, Bremen, Germany

[3] ZB MED – Information Centre for Life Sciences, Bonn, Germany

[4] University of Bonn, Bonn, Germany

[5] University of Cologne, Cologne, Germany

[6] Heidelberg Institute for Theoretical Studies (HITS), Heidelberg, Germany

[7] University Medicine Greifswald, Greifswald, Germany

[8] University Medical Center Göttingen, Göttingen, Germany

[9] German Institute of Human Nutrition Potsdam-Rehbruecke (DIfE), Nuthetal, Germany

[10] University of Potsdam, Nuthetal, Germany

[11] University Hospital Cologne, Cologne, Germany

[12] Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin, Germany

[13] TMF - Technology, Methods, and Infrastructure for Networked Medical Research, Berlin, Germany

[14] cii.research gGmbH, Frankfurt am Main, Germany

[15] Leipzig University, Leipzig, Germany

[16] Berlin Institute of Health at Charité - Universitätsmedizin Berlin, Berlin, Germany

[17] Robert Koch Institute (RKI), Berlin, Germany

[18] Max Delbrück Center for Molecular Medicine (MDC), Berlin, Germany

[19] U Bremen Research Alliance e.V., Bremen, Germany

Acknowledgements

This work was done as part of the NFDI4Health Consortium (www.nfdi4health.de). We gratefully acknowledge the financial support of the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 442326535.

Abstract

Background: Access to high-quality, FAIR health data is essential for advancing epidemiological, public health and clinical research. NFDI4Health, one of the consortia of Germany's National Research Data Infrastructure (NFDI), was established in 2020 to address this need by improving data FAIRness for the scientific community.

Methods: During the initial funding phase, NFDI4Health developed key infrastructure components and services focusing on interoperability, data sharing and research support. Our approach was guided by user needs and real-world use cases in alignment with (inter)national FAIR standards and infrastructures.

Results: We advanced findability of health data by establishing a central Health Study Hub and connecting it to local infrastructures, e.g., through the Local Data Hub software, to facilitate transfer of metadata from the local infrastructures to the Health Study Hub. Accessibility was improved by expanding the German Research Data Portal for Health (FDPG) to provide central access to study data and by providing tools to support anonymisation and synthetic data generation. We also developed an interoperable metadata schema for publishing study data and implemented it in the Health Study Hub. To further improve interoperability, a NFDI4Health FAIR sharing collection and AI support for metadata annotation and harmonisation workflows were established. To enhance reusability, data quality assessment tools were further developed, and two frameworks for federated analysis of sensitive health data were piloted. Intensive engagement with our communities through training, advisory services, and collaborative development ensured user relevance.

Conclusion/outlook: NFDI4Health has established a scalable, interoperable, and user-centred infrastructure to support FAIR data sharing in health research. Future work will focus on further developing and consolidating the infrastructure, expanding cross-domain data integration, fostering broader adoption within our research community, and strengthening national and international collaboration.

Keywords: Data sharing, Federated data analysis, Health Study Hub, Metadata, Reusability, Synthetic data, FAIR, Health data literacy, Data quality

1. Introduction

The capacity to access and reuse health data across institutions and regions is becoming increasingly imperative for clinical, epidemiological and public health research. Moreover, the global COVID-19 pandemic has emphasised the pressing need for secure health data sharing across national boundaries to facilitate rapid research, effective public health responses and evidence-based policymaking. In this context, the European Health Data Space (EHDS) [1] emerged as a strategic initiative to enable trustworthy reuse of health data for healthcare delivery, research, and innovation. As of 2031, the EHDS will be extended to encompass also health study data in addition to healthcare data. If successful, the EHDS will enable, for example, the investigation of rare diseases by pooling interoperable cohort data from all over Europe to increase the sample size and the heterogeneity of the population under study.

In parallel, the federal government in Germany initiated the National Research Data Infrastructure (NFDI). Following a bottom-up approach, 26 domain-specific consortia were funded with the objective of establishing the German NFDI [2] and integrating this infrastructure into the European Open Science Cloud (EOSC). Within NFDI, the National Research Data Infrastructure for Personal Health Data (NFDI4Health) is devoted to sensitive health data generated in epidemiological, public health, and clinical studies [3]. The NFDI4Health consortium consists of 18 partner organisations and 43 participating institutions. In addition, by enabling interoperability of and access to data from different domains, Germany's NFDI will support health researchers, e.g., to get a holistic view of the aetiology of a disease by also accounting for environmental and social factors.

The mission of NFDI4Health is not only to increase the value of epidemiology, public health and clinical trial-based medicine by making German health data findable (F), accessible (A), interoperable (I) and reusable (R) (in short: FAIR) [4], but also to boost their utilisation for health research. More detailed, the objectives of NFDI4Health are to support FAIR publication and sharing of personal health study data, to enable federated access to distributed data sources, to assist researchers in legal and ethical compliance and to promote interoperability within national and European health data ecosystems.

Our activities also consider currently existing national and international trial registries that serve as established platforms for the reporting of health studies. Among the most relevant are the German Clinical Trials Register (DRKS) [5], clinicaltrials.gov [6], the WHO International Clinical Trials Register Platform (ICTRP) [7] and the European Clinical Trials Information System (CTIS) [8]. Most clinical trials and some epidemiological studies are registered in one of these registries. However, essential information about study protocols or detailed data corpus information is often lacking. In addition, metadata are not standardised, which also hampers findability of appropriate studies. Automated workflows and application programming interfaces (APIs) are only partially available, and regular updates by data holding organisations (DHOs) are not ensured. Another obstacle is the lack of established structures to support their systematic reuse. In Germany, the Network of Coordinating Centers for Clinical Studies (KKSN), which is a member of NFDI4Health, and the Network University Medicine (NUM) study network (NUM SN) focus on supporting multi-centre clinical and epidemiological studies in Germany. On the European level, the European Clinical Research Infrastructure Network (ECRIN) supports multi-national research and provides advice as well as services for setting up and managing investigator-initiated clinical studies and those under the responsibility of small or medium enterprises (SME). One important international epidemiological registry is the Maelstrom catalogue [9]. It currently contains 451 individual studies, including collections from Swedish and Swiss cohort networks. It is obvious that such registries are extremely valuable when planning a pooling study on a certain research question that needs a large sample size and a critical appraisal of study results across different populations. Well-known examples are the

Global Burden of Disease Studies [10] and the pooling studies of the NCD (Non-Communicable Disease) Risk Factor Collaboration [11].

In close collaboration with the aforementioned initiatives, NFDI4Health develops and consolidates standards and services to support data holders and users in building a central infrastructure for all health-related studies in Germany to facilitate comprehensive and privacy-preserving reuse of health study data.

In this paper, we present the NFDI4Health infrastructure and services and discuss their integration into the broader data ecosystem of research data infrastructures.

2. Methods

To establish a national data infrastructure for personal health data, NFDI4Health develops processes, standards and tools in close alignment with community needs. We started with a use-case-driven approach to ensure that standards and services for FAIR data management are firmly grounded in the practical requirements of contemporary health research. In parallel, engagement and participatory processes were implemented to incorporate feedback into the development process and to communicate our services at an early stage. In accordance with the FAIR principles, NFDI4Health worked on standards, software, and tools to facilitate local and central searchability of health data, improve accessibility, ensure interoperability, and enhance reusability. Secure data access, guided by the Five Safes framework (see below), has been a central principle throughout these developments.

Using an iterative, community-driven approach, requirements were gathered and validated through structured stakeholder workshops and dedicated working groups, with contributions from researchers, data providers, and infrastructure specialists. A sign-off process was employed to prioritise features, balance functional and non-functional requirements, and ensure alignment with user needs, supported through a small, specialised teams. This structure enabled agile progress and responsiveness to evolve requirements, while close exchange among teams ensured interoperability across the service portfolio.

Although the boundaries between the FAIR principles are not sharp, we will describe our services across their four key areas: Findability, Accessibility, Interoperability and Reusability (see Figure 1).

Figure 1: FAIR Services provided by NFDI4Health

2.1 Guiding principle to enable data access

To comply with national (Health Data Use Act (GDNG)) and European laws (European Health Data Space (EHDS), General Data Protection Regulation (GDPR)) when providing specific services to epidemiological and clinical researchers, NFDI4Health has adopted the Five Safes framework [12]. This framework provides a unified perspective on privacy protection by considering risks and safeguards along five different dimensions: safe projects, safe people, safe data, safe settings, and safe outputs. It is particularly suited for application to federated networks [13] and serves as a basis for structuring training activities and service implementation.

- **Safe projects** ensures that only well-defined, ethically approved and legally compliant research projects gain access to sensitive data.
- **Safe people** requires that individuals accessing sensitive data are trustworthy and appropriately qualified.
- **Safe data** focuses on ensuring appropriate handling of sensitive data, e.g. by suppression or aggregation.
- **Safe settings** emphasise secure environments for analysis.
- **Safe outputs** ensure that only non-disclosive results leave secure environments.

By considering risks and safeguards across all those dimensions, a legally robust yet researcher-friendly path is provided to reuse German data from clinical, epidemiological, and public health studies.

2.2 Stakeholder involvement process

NFDI4Health defined a diverse portfolio of relevant use cases (Figure 2; ranging from large longitudinal cohort studies, clinical trials, secondary data and record linkage, to radiomics and imaging AI) to systematically capture a broad spectrum of data types, workflows, and stakeholder expectations encountered in real-world settings of epidemiological, clinical and public health research in Germany. Each use case considered specific data interoperability challenges, metadata requirements, terminological bindings, use and access procedures, and data sharing scenarios. During iterative cycles of design, prototyping, and evaluation, these use cases served as both test beds and drivers for selecting or developing suitable standards, reusable services and tools. Moreover, by engaging data producers, data stewards, and end users in collaborative workshops and pilot implementations, the use case-approach uncovered critical requirements for user interfaces, governance frameworks, and training materials. This fostered community buy-in and ensured that technical artifacts remained aligned with ethical, legal, and organisational constraints. The use case-methodology thus provided a scalable blueprint for continuous refinement of standards and services in response to evolving scientific needs.

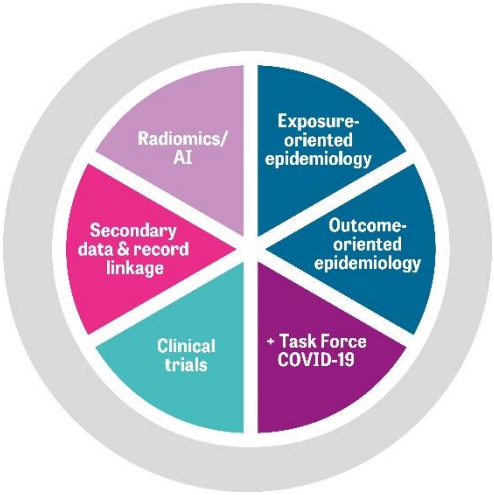

Figure 2: Use cases of NFDI4Health

To narrow down the areas of "Exposure-oriented epidemiology" and "Outcome-oriented epidemiology" (Figure 2), we focussed on "Nutritional epidemiology" and "Epidemiology of chronic diseases". These two use cases are employed throughout the paper to illustrate the use case-driven design approach, for which we translated the general FAIR-oriented framework into domain-specific pipelines that reflect the complex data structures and analytical workflows typical of dietary and long-term disease research in epidemiology.

From the outset, NFDI4Health was supported by a User Advisory Board where we discussed options for impactful community engagement. The User Advisory Board comprised representatives of the research community, professional societies, and patient organisations. We also engaged with other groups as, e.g., study participants. A portfolio of actions, comprising target group-specific workshops, network meetings, information and communication materials, as well as training offers for early career researchers (ECRs), has been provided to ensure that community needs are appropriately accounted for by our consortium. The portfolio of community engagements has been regularly reviewed and updated by a community outreach committee (COC).

2.3 Services for findability

The Health Study Hub provides the central entry point for discovering German health study data and acts as the portal to the NFDI4Health infrastructure. It is based on the open-source Dataverse platform [14], with a customised user interface that emphasises the presentation of rich metadata and domain-specific exploration and filtering features.

The Local Data Hub [15] was developed in collaboration with the FAIRDOM community using FAIRDOM SEEK [16]. It enables institutions to manage their research data decentrally within their premises by applying best practices for research data management (RDM) and FAIRness. By retaining full governance over their sensitive data, it provides granular access control to safeguard sensitive assets. It enables the comprehensive presentation of local research projects, including associated publications, funding programmes or sponsor information, study documents and datasets, serving as a complete 'showroom' for research activities.

2.4 Services for accessibility

The Research Data Portal for Health (FDPG) [17] provides a standardised workflow for requesting access to sensitive data across multiple data sources. Developed by the Medical Informatics Initiative

(MII) [18], it already serves as a central access point for research projects reusing routine hospital data. Researchers submit a single data reuse proposal via an online service, which is then distributed to the relevant data integration centres of the MII member hospitals and their governance boards. The service supports shallow integration via email notification and management via web services and custom integration with internal data request management services via its APIs. Once a data request proposal is granted, access to the data is provided.

To support the DHOs with the implementation of the Five Safes we provide guidance and tools [19] for anonymisation, synthetic data generation and privacy risk assessment. Anonymisation is an important component of responsible data sharing, ensuring the privacy of data subjects while enabling secondary use of research data. Though well-established, anonymisation can be challenging to achieve for high-dimensional data, and it requires configuration efforts. Data synthetisation offers a promising approach to reconcile research needs with privacy requirements by generating artificial data that preserve the statistical characteristics of the original dataset.

For instance, our data stewards help with using the widely used anonymisation software ARX [20], which offers a catalogue of techniques, including masking, i.e., replacing true values by fictious ones, and generalisation, i.e., replacing exact information like birth dates by broader categories. ARX is open source, runs on all major operating systems and remains under active development, making it a robust and sustainable solution for research data anonymisation and re-identification risk analysis.

In addition, our data stewards give advice on applying the synthetisation software package Variational Autoencoder Modular Bayesian Networks (VAMBN) [21] and VAMBN Memorized Time Points (VAMBN-MT) [22]. This software is particularly useful for health studies by addressing complexities such as mixed data types, irregular time intervals and missing values. VAMBN is a hybrid modelling approach, in which initially the high-dimensional feature space of a clinical study is grouped into modules, e.g., representing demographics, clinical assessments or biomarkers, and then connected via a Bayesian network.

Overall, these tools support the secure and compliant reuse of sensitive data while preserving privacy.

2.6 Standards and tools for interoperability

In addition to making data findable and accessible, NFDI4Health also ensures that information is described such that it is interoperable across systems and communities. Interoperability is essential for the structured, meaningful exchange and understanding of health data across devices, systems, institutions, and use cases. International standards play a key role by providing shared technical and conceptual frameworks that allow heterogeneous systems to process health research information consistently. Standards Developing Organizations (SDOs) such as Health Level 7 (HL7®), the International Organization for Standardization (ISO) or Systematized Nomenclature of Medicine – Clinical Terms (SNOMED) International develop, maintain, and disseminate these standards through interdisciplinary consensus-driven collaboration. Health data standards are commonly divided into syntactic or semantic categories.

- Syntactic standards define how data are structured and formatted for exchange (e.g., HL7 Fast Healthcare Interoperability Resources (FHIR®), Clinical Document Architecture (CDA), HL7 V2) ensuring reliable transfer across different technical environments.
- Semantic standards define the meaning of information through controlled vocabularies, terminologies and classifications like SNOMED CT, Logical Observation Identifiers Names and Codes

(LOINC), International Classification of Diseases (ICD), or the National Cancer Institute Thesaurus (NCIt), thus enabling precise communication, understanding, computability, and data analysis.

In this respect, an information model is needed that defines the required metadata elements and that is aligned with international standards such as DataCite [23] and HL7® FHIR®. Persistent identifiers, such as ORCID and digital object identifiers (DOIs), must be incorporated and domain-specific ontologies and terminologies need to be integrated like SNOMED CT, LOINC and Medical Subject Headings (MeSH) to ensure semantic interoperability (see also Section 3.3.1). For practical application of this information model, policies, guidelines and supporting tools are required, such as an annotation workbench and a terminology service.

2.7 Enabling reuse

The goal of the FAIR guiding principles is reuse for which ensuring high data quality is a key prerequisite. According to the ISO 8000-2:2022 standard, data quality is defined in relation to expectations and requirements, meaning it is not absolute but rather depends on the intended use. Consequently, making these expectations explicit and capturing them in a structured, machine-readable format is essential to ensure transparent, reproducible, and potentially automated data quality assessments. Such assessments also play a significant role during secondary data use, helping to evaluate their suitability for specific research questions. Therefore, the same dataset may be deemed of high or low quality depending on the context of use. The R-package dataquieR [24] provides a framework for systematically assessing data quality in terms of data integrity, completeness, and correctness that needs to be extended for our purposes.

For the secure reuse and analysis of sensitive health data, Secure Processing Environments (SPEs) are essential. They provide controlled infrastructures where researchers can analyse pseudonymised or otherwise protected data without the need for direct data download. Within NFDI4Health, SPEs also encompass frameworks for distributed analyses, in which data stay at their respective institutions and only aggregated results are shared. This approach ensures compliance with data protection regulations while facilitating reproducible, high-quality research across epidemiological and clinical studies. Distributed analysis frameworks such as Personal Health Train [25] and DataSHIELD [26] enable secure, privacy-preserving analytics by ensuring that sensitive individual-level data remain with the data owners at all times. Analytical tasks are executed locally and only non-disclosive, aggregate results are shared with research analysts.

3. Results

As a main result, NFDI4Health has established key components of an infrastructure that enables DHOs to FAIRify their epidemiological, clinical, and public health study data as illustrated in Figure 3. Sensitive data remain under the control of the DHOs, which operate their own established governance structures such as data access committees. NFDI4Health ensures central findability by the Health Study Hub, to which study metadata and related documentation are transferred by the DHOs (Figure 3, Step (1)). The Health Study Hub enables users to search for suitable datasets down to variable level to answer their research question (Figure 3, Step (2)) and directs them to the specific application portals. For DHOs lacking a suitable local infrastructure, NFDI4Health provides the Local Data Hub as a

connecting service. The FDPG will support centralised applications for health study data (Figure 3, Step (3): application; Step (5): information about decision), thus offering a one-stop procedure to simplify the application processes to dispersed health data through harmonised procedures, thereby reducing administrative burden for both applicants and DHOs. Nevertheless, final data release decisions remain with the individual DHOs (Figure 3, Step (4)). Distributed analysis systems allow for secure and privacy-compliant data use without the need for physical data transfer (Figure 3, Step (6)).

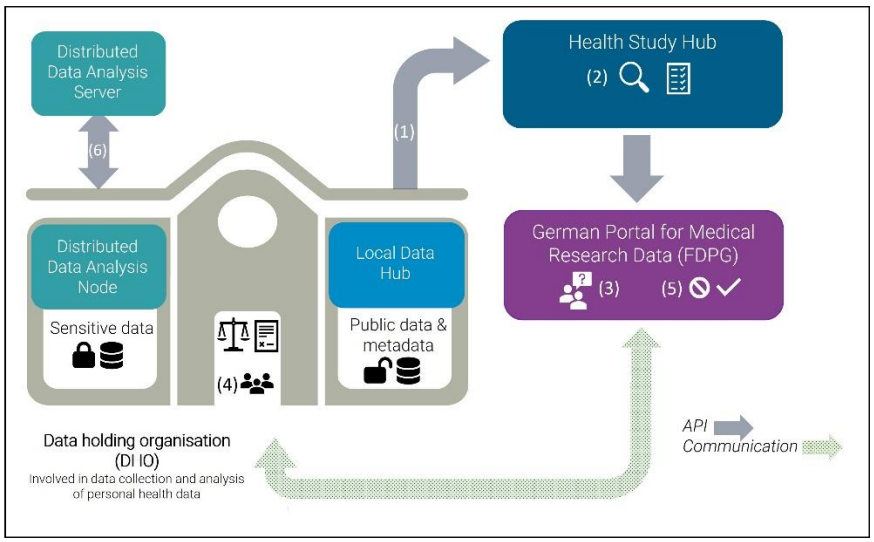

Figure 3: NFDI4Health integrated national framework: The NFDI4Health core infrastructure connects DHOs and researchers, enabling the FAIR sharing and secure reuse of health study data. Metadata are published via the Health Study Hub, while access is harmonised through the FDPG and data analysis is supported by distributed analysis systems.

3.1 Improving findability

3.1.1. Central findability via the Health Study Hub

Originally developed during the pandemic as part of the NFDI4Health COVID-19 Task Force [27], the Health Study Hub has evolved into a sustainable, community-driven infrastructure. Besides content submission via user interfaces or dedicated APIs, the service automatically integrates information from international and national registries and portals, including ClinicalTrials.gov and DRKS. Thus, the service provides data users with a national catalogue of completed, ongoing and planned public health, epidemiological and clinical studies. Further, it supports data providers to publish their research projects in a FAIR manner. For comparison of data dictionaries, the Maelstrom MICA software [28] was integrated into the Health Study Hub. It allows to filter and compare variables based on the Maelstrom taxonomy but requires the prior classification of all variables according to this taxonomy. For further implementation details and API specification, we refer to Darms et al. [29]. The Health Study Hub is currently indexing over 46,000 clinical, public health and epidemiological research studies with German involvement. On the one hand, it automatically collects and deduplicates study metadata from existing major portals. On the other hand, researchers can publish detailed study information via a web-based data capture form and receive support from professional data stewards for data curation and quality assurance.

Besides the publication of study descriptions, data dictionaries, questionnaires and other study-related resources, the Health Study Hub offers the option to create collections, e.g., to bundle all projects of

one institute or related to one topic. As one example, the Health Study Hub provides the “NFDI4Health Cohort Collection” [30]. It covers well established prospective cohort studies participating in the use cases “Nutritional epidemiology” and “Epidemiology of chronic diseases”.

As a further service, the NFDI4Health Search Hub enables direct searches via the search interface, as well as filtering by collection or data type (e.g., studies, data dictionaries or study protocols). Figure 4 shows the entry page of the Health Study Hub. On the right side, the search returns the cohort study I.Family (Investigating the determinants of food choice, lifestyle and health in European children, adolescents and their parents) [31] as an example from the “NFDI4Health Cohort Collection”. It shows an excerpt from the study description, one of the published data dictionaries and the variable classification according to the Maelstrom Taxonomy.

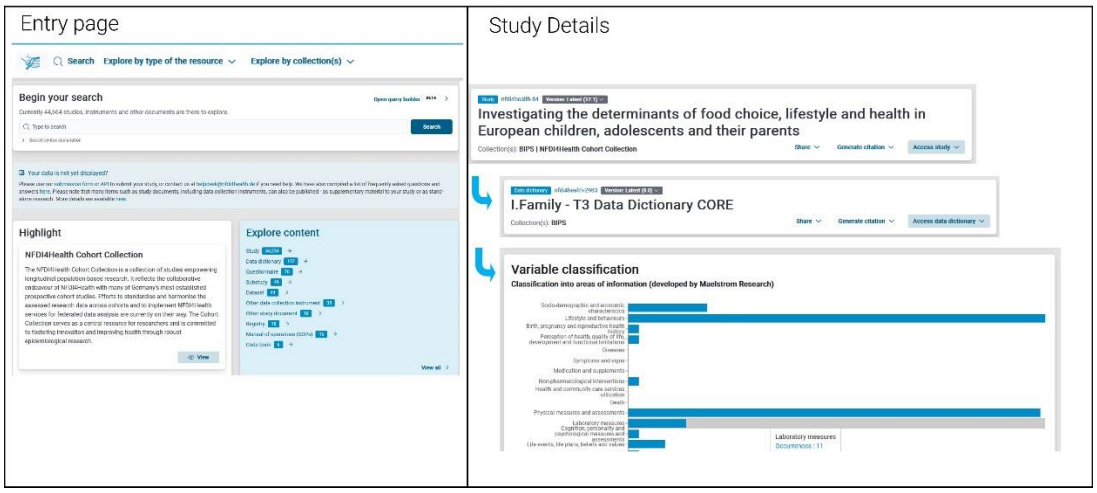

Figure 4: Entry page of the Health Study Hub and example of the NFDI4Health Cohort Collection (https://health-study-hub.de/collection/NFDI4Health_Cohort_Collection)

For the Maelstrom annotation, the DHO may use a semi-automatic approach: an AI-based classifier provides up to five different Maelstrom classes for each variable/question within the Annotation Workbench [32] and experts from the respective DHO then need to select the relevant classes. This method has already been used to annotate more than 100 dictionaries. All data dictionaries and questionnaires that have been classified by the Maelstrom Taxonomy can be viewed, filtered or compared to other dictionaries in the MICA-based variable browser of the Health Study Hub. Figure 5 presents the results of filtering for nutrition-related variables, which are currently found in 73 datasets. Opening the I.Family T3 Data Dictionary Core displays all 43 variables related to nutrition. If needed, all variable definitions can be viewed in detail.

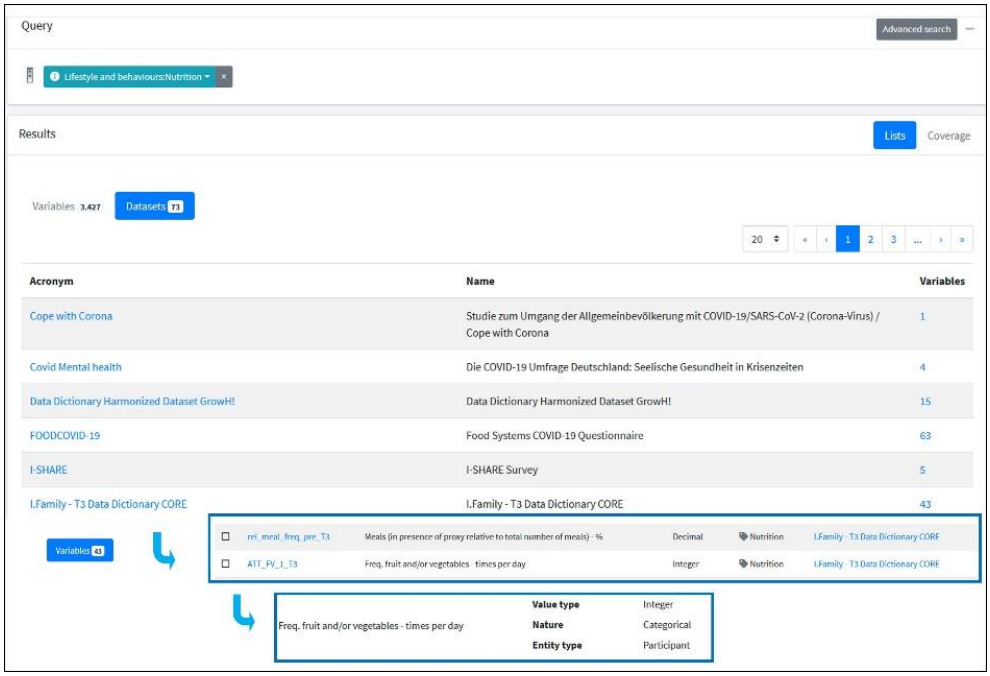

Figure 5: Filtering for variables in the Health Study Hub

3.1.2. Connecting data holding organisations

As complementary service, the Local Data Hub has been extended to provide an API that allows non-sensitive information to be published to the Health Study Hub. The service has been deployed within the community through two NFDI4Health user calls. To reduce technical and organisational barriers for DHOs adapting this system, several Local Data Hub instances (as Software as a Service incl. user support) was provided for a one-year evaluation. In total, 18 institutions, including clinical trial centres, university research units and NUM data integration centres, evaluated the Local Data Hub [33]. The transfer of hosted instances to DHO institutions is still ongoing.

As requested during the evaluation, the automatic export of metadata stored in domain-specific formats (e.g. in the HL7 FHIR standard) to the Local Data Hub is currently under development. As a further request, we now provide support to power users, such as the Robert Koch Institute and the German Centres for Health Research, in regularly publishing new studies, data collection events, and data dictionaries to their respective collections via API. Thus, the Health Study Hub's functionality has been extended to enable these power users to perform quality control of their collections on their own.

3.2 Improving data accessibility

3.2.1 Enabling central data use application

In Germany, regulatory requirements for data access are heterogeneous, resulting in fragmented, non-standardised procedures across DHOs. Researchers therefore face diverse and often complex workflows for data requests. To address this, NFDI4Health inspected existing DHO application processes and developed a harmonised, modular application model, enabling consistent workflows while maintaining flexibility for dataset-specific requirements. This application model has been integrated into the FDPG. As a pilot, we successfully incorporated the EPIC-Potsdam study, conducted

by the German Institute of Human Nutrition Potsdam-Rehbruecke (DIfE), into the FDPG. Applications for EPIC-Potsdam data are handled by DIfE staff and reviewed by the Use and Access Committee (UAC). This first inclusion of an epidemiological study in this infrastructure marked an important milestone since it demonstrates the technical feasibility of connecting NFDI4Health resources to the FDPG and establishes the basis for extending the portal to additional epidemiological and clinical studies.

3.3.2 Anonymisation and synthetisation

Legally, synthetic data may be treated as anonymised if re-identification would require disproportionate effort in terms of time, cost or resources [34]. Consequently, both anonymised and synthesised data require re-identification risk assessments. Even when these data can be shared more openly, user agreements and, where appropriate, user training remain essential to ensure responsible handling. In addition, implementing such techniques is complex. Thus, established, well-tested tools are recommended. However, the landscape of open-source anonymisation tools is diverse and often difficult to navigate due to varying capabilities and maturity levels. To assist researchers, NFDI4Health therefore compiled a comprehensive review of open-source anonymising tools for tabular data to guide the selection of the most suitable tool and anonymisation method according to the data type and context [35].

Recent advances in synthetic data generation offer promising alternatives for privacy preserving data sharing, but there is a need to account for the specific requirements of epidemiological data. Based on the use case “Nutritional epidemiology”, VAMBN has thus been extended to VAMBN Memorized Time Points (VAMBN-MT) [22] to improve the reconstruction of longitudinal dependencies, as they occur in cohort studies, via Long Short-Term Memory networks. Synthetic datasets were generated based on various patient-level datasets using different models and privacy settings and evaluated for data fidelity (similarity to original data) and privacy protection [22,36,37]. For example, synthetic data from the Dortmund Nutritional and Anthropometric Longitudinally Designed (DONALD) study generated using VAMBN-MT successfully replicated temporal and age-related trends in sugar intake in children observed in real data. Furthermore, privacy risk analysis has been applied to this synthetic dataset using the Anonymeter unified framework for quantifying privacy risk [38]. The framework was used for evaluating multiple possible vulnerabilities, which are specifically based on privacy risks that are considered by the European Data Protection Board, i.e. singling out, linkability, and attribute inference.

To evaluate the quality of synthetic data, but also potential privacy risks, we subsequently implemented the publicly available web-based tool SYNDAT [39]. The open-source tool SYNDAT enables researchers and developers to objectively assess the quality and utility of synthetic data in terms of visualisations and evaluation metrics meticulously tailored to clinical patient data studies.

3.3 Interoperability – Identifying syntactic and semantic standards

Within the NFDI4Health initiative, relevant syntactic and semantic standards for health research were identified through interdisciplinary workshops with stakeholders involved in the use cases, mapping of study questionnaires or data dictionaries, and targeted literature reviews. This effort led to the creation of the NFDI4Health FAIRsharing Collection [40], comprising seven syntactic standards, thirty-two semantic standards, and nine hybrid standards that combine both syntactic and semantic elements. In addition, 101 ISO standards from ISO Technical Committees 215 (Health Informatics) and 276 (Biotechnology) were identified as potentially relevant for health research applications [41].

3.3.1. Metadata schema (MDS)

To improve the standardised publication and interoperability of health study metadata, NFDI4Health developed a tailored metadata schema (MDS) [42]. The MDS enhances FAIRness of metadata from clinical, epidemiological, and public health research while maintaining compatibility with external metadata models of other resources and with standard terminologies. Its modular design combines generic elements with use case-specific metadata extensions (see Figure 6).

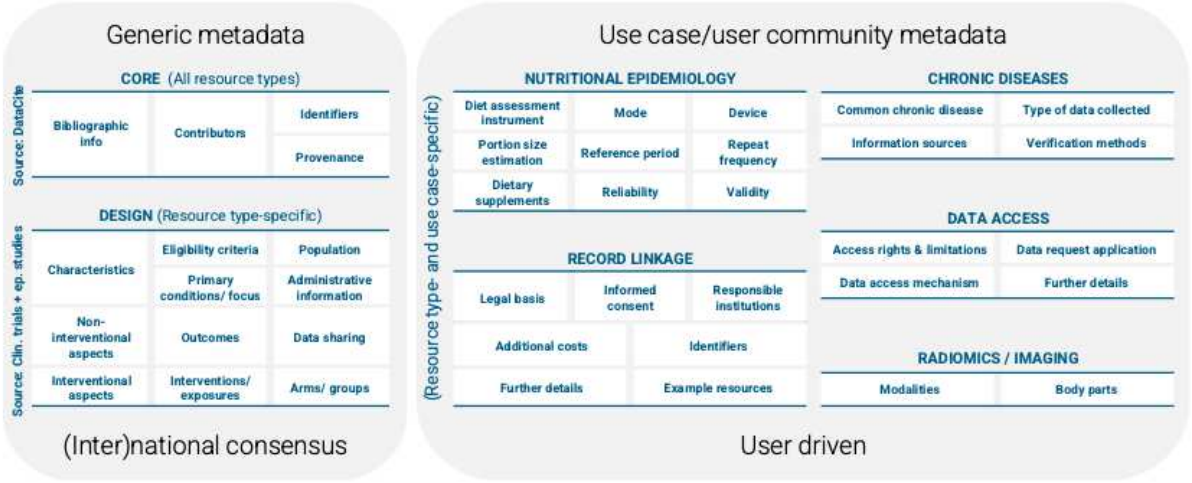

Figure 6: The NFDI4Health MDS for health research studies with its mandatory core and design modules for general metadata (left panel) and domain-specific optional modules for specific use cases (right panel).

For appropriate use case-specific metadata extensions, we conducted surveys among 16 participating epidemiological studies. The principal investigators were asked to identify the necessary metadata that are needed to thoroughly describe their data, e.g., details of dietary assessment methods (use case "Nutritional epidemiology") and standardised case definitions for conditions such as diabetes, cardiovascular disease, and mortality (use case "Epidemiology of chronic diseases") [43].

The MDS is continuously further developed using the online platform ART-DECOR for editing, maintenance, and version control [44], and is also represented in HL7 FHIR format, which supports semantic and syntactic interoperability. The corresponding MDS FHIR profile can be accessed via simplifier.net [45]. The schema builds on generic metadata consistent with DataCite and extends this with health study specific metadata aspects for interoperability with clinical trials registries and epidemiological studies. Additionally, it is being aligned with related initiatives such as the metadata schemes of the European Clinical Research Infrastructure Network (ECRIN) [46] and the German Human Genome-Phenome Archive (GHGA) [47] to enable seamless interfacing between NFDI4Health services and these external resources. Moreover, consolidation with the EHDS metadata and alignment with the corresponding HealthDCAT-AP metadata standard [48], is ongoing in close collaboration with national and international health research data initiatives.

3.3.2. Semantic interoperability and harmonisation

Direct variable standardisation is complex and resource intensive, as demonstrated by the SNOMED CT mapping example for a COVID-19 questionnaire [49]. As a practical first step towards interoperability, variables should be made machine-readable, searchable and comparable, an approach implemented in the Health Study Hub. To enhance lightweight semantic interoperability, particularly for variable and questionnaire comparability, NFDI4Health adopts the Maelstrom Research approach and employs the Maelstrom taxonomy, which classifies variables into 134 categories. The

Annotation Workbench [32], developed by NFDI4Health, enables data providers to efficiently find, compare and annotate metadata at the variable level. For this purpose, the Annotation Workbench integrates the SemLookP terminology service, which makes it easy to search for, visualise and map terms from controlled vocabularies and biomedical ontologies (e.g. MeSH, SNOMED CT and ICD) to the corresponding variables. The SemLookP terminology service [50] also provides access to terminologies recommended by the NFDI4Health FAIRsharing Collection [40], making it an ideal tool for semantic annotation of data dictionaries.

When training data are available, AI models can further support the annotation process or suggest semantic concepts for questionnaire development. A successful example is the Maelstrom classifier [51], which was trained using existing Maelstrom classifications. As one element of the Annotation Workbench, this classifier suggests suitable Maelstrom categories for each variable from which the researcher can select the most appropriate one. In the long term, the recommendation of consolidated questionnaire items combined with AI support is expected to substantially increase semantic interoperability at the variable level.

Data harmonisation remains a prerequisite for joint analysis of multiple studies, ensuring comparability of measures collected in different study settings [52]. NFDI4Health established a semi-automated, retrospective stepwise harmonisation workflow based on the Rmonize package [53] from Maelstrom Research (see Figure 7). This workflow operates within a federated environment, where participating DHOs execute harmonisation locally, while NFDI4Health provides a standardised protocol, the respective templates and study-specific R scripts. The harmonisation process is quality-controlled by NFDI4Health. The resulting harmonised datasets can then be accessed via DataSHIELD.

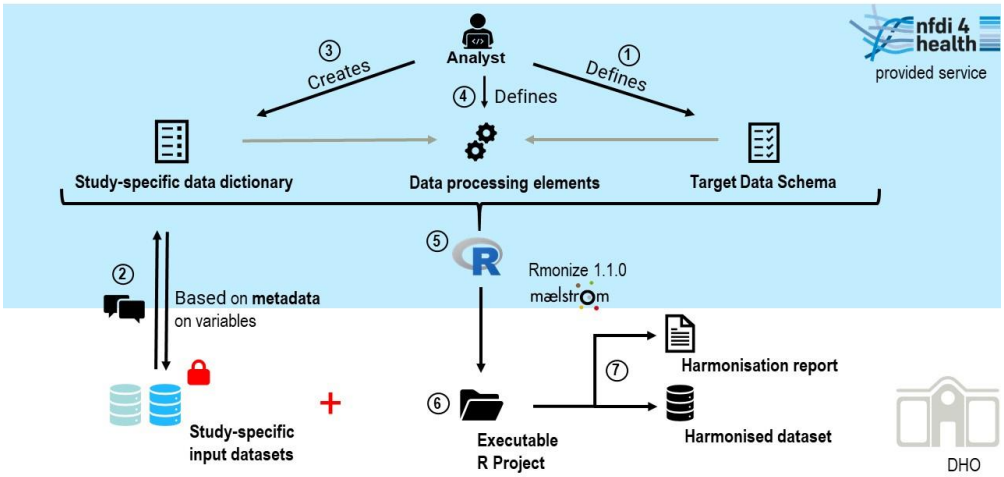

Figure 7: The data harmonisation workflow: (1) Analysts list the required research data based on their research question in a project-specific target data schema. Where available, existing standards are used to describe the data. (2) Analysts contact the data holders to check whether the required data have been collected and, if so, request detailed descriptions regarding data collection methods, formats, units, etc. (3) The analysts then create a study-specific data dictionary containing the requested variables, including descriptions, units, and categories. (4) Based on the defined variables in the target data schema and the study-specific variables, the analysts assess the harmonisation potential of each variable (complete, partial, impossible) and determine appropriate harmonisation rules. (5) The R script is slightly adapted to reflect the study's specific circumstances, keeping the workload for data holders as low as possible. (6) The templates created in steps 1, 3, and 4, along with the R script (with embedded Rmonize functions), are compiled into an R project and sent to the study. The study is asked to provide a dataset with the study-specific research data. (7) Upon successful execution of the R script, a harmonisation report and the harmonised dataset are produced. An additional file is generated for optional use with DataSHIELD.

Two pilot projects embedded in the use cases "Nutritional epidemiology" and "Epidemiology of chronic diseases" served as a testbed. Pilot project 1 aimed at systematically investigating methodological

limitations of exploratory dietary patterns, while pilot project 2 aimed at investigating the association of dietary sugar intake, sugar sweetened beverages and related foods with prospective changes in body fatness and chronic disease risk. Ten participating studies successfully harmonised the set of 214 variables, which was collected to address the two specific pilot projects. Studies uploaded their harmonised datasets to their Opal instances and thus made them accessible for DataSHIELD analyses. The harmonised datasets will be made findable via the Health Study Hub. Integration into the FDPG application process for broader use is planned.

3.4 Data reuse

As mentioned above, two aspects are crucial to ensure meaningful reuse of health research data: data quality and secure access.

3.4.1 Data quality

NFDI4Health has supported the further development of a metadata framework and a corresponding open-source tool, dataquieR [54], designed for scalable, metadata-driven quality assessment of health research data. Implemented in R, dataquieR facilitates the automated generation of standardised data quality reports with minimal programming effort. A key feature of dataquieR is its metadata-centric architecture. Data expectations are defined in structured Microsoft Excel tables. This approach lowers the barrier for non-technical users while enabling rigorous, consistent, and transparent assessments. The scope of data quality assessments depends on the provided metadata. Metadata specifications are organised into distinct sheets, each representing a different level of quality-related attributes:

- Item-level metadata describe individual variables, including permissible ranges and categories, missing value codes. This is the most used layer of information.
- Cross-item-level metadata capture expectations about relations between variables, enabling checks such as contradictions, multivariate outliers, or expected associations.
- Segment-level metadata refer to subsets of the dataset (e.g., specific examinations), including the number of observations, variable counts, and constraints such as unique identifiers or the absence of duplicates.
- Dataframe-level metadata define expectations for the entire dataset. Attributes are largely similar to the segment level.

Additional optional tables—such as missing value code or participant ID lists—may be included, further enhancing flexibility and scalability. These metadata are often developed and refined iteratively, with additional checks added based on findings during analyses. The dataquieR package is freely available via CRAN [55]. It also provides comprehensive documentation, examples, and specifications for the metadata framework [56].

3.4.2. Piloting secure processing via distributed data analysis frameworks

Within NFDI4Health, DataSHIELD and Personal Health Train (PHT) have been adopted to meet the diverse needs of different use cases. DataSHIELD is particularly well-suited for interactive, R-based statistical analyses across distributed datasets, making it ideal for NFDI4Health use cases such as “Nutritional epidemiology” and “Epidemiology of chronic diseases”. Currently, first analyses with harmonised data are conducted across ten studies based on the use cases. In contrast, PHT supports language-agnostic advanced analytics, making it more appropriate for the computationally intensive use case “Radiomics/AI”.

As part of this effort, the DataSHIELD infrastructure in Germany has been expanded. A total of eleven cohort studies is currently accessible, operated across nine active nodes (see Figure 6). To assist institutes with the initial setup, guidance via a standard operating procedure (SOP) [57] was developed and direct support through experts is additionally provided by NFDI4Health. Furthermore, a central analysis server has been established and hosted at Max Delbrück Center, serving as the entry point for analysts to perform distributed analyses. This setup is particularly important for security measures such as IP whitelisting, ensuring that only authorised users can access the system.

With the growing requirements of the use cases, we have also expanded DataSHIELD's functionality and related easy-to-use training materials. This includes the introduction of specific data analyses packages [58–60], alongside other useful tools aimed at improving user experience by reducing client-side script writing during initial stages of data analyses [61] and providing functions for DataSHIELD package development [62].

Based on the PHT concept, the PADME platform [63] has been set up, which facilitates distributed analytics and federated learning (including deep learning) by deploying algorithms to local 'stations'. The platform's design allows analytic workflows to be integrated in a language-agnostic way (via Docker images). This has been tested in multi-site demonstrator scenarios. PHT and PADME as its specific implementation are developed for flexible project support and, as a part of the NFDI4Health service ecosystem, operate on use cases. NFDI4Health primarily focused on the use case "Radiomics/AI". We have proposed solutions for mitigation of the bias caused by different CT reconstruction methods based on generative AI applied for image-to-image translation [64]. Also, the concept for cross-station data enhancement was enabled through the PHT infrastructure [65]. Working further towards trustworthy AI development beyond bias mitigation, we developed PHT-supported solutions for algorithm traceability [66].

4. Community engagement and support

NFDI4Health fosters active community engagement through diverse workshop formats, training programmes, NFDI4Health-launched project calls and targeted support measures. Workshops have been conducted on-site, online and in hybrid formats, often co-located with scientific conferences, to engage a wide range of stakeholders in FAIR data management and health research topics.

Capacity building for early career researchers (ECR) is supported through collaboration with the U Bremen Research Alliance and its cross-disciplinary RDM and data science programme *Data Train* [67]. The programme enjoys high demand, attracting between 2,400 and 2,500 annual registrations over the last three years, and resulting in an average of 700 certificates issued per year for fully completed courses (with 3,490 certificates awarded over the past five years). This initiative, co-developed with several NFDI consortia, consists of a structured curriculum covering the entire data-value chain and aims to strengthen fundamental skills in data literacy, RDM and data science, while simultaneously offering ECRs a platform to build an interdisciplinary and cross-institutional network. As of 2026, Data Train also offers ECRs the opportunity to complete a Certificate of Advanced Studies degree from the University of Bremen. As a generic training programme, Data Train has served as a blueprint for other comparable discipline-specific activities. These include the development of specific training offers for researchers and professionals in biomedicine and health research as part of NFDI4Health [68,69] and the RDMTraining4NFDI base service [70]. The latter aims to embed training activities within the NFDI as a whole and establish certification structures.

Interactive community engagement was further promoted through annual NFDI4Health-funded project calls, which encouraged small-scale collaborations that apply NFDI4Health tools and services. A pilot data steward programme demonstrated the effectiveness of local intermediaries as multipliers of FAIR data practices [3], complemented by a helpdesk that will be expanded in the consortium's second phase. User research activities, including a survey [71] and a scoping review [72], systematically captured community needs and guided the iterative development of the services, training and communication strategies. NFDI4Health also contributed to the cross-consortia NFDI Task Force on User Research to foster exchange on user-centred design and evaluation.

The consortium supports researchers with a suite of guidelines, templates and best practices that promote high-quality, legally compliant and FAIR data management. Examples include the Publication Policy [73], which guides researchers in the preparation and publication of research data on the Health Study Hub, and discipline-specific RDM plan templates for epidemiological and clinical studies [74] as well as a version aligned to the German Research Foundation checklist "handling of research data" [75]. Legal resource examples are the legal map outlining the German and European framework for research data processing [76], a template for informed consent for record linkage [77], and a compilation of use cases together with recommendations for improving record linkage practices [78,79].

5. Embedding in national and international infrastructures

Germany has a fragmented health research data infrastructure landscape that comprises among others: (1) The Health Data Lab (HDL) [80] at the Federal Institute for Drugs and Medical Devices (BfArM) providing pseudonymised administrative data from statutory health insurances; (2) The cancer registries maintained by German Federal States in Germany that transfer their data to the German Centre for Cancer Registry Data (ZfKD) [81]; the (3) MII and NUM infrastructures covering data from routine patient care documentation as well as data collected in prospective clinical and clinical epidemiological studies [82]; and NFDI4Health with a focus on epidemiological and public health, as well as clinical study data. NFDI4Health has been mandated to serve as connecting link between the NFDI as a whole and these health data infrastructures to ensure overall integration.

However, this process is further evolved alongside by European initiatives as the European Health Data Space (EHDS) that came into effect on March 26, 2025. The EHDS aims at establishing a legal and organisational framework to securely exchange and reuse health data across national borders and thus enabling improved health care and innovative health research. The amalgamation of the EHDS and German infrastructures is a major challenge for which Germany established a national Data Access and Coordination Office (DACO) at BfArM [83], among others. Furthermore, German infrastructures such as NFDI4Health also need to align their services with EHDS requirements. NFDI4Health has committed itself to support its user communities in this process and to take care of the needs of its DHOs, study participants and researchers.

The European Open Science Cloud (EOSC), which aims at creating an open, interdisciplinary, cross-border platform for scientific data access and collaboration, is another initiative on the European level concerns the NFDI [84]. NFDI is currently in the process of integration as national node of the EOSC.

Despite these challenges, the embedding of NFDI4Health in the NFDI and its alignment with European initiatives enables NFDI4Health to support interoperability across national and European infrastructures and to actively contribute to a coherent, interoperable health research data landscape.

Sustained investment in coordination, standards alignment and user support will be essential to translate infrastructural integration into routine, high-impact data reuse across disciplines and borders.

6. Discussion and conclusion

Over the past five years, NFDI4Health has demonstrated that a federated, FAIR research data infrastructure for sensitive health study data can be established and operated at the national level. The consortium has made tangible progress across all FAIR dimensions, with strengths in improving findability through the Health Study Hub, enabling first structured access to cohort data via the FDPG, and supporting secure reuse through distributed analysis frameworks. Importantly, these technical achievements are embedded in governance models that preserve institutional autonomy and legal responsibility, which is essential for sustainable data sharing.

Clearly, interoperability is hard to achieve and therefore requires a pragmatic approach. While full semantic standardisation remains a long-term objective, lightweight approaches, such as metadata harmonisation, taxonomy-based variable annotation and semi-automated harmonisation workflows, already deliver measurable benefits for data discovery and joint analyses. These approaches lower entry barriers for DHOs and allow incremental alignment with international standards and increasing syntactic and semantic interoperability over time.

Secure reuse remains a central challenge, particularly for complex analyses and high-dimensional data. To cope with this challenge, NFDI4Health offers a differentiated strategy consisting of distributed analysis frameworks as well as tools and targeted support for anonymisation and synthetic data generation combined with methods to assess data quality and re-identification risk, thereby balancing privacy with analytical utility.

Ultimately, NFDI4Health strives for enabling secure record linkage of deeply phenotyped cohort and clinical study data with large-scale care, register and administrative health insurance data. Such linkage will enhance epidemiological analyses and will support the interpretation of clinical trial findings by reflecting real-world treatment patterns and outcomes. Fully realising this potential, however, requires semantic interoperability with other relevant data types and secure processing environments to address legal and ethical constraints.

In summary, NFDI4Health´s services and software components constitute a solid community-driven ecosystem for the sustainable and interoperable reuse of health study data by combining and enhancing existing standards, tools and structured governance processes within and across national borders. Through its comprehensive engagement, NFDI4Health strengthens data competencies of the community also with respect to the emerging EHDS framework, promotes standardisation across consortia and ensures that infrastructure development remains aligned with user needs and real-world research practice. Close collaboration with data holders increases the awareness of the importance of data publication and FAIR data practices. However, achieving the comprehensive publication of questionnaires and data dictionaries in machine-readable formats remains challenging and requires further infrastructure development, continued community engagement, staff training, sustainable funding structures and a broader cultural shift in data sharing practices. Further planned developments, such as enabling variable selection within the Health Study Hub and transferring these selections directly to the FDPG, will further streamline the transition from data discovery to access. These continued improvements in data sharing, findability and accessibility are expected to significantly increase the reuse of health data in the coming years where NFDI4Health aims to monitor reuse through various indicators, including citations of our services and datasets.

References

1. European Commission. European Health Data Space Regulation (EHDS) [Internet]. 2025 [cited 2026 Jan 27]. https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space-regulation-ehds_en
2. Kraft S, Schmalen A, Seitz-Moskaliuk H, Sure-Vetter Y, Knebes J, Lübke E, et al. Nationale Forschungsdateninfrastruktur (NFDI) e. V.: Aufbau und Ziele. Bausteine Forschungsdatenmanagement. 2021;2:1–9. https://doi.org/10.17192/bfdm.2021.2.8332
3. Fluck J, Lindstädt B, Ahrens W, Beyan O, Buchner B, Darms J, et al. NFDI4Health – Nationale Forschungsdateninfrastruktur für personenbezogene Gesundheitsdaten. Bausteine Forschungsdatenmanagement. 2021;2:72–85. https://doi.org/10.17192/bfdm.2021.2.8331
4. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:1–9. https://doi.org/10.1038/sdata.2016.18
5. Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM). German Clinical Trials Register [Internet]. [cited 2026 Jan 27]. https://www.bfarm.de/EN/BfArM/Tasks/German-Clinical-Trials-Register/_node.html
6. National Library of Medicine (NLM). ClinicalTrials.gov [Internet]. [cited 2026 Jan 27]. https://clinicaltrials.gov/
7. World Health Organization (WHO). International Clinical Trials Registry Platform (ICTRP) [Internet]. [cited 2026 Jan 27]. https://www.who.int/tools/clinical-trials-registry-platform
8. European Medicines Agency (EMA). Clinical Trials Information System [Internet]. 2022 [cited 2026 Jan 27]. https://www.ema.europa.eu/en/human-regulatory-overview/research-development/clinical-trials-human-medicines/clinical-trials-information-system
9. The Research Institute of the McGill University Health Centre (RI MUHC). Maelstrom Research [Internet]. [cited 2026 Jan 27]. https://www.maelstrom-research.org/
10. Institute for Health Metrics and Evaluation (IHME). Global Burden of Disease (GBD) [Internet]. [cited 2026 Jun 5]. https://www.healthdata.org/research-analysis/gbd
11. NCD Risk Factor Collaboration (NCD-RisC). About Us [Internet]. [cited 2026 Jun 5]. https://www.ncdrisc.org/about-us.html
12. Green E, Ritchie F. The present and future of the Five Safes framework. J Priv Confidentiality [Internet]. 2023 [cited 2026 May 26];13. https://doi.org/10.29012/jpc.831
13. Wirth FN, Meurers T, Johns M, Prasser F. Privacy-preserving data sharing infrastructures for medical research: systematization and comparison. BMC Med Inform Decis Mak. 2021;21:242. https://doi.org/10.1186/s12911-021-01602-x
14. King G. An Introduction to the Dataverse Network as an Infrastructure for Data Sharing. Sociol Methods Res. 2007;36:173–199. https://doi.org/10.1177/0049124107306660
15. Meineke F, Golebiewski M, Hu X, Kirsten T, Löbe M, Klammt S, et al. NFDI4Health Local Data Hubs for Finding and Accessing Health Data: Making Distributed Data Accessible Through a SEEK-Based Platform. Proc Conf Res Data Infrastruct [Internet]. 2023 [cited 2023 Sep 9];1. https://doi.org/10.52825/cordi.v1i.375

16. Wolstencroft K, Krebs O, Snoep JL, Stanford NJ, Bacall F, Golebiewski M, et al. FAIRDOMHub: a repository and collaboration environment for sharing systems biology research. Nucleic Acids Res. 2017;45:D404–D407. https://doi.org/10.1093/nar/gkw1032

17. Prokosch H-U, Gebhardt M, Gruendner J, Kleinert P, Buckow K, Rosenau L, et al. Towards a National Portal for Medical Research Data (FDPG): Vision, Status, and Lessons Learned. In: Lovis C, et al., editors. Caring is Sharing – Exploiting the Value in Data for Health and Innovation [Internet]. IOS Press; 2023 [cited 2026 Jan 27]. 307–311. https://doi.org/10.3233/SHTI230124

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

19. NFDI4Health. Service [Internet]. 2018 [cited 2026 Jan 27]. https://www.nfdi4health.de/en/service.html

20. Prasser F, Eicher J, Spengler H, Bild R, Kuhn KA. Flexible data anonymization using ARX—Current status and challenges ahead. Softw Pract Exp. 2020;50:1277–1304. https://doi.org/10.1002/spe.2812

21. Gootjes-Dreesbach L, Sood M, Sahay A, Hofmann-Apitius M, Fröhlich H. Variational Autoencoder Modular Bayesian Networks for Simulation of Heterogeneous Clinical Study Data. Front Big Data [Internet]. 2020 [cited 2026 Jan 27];3. https://doi.org/10.3389/fdata.2020.00016

22. Kühnel L, Schneider J, Perrar I, Adams T, Moazemi S, Prasser F, et al. Synthetic data generation for a longitudinal cohort study – evaluation, method extension and reproduction of published data analysis results. Sci Rep. 2024;14:14412. https://doi.org/10.1038/s41598-024-62102-2

23. DataCite Metadata Working Group. DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs v4.7 [Internet]. DataCite; 2026 [cited 2026 May 26]. https://datacite-metadata-schema.readthedocs.io/en/4.7/

24. Schmidt CO, Struckmann S, Enzenbach C, Reineke A, Stausberg J, Damerow S, et al. Facilitating harmonized data quality assessments. A data quality framework for observational health research data collections with software implementations in R. BMC Med Res Methodol. 2021;21:63. https://doi.org/10.1186/s12874-021-01252-7

25. Beyan O, Choudhury A, van Soest J, Kohlbacher O, Zimmermann L, Stenzhorn H, et al. Distributed Analytics on Sensitive Medical Data: The Personal Health Train. Data Intell. 2020;2:96–107. https://doi.org/10.1162/dint_a_00032

26. Gaye A, Marcon Y, Isaeva J, LaFlamme P, Turner A, Jones EM, et al. DataSHIELD: taking the analysis to the data, not the data to the analysis. Int J Epidemiol. 2014;43:1929–1944. https://doi.org/10.1093/ije/dyu188

27. Pigeot I, Ahrens W, Darms J, Fluck J, Golebiewski M, Hahn HK, et al. Making Epidemiological and Clinical Studies FAIR Using the Example of COVID-19. Datenbank-Spektrum [Internet]. 2024 [cited 2024 Jun 3]. https://doi.org/10.1007/s13222-024-00477-2

28. Doiron D, Marcon Y, Fortier I, Burton P, Ferretti V. Software Application Profile: Opal and Mica: open-source software solutions for epidemiological data management, harmonization and dissemination. Int J Epidemiol. 2017;46:1372–1378. https://doi.org/10.1093/ije/dyx180

29. Darms J, Clemens V, Gonzalez-Ocanto M, Bruenings-Kuppe C, Cici S, Fluck J. The German Central Health Study Hub – A Service to Find and Publish Clinical, Public Health and Epidemiological Studies and Associated Documents. In: Ger Med Data Sci 2024. IOS Press; 2024. 129–137. https://doi.org/10.3233/SHTI240847

30. NFDI4Health. NFDI4Health Cohort Collection [Internet]. 2026 [cited 2026 Jun 5]. https://health-study-hub.de/collection/NFDI4Health_Cohort_Collection

31. NFDI4Health. Instruments for Health Surveys in Children and Adolescents [Internet]. 2019 [cited 2026 Jan 27]. https://health-study-hub.de/resource/nfdi4health-84

32. Sasse J, Fluck J. An Annotation Workbench for Semantic Annotation of Data Collection Instruments. Stud Health Technol Inform. 2023;302:108–112. https://doi.org/10.3233/SHTI230074

33. Meineke F, Hu X, Golebiewski M, Be M. Roll-Out of the NFDI4health Local Data Hubs for Medical Research Data Management. In: Intelligent Health Systems – Technology, Data, Knowledge [Internet]. IOS Press; 2025 [cited 2026 Jan 27]. 771–772. https://doi.org/10.3233/SHTI250461

34. Lettieri V, Kipker D-K. Synthetische Daten in der medizinischen Forschung. Datenschutz Datensicherheit - DuD. 2024;48:284–288. https://doi.org/10.1007/s11623-024-1931-0

35. Haber AC, Sax U, Prasser F, the NFDI4Health Consortium. Open tools for quantitative anonymization of tabular phenotype data: literature review. Brief Bioinform. 2022;23:bbac440. https://doi.org/10.1093/bib/bbac440

36. Moazemi S, Adams T, Ng HG, Schneider J, Her A-F, Fluck J, et al. NFDI4Health Workflow and Service for Synthetic Data Generation, Assessment and Risk Management. Stud Health Technol Inform. 2024 [cited 2026 Jan 27]. 21–29. https://doi.org/10.3233/SHTI240834

37. Adams T, Birkenbihl C, Otte K, Ng HG, Rieling JA, Näher A-F, et al. On the fidelity versus privacy and utility trade-off of synthetic patient data. iScience. 2025;28:112382. https://doi.org/10.1016/j.isci.2025.112382

38. Giomi M, Boenisch F, Wehmeyer C, Tasnádi B. A Unified Framework for Quantifying Privacy Risk in Synthetic Data. Proc Priv Enhancing Technol. 2023;2023:312–328. https://doi.org/10.56553/popets-2023-0055

39. Froehlich H. SYNDAT [Internet]. 2023 [cited 2023 Sep 22]. https://syndat.scai.fraunhofer.de/

40. Abaza H, Golebiewski M. FAIRsharing NFDI4Health collection [Internet]. 2023 [cited 2026 Aug 22]. https://fairsharing.org/4940

41. Vorisek CN, Klopfenstein SAI, Löbe M, Schmidt CO, Mayer PJ, Golebiewski M, et al. Towards an Interoperability Landscape for a National Research Data Infrastructure for Personal Health Data. Sci Data. 2024;11:772. https://doi.org/10.1038/s41597-024-03615-3

42. Abaza H, Shutsko A, Klopfenstein SAI, Vorisek CN, Schmidt CO, Brünings-Kuppe C, et al. Toward a Domain-Overarching Metadata Schema for Making Health Research Studies FAIR (Findable, Accessible, Interoperable, and Reusable): Development of the NFDI4Health Metadata Schema. JMIR Med Inform. 2025;13:e63906. https://doi.org/10.2196/63906

43. Schwedhelm C, Nimptsch K, Ahrens W, Hasselhorn HM, Jöckel K-H, Katzke V, et al. Chronic disease outcome metadata from German observational studies – public availability and FAIR principles. Sci Data. 2023;10:868. https://doi.org/10.1038/s41597-023-02726-7

44. Vorisek C, Klopfenstein S. ART-DECOR Metadata schema NFDI4Health V3 (3.0) [Internet]. [cited 2026 Jun 5]. https://art-decor.org/ad/#/nfdhtfcov19-/project/overview

45. Klopfenstein S, Hölter T, Saß J, Vorisek C. Simplifier Project NFDI4Health Metadata Schema [Internet]. NFDI4Health Meta Data Schema. [cited 2026 Jun 5]. https://simplifier.net/nfdi4health-metadata-schema

46. Canham S. ECRIN Metadata Schemas for Clinical Research, Version 8 (September 2023) [Internet]. Zenodo; 2023 [cited 2026 Jan 27]. https://doi.org/10.5281/zenodo.8368709

47. Deutsches Krebsforschungszentrum (DKFZ). The GHGA Metadata Model [Internet]. [cited 2026 Jan 27]. https://www.ghga.de/datenportal/metadata-model

48. European Health Information Portal. HealthDCAT AP [Internet]. [cited 2026 May 26]. https://www.healthinformationportal.eu/healthdcat-ap

49. Vorisek CN, Klopfenstein SAI, Sass J, Lehne M, Schmidt CO, Thun S. Evaluating Suitability of SNOMED CT in Structured Searches for COVID-19 Studies. Stud Health Technol Inform. 2021 [cited 2026 Jan 27]. 88–92. https://doi.org/10.3233/SHTI210126

50. Baum R, Schneider J. Semantic Lookup Platform – ZB MED [Internet]. [cited 2026 May 22]. https://semanticlookup.zbmed.de/

51. Sasse J, Fabre G, Fortier I, Zimmermann P, Fluck J. Applying AI to Support Categorization of Heterogeneous Epidemiological Datasets. Stud Health Technol Inform. 2025 [cited 2026 Jan 27]. 848–852. https://doi.org/10.3233/SHTI250479

52. Fortier I, Raina P, Van Den Heuvel ER, Griffith LE, Craig C, Saliba M, et al. Maelstrom Research guidelines for rigorous retrospective data harmonization. Int J Epidemiol. 2016;45. Advance Access publication. https://doi.org/10.1093/ije/dyw075

53. Fabre G, Maelstrom Research. Rmonize: Tools for Data Harmonization. R package version 2.0.0 [Internet]. 2025 [cited 2026 Jan 27]. https://github.com/maelstrom-research/Rmonize

54. Struckmann S, Mariño J, Kasbohm E, Salogni E, Schmidt CO. dataquieR 2: An updated R package for FAIR data quality assessments in observational studies and electronic health record data. J Open Source Softw. 2024;9:6581. https://doi.org/10.21105/joss.06581

55. Kasbohm E, Salogni E, Marino J, Richter A, Schmidt CO, Struckmann S, et al. dataquieR: Data Quality in Epidemiological Research [Internet]. 2026 [cited 2026 Jan 27]. https://cran.r-project.org/web/packages/dataquieR/index.html

56. Universitätsmedizin Greifswald. Annotation of Metadata [Internet]. [cited 2026 Jan 27]. https://dataquality.qihs.uni-greifswald.de/VIN_Annotation_of_Metadata.html

57. NFDI4Health. Standard Operating Procedure for Installation and Configuration of Opal DataSHIELD in NFDI4Health [Internet]. GitHub. [cited 2026 Jan 27]. https://github.com/nfdi4health/opal-datashield-sop/wiki/Home

58. Schwarz F. dsClusterAnalysis [Internet]. 2021 [cited 2026 Sep 22]. https://github.com/FlorianSchw/dsClusterAnalysis

59. Swenne A. dsGamlss: DataSHIELD server-side functions to fit generalized additive models for location, scale and shape. R package version 0.1.0 [Internet]. BIPS; 2025 [cited 2026 Jan 27]. https://github.com/bips-hb/dsGamlss

60. Schwarz F. FlorianSchw/dsIntestinalMicrobiomics [Internet]. 2022 [cited 2026 Jan 27]. https://github.com/FlorianSchw/dsIntestinalMicrobiomics

61. Siampani S. sofiasiamp/dsSupportClient [Internet]. 2024 [cited 2026 Jan 27]. https://github.com/sofiasiamp/dsSupportClient

62. Schwarz F. FlorianSchw/DSFunctionCreator [Internet]. 2024 [cited 2026 Jan 27]. https://github.com/FlorianSchw/DSFunctionCreator

63. Welten S, Mou Y, Neumann L, Jaberansary M, Yediel Ucer Y, Kirsten T, et al. A Privacy-Preserving Distributed Analytics Platform for Health Care Data. Methods Inf Med. 2022;61:e1–e11. https://doi.org/10.1055/s-0041-1740564

64. Li F, Wang Y, Beyan O, Schöneck M, Caldeira LL. Voxel-Wise Medical Image Generalization for Eliminating Distribution Shift. ACM Trans Knowl Discov Data. 2024;18:1–16. https://doi.org/10.1145/3643034

65. Li F, Ekaterina K, Liliana LC, Oya B. How Data Infrastructure Deals with Bias Problems in Medical Imaging. Stud Health Technol Inform. 2024 Aug 22;316:726-730. https://doi.org/10.3233/SHTI240517

66. Elwes M, Jaberansary M, Tang F-SK-B, Aswendt M, Beyan O, Kutafina E. TrainTracks - federated learning for reproducible research on sensitive medical data. BMC Med Inform Decis Mak. 2026 May 21;26(1):185. https://doi.org/10.1186/s12911-026-03553-7

67. Hörner T, Glöckner FO, Drechsler R, Pigeot I. Disziplinübergreifendes Modell zur Ausbildung von Forschungsdatenmanagement und Data Science Kompetenzen: „Data Train – Training in Research Data Management and Data Science“. Bausteine Forschungsdatenmanagement. 2021;3:56–69. https://doi.org/10.17192/bfdm.2021.3.8343

68. Dierkes J, Fürst J, Hörner T, Klammt S, Lindstädt B, Pigeot I, et al. Training concepts in research data management and data science with the focus on health research [Internet]. Zenodo; 2023 [cited 2026 Jan 27]. https://doi.org/10.4126/FRL01-006441348

69. Herres-Pawlis S, Ortmeyer J, Pelz P, Hüppe H, Stellmacher M, Lindstädt B, et al. Sektionskonzept einer Sektion im Verein Nationale Forschungsdateninfrastruktur (NFDI) e.V. - Sektion Training & Education V. [Internet]. Zenodo; 2026 [cited 2026 Jan 27]. https://doi.org/10.5281/zenodo.19666250

70. Müller R, Lindstädt B, Blümm M, Vandendorpe J, Förstner KU. Base4NFDI proposal RDMTraining4NFDI - Competence Training for Research Data Management [Internet]. Zenodo; 2024 [cited 2026 Jan 27]. https://doi.org/10.5281/zenodo.13310636

71. Depping R, Restel K. Status-quo-Erhebung 2021 - Bericht zum 1. NFDI4Health User Survey [Internet]. PUBLISSO; 2023 [cited 2026 Jan 27]. https://doi.org/10.4126/FRL01-006461780

72. Soltani N, Dietz T, Ochterbeck D, Dierkes J, Restel K, Santis KKD, et al. Digital Information Exchange Between the Public and Researchers in Health Studies: Scoping Review. J Med Internet Res. 2025; Jan 28;27:e63373. https://doi.org/10.2196/63373

73. Fürst J, Lindstädt B, Shutsko A. Publication Policy of the National Research Data Infrastructure for Personal Health Data (NFDI4Health) [Internet]. 2025 [cited 2025 Mar 13]. https://doi.org/10.4126/FRL01-006484167 https://doi.org/10.4126/FRL01-006484167

74. Fürst J, Lindstaedt B, Shutsko A, Grummich K, Gonzalez Ocanto M, Perrar I. Research Data Management Plan — Clinical studies [Internet]. 2024 [cited 2024 Aug 27]. https://doi.org/10.4126/FRL01-006431467

75. Fürst J, Lindstaedt B, Shutsko A, Grummich K, Gonzalez Ocanto M, Perrar I. Research Data Management Plan short — DFG Proposal [Internet]. 2024 [cited 2024 Aug 27]. https://doi.org/10.4126/FRL01-006483993

76. Buchner B, Lettieri V, Kipker D-K. 'Legal map' of the applicable legal framework conditions for research data processing [Internet]. Zenodo; 2024 [cited 2025 Sep 8]. https://doi.org/10.4126/FRL01-006449528

77. Intemann T, Lettieri V, Kipker D-K, Kuntz A, Ahrens W, Pigeot I, et al. Informed Consent zum Record Linkage: Best Practice und Mustertexte [Internet]. 2023 [cited 2026 Jan 27]. https://doi.org/10.4126/FRL01-006399943

78. Intemann T, Kaulke K, Kipker D-K, Lettieri V, Stallmann C, Schmidt CO, et al. Verbesserung des Record Linkage für die Gesundheitsforschung in Deutschland – wie können Defizite behoben werden? Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz. 2025;68:1440–1449. https://doi.org/10.1007/s00103-025-04153-y

79. Intemann T, Kaulke K. White Paper: Improvement of Record Linkage for Health Research in Germany [Internet]. 2023 [cited 2026 Jun 5]. Available from: https://doi.org/10.4126/FRL01-006461895

80. Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM). Health Data Lab [Internet]. [cited 2026 Jan 27]. https://www.forschungsdatenzentrum-gesundheit.de/en/

81. Robert Koch Institute. German Centre for Cancer Registry Data [Internet]. [cited 2026 Jan 27]. https://www.krebsdaten.de/Krebs/EN/Home/homepage_node.html

82. Charité-Universitätsmedizin Berlin. Network of University Medicine [Internet]. [cited 2026 Jan 27]. https://www.netzwerk-universitaetsmedizin.de/

83. Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM). Data Access and Coordination Office (DACO) [Internet]. [cited 2026 Jan 27]. https://www.bfarm.de/EN/BfArM/Tasks/Data-Access-and-Coordination-Office/_node.html

84. NFDI e.V. EOSC Node Germany [Internet]. [cited 2026 Jan 27]. https://www.nfdi.de/eosc-node-germany/?lang=en

Statements & Declarations

Funding

This work support by the Deutsche Forschungsgesellschaft (DFG, German Research Foundation) (Grant number 442326535).

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Author Contributions

All authors contributed to the study conception and design. The first draft of the manuscript was written by Juliane Fluck and Iris Pigeot and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.