DFG-Ausschreibung

Nationale Forschungsdateninfrastruktur – Ausschreibung 2019 für die Förderung von Konsortien

National Research Data Infrastructure for Personal Health Data (NFDI4Health)

Proposal

Applicant institution: ZB MED – Information Centre for Life Sciences

Spokesperson: Prof. Dr. Juliane Fluck

fluck@zbmed.de

0049 (0)228 73 60351

DOI https://dx.doi.org/10.4126/FRL01-006421856

Table of Content

1 General Information

- Name of the consortium in English and German

  National Research Data Infrastructure for Personal Health Data (NFDI4Health)

  Nationale Forschungsdateninfrastruktur für personenbezogene Gesundheitsdaten (NFDI4Health)

- Summary of the proposal in English and German

  Germany has accumulated a wealth of health-related personal data from well-designed cohort studies and health surveillance systems (healthy individuals) as well as clinical trials (patients) that are characterised by a deep phenotyping of study subjects with questionnaires, medical examinations and molecular/genetic profiling. Their longitudinal nature and high quality make these data a valuable research resource for the development of preventive and therapeutic measures on the individual and the population level.

  NFDI4Health represents an interdisciplinary research community by integrating major German institutions experienced as data holders, data analysts and methodology developers. It builds on established structures, competences and know-how and expects a rapidly growing support and participation of the research community.

  NFDI4Health aims to create the most comprehensive inventory of German epidemiological, public health and clinical trial data to date. NFDI4Health will build a centralised data catalogue with elaborate search functionalities, sophisticated data access management, and a data analysis toolbox, while respecting stringent requirements for privacy concerning personal health data. Standardisation services will ensure a high degree of interoperability. Use cases covering prototypical study types and areas of research will show the feasibility of a harmonised implementation of all infrastructures, tools and services in accordance with our user communities.

  The specific aims of NFDI4Health are (1) to enable findability of and access to structured health data from registries, administrative health databases, clinical trials, epidemiological studies and public health surveillance; (2) to implement a health data framework for centralised searching and access of existing decentralised epidemiological/clinical data infrastructures; (3) to facilitate data sharing, record linkage, harmonised data quality assessments, federated analyses of personal health data, while complying with privacy

regulations and ethical requirements; (4) to enable the development and deployment of new machine-processable consent mechanisms and innovative data access services by operationalising the FAIR (Findable, Accessible, Interoperable, Reusable) principles for scientific data management and stewardship; (5) to support cooperation between clinical trial research, epidemiological and public health communities; (6) to foster interoperability of currently fragmented IT solutions related to metadata repositories, cohort browsing, data quality and harmonisation; (7) to develop business models to secure the sustainability of structures and services.

NFDI4Health will increase the visibility and accessibility of research data, enhancing the reputation of scientists sharing their data and fostering new collaborations. The resulting infrastructure will build bridges between user communities and data holders from epidemiology, public health and clinical trials.

==========

Deutschland verfügt über eine Fülle gesundheitsbezogener Daten aus gut konzipierten Kohortenstudien und Surveillanceprogrammen (gesunde Individuen) sowie aus klinischen Studien (Patienten), die eine tiefe Phänotypisierung der Studienteilnehmenden anhand von Fragebögen, medizinischen Untersuchungen und molekularen/genetischen Profilen aufweisen. Durch ihren Längsschnittcharakter und ihre hohe Qualität sind diese Daten eine wertvolle Ressource für die Entwicklung von präventiven und therapeutischen Maßnahmen auf Individual- und Populationsebene.

Die Integration wichtiger deutscher Institute mit Erfahrung als Datenhalter, -analysten und Methodenentwickler macht NFDI4Health zu einem interdisziplinären Forschungs-konsortium, das auf etablierten Strukturen, Kompetenzen und Know-How sowie einer zunehmenden Unterstützung und Teilnahme der Forschungsgemeinschaft aufbaut.

NFDI4Health hat zum Ziel, ein umfassendes Inventar deutscher epidemiologischer, Public Health und klinischer Studiendaten aufzubauen. NFDI4Health wird unter Beachtung der besonderen Datenschutzanforderungen personenbezogener Gesundheitsdaten einen zentralisierten Datenkatalog mit ausgefeilten Suchfunktionen, anspruchsvollem Daten-zugangsmanagement und einer Analyse-Toolbox erstellen. Standardisierungsservices werden einen hohen Grad an Interoperabilität sichern. Fallbeispiele für prototypische Studientypen und Forschungsgebiete werden zeigen, wie Infrastrukturen, Werkzeuge und Services im Einklang mit den Nutzern umgesetzt werden können.

Damit wird NFDI4Health (1) die Auffindbarkeit und den Zugang zu strukturierten Gesundheitsdaten aus Registern, administrativen Gesundheitsdatenbanken, klinischen und epidemiologischen Studien sowie Public Health-Surveillance ermöglichen; (2) ein Konzept für Gesundheitsdaten implementieren, das die zentrale Suche von und den Zugriff auf dezentral verwaltete Dateninfrastrukturen ermöglicht; (3) den Austausch und die Verknüpfung von personalisierten Gesundheitsdaten, einheitliche Bewertungen der Datenqualität und verteilte Datenanalysen gemäß datenschutzrechtlichen und ethischen Bestimmungen erleichtern; (4) die Entwicklung und den Einsatz maschinenprozessierbarer Einwilligungserklärungen sowie innovativer Datenzugriffsservices durch Umsetzung der FAIR (Findable, Accessible, Interoperable, Reusable) Prinzipien für wissenschaftliches Datenmanagement ermöglichen; (5) die Zusammenarbeit zwischen klinischer und epidemiologischer Forschung stärken; (6) die Interoperabilität von fragmentierten IT-Lösungen für Metadatenrepositorien, das Suchen von Kohorten, die Datenqualität und -harmonisierung fördern und (7) Geschäftsmodelle entwickeln, die die Nachhaltigkeit der entwickelten Strukturen und Services absichern.

NFDI4Health wird die Sichtbarkeit von Forschungsdaten erhöhen, die Reputation von Forschenden stärken, die ihre Daten teilen, und neue Kooperationen fördern. Die daraus entstehende Infrastruktur wird eine Brücke zwischen Datennutzern und -haltern schlagen.

- Applicant institution

Applicant institution	Short name	Location
ZB MED – Information Centre for Life Sciences	ZB MED	Gleueler Straße 60, 50931 Cologne

- Name of the consortium spokesperson

Spokesperson	Institution, location
Prof. Dr. Juliane Fluck	ZB MED, Katzenburgweg 1a, 53115 Bonn

- Co-applicant institutions

Co-applicant institutions	Short name	Location
Leibniz Institute for Prevention Research and Epidemiology – BIPS	BIPS	Achterstraße 30, 28359 Bremen

Co-applicant institutions	Short name	Location
Charité – Universitätsmedizin Berlin, Berlin Institute of Health	Charité/BIH	Anna-Louisa-Karsch-Straße 2, 10178 Berlin
German Institute for Human Nutrition Potsdam-Rehbrücke	DIfE	Arthur-Scheunert-Allee 114-116, 14558 Nuthetal
Fraunhofer-Gesellschaft (Fraunhofer Institute for Applied Information Technology FIT; Fraunhofer Institute for Digital Medicine MEVIS; Fraunhofer Institute for Algorithms & Scientific Computing SCAI)	Fraunhofer	Hansastraße 27 c, 80686 Munich
Heidelberg Institute for Theoretical Studies (HITS gGmbH)	HITS	Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg
Network of Coordinating Centres for Clinical Trials – KKSN e.V.	KKSN	Alt-Moabit 96A, 10559 Berlin
Max Delbrück Center for Molecular Medicine in the Helmholtz Association	MDC	Robert-Rössle-Straße 10, 13125 Berlin
Max Rubner-Institut	MRI	Haid-und-Neu-Straße 9, 76131 Karlsruhe
Robert Koch Institute	RKI	Nordufer 20, 13353 Berlin
Technology, Methods, and Infrastructure for Networked Medical Research	TMF	Charlottenstraße 42/Dorotheenstraße, 10117 Berlin
University of Bonn (Institute for Agricultural Science)	U Bonn	Regina-Pacis-Weg 3, 53113 Bonn
University of Bremen (Institute for Information, Health and Medical Law (IGMR))	U Bremen	Bibliothekstraße 1, 28359 Bremen
University of Cologne (University and City Library of Cologne (USB))	U Cologne	Albertus-Magnus-Platz, 50923 Cologne
University of Leipzig (Institute for Medical Informatics, Statistics and Epidemiology (IMISE); LIFE-Research Centre for Civilisation Diseases; Centre for Clinical Trials (ZKS))	U Leipzig	Ritterstraße 26, 04109 Leipzig
University of Applied Sciences Mittweida	UAS Mittweida	Technikumplatz 17, 09648 Mittweida
University of Göttingen (University Medical Center)	UM Göttingen	Robert-Koch-Straße 40, 37075 Göttingen

Co-applicant institutions	Short name	Location
University Medicine Greifswald (Institute for Community Medicine)	UM Greifswald	Fleischmann Str. 8, 17475 Greifswald

- Names of co-spokespersons (and the task areas to which they contribute)

Co-spokesperson	Institution, location	Task area(s)
Prof. Dr. Wolfgang Ahrens	BIPS, Bremen	TA2, 3, 5, 6
Prof. Dr. Iris Pigeot	BIPS, Bremen	TA1, 2, 3, 4, 5, 6
Prof. Dr. Hajo Zeeb	BIPS, Bremen	TA2, 3, 4, 5
Prof. Dr. Fabian Prasser	Charité/BIH, Berlin	TA6
Prof. Dr. Sylvia Thun	Charité/BIH, Berlin	TA1, 2
Prof. Dr. Matthias Schulze	DIfE, Nuthetal	TA1, 2, 3, 5, 6
Dr. Oya Beyan	Fraunhofer, Fraunhofer FIT, Aachen	TA1, 3, 6
Prof. Dr. Holger Fröhlich	Fraunhofer, Fraunhofer SCAI, St. Augustin	TA1, 6
Prof. Dr.-Ing. Horst Hahn	Fraunhofer, Fraunhofer MEVIS, Bremen	TA1, 2, 5
Martin Golebiewski	HITS, Heidelberg	TA1, 2, 3, 4
PD Dr. Wolfgang Müller	HITS, Heidelberg	TA1, 2, 3, 4
PD Dr. Sebastian Klammt	KKSN, Berlin	TA1, 2, 4, 5
Prof. Dr. Tobias Pischon	MDC, Berlin	TA1, 2, 3, 5
Annette Polly	MRI, Karlsruhe	TA1, 2, 5
Dr. Thilo Muth	RKI, Berlin	TA2, 3, 4, 5, 6
Prof. Dr. Lothar H. Wieler	RKI, Berlin	TA1, 2, 4, 5
Dr. Annette Pollex-Krüger	TMF, Berlin	TA1, 2, 4, 6
Sebastian C. Semler	TMF, Berlin	TA1, 2, 4, 6
Prof. Dr. Ute Nöthlings	U Bonn, Bonn	TA1, 4, 5
Prof. Dr. Benedikt Buchner	U Bremen, Bremen	TA1, 6
Ralf Depping	U Cologne, USB, Cologne	TA2, 4
Dr. Jens Dierkes	U Cologne, USB, Cologne	TA4
Dr. Hubertus Neuhausen	U Cologne, USB, Cologne	TA1, 4
Dr. Oana Brosteanu	U Leipzig, ZKS, Leipzig	TA5
Matthias Löbe	U Leipzig, IMISE, Leipzig	TA2, 4, 5
Prof. Dr. Markus Löffler	U Leipzig, IMISE, LIFE, ZKS, Leipzig	TA1, 2, 3, 4, 5

Co-spokesperson	Institution, location	Task area(s)
Dr. Frank Meineke	U Leipzig, IMISE, Leipzig	TA3
Prof. Dr.-Ing. Toralf Kirsten	UAS Mittweida, Mittweida	TA1, 3
Dr. Harald Kusch	UM Göttingen, Göttingen	TA3, 4, 6
Prof. Dr. Ulrich Sax	UM Göttingen, Göttingen	TA1, 2, 3, 4, 6
Prof. Dr. Carsten Oliver Schmidt	UM Greifswald, Greifswald	TA1, 2, 3, 4, 5, 6
Prof. Dr.-Ing. Dagmar Waltemath	UM Greifswald, Greifswald	TA2, 3, 4
Prof. Dr. Konrad Förstner	ZB MED, Cologne	TA3
Birte Lindstädt	ZB MED, Cologne	TA2, 4

- Participants

Participants	Institution (where applicable), location
Institutions/organisations (represented by)	
AGENS – Working group on the collection and use of secondary data (Dr. Enno Swart)	Magdeburg
Brain Simulation Section at the Department Neurology, Charité Universitätsmedizin Berlin (Prof. Dr. Petra Ritter)	Berlin
Center for Biotechnology (CeBiTec) of Bielefeld University (Prof. Dr. Alfred Pühler)	Bielefeld
Center for Clinical Studies Regensburg (Prof. Dr. Michael Koller)	Regensburg
Center for Clinical Trials Essen (Prof. Dr. Karl-Heinz Jöckel)	Essen
Center for Clinical Trials Münster, University of Münster (Prof. Dr. Dr. Andreas Faldum)	Münster
Clinical Trials Centre Cologne, University Cologne (Alexandra Nieß)	Cologne
Clinical Trials Unit, University Medical Center Freiburg (Dr. Britta Lang)	Freiburg
Coordination Center for Clinical Studies in Magdeburg, Otto-von Guericke University Magdeburg (Dr. Antje Wiede)	Magdeburg
Coordination Center for Clinical Trials Düsseldorf, Düsseldorf University Clinic (Henrike Kolbe)	Düsseldorf
CoRe-Net (BMBF Infrastructure Network, University of Cologne) (Dr. Nadine Scholten)	Cologne

Participants	Institution (where applicable), location
DIN German Institute for Standardization (Rüdiger Marquardt)	Berlin
Federal Centre of Health Education (BZgA) (Dr. Heidrun M. Thaiss)	Cologne
German Association for Medical Informatics, Biometry and Epidemiology (GMDS) (Prof. Dr. Andreas Stang)	Cologne
German Association of Medical Faculties (MFT) (Dr. Frank Wissing)	Berlin
German Cancer Research Center (DKFZ), Division of Cancer Epidemiology (Prof. Dr. Rudolf Kaaks)	Heidelberg
German Consortium of Hereditary Breast and Ovarian Cancer (Prof. Dr. Rita Schmutzler)	Cologne
German Data Forum (RatSWD) (Prof. Regina T. Riphahn, PhD)	Berlin
German Institute for Medical Documentation and Information (DIMDI) (Dr. Dietrich Kaiser)	Cologne
German Network for Evidence-based Medicine (DNEbM) (Prof. Dr. Andreas Sönnichsen)	Berlin
German Nutrition Society (DGE) (Dr. Angela Bechthold)	Bonn
German Public Health Association (DGPH) (Prof. Dr. Ansgar Gerhardus)	Bochum
German Region of the International Biometric Society (IBS-DR) (Prof. Dr. Werner Brannath)	Hanover
German Society for Drug Utilization Research and Drug Epidemiology (GAA) (PD Dr. Katrin Farker)	Jena
German Society for Epidemiology (DGEpi) (Prof. Dr. Dietrich Rothenbacher)	Ulm
German Society for Occupational and Environmental Medicine (DGAUM) (Dr. Thomas Nesseler)	Munich
German Society of Medical Sociology (DGMS) (Prof. Dr. Olaf von dem Knesebeck)	Düsseldorf
German Society of Social Medicine and Prevention (DGSMP) (Prof. Dr. Susanne Moebus)	Hamburg
Gutenberg Health Study, University Medical Center of the Johannes Gutenberg-University Mainz (Prof. Dr. Philipp Wild)	Mainz
Hamburg Cancer Registry (Dr. Stefan Hentschel, Dr. Alice Nennecke)	Hamburg

Participants	Institution (where applicable), location
Hamburg City Health Study, University Medical Center Hamburg-Eppendorf (UKE) (Prof. Dr. Stefan Blankenberg, Dr. Annika Jagodzinski)	Hamburg
Heinz Nixdorf Recall and Multigeneration Study, Institute for Medical Informatics, Biometry and Epidemiology (IMIBE), University Hospital Essen (Prof. Dr. Karl-Heinz Jöckel)	Essen
Institute for Prevention and Occupational Medicine of the German Social Accident Insurance Institute of the Ruhr University Bochum (IPA) (Prof. Dr. Thomas Behrens)	Bochum
Institute of Biostatistics and Clinical Research, University of Münster (Prof. Dr. Dr. Andreas Faldum)	Münster
Institute of Epidemiology, Helmholtz Zentrum München (HMGU) (Prof. Dr. Annette Peters, Dr. Marie Standl)	Munich
Institute of Medical Epidemiology, Biometrics and Informatics (IMEBI), Medical Faculty, Martin Luther University Halle-Wittenberg (Prof. Dr. Rafael Mikolajczyk)	Halle (Saale)
Institute of Social Medicine and Health Economics, Faculty of Medicine, Otto-von-Guericke University Magdeburg (Dr. Enno Swart)	Magdeburg
Medical School OWL, Bielefeld University (Prof. Dr. Claudia Hornberg)	Bielefeld
mediStatistica Dr. Burkhard Haastert (Dr. Burkhard Haastert)	Neuenrade
NAKO Health Study, German National Cohort (Prof. Dr. Annette Peters)	Heidelberg
OFFIS Institute for Information Technology (Prof. Dr. Susanne Boll-Westermann)	Oldenburg
PMV Forschungsgruppe, University of Cologne (Ingo Meyer)	Cologne
School of Medicine and Health Sciences, Carl von Ossietzky-University Oldenburg (Prof. Dr. Hans Gerd Nothwang)	Oldenburg
Society of Epidemiological Cancer Registries in Germany (GEKID) (Prof. Dr. Alexander Katalinic)	Saarbrücken
University Computing Centre, University of Greifswald (Prof. Dr. Johanna Eleonore Weber)	Greifswald

Participants	Institution (where applicable), location
Individual participant	
Prof. Dr. Ludwig Kuntz	Seminar for Business Administration and Healthcare Management, University of Cologne, Cologne

- Names and numbers of the DFG review boards (DFG Fachkollegien) that reflect the subject orientation of the proposed consortium

  205-01 Epidemiology, Medical Biometry, Medical Informatics (I. Pigeot, member of this review board, is NFDI4Health deputy spokesperson)

  205-02 Public Health, Health Services Research, Social Medicine

  205-05 Nutritional Sciences

  205 (For clinical trials, we cannot select a certain review board; clinical trials refer to all medical subjects listed under review board 205)

2 Consortium

2.1 Research domains or research methods addressed by the consortium, objectives

NFDI4Health is targeting individual health data that are generated in clinical trials, epidemiological cohorts and public health surveillance. Compared to data harvested from routine clinical in- and outpatient care data sources, these research data already meet high quality standards, are generated according to well-defined prospective study protocols, obtain specified detailed phenotypes and are stored and documented using structured data formats. While NFDI4Health will not embrace the above routine clinical care data (covered by NFDI4MED), it will provide processes to include so-called secondary data, i.e. data from disease registries and from large administrative health databases (e.g., health insurances).

Access to such detailed information on a large and unselected number of patients and healthy subjects is pivotal to investigate the causal pathways of disease, to advance patient stratification, to support personalised medicine, to find new therapy options and to improve patient care. In the field of prevention and public health, it is in particular of utmost importance to examine

populations in observational studies and epidemiological community trials. In the field of clinical trials about 400 academic investigator-initiated trials (funded by BMBF, DFG, German Cancer Aid, etc.) are completed each year. Many of them are run by trail centres organised within the Network of Coordinating Centres for Clinical Trials (KKSN). However, neither structural meta-data describing the design, the sampling scheme and the variables of clinical trials or epidemiological/public health studies nor the content are routinely shared.

The overarching goal of NFDI4Health is to support its clinical and epidemiological research community in the best possible way to share their data with the user community in agreement with data protection/privacy regulations and ethics principles and, in the interest of improving population health, create new data analytics opportunities within the German NFDI. To reach these goals in accordance with the FAIR data principles, the key objectives of NFDI4Health are

(1) to enable findability of and access to structured health data from clinical trials, epidemiological studies, disease registries, administrative health databases and public health surveillance in Germany;

(2) to implement a health data framework for centralised searching and accessing existing decentralised epidemiological/clinical trial data infrastructures;

(3) to facilitate data sharing, record linkage, harmonised data quality assessments, federated analyses of personal health data;

(4) to enable the development and deployment of new, machine-processable consent mechanisms and innovative data access services;

(5) to support cooperation between clinical trial research, epidemiological and public health communities;

(6) to foster interoperability of currently fragmented IT solutions related to metadata repositories, cohort browsing, data quality and harmonisation;

(7) to develop business models to secure sustainability of structures and services.

2.2 Composition of the consortium and its embedding in the community of interest

Research subject, means of communication and prioritisation

NFDI4Health will focus its research method-oriented approach on personal health data where a concise evaluation of the existing research and infrastructure environment revealed a crucial gap between the huge amount of data generated and the possibilities to share these data with

other researchers. Personal health data are collected and analysed mainly by physicians, epidemiologists, public health researchers and biometricians. Researchers in epidemiology, medical biometry, medical informatics, public health, health services research, social medicine and in nutritional epidemiology as well as clinical trial researchers form a major part of our user community. Many of them have a dual role as data holders and data users. Usually data collections are stored in decentralised, autonomous data infrastructures that will remain decentralised but have to be integrated into a common framework for centralised search and access.

To identify, spell out and address the specific needs of the health research communities, NFDI4Health involves major actors in this field as (co-)applicants and participants. Since health-related primary data are mainly generated in epidemiological studies and in clinical trials, NFDI4Health focusses on these two domains. (1) To cover the epidemiological and public health expertise, research institutes such as BIPS, RKI, MRI, UM Greifswald, U Leipzig and DIfE are co-applicants that constitute the backbone of personal health data infrastructures of NFDI4Health (see Figure 1 and Table A1, Appendix 1). These institutes are well connected with the national and international research community. (2) Clinical trials are initiated by medical researchers of all medical disciplines (179 different medical societies in Germany), represented in the Association of the Scientific Medical Societies in Germany (AWMF). Publicly funded clinical trials are in general managed and supported by 24 university clinical trial coordinating centres (KKS), most of them are members of KKSN. In 2018, they performed 786 academic investigator-initiated clinical (64% randomised controlled) trials. To reach out to this broad community of data holders, NFDI4Health involves these KKS as participants and disseminators. In addition, KKSN is co-applicant in NFDI4Health to enable roll-out of this functionality and also to provide courses in data sharing in almost all medical faculties. In addition, co-applicant TMF, will stay in close exchange with MFT (German Association of Medical Faculties), where the latter contributes as a participant to NFDI4Health.

A central step taken to identify and address the needs of user communities is their direct involvement as participants in building health data infrastructures (Table A1, Appendix 1). They will be supported by NFDI4Health in the whole process by improving data access and FAIRification of health data. NFDI4Health has approached and informed the user communities through the corresponding societies (DGAUM, DGE, DGEpi, DGMS, DGPH, DGSMP, DNEbM, GAA, GEKID, GMDS, IBS-DR) who act as channels and supporters. Users have provided input via questionnaires and participating in our first community workshop in Cologne (June 2019). This helped to prioritise the needs with regard to general, more technical requirements that are the focus of Task Areas (TAs) 2 and 3 and with regard to specific requirements of our user

communities that are addressed in TA4, TA5 and TA6. Their interest in active involvement is taken up by TA4, use cases reflecting the broad range of our user communities are covered by TA5 and concerns related to hurdles and limitations of sharing individual-level data are tackled by TA6. In addition to these activities, the user communities were informed by presentations at various conferences (e.g., general assembly of the NAKO 2019 (Berlin), GMDS 2019 (Dortmund) and DGEpi 2019 (Ulm)). Current information about NFDI4Health is provided on the NFDI4Health website (www.nfdi4health.de).

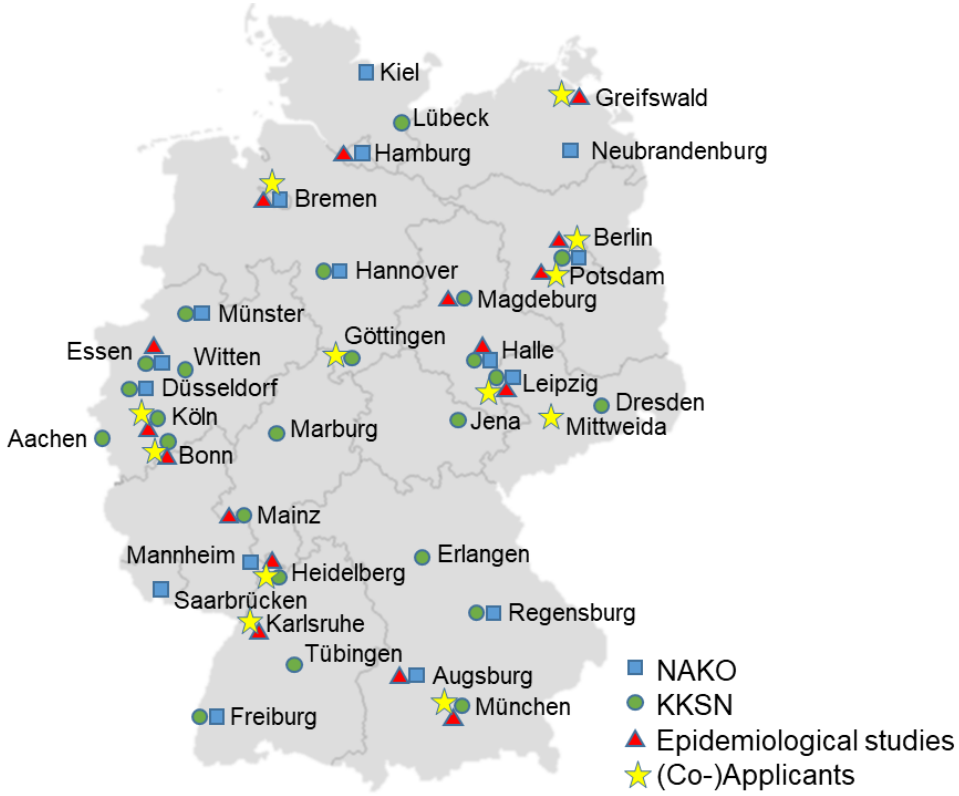

Figure 1: Overview of the entire NFDI4Health consortium (incl. all participating studies)

To ensure and formalise the long-term active involvement of users, NFDI4Health will establish a User Advisory Board as part of the overall governance structure and will launch calls among the involved user communities for additional use cases and small-scale projects supporting the NFDI4Health aims. Hence, driven by the user community, new data collections will be successively made interoperable and included in the NFDI4Health infrastructure. Furthermore, the University and City Library of Cologne (USB) will lead the development of training material, in particular e-learning tools, to be used, e.g., by the technical staff (data managers, programmers, IT-experts), researchers and students at medical faculties. Feedback loops will initiate an iterative process with users to optimise the provided educational materials. Regular workshops, datathons as well as early prototype and demonstrator releases are provided with

direct feedback possibilities to test usability of developed user interfaces and services. A further reach out to the clinical trial community will be organised by the KKSN which will for instance train trial physicians and medical data managers and constantly gather feedback from the clinical trial centres to develop education material fitting their specific needs. These activities together with all further measures for community involvement and cooperation with other NFDIs as well as information of citizens and patients as important data holders will be coordinated in TA4 which is led by the University and City Library of Cologne together with BIPS.

Strengths of the consortium

The NFDI4Health consortium covers the whole range of expertise needed to implement its work programme. This expertise will be strengthened by the broad spectrum of know-how contributed by the participants.

The mission of the applicant institute **ZB MED Information Centre for Life Sciences** is to serve as information hub and to provide services for research data management such as DOI services, PUBLISSO Life Science Repository[1] for publications, data management planning tool RDMO4Life, terminology and semi-automated annotation services. The **spokesperson J. Fluck**, *Head of the Knowledge Management Group*, has strong links and experience in semantic data integration and heterogeneous data analysis in the medical domain and is associated member of the Medicine Informatics Initiative (MII) SMITH. **B. Lindstädt** heads the *Research Management Group* at ZB MED and, in addition, has vast experience in user consulting in this area. **K. Förstner** is *Head of the Information Service Unit* which develops information services including the discovery service LIVIVO[2] and has a strong track record in the large scale analysis of life science data as well as in the development of software and services. ZB MED is responsible for the project coordination (TA1), for providing central publication and search services as well as central data access and supporting workflows for data annotation (TA3) and supporting community activities (TA4).

The **Leibniz Institute for Epidemiology and Prevention Research – BIPS** is a competence centre for epidemiology, statistics, data management, standardisation and quality control with a strong interest in counselling stakeholders, political actors and the community. Because of its longstanding expertise in planning, coordinating and conducting multi-centre, large-scale epidemiological studies BIPS serves as an infrastructure for the development and implementation of use cases (TA5). **I. Pigeot**, *Scientific Director of BIPS* and *Head of the Department of Biometry and Data Management*, will serve as deputy spokesperson (TA1, T4.4) of NFDI4Health. **W. Ahrens**, *Deputy Scientific Director of BIPS* and *Head of the Department of Epidemiological*

Methods and Etiological Research, has a longstanding expertise in the coordination of large population-based studies w.r.t. the aetiology of non-communicable diseases (e.g., pan-European IDEFICS/I.Family children cohort, NAKO Health Study). He will contribute his specific expertise in privacy requirements of population-based studies (T5.3, T6.5). **H. Zeeb**, *Head of the Department of Prevention and Evaluation*, has a strong expertise in implementation research and citizen involvement in public health research. His longstanding track record of involvement in several relevant professional societies will also serve TA4.

The **Berlin Institute of Health (Charité/BIH)** focuses on translational research to move innovative concepts from the lab to clinical practice. Within its research platform “Digital Medicine”, Charité/BIH aims at fostering the use of digital health data across different sources to improve research and patient care. **S. Thun**, *Head of the Core Facility eHealth and Inter-operability* of Charité/BIH, is a leading expert in standardisation and interoperability of health data. She has coordinated various national and international projects (e.g., epSOS cross-border healthcare in the EU, Horizon 2020 ASSESS CT, BMBF AKTIN) and has extensive connections to national and international communities and standards developing organisations. **F. Prasser**, *Head of the Medical Informatics Group* at Charité/BIH, has profound experience in medical data integration and health information privacy. Until his move to Berlin, he was technical coordinator of the DIFUTURE consortium in the MII. At the same time, he worked on setting up translational big data analytics infrastructures for two DFG-funded clinical research centres (SFB 1321 and SFB 1371). Charité/BIH will contribute to the standardisation of health data access and interoperability (TA2) and to privacy risk analyses (TA6).

The German Institute of Human Nutrition (DIfE) is devoted to nutrition research. The institute has an outstanding reputation in nutritional epidemiology as an important partner in multi-centre studies, e.g., the European Prospective Investigation into Cancer and Nutrition (EPIC) study, the NAKO Health Study, or the RODAM consortium. DIfE, due to its longstanding expertise in planning, coordinating and conducting large-scale epidemiological studies, will serve as infrastructure for the development and implementation of use cases (TA5). **M. Schulze**, *Head of the Department of Molecular Epidemiology*, will serve as co-spokesperson (TA5, T3.7). He will among others contribute his specific expertise in data harmonisation related to nutritional data and chronic disease epidemiology and federated data analyses infrastructures with DataSHIELD, the latter being based on his involvement in the EU funded InterConnect project.

Fraunhofer (**Fraunhofer FIT, Fraunhofer MEVIS, Fraunhofer SCAI**) with outstanding track records in machine learning, distributed data analysis and data mining of medical data. **O. Beyan** *(FIT)* will set up solutions based on Personal Health Train. **H. Hahn**, *Director of the*

Fraunhofer Institute for Digital Medicine (MEVIS), has profound experience in the quantitative analysis of multimodal medical imaging data and, within NFDI4Health, is in charge of implementing the Radiomics and imaging AI research paradigms. **H. Fröhlich,** *Head of Data Science & AI (SCAI*) has a long-standing experience in data science and AI with respect to applications in healthcare. He has held leadership positions in that area in academia (BMBF projects IDENTIREST and EPP) as well as pharmaceutical industry. In NFDI4Health, he will focus on simulation of synthetic health data using generative machine learning models (T6.4) that he has previously developed in the EU-IMI project AETIONOMY.

Heidelberg Institute for Theoretical Studies (HITS) is a private, non-profit research institute that conducts basic research in the natural sciences, mathematics and computer science, with a focus on the processing, structuring and analysing of large amounts of complex data and the development of computational methods and software. Co-spokespersons **M. Golebiewski** and **W. Müller** are well connected with the FAIR data community (infrastructures such as de.NBI[3], ELIXIR[4] and FAIRDOM[5]), the biological/medical terminology and standards communities, as well as with national (DIN) and international (ISO) standards organisations. **W. Müller** is scientific director and head of the SDBV group at HITS with a focus on FAIR data management solutions. SDBV is founding member of the FAIRDOM initiative and co-develops the SEEK data management platform[6] (TA3) that is widely used in many consortia in Germany and worldwide. The group is involved in the German Network for Bioinformatics Infrastructure (de.NBI) and the German node of the European ELIXIR infrastructure. **M. Golebiewski** chairs the 'data processing and integration' working group of ISO/TC 276 Biotechnology[7] and the project group *FAIR data infrastructures for biomedical informatics* of the GMDS. He is actively involved in the European network EU-STANDS4PM[8], in the German Liver Systems Medicine Network (LiSyM[9]) and is member of the coordination board of the standardisation initiative COMBINE[10].

KKS-Netzwerk (KKSN), the German network of coordinating centres for clinical studies with its co-spokesperson **S. Klammt** (managing director) provides the infrastructure for cooperation of 24 national academic centres for clinical trials. In 2018, KKSN supported more than 850 clinical trials with scientific services, expertise and specific know-how and will provide this expertise to the use case T5.4. High quality standards of each network member are ensured by external audits as a prerequisite for membership. KKSN acts as a partner of the European Clinical Infrastructures Network (ECRIN) and supports strategic moves for clinical research by cooperating with international committees and expert panels. A further essential task of the KKSN is the active involvement in harmonisation and improvement of regulatory policies for clinical trials, both at a national and European level. Training and education are essential main tasks of KKSN

as well. More than 9,500 clinical physicians, researchers and medical staff have been trained, e.g., in the basic requirements for (non-)interventional studies in 2018. KKSN will expand their training programme by contributing to the development of new teaching material (T4.3).

The **Max Delbrück Center for Molecular Medicine (MDC)** has a focus on cancer, diseases of the nervous system and cardiovascular/metabolic diseases. MDC aims at understanding the molecular basis of health and disease and translating findings as quickly as possible into clinical applications. The Molecular Epidemiology Research Group is committed to study the relationship of lifestyle, genetic, metabolic and environmental factors with risk and outcome of chronic diseases. Moreover, the research group coordinates the Cluster Berlin-Brandenburg of the NAKO Health study and the JPI-HDHL Interrelation of the INtesTInal MICrobiome, Diet and Health (INTIMIC) project, which builds on the JPI-HDHL European Nutritional Phenotype Assessment and Data Sharing Initiative (ENPADASI) infrastructure to create a (meta-)database of observational studies on diet, microbiome and health. **T. Pischon**, *Head of the Department of Molecular Epidemiology*, will lead T5.2 and contribute to T2.2, T3.7 and T5.1. His expertise in data harmonisation and use of DataSHIELD will help conduct federated data analyses of personal health data in compliance with privacy regulations and ethics principles.

The **Max Rubner-Institut (MRI)** focuses its research on consumer health protection in the field of nutrition and advises the Federal Ministry of Food and Agriculture. The Department of Nutritional Behaviour explores everyday human actions related to nutrition, making consumers and their needs the focus of its investigations. The Department of Physiology and Biochemistry of Nutrition investigates the health effects of food and nutrition based on physiological, biochemical and molecular biology methods. The objective is to detect health benefits of foods for the consumer, to identify possible risks and to make science-based recommendations for policy advice and consumer health protection. **A. Polly** heads the *Research Services and Information Management Group* and focuses on data management and publication policies (T2.1). Further contributions will be made to the use case “Nutritional Epidemiology” (T5.1).

The **Robert Koch Institute (RKI)** with its co-spokespersons **L. Wieler**, *President of RKI* and *Head of the Department of Methodology and Research Infrastructure,* and **T. Muth**, responsible for research data management in the *Department of Methodology and Research Infrastructure,* is the Federal Public Health Institute of Germany and a federal agency subordinate of the German Ministry of Health. RKI conducts research in epidemiology as well as surveillance projects for population-wide health monitoring, disease control and prevention and supports the federal states in outbreak investigations. As the first non-university research institution in Germany, RKI adopted a specific data policy that commits to the Berlin declaration for open access to

scientific knowledge. The Research Data Management service unit at RKI manages research data with the aim of making data sets (re)usable. The accredited Research Data Centre[11] (FDZ) publishes its survey data as public use files. Among others, RKI will coordinate the outreach of NFDI4Health to political decision makers (T1.4) as well as the sustainable interaction with health user communities (T4.1) and is responsible for the use case “Surveillance” (T5.5).

TMF (Technology, Methods and Infrastructure for Networked Medical Research) with its co-spokespersons **S.C. Semler**, *Executive Director of TMF,* and **A. Pollex-Krüger**, *Scientific Consultant*, is a knowledge infrastructure and the umbrella organisation for networked medical research in Germany. TMF is building on 20 years of experience in organising consensus processes in medical communities and producing generic solutions including expert opinions, generic concepts, IT applications, practical guides, training and consultation services. Currently 63 medical research projects and networks are members of TMF, building a strong medical community with, e.g., DFG-funded projects, BMBF-funded Medical Competence Centres, all German Centres for Health Research, MII consortia, Fraunhofer, Leibniz and Helmholtz institutes, the NAKO Health Study, clinical study centres and cancer registries, the German Biobank Node and the Luxembourg Centre for Systems Biomedicine. TMF operates the portal ToolPool Health Research with a continuous user feedback on metadata, research data management, data protection or secondary use of clinical data. **S.C. Semler** will provide a contact point and collaboration with the Medical Informatics Initiative (MII), which he coordinates, and to NFDI4MED, of which TMF is co-applicant, in T2.4, T6.2 and T6.5. **A. Pollex-Krüger** co-leads the interaction with health user communities (T4.1) and contributes to T4.5.

The Faculty of Agriculture at the **University of Bonn (U Bonn)** has a focus on nutrition research. Since 2012 U Bonn holds the DONALD Study on newborns of 3 months of age followed-up until adulthood. **U. Nöthlings** is *Head of the Nutritional Epidemiology Group* at the Institute of Nutrition and Food Sciences and acts as scientific director of the DONALD Study. One research focus is on the development of dietary assessment instruments for large-scale nutritional epidemiological studies. She is the coordinator of the German Competence Cluster in Nutrition Research Diet-Body-Brain (DietBB). She was actively involved in a number of JPI-HDHL-funded projects aiming at harmonising data and conducting federated data analysis (DEDIPAC, ENPADASI, HEALTHMARK). **U. Nöthlings** will contribute to NFDI4Health with respect to data harmonisation in TA2 and implementation of the use case T5.1.

The *Institute for Information, Health and Medical Law (IGMR)* with its director **B. Buchner, University of Bremen (U Bremen)**, will provide the legal expertise for the project with regard to all data protection issues (TA6). The research conducted at the IGMR focuses on the

intersections of health and data protection law. As head of the trust centre of the pharmacoepidemiological research database (GePaRD), **B. Buchner** has long-standing practical experience with the challenges posed by processing health data for research purposes. As chair of the U Bremen Ethics Committee, he also contributes expertise with the specific ethical challenges involved in processing data for research purposes.

The **University of Cologne/University and City Library of Cologne (U Cologne)** is a partner of the Cologne Competence Center for Research Data Management C3RDM[12] which bundles local RDM services and acts as an expert-network hub and a central contact point on the campus concerning research data management. **H. Neuhausen** has long experience working with different scientific advisory boards. **J. Dierkes**, who is also speaker of C3RDM, has strong experience in the training of research data management and data science. **R. Depping** is an expert in empirical social research. U Cologne will lead TA4.

The **University of Leipzig (U Leipzig)** represented by the Institute for Medical Informatics, Statistics and Epidemiology (IMISE, **M. Löffler** (*Director*)**, M. Löbe, F. Meineke**) and the Centre for Clinical Trials Leipzig (ZKS, **O. Brosteanu**) are responsible for the design, coordination, data management and biostatistical analyses of about 100 currently running prospective clinical trials. IMISE/ZKS are in particular active in the fields of cancer, sepsis, heart disease and obesity. In August 2019, the ZKS was successfully certified by ECRIN (European Clinical Research Infrastructure Network). **M. Löffler** is the principal coordinator of SMITH, one of the four consortia of the MII[13]. IMISE further leads projects for decentralised computing and phenotyping and is involved in metadata management and interoperability. The data sharing project Leipzig Health Atlas (LHA[14,15]) is led by IMISE within the framework of the BMBF program i:DSem (Integrative Data Semantics for Systems Medicine) and provides a prototype project for NFDI4Health (TA3, TA5). LHA is based on the SEEK data management platform[6], develops and operates a web platform for the presentation and exchange of a wide range of publications, (bio-)medical data, interactive models and tools. Furthermore, U Leipzig runs the LIFE Research Centre that investigates the causes and early detection of civilisation diseases in the framework of epidemiological cohorts (LIFE-Adult, LIFE-Heart) and also runs a study centre of the NAKO Health Study.

The **University of Applied Sciences Mittweida** (UAS Mittweida) is deeply involved in the design and implementation of the Personal Health Train (PHT), a distributed analysis infrastructure currently developed by MII. **T. Kirsten** has been responsible for research data management at the LIFE Research Centre (U Leipzig) and developed software systems for data use/access and for running feasibility queries, as well as for data curation infrastructure,

metadata integration and management. He is co-PI of the Leipzig Health Atlas and the LIFE-LIFT project and is involved in the MII SMITH consortium. UAS Mittweida will contribute to TA3.

The Medical Data Hub of the **University Medical Center Göttingen** (UM Göttingen) is running data services for the Medical Informatics Initiative of all Medical Schools (BMBF), German Medical Research Centers (BMBF), SFBs (DFG) and many clinical studies. It provides long-term experience in collaborative research data management spanning from early competence networks in medical research to recent FAIR data management projects with a multitude of different data types. Its curricular activities have just been reaccredited by the International Medical Informatics Association (IMIA). **H. Kusch** is deeply involved in the Göttingen eResearch Alliance. Furthermore, C. Bauer and T. Bender provide the Integrated Data Repository Toolkit (IDRT) and combine the well-established platforms SEEK[6] and tranSMART including well-established ETL tools and methods in data management pipelines for discovering, extracting, integrating and visualising complex data structures (trial, register and biomaterial data), as well as a filter and analysis tool for assessing data completeness and data quality. Furthermore, UM Göttingen provides extensive expertise in privacy concepts and deployment of privacy infrastructure. **U. Sax** will lead TA6 and will contribute to TA2 and 3.

The **University Medicine of Greifswald** (UM Greifswald) contributes long-standing experience on the conduct of population-based epidemiological cohort studies such as the Study of Health in Pomerania (SHIP) or NAKO Health Study and provides tools and standards related to data management, data quality assessments, standardisation, data integration. **D. Waltemath** is co-founder and co-chair of the world-wide umbrella organisation for coordinated development of standards for modelling in computer-aided biology and medicine (COMBINE) and member of the MIRACUM consortium (MII). Her work focuses on implementing FAIR data principles and metrics, as well as on outreach strategies and novel educational formats for medical data science. **C.O. Schmidt** heads the *Unit Quality in the Health Sciences* with experiences in tool and infrastructure building for data quality assessments (DFG funded), is partner in the H2020-funded euCanSHare consortium, the STRATOS network and heads the TMF working group *Data quality and transparency*. UM Greifswald will co-lead TA2 and contribute to TA3-6.

2.3 The consortium within the NFDI

In autumn 2017, ZB MED initiated the creation of NFDI4Life as an overarching consortium representing all disciplines in the life sciences. This broad approach turned out to be incompatible with the formal national agreement on the establishment of NFDI in Germany.

Therefore, NFDI4Life was split up into several individual consortia, including NFDI4Agri, NFDI4BioDiversity, NFDI4Health, NFDI4MED, NFDI4Microbiota, and a newly structured NFDI4Life Umbrella at the end of 2018/beginning of 2019. During the subsequent elaboration process, the two epidemiology-orientated consortia NFDI4Health and NFDI4NutEpi merged to form NFDI4Health, while NFDI4Medicine and DZG4NFDI merged to form NFDI4MED.

NFDI4Health seeks cooperation with all other life science NFDI consortia and with consortia dealing with person-related data beyond the health domain. Particularly intense consultations have taken place with the other two consortia from the medical domain, namely NFDI4MED and GHGA. Since NFDI4Health, **NFDI4MED** and **GHGA** target different types and provenances of health-related data bodies requiring different types of expertise, infrastructure and methodology, they have agreed on a close partnership in addressing processes of patient identification and record linkage, in providing metadata standards and common interoperable data models and in establishing common services building on the framework provided by the MII. This is facilitated by the fact that several NFDI4Health co-applicants and participants are also active in the MII and **NFDI4MED**. The three consortia provide complementary infrastructure components: bridging storage and management of medical research data from epidemiological and clinical studies (NFDI4Health), patient-oriented clinical routine care, clinical and basic research data (**NFDI4MED**), and omics raw data (**GHGA**) as illustrated in Figure 2**.**

Mutual consultations have led to coordinated working programmes of the three consortia and an explicit commitment to an intensive collaboration. While the different data bodies result in a clear division of labour, central tasks will be worked on jointly by all three consortia. Ethical, legal and societal issues are, e.g., synergistic cross-sectional topics to which these consortia will contribute. In detail, the following joint working activities are planned: **Collaboration data search and request brokering**: NFDI4Health and **NFDI4MED** will build central data search and access services for their communities. They will coordinate issues of data discovery and governance of data use and access processes and by this will ensure interoperability. This collaborative activity is also open to **GHGA**. **Collaboration core data set:** Like **NFDI4MED**, NFDI4Health will use the MII core data set as a starting point for, e.g., metadata standardisation which will be further specified in a joint effort. **Collaboration data publishing/archiving:** There is general agreement that the guidelines, services, etc. developed by one consortium may be used by the other consortia. For example, NFDI4Health plans to develop publication guidelines that can be used by **NFDI4MED**. **Collaboration on distributed analyses of data:** In particular, both, NFDI4Health and **NFDI4MED** will develop complementary add-ons for the DataSHIELD

framework which will be accessible by the user communities. In addition, NFDI4Health will extend the Personal Health Train (PHT) framework that will be useful for **NFDI4MED**.

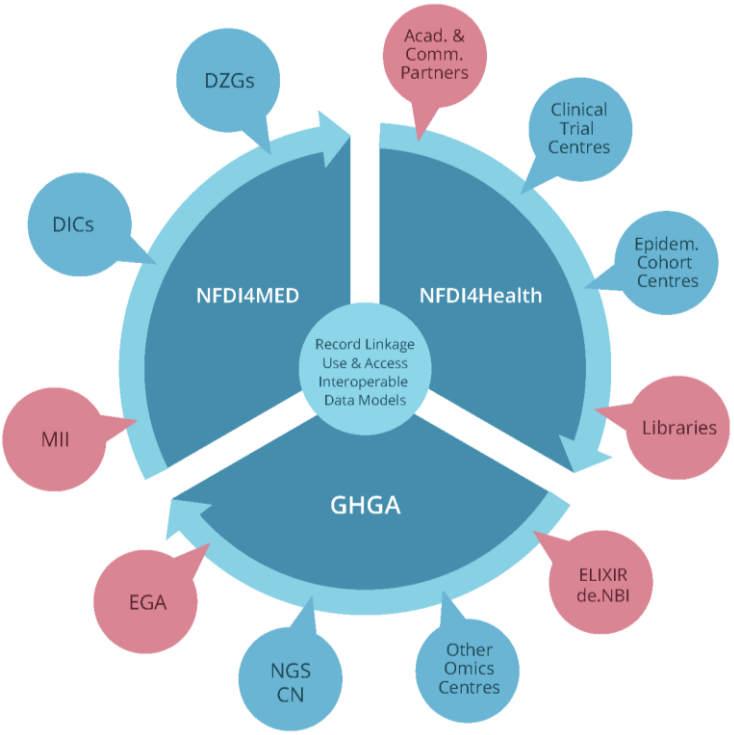

Figure 2: Interactions and responsibilities of three NFDI consortia in the medical field (blue circles: data sources, red circles: method development & transfer; DZGs: German Centres for Health Research, DICs: data integration centres, EGA: European Genome-phenome Archive, NGS CN: next-generation sequencing competence network)

NFDI4Health will also coordinate its efforts together with **KonsortSWD** in addressing challenges that are related to making sensitive cohort and survey data reusable and interoperable. BIPS has recently joined the EcoSoc Implementation Network in order to cooperate directly under the auspices of the GO FAIR initiative.

NFDI-Neuro and NFDI4Health will share standardisation policies and processes to develop common standards for (meta-)data, record linkage and interfaces. For standardisation of chemical components such as medication, dietary factors or metabolome data, NFDI4Health will explore the possibilities of standardisation and cross-sectional mapping together with **NFDI4Chem** and **NFDI4Microbiota**. NFDI4Health will further explore opportunities for data linkage with environmental data. In this respect, NFDI4Health is looking forward to a close collaboration with, e.g., **NFDI4Agri**, **NFDI4Earth**, **NFDI4BioDiversity** and **NFDI4NanoSafety**. All cross-cutting topics (see below) and standardisation with further life science consortia will be organised in concert with **NFDI4Life Umbrella**.

The above collaborations are fostered by joint memberships of (co-)applicants of NFDI4Health in NFDI4Agri (ZB MED), NFDI4BioDiversity (HITS), NFDI4Crime (UAS Mittweida), NFDI4Life Umbrella (almost all NFDI4Health (co-)applicants, planned for 2020), NFDI4MED (Charité/BIH, TMF), NFDI4Microbiota (ZB MED, planned for 2020), NFDI-Neuro (Fraunhofer MEVIS, planned for 2020). Further steps have been taken to strengthen collaboration among NFDI consortia. During a meeting in Berlin, August 2019, eleven consortia agreed on the so-called **Berlin Declaration** (https://doi.org/10.5281/zenodo.3457213) that describes a common vision in particular on cross-cutting topics and that is supported by further NFDI consortia. NFDI4Health will contribute to all cross-cutting topics that are relevant for person-related health data accounting for the specific requirements of our user communities and data holders as, e.g., overarching data standards and interoperability. Of specific importance are the following cross-cutting topics of the **Berlin Declaration**:

Technical infrastructure and concepts

Standardisation: (Co-)applicants of NFDI4Health have leading roles in relevant domain-specific (inter-)national standardisation bodies (e.g., ISO[16], IEC[17], CEN/CENELEC[18], DIN[19], HL7[20]), as well as meta initiatives such as the European network EU-STANDS4PM[8] and the European COST action CHARME[21], which will be beneficial for many NFDI consortia, especially the ones related to health and life sciences.

Monitoring of data quality and federated data analysis: Quality of data to be shared has to be closely monitored according to predefined criteria. Here, NFDI4Health will build on the DFG-funded project Standards and tools for data monitoring in complex epidemiological studies where a broad epidemiological community was involved under the lead of UM Greifswald. In addition, NFDI4Health will provide experience in the development and implementation of federated data analysis infrastructures to all other NFDI consortia.

Data management tools and data sharing models: NFDI4Health will develop tools, e.g., to create data management plans with a specific catalogue of relevant questions for data types represented by NFDI4Health, and set up processes and models for data sharing of sensitive person-related data. Appropriate models and metadata sets will be developed, e.g., by extending the generic metadata standard of DataCite. Furthermore, requirements for archiving metadata and data sets will be formulated. These tools and models can be adopted by other domains handling sensitive data, e.g., by KonsortSWD.

Legal and ethical aspects

Data privacy, data protection laws and record linkage: Given the specific requirements with respect to data privacy, NFDI4Health will provide key knowledge and expertise regarding

individual data protection and its implementation for the exchange of study data. NFDI4Health will benefit from a strong expertise provided by a legal expert with focus on data protection and long-standing experience in ethical requirements of research studies involving humans (U Bremen). NFDI4Health will develop solutions to enable sharing of personalised data in Germany and will give recommendations for necessary revisions of German law, in particular with respect to record linkage.

Community (user) involvement

Training and education: NFDI4Health will provide specific training in good practice of health data collection, data management, data access constraints and correct use/analysis of data. This does not only concern primary data collection but also the secondary use of existing databases and registries. Here, NFDI4Health will build on the strong expertise in “Good Practice of Secondary Data Analysis” and “Good Practice Data Linkage” of several co-applicants and participants, in particular U Magdeburg and PMV Forschungsgruppe. Moreover, together with NFDI4Earth, NFDI4BioDiversity, the city and state of Bremen and U Bremen, NFDI4Health has started to establish a graduate education programme on research data management and data science. Here, also KonsortSWD and NFDI-Neuro have expressed their interest to actively contribute to the development of the curriculum and its teaching modules.

From NFDI4Health perspective, the last two topics above need most urgent support from the NFDI’s collaborative framework. Changing the legislation with respect to data sharing and record linkage will only be possible as a consolidated action of the whole NFDI. Also changing the curricula of education programmes will be facilitated by a concerted action involving all federal states. Since suitable governance structures are key to ensure sustainable operations within NFDI, NFDI4Health in addition urgently asks for an appropriate legal entity to serve the interests of the consortia.

2.4 International networking

On an international level, NFDI4Health (co-)applicants will exploit their active involvement in European consortia or initiatives to ensure compatibility of NFDI4Health activities with comparable activities on a European and international level. NFDI4Health has obtained letters of support from many of these international initiatives and projects that demonstrate their interest in a mutual exchange to align standards, services and tools on the international level. An overview of the current involvement of the (co-)applicants in international initiatives and the

corresponding subject-specific interactions is given in Table A2 (Appendix 1). Selected collaborations of (co-)applicants are summarised below.

HITS and Charité/BIH are well established and actively involved in medical, health-specific and life science specific national (DIN), European (CEN/CENELEC) and international (ISO, IEC) standardisation committees, especially in ISO/TC 215 Health Informatics[22] and ISO/TC 276 Biotechnology[7] and their mirror committees in Germany. Additionally, NFDI4Health co-applicants are active in several scientific standardisation initiatives and standard defining organisations, such as HL7 International[20], DIN NA 063-07[23], as well as in European meta standardisation initiatives such as EU-STANDS4PM[8] and the COST action CHARME[21]. Fraunhofer is actively involved in GO FAIR implementation networks and RDA working groups. FIT co-leads the GO FAIR Personal Health Train implementation network and participates in the FAIR StRePo (making standards, repositories and policies FAIR) network.

BIPS, UM Greifswald and UM Göttingen participate in several (inter-)national projects focusing on epidemiological study data harmonisation and standardisation. MDC, U Bonn, DIfE, MRI and BIPS are actively involved in several European projects (e.g., ENPADASI, DEDIPAC, INTIMIC, PEN) funded by the Joint Programming Initiative (JPI) A Healthy Diet for a Healthy Life (HDHL)[24] as partners and coordinators. In addition, KKSN and U Leipzig contribute to the project ERCIN that facilitates multinational clinical research. FAIR data sharing is addressed by international activities such as GO-FAIR, FAIRsharing[25], FAIRDOM[5], ELIXIR[4] or FAIR4Health[26] with active participation of HITS, U Leipzig and Fraunhofer. The research data management platform SEEK[6], developed by HITS with partners in the UK (University of Manchester) and in South Africa (Stellenbosch University), is used in numerous research networks in Europe, became a national resource for Norwegian life-sciences in 2018 and has first installations in the US (University of Connecticut Health Center). Co-applicants are actively involved in international cooperation related to DataSHIELD[27] (University of Bristol), specifically the EU projects ENPADASI[28,29], InterConnect[30] and MyNewGut[31].

Several partners are involved in RDA working groups addressing health data, reproducible health data services and ethics and social aspects of data. Furthermore, ZB MED and HITS are connected to library organisations and data publishers.

2.5 Organisational structure and viability

The governance of the project builds on long-standing cooperation among the (co-)applicants and adapts an organisational structure that has proven its viability in several European multi-

centre projects. The overall structure is depicted in Figure 3. **The spokesperson** is the overall leader and acts as single point of contact between the Directorate and NFDI4Health. The **General Assembly (GA)** is the ultimate decision making body of the consortium; every (co-) applicant is a member. The **Steering Committee (SC)** is the supervisory body for the project execution which shall report and be accountable to the General Assembly. Every **task area (TA) leader** is a member of the SC. The **Scientific Advisory Board (SAB)** will advise the consortium in all matters related to data science and the implementation of the FAIR data principles. The **User Advisory Board (UAB)** assures that the community needs are reflected in the strategic development and goal setting of the project, co-created with the SC and approved by the GA. In addition, it will advise the spokesperson. The participants will propose and develop specific projects to be conducted after approval by the GA.

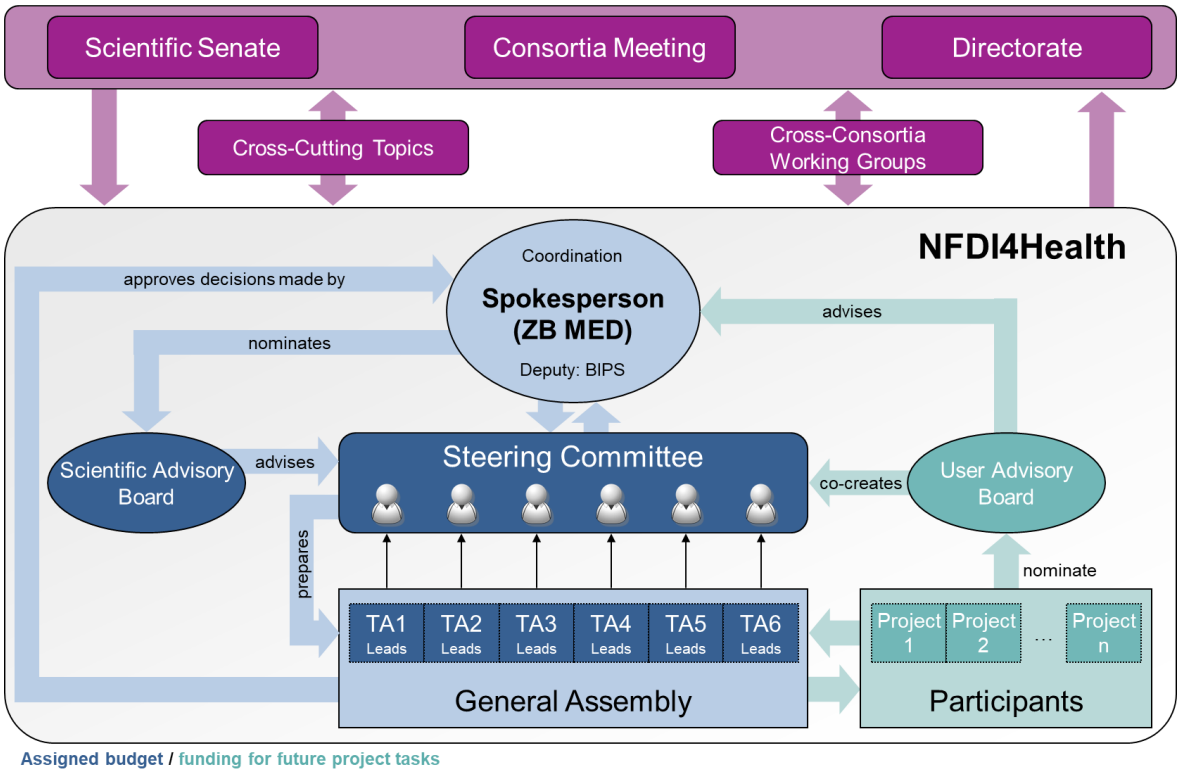

Figure 3: Overview of NFDI4Health governance

NFDI4Health will be coordinated by J. Fluck (ZB MED). She will be responsible for the orchestration of the TAs and the achievement of the overall project aims. She will ensure that all TAs complement each other, that there are no gaps in the coverage of specific measures and that all partners fulfil their responsibilities at any given time. She will coordinate all measures to ensure that inputs and outputs across TAs match correctly and that deliverables are provided in time. Where issues or delays arise, she will find a timely solution with the relevant TA teams. J.

Fluck will be supported by I. Pigeot (BIPS) as deputy spokesperson in all aspects of the project while T. Muth (RKI) will support the interaction with political decision makers.

J. Fluck will be responsible for the financial management and for legal, ethical and contractual issues. The **project office**, which will consist of the project manager (ZB MED), the financial/legal manager (ZB MED) and the communication manager (BIPS), will assist the spokesperson: The **project manager** will be responsible for administrative support incl. internal communication, progress, deliverable monitoring, reporting, quality assurance and DFG/NFDI administrative tasks. To ensure smooth and effective internal communication throughout the lifetime of the project, NFDI4Health will make use of technologies that are routinely used by the (co-)applicants, such as mailing lists, online meetings and collaborative authoring and reviewing tools. In addition, a closed-access secure extranet system, hosted at ZB MED, with a full suite of distance collaboration tools will be established. The **financial/legal manager** will be in charge of the financial management/reporting. The **communication manager** will take responsibility for public relations and external communication. In summary, the spokesperson together with her team will be responsible for the following tasks (TA1):

- prepares, with the support of the SC, the reports and documents required by the DFG;
- chairs the GA and the SC;
- is in charge of the financial management and orders the financial transactions of the grants from the DFG to the partners according to the rules as given in the contract with the DFG. The spokesperson will report and be accountable to the GA for all financial issues;
- ensures a smooth flow of communication between the partners;
- ensures prompt delivery of all software, documents and data identified as deliverables in the contract or requested by the DFG for reviews and audits, including the results of the financial audits prepared by independent auditors;
- prepares the meetings of the GA as well as the supporting material necessary for the decision-making procedure, assisted by the project office;
- performs and documents the risk assessment, assisted by the TA leaders.

TA leaders will be responsible for the scientific coordination of the respective task area, i.e.:

- quality control and timely submission of deliverables to the spokesperson;
- reporting of TA progress and of potential threats to the achievement of project aims during the regular meetings of the SC;
- provision of written half-yearly internal TA reports to the SC (Figure 4).

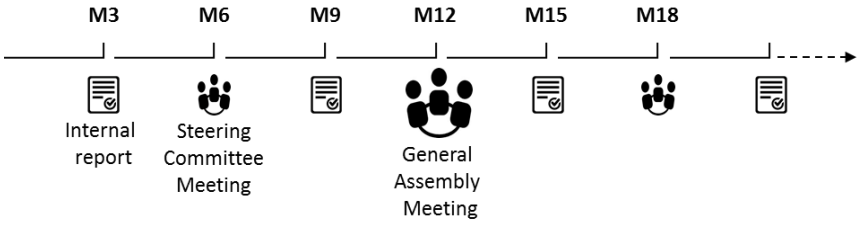

Figure 4: Timing of project meetings

The **General Assembly** (GA) will be the principal decision-making body of this project. All (co-) applicants will be represented in the GA by one representative who shall be authorised to negotiate and decide on all matters on behalf of his/her institution. The GA will be in charge of:

- all relevant budget matters;
- the acceptance of new co-applicants as well as the exclusion of co-applicants;
- any alteration of the consortium agreement (CA), subject to the approval of all institutions;
- deciding on the premature completion or termination of the project;
- approving new project proposals by participants.

The GA will meet face-to-face once a year. The preliminary agenda will be sent to all co-applicants at least 21 days prior to a meeting. Extraordinary meetings of the GA may be convened upon request of the spokesperson or upon request of one third of the co-applicants.

The **Steering Committee** (SC) will supervise the project and its dissemination during and beyond the funding period. The SC will be chaired by the spokesperson and composed of the **TA leaders**. Day-to-day routine management of the NFDI4Health project will be the role of the SC**.** The SC will prepare strategic decisions to be made by the GA. It will co-create the strategic development of NFDI4Health with the User Advisory Board, its concertation with the other NFDI consortia in the medical domain and it will assume overall responsibility for approving the deliverables. In addition, the SC will be responsible for:

- ensuring that all work is performed according to the state-of-the-art;
- ensuring that TAs interact with one another efficiently and effectively;
- reviewing and proposing to the GA budget transfers and re-allocations in accordance with the contract and the work plan;
- supporting the spokesperson in risk monitoring and risk management;
- agreeing on press releases and publication proposals.

SC meetings and internal reports are staggered to deliver a quarterly project management 'heartbeat' (Figure 4). In urgent cases, the SC can set up ad hoc panels, for specific issues or problems, e.g., to get advice in technical, scientific or quality control issues, to consult in financial matters and to support in questions concerning exploitation or dissemination of results.

The **Scientific Advisory Board (SAB)** will consist of five experts in the field of data science and research data management. The spokesperson will nominate members of the SAB who will be approved by the SC. The SAB will

- highlight critical issues and emerging global trends;
- review and advise on proposals and plans;
- assist and advise on the research data management strategy.

The **User Advisory Board (UAB)** will represent all potential user communities of NFDI4Health. Up to ten members will be nominated by the participants and approved by the GA. The UAB will consult the SC in strategic matters, selection of further use cases and preparation of new projects calls. Thus, the UAB will

- contribute user community insights and co-create strategic objectives with the SC;
- support community interaction in close exchange with TA4 (in particular T4.1);
- discuss project calls and advise (upon request) on submitted project proposals.

Both, UAB and SAB will convene annually and participate in GA meetings and feedback sessions as required.

2.6 Operating model

The operating model of NFDI4Health has to be based on (1) an organisational structure to run the collaboration among (co-)applicants and participants which is described in Section 2.5 as will be laid down in the consortium agreement (CA), (2) a structure to deliver standards and services to its user communities which is on the one hand reflected by the UAB as part of our organisational structure and on the other hand by the activities in TA4 with a strong involvement of the users, (3) a legal framework aligned with the overall organisational structure of NFDI. NFDI4Health will adapt the structure to be proposed by the NFDI Directorate, accounting for the German Fiscal Code and the German Value Added Tax Act. Moreover, according to this legal and financial framework, NFDI4Health will work out the above mentioned CA preferably based on a template provided by the DFG or the NFDI Directorate to be signed by all (co-)applicants before the official start of the project. For the time being, the (co-)applicants have committed

themselves to collaborate according to our organisational structure and according to their respective tasks described in the TAs as it is usual for DFG-funded cooperative projects.

In signing the NFDI4Health proposal, the (co-)applicants agree on collaborating in the long-term to ensure sustainability of the whole NFDI and its aims. In the future, letters of commitment will be signed by further data holding organisations.

Data holding organisations: To ensure sustainability of NFDI with respect to personal health data and to account for their high data protection level as well as the efforts needed to create and maintain such valuable and highly sensitive data bodies, NFDI4Health strongly builds upon distributed data resources provided by data holding organisations in the area of epidemiological research. These data holding organisations form the backbone of NFDI4Health. Our co-applicants, the Leibniz institutes BIPS and DIfE as well as the German federal government agency and research institute RKI committed themselves to provide institutional local data access points (LAPs) developed within NFDI4Health. Furthermore, UM Greifswald and U Leipzig agreed on providing LAPs for their public health cohorts LIFE and SHIP. In the first two years, NFDI4Health will set up these six LAPs. From the third year, NFDI4Health will integrate up to nine additional LAPs from cohorts maintained by participants.

Complementing our efforts to make personal health data reusable in the long-term, NFDI4Health will develop an according strategy for clinical trial data. Together with the KKSN and the clinical study centres, we will establish model infrastructures starting with two NFDI4Health co-applicant universities (U Leipzig and UM Göttingen) that will provide data infrastructures to allow for integrating clinical data repositories into NFDI4Health. These two data infrastructures will allow for data upload of clinical trial data and will serve as role model for other universities. NFDI4Health will invite institutions/researchers responsible for conducting clinical trials to submit their data to these two infrastructures.

From year 3-5, NFDI4Health will release the funding for future project tasks earmarked for calls on new use cases and new local data access points as well as for compensation of expenses for data upload (cf. Section 5). Selection criteria and calls will be prepared in close cooperation between the spokesperson, the TA4 Community Organisation Committee and the User Advisory Board. The necessary software and support to allow an easy set-up of new data infrastructures will be provided as service by TA3.

Furthermore, NFDI4Health will work out a business model among others to secure access to clinical trial data to be provided to university medical centres. Here, NFDI4Health will in

particular incorporate the expertise of the co-applicants U Leipzig and UM Göttingen representing the IT perspective, KKSN representing the medical perspective and the U Cologne representing the research data management at universities. Such a business model will be discussed with funding agencies to explore possibilities of re-funding data upload and preserving clinical trial data for a certain period of time. NFDI4Health aims to broker with universities and clinical trial centres to agree on long-term solutions. This business model will serve as a blueprint to be exploited by the epidemiological data holders.

Central services: Central services for search and (meta-)data access will be set up by the applicant ZB MED. Long-term perspectives of findability will be promoted through establishing (meta-)data publications in the health area (in close cooperation with the relevant international initiatives) that will allow for health data search without direct access to data. This will be facilitated by a central search hub established at ZB MED. Complementary to this hub, ZB MED will establish a central data access point (CAP) in close cooperation with NFDI4MED and GHGA. The CAP will serve as a central portal for standardised use and access applications that will be transferred to the local data access points. A persistent identifier of data publications will be set up for data access not only at the data holders themselves but also at the CAP. Standardised use and access protocols will be provided iteratively for the CAP as well as the LAPs.

Service enabling: Moreover, (co-)applicants will provide software, guidelines and training material to the consortium. For the contribution of all these resources, NFDI4Health will adhere to open source and open access policies whenever possible. In particular, sustainability of training courses and education programmes developed in course of this project will be achieved by publicly available e-learning tools and by integrating this material in routine courses, e.g., for principal investigators. To inform the user communities about all resources provided, NFDI4Health will set up a central repository but will also strive for publishing this information in public repositories with persistent identifiers. This will make those resources available for NFDI without time or license restrictions.

3 Research Data Management Strategy

Current state of research data management

There are considerable challenges for data access in the health sciences due to the high sensitivity of person-related health data. In addition, substantial efforts are necessary to bring data together from studies undertaken under different ethical, regulatory and legal arrangements, by researchers operating in multiple institutions with diverse governance structures and processes.

For some institutions, specific restrictions on sharing individual data exist and solutions which do not require transfer of data are preferable. Hence, data sharing is usually restricted to partners within multicentre studies or collaborative research projects such as the IDEFICS/I.Family children cohort, the EPIC study and ERIC (for a detailed list see Table A1, Appendix 1).

In such large multicentre studies, huge efforts are taken to implement a central research data management (RDM) that results in highly structured and curated data sets that are usually stored by a central data integration centre. This may not hold true for smaller studies which typically lack an elaborated RDM. A further challenge lies in the fact that the data sets resulting from these study types are living data bodies that are continuously curated and extended by additional variables and further follow-ups. These regular updates and the high diversity of phenotype descriptions make the usage of existing terminology standards challenging. Hence, although various health-related standards and terminologies exist (cf. Section 3.1), they are only applied in a minority of studies. Even for disease descriptions or laboratory diagnostics, standards such as ICD10 or LOINC are rarely used. Furthermore, SNOMED, the worldwide accepted medical terminology that is part of almost all standardisations efforts and tools, is not yet licensed in Germany.

The fact that there does not exist a consensus on uniform IT standards to be used in the field of clinical trials, public health surveys and epidemiological cohorts considerably impairs interoperability of these data bodies. Currently, large-scale epidemiological cohort studies such as LIFE, the NAKO Health Study, SHIP provide web-based tools to obtain an overview of the collected data and select data subsets for data applications. Yet, these implementations are not generic and search options vary widely across studies. Moreover, there is significant heterogeneity in the methods used to assess and operationalise variables across studies as strikingly demonstrated by the use cases in TA5. Further heterogeneity results from the applied IT solutions including metadata repositories, cohort browsers and interfaces that are currently fragmented and mostly disjunctive for clinical trials, public health surveys and epidemiological cohorts. Also, findability of these data bodies is hampered by the lack of searchable standardised metadata publications.

NFDI4Health will build on the existing expertise and long-stand experience of its (co-)applicants as described in detail in Section 2. For instance, the data holding organisations have already established their own RDM structures that partly allow external access to their data sets. Most of them have implemented use and access modalities as, e.g., the data integration centre of the NAKO Health Study (UM Greifswald) and its use and access committee, the central data server of the IDEFICS/I.Family cohort (BIPS), the Leipzig Health Atlas to get access to the LIFE study

(U Leipzig), the access point to SHIP (UM Greifswald), and the access strategy of RKI to its scientific use files, just to name a few. Most co-applicants are involved in joint European efforts for data harmonisation and standardisation as listed in Section 2.4 (e.g., ENPADASI). UM Greifswald will provide expertise on achieving data quality standards, gained during previous and ongoing work with community stakeholders within TMF (e.g., the 2019 initiated working group *Data quality and transparency*) or STRATOS[32], and as part of third party funded projects such as the DFG-funded project "Standards and tools for data monitoring in complex epidemiological studies"[33], or the H2020 euCanSHare project[34]. Further considerable expertise in setting up clinical trials according to a standardised protocol is provided by the KKSN. The clinical trial centres are responsible for data management for hundreds of clinical trials and have extensive experience in quality management of clinical data. A detailed list of all infrastructures and data repositories established by (co-)applicants and participants is given in Table A1 (Appendix 1). A description of services and reusable software is given in Section 3.3 and Table A4 (Appendix 1).

The data holders will team up with HITS and Charité/BIH that are well established in national and international standardisation organisations and FAIR projects in order to support the implementation of the FAIR data principles within NFDI4Health (see also Section 2.2 and 2.4). With respect to advanced IT solutions, HITS, U Leipzig and UM Göttingen have long-standing experience in setting up state-of-the-art data management and data analysis platforms. They currently work on interoperable data structures and data integration centre solutions as part of the MII and of the FAIRDOM project[5]. For more international collaborations, we refer to Table A2 (Appendix 1). ZB MED will provide important know-how and experience in setting publication standards and search services for literature and research data. Moreover, due to the legal and ethical restrictions mentioned above, in general reuse and sharing of data will be only possible if the data sets remain stored decentrally. Thus, extensive knowledge on federated techniques for data analyses is required which will be provided by DIfE, MDC and Fraunhofer FIT and which will be accounted for in the NFDI4Health RDM strategy. Finally, our RDM strategy will also be prepared for the growing application of machine learning techniques. Here, NFDI4Health will in particular build on the considerable experience of Fraunhofer FIT, MEVIS and SCAI in the application and development of artificial intelligence (AI) solutions.

The NFDI4Health research data management strategy

The NFDI4Health RDM strategy will address the core deficits in current research practice with regards to FAIRness and lack of common standards across the various personal health-related data bodies. The implementation is mainly guided through the use cases which have been selected in order to cover a broad range of needs, e.g., data curation, as expressed by the user

community during the planning phase of NFDI4Health. The use cases, which are described in detail in TA5, address key challenges for FAIRification of personal health data. NFDI4Health will start with six use cases as demonstrators addressing typical tasks, namely (T5.1, T5.2) assessment of exposure and outcome in epidemiological cohorts (→ heterogeneity in methods and variables), (T5.3) record linkage of various data sources (→ no unique identifier, different data formats, data protection/privacy requirements), (T5.4) clinical trials (→ limited search for and access to study data), (T5.5) lack of harmonisation across surveillance systems and (T5.6) application of machine learning in federated data analyses (→ growing importance of AI solutions). Further use cases will be initiated by the user community through open calls during the runtime of NFDI4Health.

Data standardisation, based on the needs-and-gaps analysis of the use case requirements, is addressed in TA2 where NFDI4Health standards for findability and interoperability will be worked out. For this purpose, international standards (see Section 2.4) will be transferred and adjusted to meet the specific requirements of NFDI4Health. In addition, gaps in data and metadata standards will be closed by adapting and – if necessary – further developing existing international standards to meet the needs of the target user communities and of the use cases of NFDI4Health (T2.2). With respect to interoperability of the data infrastructures, NFDI4Health will work on standards for interfaces (T2.4) that all data holding organisations can easily integrate into the NFDI4Health network. In addition to data standards, quality of data to be shared has to be assessed according to predefined criteria. Therefore, data quality standards, requirements and related tools for their implementation will be addressed in NFDI4Health as well (T2.3, T3.3), based on the expertise described above.

A central element of the NFDI4Health envisioned RDM strategy is to help researchers with the discovery (*Findability*) of and access to existing data. Based on the use cases, publication guidelines will be developed by making studies persistently identifiable and providing study metadata (T2.1). They will be based on publication standards such as the DataCite Schema[35]. NFDI4Health will provide a best practice blueprint for the publication of metadata that can be searched and that is linked to distributed data sets hosted by different infrastructures. This will ensure permanent findability of these resources independently of any NFDI4Health service.

NFDI4Health will establish a central search hub according to internationally accepted standards for data and metadata (e.g., CDISC[36] standards and HL7 FHIR[37], see Section 3.1) and open source components (e.g., SEEK[6] and Mica, see Section 3.3) in order to increase the impact of own developments. The search functionality of the NFDI4Health search hub (T3.1) will be independent of data access. This is extremely important for personal health data, as often the data

themselves are *not* openly accessible. In close interaction with international partners, our final goal is to develop generic search environments for health data similar to search functionalities for literature in PubMed[38] and the ZB MED search portal LIVIVO[39]. In addition to semantic search and filtering functionalities, cohort browsing will be an innovative component of the NFDI4Health search hub.

As further component of the RDM strategy, NFDI4Health will establish a central data access point (CAP) to support researchers with one standard interface and to guide the users through a common access protocol. The CAP will provide a data and consent broker service and facilitate the user contact with the various data holders (LAPs). The corresponding workflow is described in detail in TA6.

Since the NFDI4Health data bodies comprise highly sensitive personal health data, its RDM strategy has to comply with data protection and data privacy regulations. If processing of such data requires the informed consent of study participants any reuse of their data has to comply with the original consent. NFDI4Health will develop an automated machine-processable informed consent mechanisms to facilitate the extraction of the structured consent information from the primary data source. Since such a machine-processable mechanism will not be available for historical data sets that will be made accessible by NFDI4Health, the CAP will also offer the functionality of a manual check as an alternative. NFDI4Health envisions and will work on the legal instrument of broad consent, which allows for consent to data processing for a broad range of health-related research questions by referring to certain areas of scientific research in general terms (cf. Recital 33 of GDPR).

In accordance with data protection regulations, federated data analysis is a major opportunity to analyse data distributed in different infrastructures without retrieving all data. NFDI4Health will build on first experiences with infrastructures for data sharing and federated meta-analyses as, e.g., DataSHIELD in the context of the EU projects ENPADASI and InterConnect. NFDI4Health will extend the current DataSHIELD functionalities inspired by the use cases in TA5. While DataSHIELD is restricted to R-programming language, Personal Health Train (PHT), developed in the context of the GO-FAIR initiative and piloted within MII, offers a more generic distributed analysis infrastructure suited for all applications. In NFDI4Health, we will set up the Personal Health Train to explore two AI-based applications: We will implement (1) federated radiomics and imaging AI workflows (T5.6) based on this infrastructure, such that learned image analysis and radiomics models can be deployed, validated and further trained at multiple locations. (2) Eventually, NFDI4Health will explore whether synthetic data sets can be generated based on the wealth of data available in this consortium that will serve as a good proxy of the original data

sets. Dedicated software (cf. T6.4) will be used in PHT to generate such synthetic personal health data. The tool will allow for synthetic health data while specifying a risk threshold for re-identification in order to ensure differential privacy, i.e. to withhold information about individuals in the data set. This approach will provide a mechanism to share personal health data across different organisations while providing guarantees for preventing re-identification of study subjects and thus in full agreement with any data protection regulations.

User involvement

To foster acceptance and use of the above concepts, NFDI4Health already involved the user community during the preparation of this application by conducting a survey to learn about the needs and by inviting the users to a community workshop (June 2019). NFDI4Health will continue this commitment and will actively involve the user community in its overall governance, launch project calls, offer training and dissemination activities for various target groups and establish feedback loops. NFDI4Health explicitly devotes two task areas, namely TA4 and TA5, to facilitate user involvement, where a detailed description of our activities can be found.

As already described in Section 2.5, a User Advisory Board (UAB) representing the user community will be involved in the strategic development of all NFDI4Health activities by close interaction with the Steering Committee and, in particular, by discussing and advising on project calls. Two types of project calls will be launched among participants: (1) calls for additional use cases to further adapt our RDM strategy and resulting NFDI services to user requirements and (2) calls for integration of new data bodies and setting up additional local data access points to further integrate existing and newly generated personal health data. Selected projects will be funded by the NFDI4Health budget earmarked as “funding for future project tasks”. In addition to the project grants, especially participants holding clinical trial data, will be invited to upload their data in standardised formats into the existing local data access points. Costs incurred by this activity will also be reimbursed by the NFDI4Health funding for future project tasks.

Back to back with the annual General Assembly meeting, NFDI4Health community workshops will be organised to offer training for participants and to strengthen mutual exchange on standardisation, guidelines and services. NFDI4Health will also organise thematic symposia with various stakeholders from academia, scientific societies, health services and policy to increase the outreach of NFDI4Health and to obtain input from outside the consortium. The training material will be developed in close exchange with the user communities and continuously revised according to their feedback where NFDI4Health will in particular benefit

from the rich training experience of the KKSN network where in 2018 alone, more than 9,500 clinical physicians and other medical researchers were trained.

To monitor the user needs and to implement a change process we will establish various feedback loops. Among others we will implement usability tests for web-based services, qualitative in-depth interviews, focus group discussions about specific service aspects and crowd-sourcing for user-generated services. As a central element for collecting information on requirements and user feedback NFDI4Health will use the concept of PALs (Project Area Liaisons) adopted from FAIRDOM[40]. PALs will be a smaller number of “front line” experts from research projects and from data holders. They act as data management advocates and multipliers and communicate new developments back to their projects. NFDI4Health will select PALs from the user community and work with them on: (a) gathering requirements and consultation, (b) reviewing of ideas and prototypes, (c) testing and reporting on solutions and (d) gathering intelligence from research projects. NFDI4Health can build on extensive experiences such as the PALs concept, introduced by FAIRDOM more than 10 years ago, which has already been used by several co-applicants.

3.1 Metadata standards

In NFDI4Health, we define a semantic annotation as a computer-accessible metadata item that captures, entirely or in part, the meaning of a data set, data set component or data element. Semantic annotations are a critical feature of the vision of the semantic web, where documents are linked to metadata describing the documents’ contents, thus facilitating search and retrieval as well as data interoperability. Common metadata standards for health research that will be used by NFDI4Health include: (a) **ICD** (International Classification of Diseases), a standard diagnostic classification for diseases, (b) **LOINC** (Logical Observation Identifiers Names and Codes), a universal code system to identify laboratory test results and clinical observations to enable exchange and aggregation of electronic health data from many independent systems and, as soon as possible in Germany, (c) **SNOMED CT** (Systematized Nomenclature of Medicine, Clinical Terms), a comprehensive nomenclature to accurately store and/or retrieve records of clinical care in human and veterinary medicine, which will presumably be licensed in Germany by the BMBF in 2021 and by the BMG in 2025.

In addition, the metadata concept of NFDI4Health, that will be worked out in TA2, will build on domain-specific standards, such as **HL7**[20] (Health Level Seven International) and **CDISC**[36] (Clinical Data Interchange Standards Consortium) (for details see below) as well as the future

ISO 20691[41] (Requirements for data formatting and description in the life sciences for down-stream data processing and integration workflows) that is currently developed under the lead of HITS in ISO/TC 276/WG5. In particular, the provided services and resources will comply with metadata recommendations from ISO 20691 regarding domain-specific terminologies and reporting guidelines for medicine, health and disease. For an overview see Table A3 (Appendix 1).

With respect to metadata and data formats of technology-specific datatypes in the life sciences that are not exclusively used in health research, such as data from genomics, transcriptomics, translatomics, proteomics, metabolomics, lipidomics, glycomics, enzymology, immunochemistry, imaging, synthetic biology, systems biology, systems medicine and others, NFDI4Health will work closely together with **GHGA** and **NFDI4MED.** By using international and national standards, these data will be rendered interoperable with comparable data across these consortia. Comprehensive lists of corresponding data and metadata formats, minimal information guidelines and terminologies and ontologies for these technology-specific datatypes can be found in the information portal of the international initiative FAIRsharing[25] (see below) that will be used as an information resource on data and metadata standards for users of the services provided by NFDI4Health.

Furthermore, we will establish strategic collaborations with international partners that develop and provide such domain-specific solutions, such as European Bioinformatics Institute (EMBL-EBI[42]), ELIXIR[4], the European bioinformatics infrastructure, FAIRsharing[25] and FAIRDOM[5]. All these initiatives and institutions have provided letters of support for NFDI4Health.

FAIRsharing, in particular, is an informative and educational resource that describes and interlinks community-driven standards, databases, repositories and data policies, as well as bundles information on metadata standards, minimal information guidelines (checklists) and links to domain-specific terminologies and ontologies. As of September 2019, FAIRsharing has 1,323 standards, 1,281 databases and 121 data policies, covering natural sciences (incl. biomedical and life sciences), engineering, humanities and social sciences. NFDI4Health will work closely together with FAIRsharing to establish a collection of data standards, resources and policies specific for the health research domain. Based on this collaboration, NFDI4Health will serve as bridge between FAIRsharing and other NFDI consortia.

The **EMBL-EBI** (European Bioinformatics Institute) maintains a service at https://identifiers.org for resolving Uniform Resource Identifiers (URIs) used in annotations[43]. This service will be used by NFDI4Health for resolving annotations of data sets and data set elements, when these annotations are defined in the corresponding format.

As standards of CDISC we will use: (1) **ODM-XML**[44] which is a platform-independent format for exchanging and archiving clinical and translational research data (incl. metadata, administrative data, reference data and audit information). ODM-XML facilitates the regulatory-compliant acquisition, archival and exchange of metadata and data. (2) **BRIDG**[45] which is a domain analysis model that represents the realm of protocol-driven clinical, pre-clinical, translational and basic research.

As standards of HL7, we will exploit **FHIR**[37] (Fast Healthcare Interoperability Resources)**,** a widely used international standard for exchanging digital health data in health information technology. FHIR relies on so-called resources, which define typical healthcare concepts for common use cases and which can be extended where necessary. The increasing adoption of FHIR in medicine and healthcare can make data accessible to large-scale research and analytics projects. Using FHIR in NFDI4Health will therefore not only enable data exchange within NFDI4Health and related consortia but also follows interoperability efforts at international levels. For NFDI4Health, FHIR will serve as a tool to define standard profiles for consistent data structures. Importantly, the definitions of these FHIR profiles within NFDI4Health will be performed in close coordination with national and international activities that aim to use FHIR for interoperable data exchange (e.g., the MII or the basic FHIR profiles defined by HL7 Germany).

With regards to data quality standards, strong community-driven activities have evolved over the last years to improve data quality with many co-applicants participating, particularly from epidemiology and clinical studies. The development of standards and tools has been supported by DFG, as outlined above, and plays a key role in the euCanSHare initiative, which establishes a cross-border data sharing and multi-cohort cardiovascular research platform. euCanSHare strongly supports the NFDI4Health activities (see letter of support) and will collaborate on the developments. Our consortium will proceed further on this path to align available standards and tools with community demands and to integrate into prospected NFDI4Health services. Developments and community exchange will also be supported by and coordinated with the TMF working group *Data quality and transparency*, led by C.O. Schmidt (UM Greifswald).

Tools and methods to further improve reproducibility and reusability of data belong to the key issues of the TMF working group *Information technology and quality management (ITQM)*, led by U. Sax (UM Göttingen) for many years. NFDI4Health will implement basic data and workflow provenance services. This includes the regular HL7 FHIR Provenance Information like the initial cause for data capture, the technical data source and the workflow describing the data extraction and integration[46,47,48]. The resulting service will provide standardised data provenance

and workflow provenance information and will enable the retrieval of the provenance trail of any data item stored in a standardised format.

NFDI4Health will also align its work with the core data set defined by the interoperability working group of the MII (led by U. Sax, UM Göttingen). This core data set defines a standard data model for German university hospitals and consists of a basic set of modules including patient demographics, cases, diagnoses, procedures, laboratory results and medication. This basic set is supplemented by use case-specific extension modules, e.g., in the fields of oncology, imaging or genomics. Wherever applicable, NFDI4Health will base its own standardisation efforts on this core data set to ensure interoperability with the MII-consortia.

3.2 Implementation of the FAIR principles and data quality assurance

NFDI4Health will empower clinical, epidemiological and public health research communities to develop capabilities to mature processes for FAIR data sharing and reuse. Although health research communities are quite advanced in standardisation of data and documentation of studies, there is a well-recognised need to improve FAIR data generation, access and linking. NFDI4Health will outline incremental maturation paths for FAIRification of data. In alignment with the RDA FAIR maturity model, NFDI4Health will support data holding organisations in developing capabilities to implement the FAIR data practice while accounting for their priorities and restrictions. The FAIR maturity pathway starts with an assessment of the current state and will set a target level of FAIR capabilities that the data holder wants to achieve. NFDI4Health will provide technical, organisational, legal and ethical guidance to enable capability building with respect to the following FAIR maturity paths, namely “Findability and Discoverability”, “Accessibility and Distributed Data Analytics”, “Data Reuse and Patient Consent”, “Interoperability and Record Linkage” and “Data Quality and Provenance” (see Figure 5).

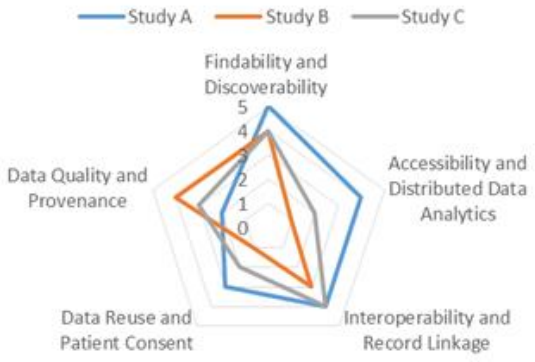

Figure 5: Different maturity levels of studies according to NFDI4Health Fair Maturity Pathways

Findability and Discoverability of data is the first dimension of the NFDI4Health FAIR maturity paths. All data holders will be enabled to develop FAIR data management capabilities to create and register rich metadata of their data sets, based on publication policies to be created within NFDI4Health. For this purpose, NFDI4Health will establish metadata standards based on existing community practices and expertise of the (co-)applicants and provide enabling services and tools to create metadata according to these standards. (Co-)Applicants will also offer training programmes in health research data management and data science. ZB MED will provide sustainable services for persistent identification and health data search and will establish a central search hub as repository for metadata which will ensure long-term storage of metadata and will serve as a single gateway to find personal health data sets.

Accessibility and Distributed Data Analytics is the second maturation dimension of NFDI4Health. Since data access to personal health data is usually restricted by data protection and privacy regulations as well as by the given informed consent of study participants, NFDI4Health will offer solutions for data access and data sharing that build on distributed data infrastructures. Thus, NFDI4Health will implement – as described above – a central data access point (CAP) at ZB MED and local data access points (LAPs) at all data holders. All requests for data access will be enabled by the CAP and then transferred to the LAPs. In addition, NFDI4Health will offer FAIR distributed data analytics approaches to allow joint analyses of various data sets without physically extracting the data from the data holding organisations. In this respect, NFDI4Health will extend the methodical arsenal of DataSHIELD and will enable the application of innovative machine learning techniques by exploiting the Personal Health Train.

Data Reuse and Patient Consent is a challenging maturation dimension, in particular, for highly sensitive data gained from human populations. To enable reuse of personal health data while accounting for given informed consent NFDI4Health will apply the approaches described above and exploit the concept of differential privacy. In particular, NFDI4Health will develop machine-processable consent mechanisms and procedures for privacy risk assessment. In addition, NFDI4Health will work on the legal instrument of broad consent.

Interoperability and Record Linkage is a major maturation dimension of NFDI4Health. Terminologies, ontology and vocabulary annotation services will be offered centrally and will be used for creating rich metadata about data. Core data sets, profiles, data elements and value sets will be defined for improving machine interpretability and for facilitating interoperability of data sets. Interoperability of the central and the LAPs will be achieved by standardisation of interfaces. As an ultimate goal with respect to interoperability of data sets, NFDI4Health will

strive for record linkage of primary and secondary data sets. Respective legislative demands will be formulated.

Data Quality and Provenance is the fifth maturation dimension of NFDI4Health FAIR maturity paths. This dimension targets “fitness for use” of a data set and precise knowledge about how the data set was generated and by whom (instruments and origin). A data set of high quality and a deep understanding of the data generation process are key prerequisites for valid scientific data analyses. NFDI4Health will offer functionalities and tools to support decentral as well as central data quality assessments (e.g., missing or inconsistent data) according to harmonised standards (SQUARE², tranSMART, Opal, R libraries on data quality assessments). One key objective within this maturity dimension is the transparent handling of data quality-related findings to guide researchers about potential pitfalls when conducting scientific analyses within and across data bodies. For this purpose, harmonised data quality-related findings will become visible through central data search services, even if they have been computed locally. Any work on data quality standards will be coordinated with other NFDI initiatives as well as other networks, projects and community networks (e.g., STRATOS, euCanSHare, TMF, RDA FAIR Data Maturity Model group[49]).

We will ensure compliance of our technical implementation with the quality requirements in health research based on individual-level data by (1) building upon infrastructures, services, tools and methods that have been developed within our community and that have already reached some degree of routine use and technical maturity (see Section 3.3) and (2) implementing feedback loops with representatives of the target community on developments within NFDI4Health (see above).

3.3 Services provided by the consortium

NFDI4Health will provide services (for details see TA3) that target two different user groups, namely (A) data analysts who want to gain an overview over existing data and to get access to these data sets and (B) data holders who need to make their data FAIR. The following minimum services will be provided to user group (A) to foster their acceptance and use:

- a centralised search hub (a) to enable the findability of studies and data sets, (b) to obtain details on study characteristics, (c) to enable cohort browsing at the variable level, (d) to get information on modes and conditions of data access, (e) to get information about FAIR metrics, data quality and data usage;

- a central data access point (CAP) (a) to submit applications for data access, (b) to check the compliance of the research purpose with the consent given by study participants, (c) to check the compliance of the application with study-specific conditions for data reuse and (d) to provide templates for use and access agreements between data holders and users;
- local data access points (LAPs) with facilities (provision depending on the study-specific conditions and risk of re-identification) (a) to transfer a tailored data set to the user, (b) to analyse data locally in a secured environment with standard statistical software, (c) to enable distributed data analyses with tools certified by NFDI4Health;
- training modules, help desk and FAQs.

The following minimum services will be provided to user group (B)

- quick user guide on how to make the data fair;
- template for a generic research data management plan;
- publication policy and guideline to create metadata of a given study;
- central portal to upload and publish study metadata (based on generic standards such as the DataCiteSchema[35]) (central search hub);
- guidelines on harmonised data quality assessments;
- assignment of a permanent identifier to the published data set (e.g., DOI);
- assignment of a FAIR metrics and a data quality label;
- data usage information;
- optional data storage at an existing LAP;
- downloadable reference solution to establish an own LAP;
- management for machine-processable broad consent;
- provision of survey instruments and examination modules;
- training modules, help desk and FAQs.

An overview of the NFDI4Health services is shown in Figure 6. These services will require structured maintenance which will be supported by a centrally managed software repository which will provide all software modules and a repository for certified services and documents.

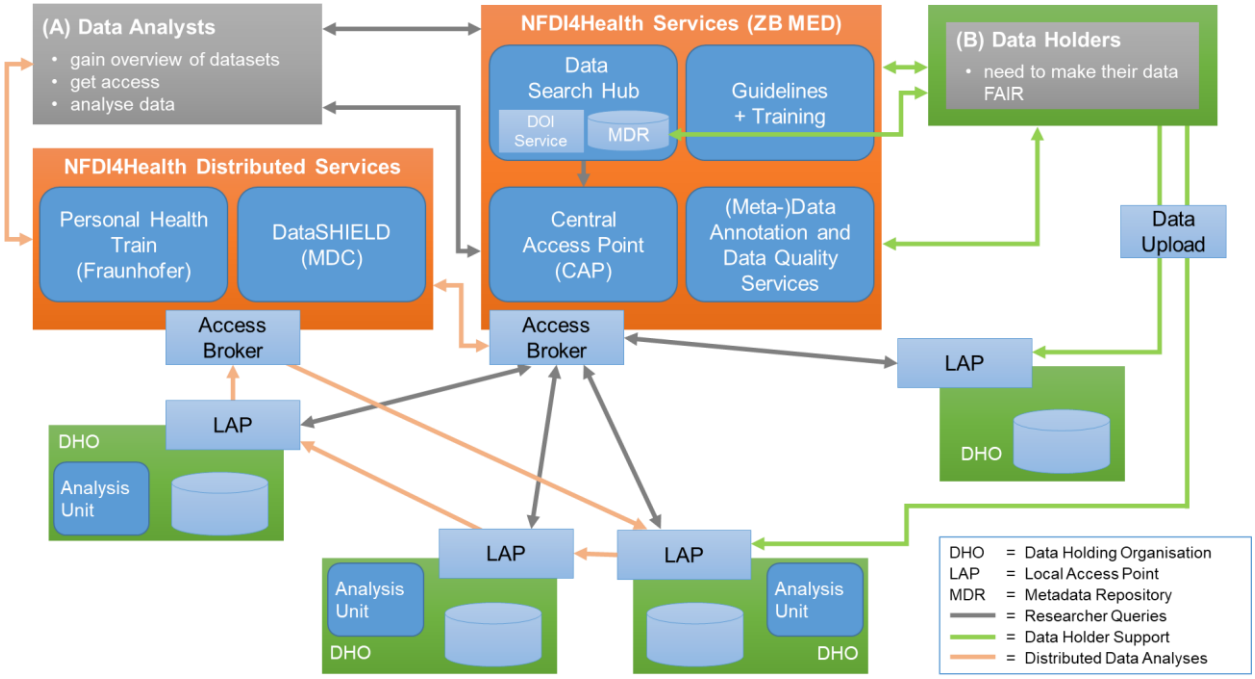

Figure 6: Overview of NFDI4Health services for (A) data analysts and (B) data holders. The local data access points (LAPs) may vary in complexity (e.g., not all data holding organisations (DHOs) have external accessible analysis unit or take part in distributed computing).

To develop the above services, NFDI4Health will build on a broad range of software that have been developed by (co-)applicants with institutional funds as part of their institute's mission or by scientific developer communities mostly with participation of (co-)applicants and predominantly as open source, both ensuring permanent availability (see Table A4, Appendix 1, for the most important services). Active participation in community-based open source software development is an accepted way to ensure maintenance, method advancement and availability of academic software based on open standards. As a prominent example, the web-based SEEK platform[6,50] has been developed by HITS together with partners at the University of Manchester, UK, in the framework of the FAIRDOM project. SEEK supports management, sharing and exploration of data in life science consortia (e.g., Virtual Liver Network[51,52], Liver Systems Medicine LiSyM[9], ERASysApp[53]). It provides an access-controlled, web-based environment for scientists to share and exchange data, information and models. A plug-in architecture allows linking of experiments, their protocols, data and models. Through the representation of metadata in RDF (Resource Description Format), SEEK enables rich semantic queries. UM Göttingen and U Leipzig customise the SEEK platform to build up additional services as, e.g., in MyPathSem[54], UMG MeDIC[55] and the Leipzig Health Atlas (LHA[15]). The data sharing concept of the LHA has been developed in accordance to FAIR and OAIS (Open Archival Information

System). LHA is already in use as storage platform (incl. rights management, (meta-)data standardisation, data explorations tools) for epidemiological and clinical trial data.

Intensive interaction with a growing user community (for details see TA4) will further ensure the continuous adaptation of those tools towards user needs. All services will be updated continuously according to the feedback of the users, gathered by appropriate feedback loops. All feedback loops running via PALs or via direct communication with service providers of NFDI4Health will be maintained by the institutions concerned beyond the duration of this project to adapt the above services to new needs of the community. In addition, LAPs for epidemiological data such as those provided by NAKO Health Study, RKI, U Leipzig, UM Greifswald and BIPS have established strict access rules and control mechanisms to guarantee data security. These local data access points will be maintained beyond the funding period of NFDI4Health. Corresponding LAPs for clinical trials will be established by U Leipzig and UM Göttingen. With the support of KKSN, roll-out to further clinical trial centres is planned from year three onwards. The same is true for the central search hub and the CAP as well as the software repository and the functionalities of study identifier assignment that will be integrated in the ZB MED search environment. All co-applicants conducting surveys, epidemiological studies and clinical trials will offer their questionnaires and examination modules as well as the corresponding standard operating procedures (SOPs) developed in course of their research projects to the user communities as, e.g., the IDEFICS/I.Family instruments[56] provided by BIPS. The training modules will be maintained by training providers, in particular the USB Cologne and the KKSN but also by others as, e.g., the graduate programme on research data management and data science at U Bremen.

4 Work Programme

NFDI4Health will focus on structured personal health data (epidemiological, public health and clinical trial data) comprising precise and deep phenotyping of study subjects. In order to reach the NFDI4Health objectives, we will build upon existing solutions according to national and international standards (cf. Section 2.4). NFDI4Health will comprise six interacting task areas with the user community taking a central part:

- TA1 focuses on project coordination measures including overall project financial controlling, dissemination, public relations and outreach to political decision makers.

- TA2 will develop standards for findability and interoperability. This concerns data management and publication policies, (meta-)data standard harmonisation, data quality assessment and data provenance identification as well as standardisation of health data access.
- TA3 will use the standards provided by TA2 to deliver NFDI4Health services, service enabling tools and the necessary technical support. These include a central search service, tools supporting data quality assessments and metadata annotations. They also include services for data use and access as well as data/software repositories and archiving services. Moreover, TA3 will offer advanced distributed data analysis services.
- TA4 will address the needs and demands of the user community, inform and train our user community and organise community participation by disseminating tools and knowledge to and triggering feedback from the community.
- TA5 will implement the standards and services developed in TA2 and 3 in thematic use cases which cover a broad range of needs expressed by the user community during the planning phase of NFDI4Health. Each use case will serve as an example for a specific challenge for FAIRification of personal health data. Based on these use cases we will develop and apply practical solutions that will facilitate the whole process from findability to reuse.
- TA6 is concerned with the legal framework conditions (EU GDPR) of data protection and data privacy, record linkage and solutions such as broad consent and differential privacy, e.g., via simulation of synthetic health data.

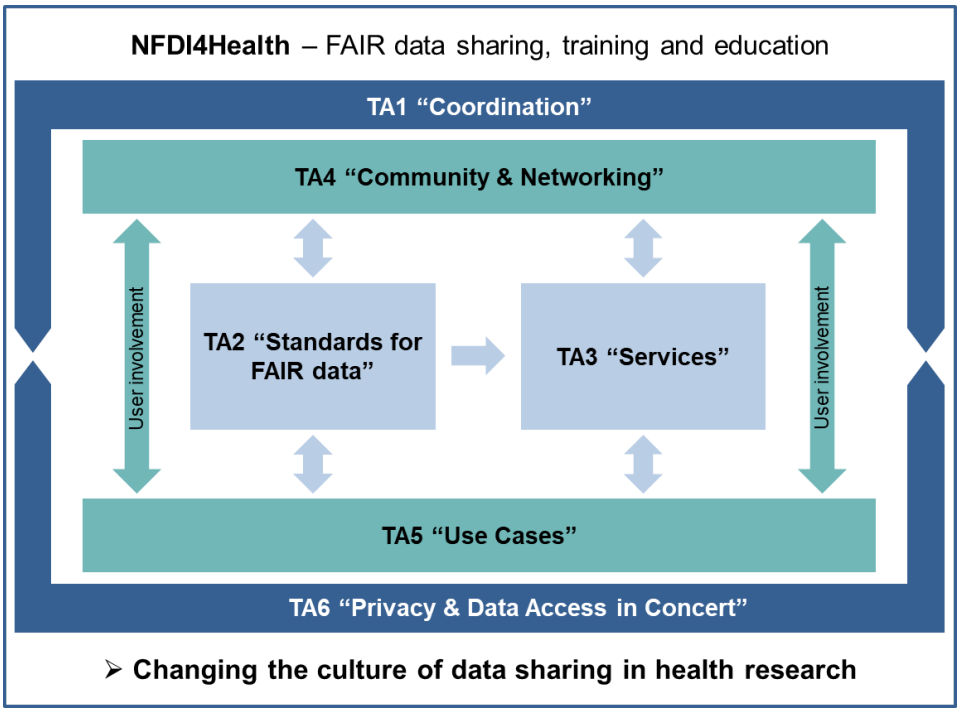

Figure 7. Task area structure of NFDI4Health

As indicated in Figure 7, TA4 and TA5 will actively involve the user communities to learn about their needs and requirements which will then feed the development of standards for FAIR data and of services to enable the process of FAIRification in TA3 where the latter builds on TA2. TA4 will set up facilities that will allow a continuous feedback of the user community on, e.g., early releases of standards and software demonstrators to be incorporated in further versions. Further use cases suggested by the user community will form a central building block of the participatory user involvement. The whole work programme will be coordinated and monitored by the NFDI4Health management (TA1) and embedded in the framework of data privacy and data access regulations (TA6).

The tables given at the beginning of each TA list all (co-)applicants committed to this TA regardless whether they request funding or contribute their own resources. In addition, contributing participants who are listed for each measure separately will be refunded by the amount earmarked as funding for future project tasks. The work progress of each TA will be documented by its deliverables (see table attached to each TA description) and by its respective milestones (Table 1). A detailed Gantt chart is shown in Figure 8. Eventually, NFDI4Health will establish a network enabling close interaction of data holders in public health, epidemiology and clinical research with their user communities. This will lead to a change in the culture of data sharing in health research.

Table 1: Milestones of NFDI4Health

No.	Description (month)
M1.1	Kick-off meeting (month 2)
M1.2	Critical midterm review by SAB (month 28)
M1.3	Follow-up review by SAB (month 55)
M1.4	Successful roll-out of all services developed by NFDI4Health (month 60)
M2.1	First consented cohort data model (month 12)
M3.1	All NFDI4Health services with basic functions are publicly available (month 36)
M3.2	Data access of NFDI4Health data can be requested through the CAP (month 54)
M4.1	Training programme on research data management and data science launched (month 24)
M4.2	NFDI4Health congress („New Horizons“) (month 30)
M4.3	FAIR sharing collection of guidelines and standards published (month 36)
M4.4	NFDI4Health congress („Continuation“) (month 55)
M5.1	First local data access point (month 24)
M5.2	Users can search and browse NFDI4Health data of co-applicants based on published (meta-)data sets according to the publication standard (month 58)
M6.1	Generic data access workflow developed (month 24)

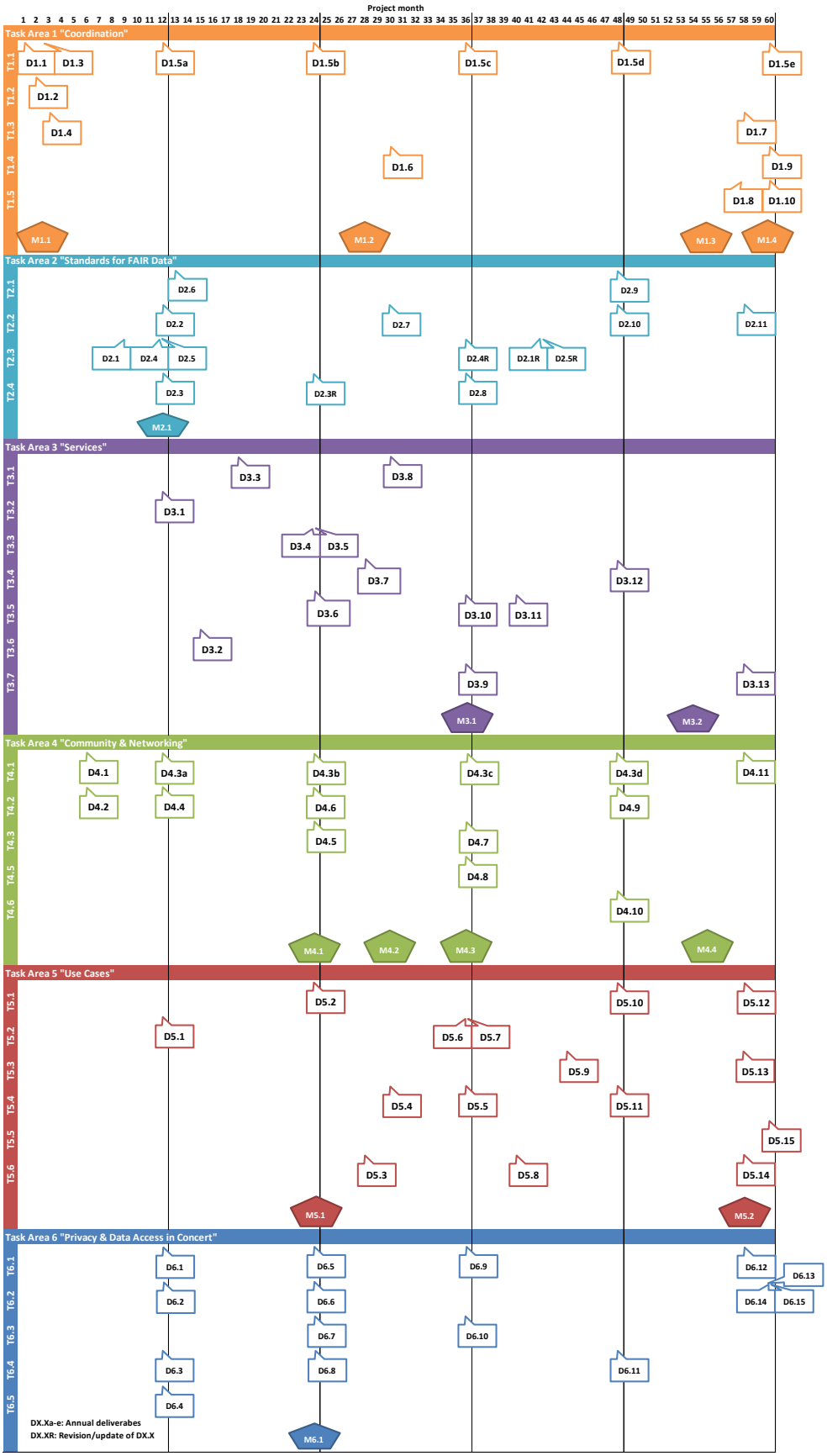

Figure 8: Detailed Gantt chart

4.1 Overview of task areas

Task Area	Measures	Responsible Co-Spokesperson(s)
TA1 “Coordination”		J. Fluck (ZB MED), I. Pigeot (BIPS)
	T1.1 “Project governance”	J. Fluck (ZB MED)
	T1.2 “Project financial controlling”	J. Fluck (ZB MED)
	T1.3 “Communication and public relations”	I. Pigeot (BIPS)
	T1.4 “Outreach to political decision makers”	L. Wieler (RKI)
	T1.5 “Business model”	J. Fluck (ZB MED), M. Löffler (U Leipzig)
TA2 “Standards for FAIR Data”		M. Golebiewski (HITS), C.O. Schmidt (UM Greifswald)
	T2.1 “Data management and publication policies”	B. Lindstädt (ZB MED), A. Polly (MRI)
	T2.2 “Data and metadata standards and integration”	M. Golebiewski (HITS)
	T2.3 “Data quality and data provenance”	C.O. Schmidt (UM Greifswald)
	T2.4 “Standardisation of health data access and interoperability”	S. Thun (Charité/BIH)
TA3 “Services”		M. Löffler (U Leipzig), K. Förstner (ZB MED)
	T3.1 “Central search hub”	K. Förstner (ZB MED)
	T3.2 “Terminology and metadata annotation services”	J. Fluck (ZB MED), D. Waltemath (UM Greifswald)
	T3.3 “Data quality and provenance services”	C.O. Schmidt (UM Greifswald)
	T3.4 “Central data access point (CAP)”	U. Sax (UM Göttingen), T. Kirsten (UAS Mittweida)
	T3.5 “Data repository and archiving services / local data access points (LAPs)”	M. Löffler (U Leipzig), F. Meineke (U Leipzig)
	T3.6 “Software repository”	F. Meineke (U Leipzig)
	T3.7 “Distributed data analysis infrastructure”	M. Schulze (DIfE), O. Beyan (Fraunhofer)
TA4 “Community & Networking”		H. Neuhausen (U Cologne), H. Zeeb (BIPS)
	T4.1 “Sustainable organisation for outreach interaction and dissemination to the community”	T. Muth (RKI), A. Pollex-Krüger (TMF)
	T4.2 “FAIR data sharing – community aspects”	M. Golebiewski (HITS), W. Müller (HITS)
	T4.3 “Training and education”	J. Dierkes (U Cologne), S. Klammt (KKSN)
	T4.4 “Networking with NFDI and beyond”	I. Pigeot (BIPS), J. Fluck (ZB MED)
	T4.5 “Citizen and patient involvement”	H. Zeeb (BIPS)
	T4.6 “Service evaluation and user research”	R. Depping (U Cologne)

Task Area	Measures	Responsible Co-Spokesperson(s)
TA5 “Use Cases”		I. Pigeot (BIPS), M. Schulze (DIfE)
	T5.1 “Use case ‘Nutritional epidemiology’”	M. Schulze (DIfE), U. Nöthlings (U Bonn)
	T5.2 “Use case ‘Epidemiology of chronic diseases’”	T. Pischon (MDC), H. Zeeb (BIPS)
	T5.3 “Use case ‘Secondary data and record linkage’”	W. Ahrens (BIPS), I. Pigeot (BIPS)
	T5.4 “Use case ‘Clinical trials’”	O. Brosteanu (U Leipzig), M. Löbe (U Leipzig)
	T5.5 “Use case ‘Surveillance’”	T. Muth (RKI)
	T5.6 “Use case ‘Radiomics / imaging AI’”	H. Hahn (Fraunhofer)
TA6 “Privacy & Data Access in Concert”		U. Sax (UM Göttingen), B. Buchner (U Bremen)
	T6.1 “Generic concept and services for data sharing and data protection”	U. Sax (UM Göttingen), H. Kusch (UM Göttingen)
	T6.2 “Legal framework, esp. common consent standards and data access procedures”	B. Buchner (U Bremen)
	T6.3 “Development of concepts and methods for privacy risk assessment”	F. Prasser (Charité/BIH)
	T6.4 “Synthetic health data”	H. Fröhlich (Fraunhofer)
	T6.5 “Best practice of record/data linkage with regard to data privacy and data protection requirements”	W. Ahrens (BIPS), S.C. Semler (TMF)

4.2 Task Area 1 “Coordination” (lead: ZB MED, BIPS)

Key activities of TA1 concern the establishment of functional bodies and of the project governance. The interaction between partners and functional bodies will be facilitated via an internal web platform (extranet) for documents, meetings and telephone/video conferences. Regular meetings of the Steering Committee (SC) in combination with close monitoring of work progress (defined by milestones and deliverables) will enable early identification of bottlenecks and initiation of necessary adaptations to ensure a smooth workflow. The administrative management is responsible for the preparation of the periodic activity report and the summary financial report including justification of resources and amendments. Eventually, TA1 will deliver a sustainability plan for the accomplishment of FAIR principles in health data management systems (**D1.10, M60**). Any changes to the work plan or to the financial plan, subject to approval by DFG, can only be made by the General Assembly. Legal issues will be settled in close contact with the Directorate.

Involved (Co-)Applicant Institutions in Task Area 1				
ZB MED	BIPS	Charité/BIH	DIfE	Fraunhofer
HITS	KKSN	MDC	MRI	RKI
TMF	U Bonn	U Bremen	U Cologne	U Leipzig
UAS Mittweida	UM Göttingen	UM Greifswald		

T1.1 "Project governance" (**ZB MED (lead), all co-applicants**; participants represented in UAB)

Based on the consortium agreement (**D1.1, M1**) and the grant notice, the spokesperson will carry out the **overall project management** and provide the periodic reports and documents to the DFG with support by the SC. The spokesperson will establish a central point of contact for the **consultation of partners in legal and contractual issues** and **the protection of intellectual property rights**. If necessary, external legal advice will be sought. Proposals for decisions by the General Assembly (GA) will be prepared.

To ensure transparency, ZB MED will communicate all plans, activities, rules and decisions to all members of the consortium and will organise the communication with the advisory boards (UAB, SAB). ZB MED will report all relevant outcomes to the DFG/NFDI committees. This will be achieved **(a)** by **meetings** of the **project's bodies** (kick-off meeting (**D1.3, M3**), half-yearly face-to-face meetings of the SC, annual meetings of the GA (**D1.5, annually**), meetings with the UAB and SAB), regular telephone/video conferences and on demand communication for conflict management which will all be planned, prepared and organised by ZB MED. In addition, **(b) a closed-access web platform (extranet) for internal communication**, i.e. a consortium's web community and tools for distance collaboration, will be established to facilitate cooperation and information exchange and save travel costs. All relevant documents will be made available through this extranet (consortium agreement, deliverables, rules, publication plans, presentations, minutes of meetings, standard operating procedures (SOPs), tools, templates, press releases, announcements, etc.). The project management structure is described in Section 2.5.

T1.2 "Project financial controlling" (**ZB MED (lead), all co-applicants**)

ZB MED will set up the project financial controlling (**D1.2, M2**) and further serve as the central contact point to provide **consultation in financial matters.** It will prepare **management level justification of resources deployed** and the summary financial report. The spokesperson is responsible for the financial reporting towards the DFG. Necessary revisions of the financial

plan will be prepared for decision by the GA. The spokesperson will transfer the financial resources received from the DFG to the project partners based on their cost statements. In addition, it will organise financial planning and all transactions for the internal NFDI4Health project calls.

T1.3 "Communication and public relations" (**BIPS (lead), all (co-)applicants**; all participants)

NFDI4Health will develop and deliver strategic, consistent and measurable communications plans and actions to raise awareness, establish the project's reputation and drive engagement with all stakeholder groups and the media to foster utilisation of the services and strategies developed (**D1.7, M58**). Communication about NFDI4Health and its results will address three target groups: (1) key opinion leaders and researchers in the scientific community as they need to use the standards and services developed by NFDI4Health; (2) political decision makers as they need to revise the legal and regulatory framework for data sharing and record linkage (see TA 1.4); (3) the general public as to promote understanding of the benefits of data sharing/reuse as well as to strengthen their trust in a high level of data protection while handling their person-related data for research purposes.

Stakeholders shall be able to access reliable information about NFDI4Health rapidly through pre-established channels. Therefore, the project website will provide updates and briefing material suitable for use by all stakeholders. The website will include a description of the project and the partners, a news and events page, links to related projects (e.g., other NFDI consortia) and contact and media details for the project. Overall, web material will be aimed at the general public, with additional pages for specialist audiences (clinicians, researchers, other projects, potential collaborators) that will also enable community building – this will happen in close collaboration with TA4. A logo will be designed and used on the website and all other publications of NFDI4Health to increase its recognisability (**D1.4, M3**).

Furthermore, the communications and public relations team will manage social media, write newsletters for specific audiences, develop a blogging community, as well as organise hands-on workshops, collaborate on cross-project concertation events and create communication and dissemination assets such as videos, animations, leaflets or posters to attract our user community, a range of stakeholders, data holders and end-users from the public health, academic and industrial sectors. Additionally, press releases will be issued over the course of the project. Existing communication channels of the NFDI4Health consortium will be used for communications purposes. Emphasis will be put on establishing sustainable cooperation between the public relations officers of the institutions involved in NFDI4Health and the media.

This will focus on measures to improve public confidence in the newly developed standard services, by transparent communication of potential risks and opportunities.

T1.4 "Outreach to political decision makers" (**RKI (lead), U Cologne, all (co-)applicants**)

The rationale behind this measure is that specific issues (e.g., concerning data protection in the context of data sharing) arising during NFDI4Health data infrastructure building may not be solvable by the (co-)applicants, because regulatory adjustments need to be settled on a legislative level. For this purpose, the German Council for Scientific Information Infrastructures (RfII) will be approached that consists of representatives from scientific institutions, information facilities, as well as federal and state governments. The Council is in charge of providing suggestions for NFDI development and its further integration into European and international scientific systems. The stakeholders and decision makers will be informed about the progress and potential issues by annual summary reports. In addition, representatives from the Council and selected experts and representatives from the Federal Ministry of Health (BMG), the Federal Ministry of Food and Agriculture (BMEL), the Federal Ministry of Education and Research (BMBF), the German Joint Science Conference (GWK), the Alliance of Science Organisations in Germany, the German Association of University Hospitals (VUD) and representatives of epidemiological and medical societies will be invited to dedicated meetings, user community workshops (see T4.1) or NFDI4Health symposia (**D1.6, M30)**. A public event presenting the key achievements of NFDI4Health will be organised at the end of the first funding phase (**D1.9, M60**). These liaison meetings will fulfill three purposes: (i) to convey the actual project progress, (ii) to highlight and discuss issues encountered and (iii) to formulate demands for the establishment of an efficient NFDI in the health arena.

T1.5 "Business plan and sustainability" (**ZB MED (lead), U Leipzig (co-lead), BIPS, KKSN, U Cologne, UM Göttingen, UM Greifswald**)

NFDI4Health is a consortium of major players in public health, epidemiology and clinical trial research. Most of the (co-)applicant institutions are able to sustain NFDI infrastructures beyond the project duration. For instance the ZB MED and the Leibniz-institutes (BIPS, DIfE) may allocate institutional resources for maintenance activities. Clinical trials coordinating centres, established at many universities over the last 20 years, have proven to be stable research service infrastructures offering a growing wealth of research data. Data holders of large epidemiological cohort studies are expected to guarantee provision of epidemiological data for many years to come since these studies are designed as long-term projects running over several decades. RKI and MRI are government's central scientific institutions responsible for

regular surveillance and health reporting to safeguard public health in Germany. According to their mission they allocate institutional resources to the sustainable provision of their study data to the research community.

NFDI4Health will derive a prototype of a business model with two closely related major components. First, NFDI4Health will explore options to attract data holders to use central resources to set up their metadata and data sets according to NFDI4Health standards and to rely on central resources to guarantee sustainable FAIRness. Second, the business model will strive for an upscalable provision of central services based on NFDI4Health developments (e.g., central search hub, central and local data access points, metadata and data quality standards, distributed data analysis, use and access brokerage as well as related trainings). These approaches include (1) reimbursements for the upload of standardised data provided to data holders; the corresponding budget is requested as part of the funding for future project tasks. The amount reimbursed will depend on the complexity of the respective data set to be standardised and uploaded. NFDI4Health will develop a cost model to calculate the reimbursement as a function of size and complexity of data items, sample size, data quality, as well as scope of requested services. Also, the reimbursement of providing extensive services to data holders has to be considered. (2) In the long run, the resources needed to reimburse data holders for their efforts to make their data FAIR may become a cost item of project grants by DFG, BMBF and others, of course conditional on the obligation to use the NFDI services. (3) As an alternative option, data holders may also charge fees for data use and access as, e.g., currently done by the German cancer registries where expenses for data provision have to be reimbursed by data users. We will explore the acceptance of different options in tight feedback with the user community while also considering corresponding practices in other countries to propose a prototypic business model. Ultimately, we may explore the feasibility of our vision for a long-term archive as core of an NFDI that is sustainably funded by the federal government and/or by the federal states (**D1.8, M58**).

Role of members and participants: ZB MED leads T1.1 and T1.2 and supervises the project office team. It is responsible for money transfer and project controlling. BIPS will support ZB MED in coordination and leads the measure dissemination and public relations (T1.3). RKI leads T1.4 and coordinates the outreach to political decision makers. All (co-)applicants will be involved in decision making as voting members of the decision-making GA. TA leaders BIPS, DIfE, HITS, U Bremen, UM Göttingen, UM Greifswald, U Cologne, U Leipzig and ZB MED will contribute to the project management as representatives of the SC.

Cooperation with other task areas: All task area leads are members of the SC, report directly to the project office and support the spokesperson as described in Section 2.5. For the communication with the community and networking with other NFDI and beyond, TA1 will closely cooperate with TA4.

Risks of implementation: (1) Failure of a partner regarding deliverables, i.e. non-provision, insufficiency or delay, may be compensated by other partners. If one partner is, e.g., unable to deliver, re-allocation of resources will enable other partners to compensate. (2) The risk of a delayed approval of procedures and policies will be minimised by the fact that all partners are bound by the consortium agreement that prescribes necessary actions in case of non-compliance. (3) To mitigate the risk of delayed approval of procedures by the local ethical review committees, TA6 will prepare risk assessment documentation for all automatisation steps, including standard templates, and organise workshops/discussion fora. (4) To prevent a partner from failing to deliver a cost statement, i.e. non-provision, insufficiency or delay, ZB MED will be in close contact with partners' administrations to support the preparation and submission of cost statements and will prepare templates in accordance with DFG funding guidelines. (5) Conflicts regarding budgets will be resolved by direct communication between the spokesperson and concerned partner(s). Any necessary changes to the financial plan will be prepared by the spokesperson for decision by the GA.

Table 2: Deliverables of Task Area 1

No.	Type*	Description (month; measure)
D1.1	DOC	Consortium agreements are signed (M1; T1.1)
D1.2	SVC	Financing controlling and money transfer is established (M2; T1.2)
D1.3	EV	Governance structure is established and first GA has taken place (M3; T1.1)
D1.4	DISS	Corporate design (M3; T1.3)
D1.5	EV	GA with NFDI4Health status report (M12; annually; T1.1)
D1.6	EV	Midterm outreach to decision makers by invitation to meetings (M30; T1.4)
D1.7	DOC	Report on all PR activities (M58; T1.3)
D1.8	DOC	Prototype of a business model (M58; T1.5)
D1.9	EV	Public presentation of key achievements of NFDI4Health (M60; T1.4)
D1.10	DOC	Sustainability plan for the accomplishment of FAIR principles in health data management systems (M60; TA1)

* DOC: document, INT: interface definitions, APIs, plan designs, architectures, SVC: service, SW: software, EV: event, DISS: dissemination

4.3 Task Area 2 “Standards for FAIR Data” (lead: HITS, UM Greifswald)

TA2 targets core deficits in medical sciences as already outlined in Section 3 on research data management, i.e. the lack of harmonised standards for data and data quality management in clinical trials, public health surveys, and epidemiological cohorts, as well as the lack of information on and access to relevant standards. By making standards available, TA2 will improve the findability, accessibility and interoperability of existing and novel data bodies. For this purpose, guidelines, standards and policies on data management and publication (T2.1), data and metadata standards and integration (T2.2), data quality and provenance (T2.3), health data access and interoperability (T2.4) will be elaborated and disseminated specifically for the use cases (TA5), with the involvement of the user community (TA4). Thus, TA2 will provide the conceptual background for NFDI4Health services to be implemented in TA3.

To make person-related health data findable, accessible, interoperable and reusable (FAIR), TA2 will build on existing domain-specific standards, guidelines and corresponding resources developed and published by well-established standardisation initiatives and specific committees of ISO, CEN and DIN. Primarily, these standards will be adapted and customised to domain-specific requirements of the use cases and user communities, to be bundled into a tailored standardisation toolkit for research data in the public health domain.

To ensure interoperability of heterogeneous and complex data, TA2 will also liaise with pan-European activities aiming at harmonising health-related data standards and defining guidelines for implementing data interoperability across subdomains and across borders, such as the H2020 project EU-STANDS4PM[8] or eStandards[57], as well as euCanSHare[58], a joint EU-Canada project to foster cross-border data sharing. Such initiatives support the development of domain-specific recommendation guidelines, in close collaboration with the European Commission. This will help to prepare European guidelines that also consider the needs of the NFDI4Health user communities.

Nationally, TA2 will facilitate the harmonisation of data standards and data interoperability across NFDI consortia (including NFDI4MED and GHGA) from the NFDI4Health perspective as well as with domain-specific initiatives and networks developing standards and guidelines. This comprises the close collaboration with the MII working groups on interoperability and data sharing, as well as with the project group on FAIR data infrastructures of the GMDS. For this purpose, domain-specific cross-consortia working groups will be established.

To maximise the usability of work results, an agile interactive process involving all (co-) applicants and the user communities will be initiated in all TA2 measures as follows: (1) Based on existing standards and guidelines, an initial set of guidance documents and recommended standards will be issued during the first year (**D2.1**, **D2.2**, **D2.3**, **D2.4**, **D2.5**, **D2.6**) to be used as reference for service developments in TA3 and to guide the use cases in TA5. (2) Experiences of TA3 and TA5 as well as feedback from the user community (TA4) will guide the incremental revision of the documents, supported by standardisation initiatives and organisations. This will be realised through web-based feedback surveys and joint workshops with TA4. (3) The refined guidelines, policies and standard recommendations will be released during the third and fourth project year (**D2.1**, **D2.4**, **D2.5**, **D2.7**, **D2.8**, **D2.9**, **D2.10**), disseminated to the user community (TA4), and implemented by the use cases (TA5). Finally, by combining all the guidelines, policies and standard recommendations, an integrated NFDI4Health standardisation and quality recommendation document will be published in month 58 (**D2.11, M58**).

TA2 will address the Key Objectives (1), (3), (5) and (6) listed in Section 2.1 of this proposal. In addition, the standardisation efforts and harmonisation of standards addressed in TA2 will provide a basis for Key Objectives (2) and (4). The deliverables of TA2 are listed in Table 4.

T2.1 "Data management and publication policies" (**ZB MED (lead), MRI (co-lead), BIPS, Charité/BIH, DIfE, HITS, RKI, U Cologne, UM Greifswald**)

T2.1 aims for policies for data management and publication in order to make data findable and interoperable. To find information about studies and (meta-)data and to ensure their interoperability, it is necessary to document the descriptive core elements in a structured way already when planning projects. This applies both to data management and to the subsequent publication of research results and data and is particularly important for research projects handling person-related data.

In the past, research and infrastructure projects already defined domain-specific data management policies for health (meta-)data defined for structured data capture and data safety, e.g. the data management checklist developed by FAIRDOM[59], that defines guidelines and rules for the handling of research data (including sensitive patient-related data), or the data use and access policy defined by the MII, valid for all 33 university hospitals involved[60].

T2.1 will lay the foundation for a structured collection of all (meta-)data items (e.g. study design, survey instruments, methods, questionnaires, data sets, variables, concepts and publications).

Therefore, policies for publication of (meta-)data encompassing the use of metadata standards, terminologies and ontologies, file types, licenses for second use, language and data versioning will be developed. The standardisation results of T2.2 - T2.4 and a use case-driven definition of minimal common datasets for standardisation and publication will constitute the foundation for the guidelines. In addition, guidelines for the registration of persistent identifiers (PID) for data, metadata and documents will be incorporated, encompassing PIDs for unpublished data. Another important point are guidelines and recommendations for the use of publication infrastructures (e.g. repositories) developed in NFDI4Health in close cooperation with TA3.

Agreeing on common guidelines and policies is the first step to reach this aim. The work will be performed following the interactive process outlined above: First initial data publication guidelines will be developed by month 13 (**D2.6, M13**) and evaluated with the community to receive early feedback and to ensure that the guidelines lead to interoperability. Exemplarily, the guidelines will be implemented first by two use cases (T5.1 and T5.4). Finally, elaborated policies will be created and continuously updated following the requirements of legal regulations (TA6) and of the user community (TA4), to be tested by all use cases (**D2.9, M48**), and incorporated in a NFDI4Health standardisation and quality recommendation document (**D2.11, M58**).

T2.2 "Data and metadata standards and integration" (**HITS (lead), Charité/BIH, DIfE, Fraunhofer MEVIS, RKI, U Leipzig, UM Greifswald, ZB MED**; participants: DIMDI, DIN, GMDS; additional partners: CDISC, COMBINE, EU-STANDS4PM, FAIR4Health, FAIRDOM, FAIRsharing, GA4GH, HL7)

To achieve data "FAIRification by standardisation" and enable the user community to integrate heterogeneous and complex data, recommendations and guidelines will be developed for the consistent use of domain-specific standards for data formats, as well as for consistent data descriptions based on established metadata standards and terminologies. This standardisation concept will be based on existing standards, such as ISO 20691[41] (ISO/TC 276[7]) and will include the definition of a minimal metadata set for discovering, specifying and assessing data in the NFDI4Health services (implemented in TA3).

To develop a data standardisation roadmap (**D2.2, M12**) for clinical trial, public health and epidemiological data, domain-specific data standardisation requirements derived from the use cases (TA5) will be aligned with existing data and metadata standards at the beginning of the project. To fill the gaps identified as well as to adapt and customise existing standards for the user communities and use cases of NFDI4Health, co-applicants will intensify their engagement in international standardisation organisations (ISO/TC 276, ISO/TC 215 and CEN/TC 251),

nationally supported in Germany by DIN, and in standardisation initiatives (e.g. HL7, CDISC, SNOMED, GA4GH, openEHR, COMBINE[10], DICOM, etc.). Relevant standards of these and other standardisation efforts will be bundled and made available to the user community in a FAIRsharing[25] collection (T4.2) for clinical and health research (**D2.7, M30**).

To create an NFDI4Health core metadata set, a synopsis of overarching metadata elements in these domain-specific standards will be elaborated and aligned with the requirements of the user communities in the use cases (TA5), taking account of the core data set specifications of the MII and feedback from the community (TA4), following the approach outlined above. This will be used to formulate a generalised set of common metadata in the served communities (**D2.10, M48**), as preparation of a joint metadata repository (MDR) to bundle and host data elements that will be collected by our user communities. This may include the following parameters (Table 3):

Table 3: Overview of building blocks of the NFDI4Health minimal metadata set

Metadata describing…	Means
the project/study identifier of the data source (provenance)	e.g. WHO universal trial number (UTN)[61]
quality and provenance of the data, purpose, audit trail	HL7 FHIR provenance resource
consent information concerning each data element	Date and check for revocation
study design	Practices and processes, SOPs
observed parameters	SOPs, units, standard range
vocabulary and classification used for the dataset and its items (observations and measurements)	Domain-specific taxonomies, ontologies, ISO 20691, core dataset of MII
temporal resolution of the data	Check time stamp synchronisation
data cleaning and data integration	R-scripts
results of analyses, publications	DOI, PubMed ID

By the above activities, T2.2 will promote and coordinate further harmonisation of data standards for clinical and health research, based on findings from the use cases (TA5), as prerequisite for a “FAIRification by standardisation” guideline (TA4). This will be accomplished by the definition of recommendations and standards related to nutritional epidemiology (T5.1), chronic diseases (T5.2) and clinical trials (T5.4), as well as by their adaptation to meet the challenges for the creation of metadata of secondary data sources (T5.3). These activities will be merged with ongoing standardisation efforts regarding overarching metadata standards and recommendations of data format and metadata use, e.g. in ISO 20691[41].

All recommendations and documents will be bundled for the preparation of guidelines for FAIR data sharing and integration (T4.2) and for the data publication guidelines (T2.1), as well as to prepare a final NFDI4Health standardisation and quality recommendation document (**D2.11, M58**) that will also include guidelines and recommendations from T2.1, T2.3 and T2.4. This will be developed in close exchange with FAIR data standards initiatives (GO FAIR, FAIR4Health DataCite, ORCID, Research Data Alliance (RDA), FAIRDOM, EOSC Life etc.) and expert groups (e.g. the GMDS project group *FAIR data management*). T2.2 will consult TA6 on all data privacy and data protection issues.

T2.3 "Data quality and data provenance" (**UM Greifswald (lead), HITS, KKSN, U Leipzig, UM Göttingen**; participants: DIN, GMDS; additional partners: CDISC, COMBINE, FAIR4Health, GA4GH, EU-STANDS4PM, HL7, MAELSTROEM)

Following ISO 8000 T2.3 will target data quality as the "degree to which a set of inherent characteristics of data fulfils requirements" and provide consented standards and metrics to assess the data quality at different stages of the scientific data lifecycle. T2.3 will first consider FAIR standards in collaboration with the FAIRMetrics group[62], FAIRsharing[25], and RDA FAIR Data Maturity Model group[49]. If necessary, adaptions will be made to suit data bodies of relevance within NFDI4Health (**D2.1, M9, M42**). The second focus will be on adherence to defined data and metadata standards as recommended by T2.2 and as developed in several initiatives (e.g. HL7, CDISC, SNOMED CT, LOINC, openEHR, EU-STANDS4PM[8], COMBINE[10], MAELSTROEM, MII, etc.) (**D2.4, M12, M36**). The first and second topic share the focus on formal aspects of data, data accessibility, and processability. The third focus will be on the adequate representation of metadata related to provenance (**D2.4, M12, M36**) – both of the data sources and of the data workflow (extraction, integration etc.). The PROV-DM model, developed by the World Wide Web Consortium (W3C), can serve as a basis for the storage of domain-agnostic provenance information[63]. HL7 FHIR implements provenance resources based on the PROV-DM model[64] tailored to the medical research domain. Both W3C PROV-DM and HL7 FHIR allow for a modern clinical data management infrastructure, which will be directly implemented in the NFDI4Health test bed (T3.3). This work will be complemented by factoring in rules and recommendations from ISO 23494, a series of standards for provenance information management in the life science domain, that is currently developed by ISO/TC 276/WG5.

The fourth focus will be on data quality in the measured clinical data itself to provide a solid basis for data management, data harmonisation, and subsequent scientific analyses (**D2.5, M12, M42**). References are data quality frameworks and guidelines for primary as well as secondary data collections and community-based projects and initiatives to define common data

quality standards, e.g. within the DFG-funded project on standards and tools for data monitoring in complex epidemiological studies[33].

The documents produced in T2.3, together with those from T2.2, will provide a *conceptual* (meta-)data standardisation background for services in TA3, for the alignment with other FAIR data initiatives (T4.2) and for the use cases in TA5 and will finally be incorporated in a NFDI4Health standardisation and quality recommendation document (**D2.11; M58**).

T2.4 "Standardisation of health data access and interoperability" (**Charité/BIH (lead), DIfE, Fraunhofer MEVIS, HITS, MDC, RKI, TMF, U Leipzig, UM Göttingen, UM Greifswald, ZB MED**; participants: DIMDI, DIN, GMDS; additional partners: CDISC, EU-STANDS4PM, GA4GH, HL7)

T2.4 will define standardisation requirements and develop guidelines, as well as standard-based solutions for data access and interoperability in the defined use cases. To ensure compatibility with existing efforts aiming to improve data interoperability in medicine and healthcare, T2.4 also will coordinate its work closely with the same standardisation initiatives and technical committees of standardisation organisations as T2.2.

To enable a seamless access and exchange of health data within the consortium, T2.4 will focus on (1) the development of an interoperability roadmap that identifies and defines interoperability requirements relevant to the use cases (coordinated with T2.2) within the first year (**D2.3, M12, M24**) that will be revised and adapted to the use cases (TA5) and user communities (TA4) of NFDI4Health in the second year. (2) To achieve semantic interoperability, relevant standards, terminologies and ontologies will be identified and applied to the use cases (this includes the development of use case-specific value sets based on international terminologies, such as SNOMED CT or LOINC). (3) Guidelines and implementation guides will be specified that define information models and core datasets for the use cases within the first 36 months (**D2.8, M36**). These models will provide a data interface to connect different systems based on jointly used data and metadata standards. A first preliminary consented cohort data model (**M2.1**) will be developed within the first year and incrementally refined in later stages of the project, both with support from T2.2 and T2.3. Where possible, guidelines will take into account existing common data models such as the International Patient Summary (IPS) to ensure international interoperability. Work in this area will rely on tools used in the international community (e.g., as ART-DECOR[65] for defining data models as well as Forge and Simplifier[66] for defining and sharing HL7 FHIR profiles).

Finally, the recommendations of standards and developed guidelines for data interoperability from T2.4 will be merged with results from T2.1, T2.2 and T2.3 into an integrated NFDI4Health standardisation and quality recommendation document (**D2.11, M58**).

Role of members and participants: Task area leaders (HITS, UM Greifswald) will coordinate all measures in TA2. ZB MED and MRI will coordinate data management and publication policies in NFDI4Health. HITS will serve as contact to standardisation and FAIR data standards initiatives, as well as to ISO/CEN/DIN committees (together with BIH) and will coordinate data standard harmonisation and preparation of standard recommendations. UM Greifswald will coordinate all activities related to data quality and provenance standards, including a FAIR metrics and the contact to relevant initiatives. BIH will coordinate standardisation for systems interoperability and recommendations across the subdomains. All (co-)applicants will support the data standardisation activities, contribute to the drafting of deliverables, recommendations and standards, as well as support the gathering of requirements from the user community. All additional participants and partners will contribute to the respective measures and help to produce the deliverables.

Cooperation with other task areas: As TA2 elaborates on a conceptual standardisation background and defines guidelines relevant for the whole NFDI4Health consortium, it cooperates basically with all other TAs: (1) Use cases (TA5) and user community (TA4) will be analysed for their data standardisation requirements; (2) experiences from service implementations (TA3), feedback from the user communities (TA4) and from the use cases (TA5) are used to define gaps in existing guidance documents and in the available landscape of domain-specific standards; (3) guidelines and recommended standards will serve as a basis for services (TA3), for training and teaching (TA4), and for the implementation in use cases (TA5). TA2 will accompany all TA5 use cases to harvest feedback and further elaborate the development of guidelines and standards; (4) TA2 will closely collaborate with TA6 on all aspects of data privacy and data protection.

Risks of implementation: T2.2 and T2.4 depend on collaboration with external partners (standardisation initiatives and organisations) and therefore carry the potential risk of possible divergent interests. The leading roles of HITS and BIH in key committees and initiatives, however, minimises this risk. Potential gaps in the translation from FAIR principles to practical and tangible solutions can be minimised by collaboration with FAIR initiatives. The possible lack of acceptance of guidelines and standards in the user community will be addressed by constant alignment with the requirements from the use cases (with TA5) and recurrent user involvement

(with TA4). Possible ethical, legal, data security and data protection issues will be tackled together with TA6.

Table 4: Deliverables of Task Area 2

No.	Type*	Description (month; measure)
D2.1	DOC	Guidance documents on FAIR metrics (initial M9, revised M42; T2.3)
D2.2	DOC	Standardisation and interoperability roadmap (M12; T2.2)
D2.3	DOC	Identification of relevant data formats, standards and terminologies for interoperability (initial M12, revised M24; T2.4)
D2.4	DOC	Guidance documents on adherence to defined data and metadata standards and data provenance (initial M12, revised M36; T2.3)
D2.5	DOC	Guidance documents on data quality assessment and results handling (initial M12, revised M42, T2.3)
D2.6	DOC	First initial guidelines for the publication of health research data (M13; T2.1)
D2.7	SVC	Collection of recommended standards prepared for a domain-specific FAIRsharing collection (implemented in TA4.2) of data standards and policies (M30; T2.2)
D2.8	INT	Implementation guides (IG) and value sets for core data sets and use cases (M36; T2.4)
D2.9	DOC	Revised guidelines for the publication of health research data (M48; T2.1)
D2.10	INT	Generalised set of common metadata to prepare for an MDR (M48; T2.2)
D2.11	DOC	NFDI4Health standardisation and quality recommendation document (M58; T2.2, with contributions of T2.1, T2.3 and T2.4)

* DOC: document, INT: interface definitions, APIs, plan designs, architectures, SVC: service, SW: software, EV: event, DISS: dissemination

4.4 Task Area 3 “Services” (lead: U Leipzig, ZB MED)

TA3 focusses on the services, service enabling tools, and software that NFDI4Health will provide to the user community. Most services and tools will be based on open source software that has already been developed by the (co-)applicants or by the broader scientific developer community. In close cooperation with TA4 and TA5, use case requirements and community feedback will help to further develop these tools and to foster interoperability of currently fragmented IT solutions for storage of metadata, cohort browsing, assessment of data quality and data harmonisation. Services and tools that will be developed or advanced during the project will be published as open source software. Service developments will be based on specifications, processes and guidelines established in TA2. NFDI4Health services will target two main user groups: (A) data analysts and (B) data holders. An overview of their requirements and of the minimum services to serve their needs is presented in Section 3.3 (see also Figure 6). These correspond to the following services to be developed by TA3:

The central search hub will be developed in T3.1, services for terminology and metadata annotation and services for data quality and provenance will be provided by T3.2 and T3.3 respectively. These services will support data holders in their obligation to make their data collections FAIR. The services will be employed for harmonisation and quality enhancement of already existing data bodies as well as data bodies to be newly collected. They will also support the annotation of metadata. T3.4 will set up a central data access point (CAP) as an entry point to all data available in NFDI4Health, where data analysts will submit their applications for data access. Each application will be pre-checked and, if positive, forwarded to the respective data holding organisation via the local data access point (LAP) where the conditions for data access and analysis will be settled. T3.5 will develop a software solution to set up LAPs with a uniform interface by the data holders in NFDI4Health (e.g., U Leipzig, UM Göttingen, DIfE, BIPS, RKI, UM Greifswald). From year three on, the roll-out will be extended to further data holders such as clinical trial and epidemiological study centres. The provided IT tools also require structured maintenance which will be supported through a centrally managed software repository (T3.6). In T3.7, two different distributed data analysis solutions will be applied to several use cases (T5.1, T5.2, T5.4, T5.5, T5.6) and functionalities will be extended accordingly.

TA3 will seek close collaboration with NFDI4MED and GHGA for the development of complementary DataSHIELD extension packages and respective training modules as well as for the set-up of the CAP. Overall, TA3 will contribute to the Key Objectives (1) to (4) listed in Section 2.1 of this proposal.

T3.1 “Central search hub” (**ZB MED (lead), HITS, U Leipzig, UM Göttingen, UM Greifswald**)

This measure will develop the NFDI4Health central search service. Its core functionalities will (1) allow the search for data bodies in health research, e.g. cohort studies, clinical trials; (2) enable the search for data elements within these data bodies.

The hub will contain only metadata and selected data summaries with unrestricted access. In addition, data use and access information is provided and the data themselves are referenced with persistent identifiers. Users can retain their search result for their data access application at the CAP or at the LAPs. The search functionalities developed in T3.1 will build upon the policies and standards for publication of (meta-)data for targeted data bodies and study types developed in T2.1.

The development of the hub will be guided by community feedback while using common tools and platforms of the data sharing community. Our initial approach will be based on previous experiences with the SEEK platform[6], developed by HITS (with partners in the UK) within the FAIRDOM project and adapted further by U Leipzig and UM Göttingen. Wherever applicable, available analysis tools (e.g. i2b2, tranSMART) will be integrated for search and visualisation. Furthermore, there will be a collaboration with euCanSHare[58] and Maelstroem[67] to harmonise the NFDI4Health implementation of the central search hub with the existing Mica metadata portal[68] that offers elaborated search functionalities to gain an overview on available studies through a study description module and accessible data elements. For a description of the software listed above we refer to Section 3.3.

We will establish a first version of the search hub, a so-called MVP (minimum viable product), at an early stage (**D3.3, M18**) and will collect user feedback for further refinements. The search hub will be publicly accessible in month 30 (**D3.8**) and will be updated regularly until the final release at the end of the project (**D3.13, M58**). Because SEEK and Mica have been developed by different scientific communities, their functionality will be assessed across different data bodies to ensure that the final implementation will serve the demands of the diverse user communities.

T3.2 "Terminology and metadata annotation services" (**ZB MED (lead), UM Greifswald (co-lead)**)

For semantic search as well as for metadata annotation, (1) a domain-specific lookup service providing access to the NFDI4Health terminology resources (as agreed in T2.2) and (2) a metadata annotation service will be set up. First, the terminology service will be based on the ELIXIR Ontology Lookup Service (OLS) and customised for NFDI4Health. For customisation, only domain-specific terminologies, specified in T2.2, will be included to ensure that all necessary metadata are available for later standardisation, search and documentation. Major terminologies, such as ICD or LOINC, will be transformed into OLS-compatible formats and included in the lookup service. We will build on experience made in the BMBF-funded project IDSN[69] that has developed software extensions of the original OLS and that will provide a preliminary health terminology set (publication incl. software code and a public web service in preparation). The extended OLS will be integrated as a widget into different web services such as the metadata annotation service, the central search hub (T3.1), the CAP (T3.4) and potentially also the LAPs (T3.5). Second, support for standardised metadata annotation is necessary for the upload of new data sets as well as for the FAIRification of data collections. Two tools, the CEDAR workbench and Rightfield, will be evaluated based on the requirements

of data holders to allow for easy integration of standardised data items within data models and curation of existing data collections.

Within NFDI4Health, a first version of the extended OLS and early demonstrators of the metadata annotation service will be set up after six months to support co-applicants in harmonising their data sets, e.g., in the use cases in TA5. First public NFDI4Health versions will be available after one year (**D3.1, M12**). Regular updates based on standardisation requirements of T2.2 and on user feedback will be provided including a crowdsourcing functionality of the metadata annotation service (final release: **D3.13, M58**).

T3.3 "Data quality and provenance services" (**UM Greifswald (lead), HITS, Fraunhofer MEVIS, UM Göttingen**)

The conceptual basis of data quality assessment services will be provided by T2.3. A core result will be a harmonised data quality assessment workflow which will be based on three steps with increasing complexity. Core functionalities for this purpose will be accessible through central NFDI4Health infrastructures and LAPs which may be also used for data bodies outside the NFDI4Health infrastructures. By this, we will target data bodies which, for data privacy reasons, may not be uploaded to the NFDI4Health LAPs.

First, templates and tools provided by the FAIRMetrics group[62] and RDA FAIR Data Maturity Model group will be adapted to generate a metric for the degree of compliance with FAIR principles. The second step will be the assessment of the quality of metadata. Results from metadata annotation services (T3.2) will be used to improve assessments. The third step will be the standardised assessment of the quality of study data themselves. Information on study data quality will support the decision on an appropriate statistical analysis strategy, in particular, for federated data analyses.

Beyond the provision of software through a central repository this comprises services on their implementation and use. Because of the diversity of data bodies, multiple tools will be needed. These are R data quality libraries on standardised data quality assessments developed in previous projects (e.g., in a recently completed DFG-project[33]) as well as other software (e.g. Square[270], tranSMART[71]). Maximising the interoperability of analysis routines will be key to harmonise the conduct, output, and comparability of data quality assessments. These tools will be applied in the use cases in TA5. A first version will be available in month 24 (**D3.5**), additional functionality will be integrated in regular updates and a final version will be delivered in month 58 (**D3.13)**.

Furthermore, T3.3 will extend SEEK by inclusion of provenance services. T3.3 will assist to implement such services for data holding organisations that do not rely on SEEK. These services will include the HL7 FHIR provenance information as, e.g., the initial cause for data capture, the technical data source and the workflow how data were generated[46,47,72,73,74]. The resulting service will provide: (1) A simple way of augmenting stored data with standardised data provenance and workflow provenance information. (2) Retrieving the provenance trail of any data item stored in SEEK in a standardised format. Initial provenance services will be available in SEEK and tested early in the project (**D3.4, M24**), e.g. by the use cases in TA5. These services will then be refined according to the users' demands and regulatory requirements (**D3.13, M58**).

T3.4 "Central data access point (CAP)" (**UM Göttingen (lead), UAS Mittweida (co-lead), HITS, ZB MED**)

The central data use and access process in NFDI4Health will rely on existing local infrastructures at the data holding organisations (DHOs). We will design, implement and evaluate central services that are needed to (a) manage consortium-wide data access requests, (b) provide a data access broker service and (c) support data access and use within the common NFDI4Health research infrastructure.

Following the framework of data use and access, addressed in TA6 (see Table 9), our service will guide the user through the data application process. We will provide a central data access point (CAP) at an early stage (year two) with basic functionality (**D3.7, M28**) which will be extended over time. Preventing a central data storage, the CAP will forward the data use requests to the data holding organisations, using their LAPs. All connections to the LAPs will be ready at month 48 (**D3.12**). Community feedback as well as requirements of other NFDI consortia will be incorporated during the development of the CAP. To establish the functionality for the CAP, we will develop

- registration and user authentication;
- interface to central search hub;
- standard operating procedures (SOPs) for data use and access requests covering all data collections as well as usage scenarios within NFDI4Health (see TA6) and SOPs for metadata formats (see T2.2);
- simple upload function for data use and access conditions;
- online form for central data requests with select boxes for planned data analysis;
- transfer of data request to local data access points (LAPs);

- feedback on the data request status;
- integration of automatic consent mechanisms;
- forwarding process (when access is granted) to analysis platform of LAPs directly or to the distributed analysis platforms (see T3.7).

In project year 1-4, complex data requests that require the integration of data will directly use the distributed data analysis platforms (see T3.7). In the last project year, the broker will guide the user directly to certified services available for distributed data analysis (**M3.2**).

T3.5 "Data repository and archiving services / local data access points (LAPs)" (**U Leipzig (lead), BIPS, DIfE, HITS, RKI, UM Göttingen, UM Greifswald**)

This measure will develop and provide a reference solution for an LAP for the decentralised provision of annotated data bodies of the respective DHOs. An LAP will contain a local data repository with upload functionality for data bodies prepared according to the specifications in T2.2 and T2.3. It will offer interfaces (specified in T2.4) with the following functionalities: (a) connecting to the metadata annotation workflows provided by (T3.2); (b) metadata publication for the central search hub (T3.1); (c) search options based on the developments in T3.1; (d) data use and access requests via the CAP (T3.4); (e) data interfaces for the connection of distributed computing solutions (T3.7); (f) data analysis platforms; and (g) direct exports within the framework of the data use and access regulation that can be routed via the CAP (T3.4). A first LAP prototype for usability tests will be provided at month 18 (**D3.3**) and the first implementation (U Leipzig) connected to the CAP will be available six months later (**D3.6, M24**). From month 25 on, the LAP solutions will be rolled out and training will start in close coordination with TA4 (**D3.10, M36**).

A preferred solution for the LAPs is based on SEEK, a powerful data sharing web platform. In T3.5 the following adaptations will be performed: (i) the SEEK ISA (Investigation – Study – Assay) data model will be extended for special characteristics of clinical, public health and epidemiological data bodies; (ii) the SEEK API (application programming interface) for entity management will be extended to support more complex search queries such as counts of patient groups as, e.g., needed in trial feasibility queries. Direct visual analytics functions will be offered by a SEEK connected i2b2/tranSMART (**D3.11, M40**). DHOs will obtain the LAP software from the repository (T3.6) and a locally usable data sharing solution with all necessary interfaces to the central services of TA3.

T3.6 “Software repository” (**U Leipzig (lead), ZB MED, Fraunhofer MEVIS**)

All software solutions developed in TA3 will be made available to the user community via a public software repository. This will include appropriate traceability, versioning, change tracking and documentation. The repository will hold pre-configured solutions, pipelines and tools developed in TA3 as well as the software for generating synthetic cohort data (T6.4) and analysis tools necessary for data analysis in the use cases (TA5). A landing page (integrated in the NFDI4Health central services) will guide the user, will give information about available software and links to the NFDI4Health instances of the corresponding community standard solutions (Docker registry, GitLab available by month 3) implemented here as well as to existing external repositories (e.g. Docker Hub, TMF Tool Pool, CRAN for R programs). The archive solution will be set up technically at U Leipzig and a first version including the landing page and corresponding repositories will be internally available at month 15 (**D3.2**). A medium-term transfer to ZB MED for sustainable operation is planned. Once set up, the service will be permanently operated beyond the funding period. For all software uploaded, the landing page will be updated and the documentation will be revised accordingly.

T3.7 “Distributed data analysis infrastructure” (**DIfE (lead), Fraunhofer FIT (co-lead), Fraunhofer MEVIS, MDC, UAS Mittweida**)

This measure will focus on establishing and enabling services within NFDI4Health for distributed data analytics. Personal Health Train (PHT[75]) and DataSHIELD[76,77] are infrastructures supporting inclusion of data from several local data access points into a seamless overarching data analysis in a privacy preserving way. They allow for analysing data without transferring individual data items such that only intermediate, aggregated results are transferred. Both infrastructures have previously been applied in other projects by the co-applicants, e.g., DataSHIELD in the EU projects ENPADASI[28,29] and InterConnect[30] and PHT in the MII consortia SMITH[78] and DIFUTURE[79] as well as in the MII consortia overarching projects CORD and POLAR. However, there is still the need to create secure interfaces for data access at each site and distributed data analysis which becomes obvious in the use cases in TA5. Currently, data are managed by different systems, which use heterogeneous structures and semantics to describe study data. Moreover, statistical modelling capacity of PHT and DataSHIELD is limited, e.g. in terms of modelling longitudinal data or implementing machine learning. T3.7 therefore aims to develop and validate statistical analysis extensions for distributed data analysis and provide central services to set up the DataSHIELD infrastructure to fulfil the specific needs of the use cases in T5.1 and T5.2 that include major epidemiological studies in Germany. Furthermore, we will set up a analysis software repository containing analytics software

containers for use by the PHT and will develop interfaces for the SEEK platform to integrate data centres into PHT distributed analytics services. With PHT, the use case in T5.6 and the generation of synthetic health data (T6.4) will be addressed.

T3.7 will involve the following activities: (1) adaptation/development of statistical modelling features for distributed data analysis based on identified needs of community and use cases (TA5); (2) set-up and maintenance of a central server infrastructure that will connect to the DataSHIELD servers of DHOs involved in TA5 and will serve as LAPs for distributed data analysis, including networking, storage, processing, access control, audit and performance logs, incident management (**D3.9, M36**); (3) set-up of PHT connectors to interact with data and execute analytics tasks; (4) provision of information for set-up and maintenance of infrastructure for use cases, e.g., installation and maintenance manuals, and (5) support for development of necessary training material for distributed data analysis (TA4).

Role of members and participants: ZB MED will lead the development of the central search hub (T3.1) supported by the partners HITS, U Leipzig and UM Greifswald. The terminology service (T3.2) will be adapted and set up by ZB MED, the metadata annotation service will be advanced by UM Greifswald and ZB MED. UM Greifswald will lead the development of data quality assessment (T3.3) with contributions of Fraunhofer concerning image data machine learning models. HITS and UM Göttingen will set up provenance services. The CAP will be developed by UM Göttingen with support of UAS Mittweida, HITS and ZB MED. The latter will host the central service. U Leipzig will be responsible for the data repository, supported by HITS (T3.5) and the software repository, supported by ZB MED (T3.6). DIfE will be responsible for the extension of DataSHIELD and MDC for setting up the central server infrastructure for distributed data analysis, while Fraunhofer together with UAS Mittweida will be responsible for the establishment of the PHT (T3.7).

Cooperation with other task areas: TA3 builds upon standardisation efforts achieved in TA2. Agile development in TA3 will bemental releases of TA2. The services are the main gateway to the users, therefore a close cooperation with TA4 and the provision of early prototypes and feedback harvested in TA4 as well as fast incorporation into the software development is necessary. Similar is true for the services provided to the data holders. They are represented in TA4 as well as in the use cases conducted in TA5. Therefore, close collaboration with TA5 is planned. Furthermore, the CAP is based on the framework provided by TA6.

Risks of implementation: Implementation of NFDI4Health services is largely dependent on progress of TA2 and TA6 to define standards for implementation and setting up a data use and access framework. A delay in these task areas might hamper the service development. Through participation of the co-applicants in TA2 and 6 and preliminary releases by TA2 and 6, the risk will be minimised. Another risk is missing acceptance of our services by the user community. To minimise this risk, we will release early demonstrators and prototypes to test usability of our services by (1) NFDI4Health co-applicants in correspondence to our use cases (TA5) and (2) the wider user community (TA4). Feedback by those will guide our agile development and assure acceptance. Furthermore, there are regulatory, ethical and legal challenges (e.g., various types of informed consent, acknowledgement of intellectual property) to implement the CAP which will be tackled in consultations with TA6. Regarding data element harmonisation, heterogeneity of data might interfere with search functions, metadata annotation and data quality services. Adherence to minimal standards (T2.1 and T2.2) will ensure basic functionality.

Table 5: Deliverables of Task Area 3

No.	Type*	Description (month; measure)
D3.1	SVC	Customised terminology service V1 and annotation workbench V1 are available (M12; T3.2)
D3.2	SVC	NFDI4Health repository landing page, GitLab instance and Docker registry instance set-up and running (M15; T3.6)
D3.3	SW, SVC	First prototypes of central search hub and LAP are available for usability tests (M18; T3.1, T3.5)
D3.4	SVC	Provenance service for prioritised datatypes and pipelines integrated in SEEK (M24; T3.3)
D3.5	SW	Data quality assessment routines with focus metadata for central use from within NFDI4Health resources available (M24; T3.3)
D3.6	SVC	First LAP for clinical trial data upload available (M24, T3.5)
D3.7	SVC	First prototype of CAP is available (M28; T3.4)
D3.8	SVC	NFDI4Health central search hub is accessible (M30; T3.1)
D3.9	SVC	Set-up of DataSHIELD server infrastructure completed (M36; T3.7)
D3.10	DISS	First LAP solution rolled out and trained (M36; T3.5)
D3.11	SVC	Data analysis platform is available at least at one LAP (M40; T3.5)
D3.12	SVC	CAP with connections to the NFDI4Health resources is available (M48; T3.4)
D3.13	SW, SVC	Final releases of community approved services are available (M58; TA3)

* DOC: document, INT: interface definitions, APIs, plan designs, architectures, SVC: service, SW: software, EV: event, DISS: dissemination

4.5 Task Area 4 “Community & Networking” (lead: U Cologne, BIPS)

With its focus on interaction, networking and exchange, TA4 addresses the overall NFDI4Health Key Objective (5), i.e. to support cooperation between clinical research, epidemiological and public health communities. TA4 also contributes to Key Objectives (3) and (4) in that it provides training and education for the health research community and beyond, focusing on FAIR data principles.

Designing NFDI4Health services to the needs of the users is central for the whole initiative. All (co-)applicants are in close contact with users, and all our TAs explicitly connect with the scientific community by involving it in the development and implementation of their services. The strong focus on systematic involvement of the user community supports the consortium on its way to a successful service. Four measures are devoted to the involvement of the user community, namely “Outreach and dissemination” (T4.1), “Data sharing” (T4.2), “Training and education” (T4.3) as well as “User research” (T4.6). TA4 is engaged with two further groups: (a) with the entire interdisciplinary NFDI-community (T4.4), and (b) with citizens and patients, whose continuous active participation (T4.5) is essential in order to increase the public acceptance of sharing sensitive personal data for health research, to improve the understanding of the benefits of data reuse and to enforce the trust into the secure handling of these data.

T4.1 “Sustainable organisation for outreach interaction and dissemination to the community” (**RKI (lead), TMF (co-lead), DIfE, HITS, KKSN, U Cologne, UM Göttingen, UM Greifswald**; participants: all scientific societies)

NFDI4Health will provide a sustainable organisation for outreach interaction with the user communities and stakeholders. We will establish a Community Outreach Committee (COC) at an early stage (**D4.1, M6**), consisting of all (co-)applicants in TA4, with the following objectives: (1) **to coordinate information dissemination about NFDI4Health**, its objectives, methods and services amongst the target audience of research communities, medical societies, infrastructure providers and political decision makers; (2) **to collect information** on user requirements and feedback; (3) **to stimulate information exchange and collaborations** between health research and infrastructure communities; (4) **to establish two-way communication channels** with communities, societies, infrastructure providers and policy-makers for disseminating deliverables and retrieving timely feedback; (5) **to inform the steering committee (SC)** (TA1)

in due time **on critical issues** in NFDI4Health and to develop recommendations for the SC that allow swift risk management.

The scientific medical societies, including those in the area of epidemiology and public health, are important stakeholders. The COC will connect with the members of these societies via the NFDI4Health community mailing list (in cooperation with T1.3). We will involve the Association of the Scientific Medical Societies (AWMF[80]), consisting of 179 scientific medical societies. Further, a special call will be directed to all scientific societies with a focus on large-scale cohort data (e.g., DGEpi, DGSMP, GMDS, IBS-DR) to ensure their involvement in the governance; cooperative links have already been established during the preparation of this proposal. In addition, the COC will stay in close contact with the User Advisory Board that represents the user community (UAB, TA1).

Annual NFDI4Health community workshops (**D4.3, annually**) open to national and international communities will be organised jointly with the communication manager (T1.3). The topics of the workshops (e.g. standardisation, guidelines and services) will be elaborated together with TA2 and TA3. Through these workshops, actors that would not usually interact with each other will get into contact, which is likely to enhance NFDI4Health impact among various stakeholders (academia, societies and policy makers). The first NFDI4Health community workshop was already held on June 25, 2019 in Cologne (55 participants). It focused on the user needs regarding the generation, provision, management and analysis of clinical trial/epidemiological data as well as data from, e.g., disease registries and health insurances. A report incorporating key results on the interaction with and feedback from the user communities obtained by user surveys before and during the annual workshops of NFDI4Health will support the further development of the consortium for future funding periods (**D4.11, M58**).

T4.2 “FAIR data sharing – community aspects” (**HITS (lead), U Leipzig, UM Göttingen, UM Greifswald**; GMDS; additional partners: ELIXIR, FAIRDOM, FAIRsharing)

Despite the increasing support by open science frameworks and FAIR data initiatives in the health domain, the “FAIRification” level of the data still is low, and novel reward mechanisms for scientists in the domain still have to be established. We will focus on the obstacles to share research data in the health domain. More general aspects of this cultural change towards a spirit of data sharing will be addressed jointly with other NFDI consortia (via T4.4).

The overall objective of T4.2 therefore is the **adaptation and promotion of FAIR data principles** in the health data domain. Several activities will support this objective, notably (a)

facilitating and bundling the access to information about “FAIRification” of health-related data, including the access to domain-specific standards, guidelines and terminologies, (b) adapting a concept of FAIR metrics with relevance to the health research community, and (c) providing a high-level community platform for FAIR data sharing within NFDI4Health. The development and testing of novel reward mechanisms for quality-controlled research data sharing will be considered an additional meta-objective to be pursued together with other NFDI consortia.

As an initial measure, the SEEK-based FAIRDOMHub[81] will be adapted and established as internal community exchange platform (**D4.2, M6**). Secondly, together with T4.6, a survey in the health research community augmented by semi-structured interviews will be conducted in order to investigate what would be needed to enhance reputation through data sharing and what the main obstacles for research data sharing in the health research domain are (**D4.4, M12**). In order to adapt and promote FAIR data principles, existing contacts to FAIR data initiatives throughout Europe and beyond will be intensified (coordinated with other NFDI consortia), including GO FAIR, FAIRsharing, EOSC-Life, ELIXIR, FAIRDOM, etc.. New strategic partnerships will be established beyond the already existing networks. This networking will facilitate international alignment of FAIR concepts and standards (feedback to T2.3).

Implementing a concept of FAIR metrics (based on developments in T2.3) with relevance to the health research community is another major activity of T4.2 (**D4.6, M24**). This will build on work of the FAIRMetrics group[62] where several of the group´s indicators will need to be reviewed and adapted. In partnership with the FAIRsharing initiative (see the corresponding LoC), a collection of recommended data standards, guidelines and policies for health research will be compiled[82]; based on T2.2 and T2.4) and made public by month 30 (**D2.7**).

NFDI4Health will evolve around large current and future studies that contain individual, person-related medical and other sensitive data. The sharing of such health data poses particular challenges. The results from TA2, TA3 and TA6 will be brought together to develop and disseminate a modular concept for data sharing in health research, which needs to be flexible with respect to various scenarios in terms of complete or partial sharing of data sets. This concept, together with the FAIRification mechanisms adopted and developed in this measure, will be published and distributed as a guidance document (with T4.3) on “How to make your data FAIR in the health domain” (**D4.9, M48**). Once the modules of the concept have been developed, T4.2 will collaborate closely with T4.6 on distribution and gathering feedback for refinements, as well as requests for further standards development if needed, through the NFDI4Health PALs (see T4.6).

T4.3 “Training and education” (**U Cologne (lead), KKSN (co-lead), BIPS, DIfE, U Bonn, U Leipzig, UM Greifswald, ZB MED**)

The **objective** of this task is to address the lack of a broad basic education in data science and data management skills specific to emerging use cases in clinical trials and epidemiology (TA5) by adapting and developing appropriate **teaching material, graduate curricula, prototypes**, and **dissemination strategies**[83,84]. Our training activities will concentrate on health research data including legal issues concerning sharing and publishing of data and data integration. The three primary target groups are: (1) young researchers, i.e. mainly PhD students, (2) clinical trial physicians and epidemiologists, and (3) technicians such as programmers, IT-experts and data managers. The latter two groups will be called “professionals”.

Young researchers will be addressed by an initiative starting in December 2019 at the University of Bremen, where, the Federal State of Bremen, the U Bremen Research Alliance (UBRA), NFDI4Health, NFDI4Biodiversity, and NFDI4Earth will establish a **cross-domain graduate education programme on research data management and data science**. The curriculum and modules will be developed, pre-tested and refined following student feedback. Subsequently it will be provided to all NFDI consortia and evaluated by the NFDI4Health consortium and our user communities (year two) (**D4.5, M24**). After further revision of the programme in year three, the NFDI-wide refinement and roll-out is planned for years four and five.

Advanced training in research data management (RDM) for professionals will be developed in parallel to the graduate programme and will be put into practice in close interaction with the graduate education. Recently, the role of data stewardship in providing RDM guidance and the role of training services in acting as multipliers was highlighted[85]. Research libraries at universities can serve as hubs to connect local communities of researchers, professionals, and infrastructure providers with the NFDI. Together, ZB MED and U Cologne plan to explore the interface of NFDI4Health with a university by building up a **pilot RDM-training-service** within the first three years at the Medical Faculty of U Cologne. We will explore the **role of a data steward** in his/her function of transferring NFDI4Health outputs (services, educational materials, reports, etc.) into the university environment. In the light of division of labour and finding local multipliers (scalability), the (co-)applicants will develop a concept for establishing a carefully balanced cooperation between a specialised steward at the faculty and the library staff offering more generic training services. A corresponding pilot focusing on epidemiological research will be run by ZB MED in cooperation with U Bonn.

After the pilot phase, the data steward concept (**D4.7, M36**) will be disseminated by the KKSN together with the MFT and the DINI/nestor AG *Forschungsdaten* [working group *Research data*]. Special attention will be paid to the involvement of epidemiological research institutes. Close cooperation is foreseen with institutions involved in the use cases (TA5) and with professional societies. The concepts and implementation approaches of the education programme and the data stewardship pilot will be published on the NFDI4Health website for reuse. Training material will continuously be distributed by the KKSN. Here, T4.3 will benefit from the rich training experience of the KKSN. Of note, in 2018 alone, more than 9,500 clinical physicians and other medical researchers were trained e.g. in data requirements for interventional and non-interventional studies. Results of other TAs of NFDI4Health, NFDI4MED and other NFDIs may feed into these training programmes as well as into newly developed courses.

For professionals in particular, NFDI4Health will also offer three annual **summer schools** on specific aspects of data stewardship, hosted at the institutions of the T4.3 (co-)applicants, starting in project year three.

All training material will be offered as open educational resources and will be accessible via the website established in T1.3, together with information about training events. Additionally, U Cologne will develop a concept and a pilot implementation (for reuse) to make the material available to **students for asynchronous self-study** or to be **integrated** in individual lectures (e.g., e-learning platforms like ILIAS).

T4.4 "Networking with NFDI and beyond" (**BIPS (lead), ZB MED (co-lead), HITS, RKI, U Cologne**)

All external and internal collaborations will be coordinated and monitored by T4.4. Networking activities of NFDI4Health (co-)applicants will be documented to enhance transparency of ongoing interactions. As described in Section 2.3, NFDI4Health will closely cooperate with all NFDI consortia that address management and sharing of highly sensitive, person-related data collected in the health domain (e.g. **NFDI4MED**, **NFDI-Neuro**, **GHGA**) or in the social domain (e.g. **KonsortSWD**). Data protection is a joint and overarching challenge. NFDI4Health will also collaborate with other NFDI consortia that address data of high relevance for health (e.g. **NFDI4Agri**, **NFDI4Earth**, **NFDI4BioDiversity**, **NFDI4NanoSafety**). We already started collaboration with these consortia during the preparation of the application in order to identify and discuss cross-cutting topics. Cooperation will be fostered further by joint working groups. Eleven consortia discussed these issues during a first meeting in Berlin (2019), and agreed on the so-called Berlin Declaration. The cross-domain graduate education programme described in T4.3

may serve as an example for such joint activities, In future, all overarching topics may be addressed in close cooperation with the **NFDI4Life Umbrella** consortium, e.g. networking and governance, development of new reward systems for data sharing.

Networking beyond the NFDI consortia will build on the ongoing commitment of NFDI4Health (co-)applicants in a number of infrastructures. For example, at the national level (1) several (co-)applicants are actively involved in setting standards for biomaterial management including biomaterial (meta-)data (German Biobank Alliance, Biobanking and Biomolecular Resources Research Infrastructure – European Research Infrastructure Consortium); (2) many (co-) applicants are members of relevant scientific societies such as the DGPH, the DGEpi and the GMDS, notably the GMDS project group on FAIR data infrastructures (led by HITS, UM Göttingen and U Leipzig); (3) co-applicants are involved in the German Research Foundation (DFG) collaborative research centres (CRC) transregio CRC (TRR). Close exchanges with (4) the German Network for Bioinformatics Infrastructure (de.NBI) with respect to their cloud resources for computation and as central hub for distributed data access and with (5) the Research Data Alliance Germany are planned.

Internationally, NFDI4Health partners will draw on their active involvement in European (and worldwide; Table A2, Appendix 1) initiatives such as EHDEN, EOSC-Life, FAIRDOM, RDA, GO FAIR, FAIR4Health, FAIRplus and euCanSHare, EU-STANDS4PM, and COMBINE to ensure compatibility of NFDI4Health activities with actions at European and international level. By this, we will foster the international exchange of approaches, infrastructures and technical solutions.

T4.5 “Citizen and patient involvement” (**BIPS (lead), Fraunhofer FIT, KKSN, U Cologne**)

Involving citizens and patients in the overall work of the consortium will be the core objective of this measure. Participatory processes form a key component of integrated knowledge translation (IKT). This has become a cornerstone of implementation-oriented health research[86]. For NFDI4Health, a specific focus will lie on capacity building towards citizens´/patients´ use of personal data for research purposes, with the goal of the establishment of a web-based *citizen research data portal* (**D4.8, M36**). Through such a portal, citizens and patients who have contributed personal data may get an overview of their data and view the conditions under which their data may be (re-)used (e.g. privacy regulations, application procedures, overview of studies that have used their data). In addition, there will be the opportunity for citizens and patients to explore their data and generate own (public-driven) research questions that may then develop into cooperative research studies (see e.g., “Cloudy with a Chance of Pain”[87]).

A wide spectrum of approaches will be applied to support citizen/patient involvement and participation. NFDI4Health will enable (1) participation of citizen/patient representatives in the UAB. The representatives should have some experience as research participants. They will regularly interact with the consortium and contribute citizen/patient views and insights into relevant issues around data provision and usage. (2) The citizen research data portal will be developed and implemented once study data sets have been identified and integrated that can be used for this task. (3) NFDI4Health will provide an interactive webpage with components specifically designed for patients and the public, which not only serves to inform about the consortium´s work but also includes at least bi-annual user quizzes, surveys and/or new video sequences supporting active engagement in NFDI4Health activities. Close cooperation with T1.3 is foreseen, e.g. in use of social media. (4) An evaluation framework regarding citizen/patient involvement will be developed in order to assess, inform and adapt the consortium´s work from the perspective of citizens and patients.

T4.6 "Service evaluation and user research" (**U Cologne (lead), HITS, RKI, UM Greifswald**)

This measure will conduct **user research** and **evaluation** to obtain a better understanding of user demands and needs. Mapping these demands to the existing services will reveal service gaps and thus guide the tailoring of NFDI4Health services to user needs. Evaluating the existing and planned services is a key requirement for the continuous improvement of services and products. We will use a method mix of different tools of empirical social research. A widely disseminated online survey right at the beginning of the project will assess the status quo. A second online survey in project year four will inform about the progress and collect recommendations for the next funding period (**D4.10, M48**). Additional studies (using different methods of qualitative social research) will offer deeper insight into critical aspects and individual services. These include in particular usability tests for web-based services, qualitative in-depth interviews and focus group discussions, crowd-sourcing for user-generated services and non-reactive methods of user research.

A central element for collecting information on requirements and user feedback is the concept of PALs (Project Area Liaisons) adopted from FAIRDOM[40]. PALs will be a small number of "front line" experts from research projects of the user communities addressed by NFDI4Health. They act as data management advocates and multipliers, and communicate new developments back to their projects. By working with a small team of committed representatives, NFDI4Health can ascertain to fulfil requirements for some exemplary projects and disseminate new material and recommendations to the wider community. The PALs will (a) gather requirements and offer consultation, (b) review service development plans and first prototypes, (c) test and report on

solutions and (d) collect knowledge from research projects that is relevant for NFDI4Health services. NFDI4Health can build on extensive experiences as the PALs concept was introduced by FAIRDOM with HITS as co-founder more than 10 years ago. Since then it has been applied to many different FAIRDOM-related research consortia in Germany and Europe.

Role of members and participants: U Cologne and BIPS will coordinate TA4. All (co-)applicants and participants will contribute to user community surveys, training and education, citizen involvement (T4.5) with RKI in charge of coordinating outreach efforts (T4.1). BIPS and ZB MED will coordinate all networking activities and the exchange with other NFDI consortia (T4.4), with support by TMF as umbrella organisation for networked medical research in Germany and by MII. U Cologne will coordinate the user research programme (T4.6) and explore prototypical linking of NFDI4Health to its university campus. ZB MED and U Cologne will support the roll-out to other universities (T4.3). HITS will offer its FAIRDOMHub as the NFDI4Health community exchange platform and contribute to FAIRsharing (T4.2) and to the PALs concept (T4.6).

Cooperation with other task areas: Outreach and dissemination of the project results requires close cooperation with all other TAs, in particular with T1.3, T1.4 and with the use cases in TA5. T4.2 will build on the standards bundled and developed in T2.2/T2.4 and base its FAIR metrics concept on T2.3. Training and education (T4.3) will be closely linked to almost all other areas dealing with the different teaching contents. Similarly, the exchange and cooperation efforts in TA4 will be linked to activities in the respective TAs across NFDI4Health. The service evaluation and user research (T4.6) will provide a continuous feedback to all other TAs and will thus be of high relevance for the design of the NFDI4Health services.

Risks of implementation: (1) Developments from NFDI4Health may not be applicable to the user community. To address this risk, community workshops and the UAB will provide feedback and report potential issues to the spokesperson and the SC (TA1). The implementation of the FAIRsharing collection for public health data (T4.2) depends on the long-term commitment of additional partners, as, e.g., FAIRsharing that has committed itself to support NFDI4Health (see respective letter). (2) Training and education is a core measure of TA4. We do not expect that these activities will satisfy all training demands. A flexible adaptation of training offers at later stages will be one way to account for this. (3) In the field of citizen and patient involvement (T4.5) there is uncertainty regarding the interest that our topic may raise in the target group. Lack of interest may require further intensification of communication and interaction efforts. (4) The user involvement (T4.6) strongly depends on the anchoring of the (co-)applicants in the

user community, especially regarding the PALs concept. Confining the initial selection of PALs to users already dedicated to data management and committed to the FAIR principles might help to overcome this risk. At a later stage the user involvement can be extended.

Table 6: Deliverables of Task Area 4

No.	Type*	Description (month; measure)
D4.1	DISS	Establishment of Community Outreach Committee (COC) (M6; T4.1)
D4.2	SW	FAIRDOMHub (SEEK-based) as high-level NFDI4Health community platform (M6; T4.2)
D4.3	EV/DOC/ DISS	Organisation of user community workshops with surveys and follow-up community publications (M12, annually; T4.1)
D4.4	DOC/ DISS	Community survey on reputation gain through data sharing (M12; T4.2)
D4.5	DOC/DISS	Cross-domain graduate education programme on research data management and data science (M24; T4.3)
D4.6	DOC/ DISS	Implementation of a FAIR metrics concept (based on T2.3) for health data (M24; T4.2)
D4.7	SVC	Demonstrator of the data stewardship service (M36; T4.3)
D4.8	SVC	Web-based citizen research data portal (M36: T4.5)
D4.9	DOC	Guidance document: “How to make your data FAIR in the health domain” (M48; T4.2)
D4.10	DISS	Community survey on service developments of NFDI4Health (M48; T4.6)
D4.11	DOC	Report on overall interaction with and feedback from user communities during project progress (M58; T4.1)

* DOC: document, INT: interface definitions, APIs, plan designs, architectures, SVC: service, SW: software, EV: event, DISS: dissemination

4.6 Task Area 5 “Use Cases” (lead: DIfE, BIPS)

Structured, quality assured individual phenotype and exposure-related health data are generated in clinical trials and epidemiological and public health studies. Clinical trials and (prospective) epidemiological studies are the most appropriate study designs to investigate (1) the onset and dynamics of measurable phenotype profiles and diseases on individual and population level and (2) the efficacy of diagnostic and therapeutic interventions. Access to detailed information on a large and unselected number of participants and patients is pivotal to advance patient risk stratification, to support personalised medicine, to find new interventional options and to improve patient care. Nowadays, even more data become available through large primary data collections (e.g., for cohort studies) and secondary data sources such as disease registries or health insurance data. The sharing and reuse of such data for research purposes are currently limited for several reasons: (1) lack of standardised metadata and

methodological diversity in the collection of health information across studies, (2) incomplete knowledge of relevant studies and the nature of data collected in the research community, (3) regulatory, governance, ethical and legal challenges of data access and of bringing data together across populations or data sources and (4) lack of knowledge and skills of potential data users and limitations of data analytical infrastructures to use study data.

While TA2, TA3 and TA6 will develop necessary concepts and services related to data and metadata harmonisation and standardisation, data and metadata repositories, data access and data analyses, the main objective of TA5 is to implement or to at least explore the possibilities to implement these infrastructure components in specific use cases which reflect core needs of the scientific community related to health data research. The use cases will address a range of areas which will lay the ground for future expansion for full coverage of the broad range of data collected in health research. The use case “Nutritional epidemiology” (T5.1) addresses the needs to standardise complex exposure data across a variety of data sources with the example of dietary data, while similarly the use case “Epidemiology of chronic diseases” (T5.2) will do so for a variety of important health outcomes. Both use cases will implement NFDI4Health infrastructure components in major epidemiological studies in Germany, including the NAKO Health Study, the EPIC Potsdam and Heidelberg studies, KORA, SHIP, LIFE and others (see Table A1, Appendix 1). The use case “Secondary data/record linkage” (T5.3) will explore access and linkage procedures which will foster specifically the enrichment of primary data from epidemiological and clinical studies, e.g., by registry and health insurance data. Furthermore, clinical trial data will be the focus of the use case “Clinical trials” (T5.4) where infrastructures for data repositories as well as search and access will be implemented for publicly funded clinical trials based on the KKSN. The use case “Surveillance” (T5.5) will implement NFDI4Health infrastructure components to support continuous, systematic collection, analysis and interpretation of health-related data for planning, implementation and evaluation of public health practice. Finally, the use case “Radiomics / Imaging AI” (T5.6) will focus on multidisciplinary predictive pattern recognition based on combined imaging and structured health data implementing infrastructure components related to distributed learning, quality assurance and data harmonisation.

A central element of TA5 is to help researchers with the discovery (*Findability*) of existing data by making data sources clearly and persistently identifiable and providing metadata. With respect to data standards, there is considerable heterogeneity in the methods used to assess and operationalise exposures and outcomes across studies as there is heterogeneity in secondary data sources. Thus, TA5 will contribute to the definition of data standards and will

implement harmonisation steps (*Interoperability*) and linkage procedures. Furthermore, TA5 will implement infrastructures to support data access and analyses (*Accessibility, Reusability*).

Consortium members of NFDI4Health TA5 are primary data holders reflecting a broad range of health-related epidemiological and clinical research in Germany, in particular major epidemiological studies, the KKSN and the German health surveillance system. Furthermore, they are well established and active in relevant national and international activities related to data standardisation, data accessibility and joint data analyses (e.g., DEDIPAC consortium[88], ENPADASI[28,29], InterConnect[30], SOS[89], ARITMO[90]).

Summarising, TA5 is an essential component of NFDI4Health and will in particular contribute to the Key Objectives (1), (2) and (3) listed in Section 2.1 of this proposal, while it will support achieving the Key Objectives (4) and (5). The deliverables of TA5 are listed in Table 8.

T5.1 "Use case 'Nutritional epidemiology'" (**DIfE (lead), U Bonn (co-lead), BIPS, MDC, MRI**; participants: DKFZ, HMGU, NAKO, UKE)

There is considerable heterogeneity in the methods used to assess and operationalise dietary exposures across epidemiological studies. First, different dietary assessment methods (food frequency questionnaires, dietary recalls, dietary record, diet history questionnaires) are applied. Second, the specific design of assessment instruments varies, to some extent reflecting the variation in (regional) diet composition. Third, dietary intake is characterised by different levels (meal intake, dietary intake, dietary pattern) and corresponding dimensions (e.g., food intake, nutrient intake) and subdimensions (e.g., specific nutrients). To overcome these limitations, this measure will implement the data infrastructure developed in NFDI4Health in the context of nutritional epidemiological studies to facilitate population-based research on dietary factors and their association to health-related outcomes. Our focus will be on diet-related data and information in existing infrastructures and studies, such as NAKO Health Study, LIFE-study, EPIC-Potsdam, DONALD, IDEFICS/I.Family (full list and details of data sources: Table A1, Appendix 1).

A central element of T5.1 is to help researchers with the discovery (*Findability*) of existing data from observational studies related to the assessment of nutritional behaviour (exposure), by making studies clearly and persistently identifiable and providing study metadata. T5.1 will implement data harmonisation and annotation steps (*Interoperability*), which are specifically informative for dietary data. Furthermore, it will build on experiences with infrastructures for data sharing and federated meta-analyses across different studies as, e.g., DataSHIELD in the

context of the EU projects ENPADASI and InterConnect. Thus, T5.1 will implement infrastructures for data access (*Accessibility*) and distributed data analyses (*Reusability*) which allows analysis of individual-level data of (longitudinal) observational studies using state-of-the-art epidemiological methods related to dietary data. With respect to interoperability and reusability, knowledge and skills to analyse dietary data are, in the light of the inconsistency of dietary data, a major limitation for researchers who want to make use of study data, specifically for those without a nutritional background. Thus, T5.1 will also work out a draft framework for a description of the instruments applied in the respective study, a data set description including potential pitfalls and limitations when using these data.

This use case pursues **five major objectives**: (1) adoption of **(meta-)data publication** based on the publication guidelines developed in T2.1, (2) compilation of methods for **data classification and harmonisation** used in existing studies (to feed T2.2) (**D5.2, M24**), (3) implementation of **data standards** (from T2.2) (**D5.12, M58**), (4) implementation of **data access infrastructures** based on specifications in T2.4 and in cooperation with T3.4, (5) implementation of **distributed data analyses infrastructures** (T3.7) specifically for nutritional epidemiological studies (**D5.10, M48**).

T5.2 "Use case 'Epidemiology of chronic diseases'" (**MDC (lead), BIPS (co-lead), DIfE**; participants: DKFZ, HMGU, NAKO, U Leipzig LIFE)

Chronic diseases (also termed non-communicable diseases, NCD) represent the major disease burden in Germany and globally. There is a strong epidemiological research focus on disease distribution and determinants, as well as on clinical outcomes. Combined analyses of disease data from different data sources are an important cornerstone for comprehensive research. However, this approach poses numerous challenges at present since, e.g., cohort studies with cardiovascular and metabolic disease endpoints in Germany show considerable heterogeneity in disease classification[91]. Similar conditions are expected for other NCD endpoints. For example, diseases such as stroke exist in different manifestations that represent different aetiologies and thus require differentiated assessment and verification methods which may or may not be reflected in study data. Furthermore, classification schemes change over time, moving from a phenotypic description to an aetiologically-oriented grouping, often based on molecular and genetic information. Changes in diagnostic criteria also occur over time and hamper comparability. The assessment of NCD varies substantially between studies (e.g., self-report versus medical instrument and physician-based assessment) thus affecting validity of disease data which needs to be made transparent for research across different studies. Additionally, secondary data, e.g., from disease registries or derived from insurance data, will

be an important source with potential to enhance data breadth, depth and quality (cf. T5.3). To overcome these challenges, T5.2 will implement the data infrastructure developed in NFDI4Health in chronic disease epidemiological studies to facilitate population research on distribution of NCD and their determinants. Our focus will be on NCD-related existing infrastructures and studies, such as NAKO Health Study, EPIC-Potsdam, LIFE and SHIP/SHIP Trend (full list and details of data sources: Table A1, Appendix 1).

T5.2 aims to help researchers with the targeted retrieval (*Findability*) of existing data from observational studies on NCD, by making studies clearly and persistently identifiable and providing study metadata. Together with T5.3, the use case will strive to test and implement *data linkage* according to the standards developed in that measure. T5.2 will implement data harmonisation steps (*Interoperability*) related to chronic disease outcomes, specifically for cancer, cardiovascular diseases and diabetes. Here, T5.2 will build on experiences gained in previous joint analyses of NCD studies[92], aiming to extend these approaches with NFDI-wide new infrastructure components such as DataSHIELD. T5.2 will also implement infrastructures for data access (*Accessibility*) and distributed data analyses (*Reusability*), using a step-wise approach that initially builds on a small set of studies and disease outcomes. T5.2 will also develop a draft framework for a description of the instruments and approaches applied in outcome data collection in NCD studies.

Similar to T5.1, the chronic disease use case pursues **six major objectives**: (1) adoption of **(meta-)data publication** based on the publication guidelines developed in T2.1 (**D5.1, M12**), (2) contribution to the development of standards for **data classification and harmonisation** based on existing studies (in conjunction with T2.2), (3) implementation of **data standards** (from T2.2) (**D5.12, M58**), (4) implementation of **data access infrastructures** based on specifications in T2.4 and in cooperation with T3.4 (**D5.6, M36**), (5) implementation of **data linkage approaches** explored in T5.3 (**D5.7, M36**), (6) **implementation of distributed data analyses infrastructures** (T3.7) (**D5.10, M48**). As indicated, the focus is on epidemiological studies of NCD but may be extended over time.

T5.3 “Use case ‘Secondary data and record linkage’” (**BIPS (lead), RKI, UM Greifswald**; participants: GEKID, IMIBE, NAKO, U Magdeburg)

In epidemiological studies, a vast amount of information has to be gathered on each individual participant, which in turn results in demanding examination protocols and questionnaires. Also, information may not be correctly recalled or not at all known by study participants. It is therefore desirable, to tap external data sources (secondary data) that record objective data and that – if

such data are made available – reduce the amount of information to be provided by study participants (primary data). Besides disease registries, most secondary databases store individual level data for other purposes than research as, e.g., claims data collected by health insurances for reimbursement purposes. But these data are nevertheless a powerful resource for research. Making these data accessible for research is a huge challenge because access is restricted by rigid privacy and data protection regulations. **Record linkage**, i.e. linkage of primary data or linkage of different secondary/registry data sources with each other on an individual level creates synergies. That means, it offers most efficient options for health research where highly informative but yet under-used data are shared and combined to answer research questions that cannot be addressed by traditional single-data-source approaches. These data would also enrich clinical trial data by medical care and outcome data and even by other social data and by this allow for a more holistic approach when, e.g., analysing the efficacy and long-term toxicity of drugs. However, any record linkage is hampered by the fact that no unique identifier exists in Germany as opposed to the Nordic countries like Denmark. Different alternative linkage approaches have been tried in Germany, each having its specific limitations[93] and quality issues like linkage error[94].

This measure aims to improve the *findability* and *accessibility* of secondary/registry data and to offer feasible solutions for record linkage of primary and secondary data. First, it will formulate requirements for the NFDI4Health search interface (developed in T3.1) for the identification and classification of secondary data sources and the data that they hold. For this purpose, we will describe the specific challenges for the creation of metadata of secondary data sources needed for the framework for data standardisation and harmonisation worked out in T2.2. These metadata standards, in turn, will be adopted for the **metadata publication** based on the publication guidelines developed in T2.1. As the owners of secondary data typically have no resources to implement data standards other than those needed for their administrative duties, such service has to be provided by the co-applicants. Therefore, BIPS and RKI will pilot the **implementation of data standards** provided by TA2 on claims data hosted by BIPS (German Pharmacoepidemiological Research Database, GePaRD) and cancer registry data hosted by RKI (German Centre for Cancer Registry Data, ZfKD) and explore the options for upscaling to other secondary data sources. Second, this measure will assess whether it is feasible to create a “common” record linkage algorithm for all person-related data sources relevant for health research in consideration of data privacy and data protection regulations for record linkage as worked out in T6.4. To this aim, we will evaluate the applicability of different linkage methods to typical data sources in this research field, building on the experiences of co-applicants and participants as listed in Table 7 and the expertise provided by the Medical Informatics Initiative (MII) Germany

which explored the application of methods introduced by Schnell et al.[95], Schnell & Borgs[96] and Contiero et al.[97]. Depending on this assessment, the following further steps will be taken: (a) specification of requirements for an "ideal" record linkage procedure will be drafted, (b) the feasibility to develop a record linkage procedure according to this specification will be assessed in close collaboration with data owners, (c) recommendations regarding the practical implementation of specific procedures depending on the data sources to be linked will be provided and (d) needs regarding the specification of the data items needed for linkage will be formulated.

Table 7: Epidemiological databases with record linkage to or between secondary/registry data

Study	Linked secondary/registry data	Responsible co-applicant/ participant	Reference	Consent for linkage
GePaRD	Hospital records Mortality registry	BIPS	Ohlmeier et al.[98] Ohlmeier et al.[99]	No
NAKO Health Study	Health insurances, pension fund, unemployment insurance, cancer registries	NAKO	Stallmann et al.[100]	Yes
SHIP	Hospital records, health insurances	UM Greifswald	Schmidt et al.[101]	Yes
LIFE	Health insurances	U Leipzig	initiated	Yes
Heinz-Nixdorf Recall Study*	Unemployment insurance	IMIBE	Wahrendorf et al.[102]	Yes
lidA*	Unemployment insurance, health insurances	U Magdeburg	March et al.[103], March[104]	Yes

* Databases held by participating institutions

In summary, this use case pursues **four major objectives**: (1) adoption of standards for **publication of metadata of secondary data** based on the publication guidelines developed in T2.1, (2) development and **pilot test of data access infrastructures** for at least two data sources (T3.4), (3) **assessment of the usefulness and feasibility** of various record linkage options (in collaboration with T6.4) (**D5.9, M44**), (4) **provision of solutions to improve their implementation** (**D5.13, M58**).

T5.4 "Use case 'Clinical trials'" (**U Leipzig (lead), KKSN**; participants: clinical trial centres)

For **clinical trial planning** (design and power calculation), information is needed on the disease, the target population and the biometrical properties of the endpoints used. Clinical trial planning should be based on best available evidence. However, in many cases published results of prior or similar trials are available, but relevant details are not provided. Examples of such information are statistical properties of certain variables in a target patient population, time course of response variables, statistical properties of before-after differences, effect sizes of

treatment differences in trial endpoints, distribution of prognostic factors in the target population, frequency of conditions considered as exclusion criteria in the intended trial, and many more.

Clinical trials prospectively collect data based on a predefined documentation concept derived from the study protocol which often includes several hundred variables in structured case report forms. Great efforts are made to specify and enforce quality metrics for each data element (e.g., mandatory fields, formatting requirements, units of measurement, reference intervals, query actions). While the exact meaning of variables and measurement methods are known to local investigators through the study protocol, specific training or standard operating procedures (SOPs), it is difficult for external researchers to understand the exact meaning because of lacking or limited descriptions. A detailed description incl. concepts from medical terminologies, plausibility checks and information on the origin of the data (inquired/measured/calculated/ transferred from secondary system), would increase the quality of the documentation concepts and support the reuse of the data. In the long run, **harmonisation and standardisation of clinical trial data elements** would be helpful both in the creation of new study concepts (since one could refer to gold standard data elements) and in meta-analyses comparing studies on the same topic.

Clinical trials centres (CTC) at German medical faculties support a wide range of academic, publicly funded clinical trials. About 400 academic clinical trials are activated in Germany each year. Once a clinical trial is completed and its results published, data should be made available for scientific reuse. This is strongly encouraged by public funders of clinical trials such as DFG or BMBF. Currently, no infrastructure exists to support **clinical trials data sharing** with regards to *findability*, *accessibility* and *reusability* and in compliance with the legislative and regulatory requirements for personal medical data.

To overcome these limitations, this use case will first develop a catalogue of typical characteristics of clinical data sets (e.g., disease area, type of clinical trial, trial unique identifier, trial design characteristics, type of intervention, full text search in the trial synopsis, target population, outcome variable, type of therapy) in collaboration with T2.1 and T2.2. The characteristics will be implemented as searchable and filterable facets in T3.1. Together with TA6, concepts for different use and access mechanisms will be developed and thus implemented in T3.4 or as distributed analysis in T3.7. As part of this activity, the implemented service and their orchestration will be evaluated in real-life scenarios in a CTC.

Second, we will upload clinical trial data, metadata and data set descriptions, and additional documents (e.g., clinical trial protocol including all amendments and patient consent forms) to

the existing infrastructure of NFDI4Health (T3.5). To allow sharing of clinical trial data, we will establish a workflow including the following steps: permission of the principal investigator (PI) and/or sponsor, evaluation of informed consent documents (T6.2), anonymisation (if required), specification of access modalities (T6.1). After successful completion of all these governance tasks, the actual upload of data and documents will take place. The entire workflow will be paradigmatically tested by three to five CTCs with ten clinical trials each. The results will be documented to provide a more accurate estimate of the effort required for future uploads.

Third, we will define and consent a metadata catalogue for top 1,000 most used data elements in typical academic clinical trials. For this, three to five CTCs will provide the annotated CRFs – preferably in a machine-readable format like CDISC ODM – for ten different trials each. The metadata definitions will be extracted in collaboration with T2.2 as recommendations for the creation of data dictionaries. Then metadata definition will be curated and completed based on best practices specified by T2.2 utilising T3.2. Based on the standards defined in T2.2, to support semantic interoperability, data elements will be annotated with concepts from medical terminologies (LOINC, SNOMED, CDASH, UMLS, MeSH) and local metadata will be mapped to a central catalogue consisting of different domains (as defined in CDISC CDASH). Metrics for computing similarity of data elements beyond trivial comparisons of lexical labels will be implemented. Finally, central metadata will be harmonised (in collaboration with T2.2). This activity will be undertaken in alignment with the interoperability working group of the MII.

In summary, this use case pursues three **main objectives**: (1) definition of requirements and use of elaborate **publication metadata** to improve findability and reusability of scientific result publications of clinical trials (**D5.5, M36**), (2) support of **data sharing** to improve the accessibility of clinical trial data (**D5.4, M30**), (3) **harmonisation** of data elements to improve the interoperability and reusability of clinical trial (meta-)data (**D5.11, M48**).

T5.5 “Use case ‘Surveillance’” (**RKI (lead), BIPS**)

The degree of standardisation, harmonisation and reusability of existing surveillance studies and health surveys is currently limited because they have been designed and conducted independently of each other. To overcome this limitation, we will develop a reproducible, validated and robust framework of instruments, methods and databases for health survey and monitoring data concerning data collection, harmonisation, transfer, linkage, management and privacy (**D5.15, M60**). This framework will be developed, evaluated and adapted using health monitoring studies, metadata and databases covering populations of all ages. This methodological surveillance framework will be discussed with and offered for future implementation to

already established European initiatives, such as WHO-COSI, HBSC, the European Health Interview Survey (EHIS) and the Nordic Monitoring of Diet, Physical Activity and Overweight. This consenting approach will parallel with BIPS and RKI activities in the Policy Evaluation Network (PEN) to develop a consolidated approach to policy evaluation across Europe that can be used by existing monitoring and surveillance systems[105].

In particular, this use case will focus on monitoring and surveillance of overweight and obesity as one example of diet- and lifestyle-related disorders in (nationwide) population studies which may serve as blueprint for other health outcomes. Heterogeneous data sets from different surveys will be harmonised with a specific focus on overweight and obesity information. To build up a metadata repository in NFDI4Health, existing metadata models for health surveys will be extended to include additional information, such as survey description, periodicity, sources and technical definitions. T5.5 will facilitate access to data inventories from public surveys and surveillance studies, such as the BGS98, DEGS1, GEDA surveys and COSI (full list and details of data sources: Table A1 Appendix 1). In this context, RKI already has broad experience and it provides data access via public use files by its RatSWD-accredited research data centre[11] (FDZ). The survey data sets will be made *findable* and *accessible* using dedicated local data access points and webservices (TA3). If applicable, the data sets and metadata will be adopted for (a) publication into the central search hub (T3.1), (b) website download (e.g., public/scientific use files) and/or (c) programmatic remote access through the CAP (T3.4). The data sets will be made *interoperable* with other systems (T2.4) and *reusable* for distributed data analysis (T3.7). Requirements for automated and data privacy-controlled data access will be specified in accordance with TA6.

This use case therefore pursues **four major objectives**: (1) adoption of **standards for data access and interoperability** in human and machine-readable formats (T2.4) and **set-up of LAPs for data access**, (2) implementation of (**meta-)data standards** of health monitoring and surveillance data sets (T2.2), (3) specification of requirements for human and machine-readable **data privacy-related consent mechanisms** (T6.1), (4) development of **a harmonised health survey and surveillance framework,** i.e. establishment of best practice guidelines and consulting expertise (**D5.15, M60**).

T5.6 “Use case ‘Radiomics / imaging AI’” (**Fraunhofer MEVIS (lead), BIPS, U Leipzig, UM Greifswald**; participant: NAKO)

Biomedical imaging has become an important field of data collection in epidemiological and clinical studies due to its sensitivity to both localised and systemic alterations, large information

density and broad availability. Radiomics is defined as a method that extracts large amounts of quantitative features from medical images using pattern recognition and machine learning algorithms. However, radiomics faces several challenges: (a) variability in image acquisition, reconstruction and post-processing, (b) lack of reliable and automated image segmentation and identification of anatomical and pathological structures of interest, (c) problems to robustly extract features that describe the image content and to transform unstructured image information into a set of well-defined descriptors and (d) problems to create large representative databases and data exchange infrastructures that provide the means to use statistical analysis and machine learning to explore new patterns and create new hypotheses in the respective area[106]. Modern machine learning techniques combined with conventional data analytics within multi-disciplinary health data promises to bring medical imaging in a quantitative fashion closer to understanding of complex diseases through integrating clinical, structured and imaging-based features[107]. Today, with an appropriate network architecture and often including a transfer learning component, 4th generation radiomics uses supervised end-to-end learning to predict clinical outcomes from complex multi-source data[108]. In addition to single-time-point analyses, the concept of delta radiomics includes temporal change at the feature computation stage and was assessed to predict outcome in lung cancer patients undergoing chemotherapy[109]. Overall, radiomics provides an appealing approach to comprehensively modelling epidemiological, diagnostic and outcome-related data, such that new knowledge can be found and available hypotheses can be systematically expanded based on the data-driven approach.

First, this use case will implement NFDI4Health infrastructure components as a prototypical AI analysis test bed along the radiomics paradigm with the ultimate goal of translating AI to standardisation and validation for broader use. Where applicable, the DICOM standard will be employed to harmonise image data and associated metadata (coordinated with T2.2). We will implement a proof-of-concept for automated AI enabled quality assurance of imaging data as a DICOM service layer within NFDI4Health using the NAKO and SHIP whole-body MRI sub-cohorts (details of data sources: Table A1 Appendix 1). Second, a demonstrator will be carried out to enable predictive pattern recognition in combined imaging and structured epidemiological data. Trained AI modules will be encapsulated in order to be accessible for standardised deployment in variable environments, including combined MRI and structural data from another provenance. Co-applicants already cooperate in the context of two large population-based cohorts (NAKO, SHIP) and the German Consortium for Hereditary Breast and Ovarian Cancer (GC-HBOC[110]), the latter to derive radiomics signatures from combined imaging and structured data to predict the onset of breast cancer within the DFG Priority Programme 2177 “Radiomics: Next Generation of Biomedical Imaging”.

This use case pursues three **major objectives**: (1) adoption of **interoperability and standardisation standards** (T2.2, T2.4) to integrate radiomics/imaging AI, (2) implementation of **data quality services** (T3.3) for population-based imaging studies (**D5.3, M28**), (3) provision of a showcase to extend the **software repository** (T3.6) (**D5.8, M40**) and **distributed data analyses services** (T3.7) towards quantitative image analysis and radiomics (**D5.14, M58**).

Role of members and participants: The leaders of the respective measures will guide the processes described above with regard to metadata publication, data classification and harmonisation, implementation of data standards, data access infrastructures, data linkage approaches and of distributed data analyses infrastructures. The co-applicants and participants lead or participate in population-based studies and clinical trials, maintain monitoring surveys, secondary and registry databases or have experience with record linkage of different data sources and the processing of biomedical image data. They are among others responsible for data management and statistical analysis for a large number of clinical trials and epidemiological studies mainly in academic, but also in commercial settings. Co-applicants will take the responsibility to support the implementation of the NFDI4Health infrastructures and will closely cooperate with participants to disseminate the results, explore further applicability of these infrastructures and improve the performance and scientific value of NFDI4Health measures and infrastructures.

Cooperation with other task areas: Publications of (meta-)data will be performed in cooperation with T2.1 and will be submitted to the central search hub provided in T3.1. Furthermore, the process of retrospective harmonisation of the various data bodies requires expert knowledge about assessment methods and the creation of an analytical framework – TA5 will therefore closely cooperate with T2.2 and T2.3 with regard to data standards and quality. Implementation of data standards will be carried out with supporting tools and based on metadata repositories provided by T2.2, T3.2 and T3.3. Based on specifications provided by T2.4, participating studies and infrastructures will implement standardised interfaces defined for central data access in close cooperation with T3.4. Furthermore, data sets and infrastructures will be adopted to be integrated into distributed data analyses environments set up in T3.7. To address the limited knowledge and experience of potential users and to contribute to the outreach to political decision makers and public health communities, TA5 will also collaborate with TA1 and TA4 in terms of outreach and interaction with users and training and education. With regard to record linkage and data protection issues TA5 will closely collaborate with TA6.

Risks of implementation: Implementation of NFDI4Health infrastructures in individual studies and in secondary/registry data for joint analyses, record linkage and public health monitoring is largely dependent on progress of TA2 and TA3 to define and develop the necessary infrastructures. There might also be governance, ethical and legal challenges (e.g., variants of informed consent, acknowledgement of intellectual property) to implement specific infrastructure components and to integrate the developed services into routine processes and SOPs of CTCs which in turn requires close exchange with TA6 and TA4. Regarding data element harmonisation, not all data elements can be well mapped to medical terminologies, since often clinical assessments and corresponding data collection are specifically designed to match the needs of the corresponding trial.

Table 8: Deliverables of Task Area 5

No.	Type*	Description (month; measure)
D5.1	DOC	Report on outcome metadata publication for participating NCD studies (M12; T5.2)
D5.2	DOC	Compilation of methods for data classification and harmonisation related to dietary data completed (M24; T5.1)
D5.3	SVC	Implementation of automated image quality assurance service successfully applied to two epidemiological cohorts (M28; T5.6)
D5.4	INT	30-50 clinical trials are uploaded to LAPs (M30, T5.4)
D5.5	INT	Catalogue of 1,000 curated and annotated common data elements from different health domains (M36, T5.4)
D5.6	DOC	Protocol for data access infrastructure implementation for at least five NCD studies (M36; T5.2)
D5.7	DOC	Report on exploration and test of linkage approaches in 3 NCD studies (M36; T5.2)
D5.8	SVC	Availability on encapsulated radiomics/imaging AI module (M40; T5.6)
D5.9	DOC	Manuscript on the results of the assessment of record linkage methods (M44; T5.3)
D5.10	DOC	Report on preliminary implementation of distributed data analyses infrastructure (M48; T5.1 and T5.2)
D5.11	DOC	Report on the list of typical characteristics of clinical data sets and experiences made in the implementation and validation (M48, T5.4)
D5.12	INT	Standardised metadata of studies published (M58; T5.1 and T5.2)
D5.13	DOC	Practical guidance on implementation of record linkage in Germany (M58; T5.3)
D5.14	DOC	Report on implementation and test of radiomics / imaging AI use case including survey of international infrastructure landscape to-date (M58; T5.6)
D5.15	DOC	Publication manuscript describing the harmonised framework and metadata models for public surveys and surveillance studies (M60; T5.5)

* DOC: document, INT: interface definitions, APIs, plan designs, architectures, SVC: service, SW: software, EV: event, DISS: dissemination

4.7 Task Area 6 “Privacy & Data Access in Concert” (lead: UM Göttingen, U Bremen)

Data protection regulations have to be taken into account on many levels of the infrastructure developed by NFDI4Health. Moreover, the health data managed by NFDI4Health belong to the so-called special categories of personal data, the processing of which is subject to particularly strict data protection requirements. But nevertheless, data protection law contains a variety of regulatory approaches of data processing for scientific research purposes, which are all aimed at a privileged treatment of such data processing. They include a relaxation of the principle of purpose limitation, legal exceptions from the general prohibition of data processing, the option of so-called broad consent as well as numerous exceptions with regard to the rights of the persons concerned.

Data protection and data security are overarching challenges for all consortia which process personal data and thus synergies can be exploited by close coordination with other consortia (e.g. NFDI4MED and GHGA). Data protection and data security are also cross-sectional issues within NFDI4Health that have to be observed by all TAs. A main focus of TA6 will therefore lie on ensuring that data protection concerns are adequately addressed on all levels of the consortium. This applies in particular to the work in TA3, TA4 (T4.2, T4.5) and TA5.

Furthermore, TA6 will analyse state-of-the-art organisational structures and technical tools. It will develop concepts and solutions required to implement the use cases and adequately protect the various types of data provided by the local data access points (LAPs) (see T3.5). For this, we will subsequently evaluate and roll out governance structures and technical solutions in close collaboration with the MII, NFDI4MED and other consortia. The use cases will serve as test beds for the design and implementation of solutions, in particular T5.4 concerning record linkage. TA6 will primarily address Key Objectives (1), (3) and (4) listed in Section 2.1.

T6.1 “Generic concept and services for data sharing and data protection” (**UM Göttingen (lead), U Bremen**)

We will provide a concept and solutions for managing data use and access processes in a GDPR-compliant manner (General Data Protection Regulation) for ingesting, storing, displaying and sharing data for the use cases in TA5. Each local data access point (LAP) will face the **challenge of confidentiality, integrity and availability of data** which have to be assessed and handled locally, e.g. by using existing information security management processes. Analogously, the central data access point (CAP) in NFDI4Health will rely on existing local

structures, e.g. when forwarding data access requests to the LAPs. Hence, we will **design, implement and evaluate additional services** that are needed to (a) make data discoverable, (b) manage consortium-wide data access requests, (c) design a data and consent broker service and (d) support data usage within the common NFDI4Health research infrastructure. We expect a workflow which is similar to the one illustrated in Table 9 depicting a broker function accessing the well-established data access committees (DACs) procedures at the data holder.

Table 9: Steps in applying for NFDI4Health data: NFDI4Health is acting as a brokering service accessing the well-established local data transfer offices at the data holder site. This reduces complexity massively, as the local data access committees are still in charge.

Step	Description	Actor	Action
1	Discover interesting article/data set in the central search hub	Applicant	Register applicant and search query
2	Check feasibility using the openly available metadata in the central search hub	Applicant	More functionality for registered users
3	Submit data access application	Applicant	Applicant files data request
4	Check availability	Broker access to relevant LAP	Check consent
5	Register application and LAP statement on availability	LAP and CAP	Register application
6	Review the application according to SOPs	LAP (manually, automation intended) and consent broker:	Identify responsible DAC, identify order in case of linked data sets
7	Transfer application to responsible DACs	LAP (manually, automation intended) and CAP	Use workflows, assemble decisions by DACs
8	Check DAC recommendation	Applicant	Accept/revise and resubmit
9	Check if veto against DAC recommendation	Diverse actors, e.g. study PIs, DAC	Approve, reject, specify conditions, correct (error handling)
10	Review DAC decisions	CAP	Inform relevant LAPs, request data transfer agreements
11	Conclude data transfer agreements	LAP, applicant	Inform CAP
12	Final check of consent	LAP	Involve Trusted Third Party (TTP) if needed
13	Pseudonymise data	LAP	TTP ID needed, use NFDI P
14	Facilitate access to selected data sets in adherence to EU GDPR	LAP	Transfer data, enable remote data access, enable distributed data analysis, or offer guest researcher workstation
15	Analyse data according to procedure defined by transfer agreement (using certified algorithms as prescribed)	Applicant	Distributed data analysis, on-site/remote data access, data download/transfer
16	Analyse, evaluate, publish	Applicant	Adhere to Good Scientific Practice and to Good Practice Data Linkage

Step	Description	Actor	Action
17	Return results and documentation to LAP	Applicant	Report back to LAP, delete transferred data
18	Archiving	LAP	Store provided analysis data set and documentation for at least 10 years

The core of the concept will be a detailed description of the data sharing processes in NFDI4Health with a specific focus on the use cases including a Business Process Model and Notation (BPMN) (**D6.1, M12**). From this, we will derive key services which will be implemented prototypically by adopting proven methods e.g. from the German Centre for Cardiovascular Diseases (DZHK)[111] and several MII-projects[112] (**D6.5, M24**). Our framework will cover solutions for assessing protection needs for different data types considering structured consent information provided by the primary data source along with further metadata.

To support this assessment, we plan to create a plugin for the NFDI4Health CAP (see D3.4). The plugin will support data holders in selecting an appropriate set of variables depending on legal conditions, scientific requirements, data types/formats and re-identification risks (see T6.3; **D6.9, M36**). Depending on data protection requirements, the concept and plugin will support different protection methods, including the separation of data according to the well-established TMF generic data protection guideline[113] (**D6.12, M58**).

T6.2 "Legal framework, esp. common consent standards and data access procedures" (**U Bremen (lead), TMF, UM Göttingen**)

T6.2 carves out the **legal framework conditions** for passing on research data in compliance with data protection regulations. As illustrated in TA5, health data can be generated in different ways (clinical, epidemiological, public health studies). Accordingly, depending on the nature and origin of the data, the legal requirements for sharing data differ as well. To begin with, the criteria for a differentiation between anonymised, pseudonymised, and person-identifying data will therefore be determined in accordance with T6.1 (**D6.2, M12**). In addition, a concept for the differentiation of data depending on legal provisions permitting their further use will be developed (**D6.6, M24**). Based on this, the corresponding common consent standards will be assessed.

In so far as the processing of data requires the consent of the person concerned, the structured consent-information which has been extracted from the primary data source (T6.1) has to be taken into account. In general, NFDI4Health data are always covered by a specific trial consent.

In the future, more emphasis will be placed on the legal instrument of broad consent for reusing data beyond the scope of the initial consent. This allows for data processing for scientific purposes in less definite terms, and instead refers to certain areas of scientific research in general terms (cf. Recital 33 of GDPR).

Otherwise, if the sharing of data is based on the data protection privileges established by the GDPR and national data protection regulations, it is above all the requirement to provide "**suitable and specific measures** to safeguard the fundamental rights and the interests of the data subject" (Art. 9 (2) lit. j GDPR) which has to be implemented within the framework of the development of data access procedures. In particular, measures of anonymisation and pseudonymisation will have to be taken into consideration in this context (if compatible with the purpose of the research). Data protection law provides a number of additional protective measures which can ensure that health data are processed in compliance with regulations, e.g. the revisionability and transparency of data processing as well as access restrictions (see § 22 (2) BDSG). T6.2 will act as an enabler for exploring new ways to reuse research data in a privacy-protecting manner (**D6.14, M60; D6.15, M60**). Therefore, T6.2 will set up a CAP streamlining the application and decision processes. We will closely collaborate with the MII ZARS (Central Application and Registry[114]) and NFDI4MED.

In principle, the new European data protection law is supportive of research and thus also allows the processing of health data in accordance with the FAIR principles by a consortium such as NFDI4Health. However, this research-supportive approach of European law has hardly been reflected by national law up to now. The legal regulations on research data processing are mostly patchwork, composed of various specific provisions of Federal and State law (medical law, hospital law, higher education law, social law etc.). One of the long-term objectives of NFDI4Health is therefore the promotion of **harmonised regulation of research data processing**. This includes the promotion of a change of perspective away from project-oriented and towards institution-oriented privileges for research data processing. Besides that, any kind of concept which is based on the FAIR principles will also have to take into account intellectual property rights (IPR) of data holders. Thus, future frameworks regulating the use of and access to research data should pursue a comprehensive approach covering not only issues of privacy and data protection but also general questions of rights in data (**D6.13, M60**).

T6.3 "Development of concepts and methods for privacy risk assessment" (**Charité/BIH (lead), U Bremen, UM Göttingen**)

We will develop a concept and provide tools enabling the NFDI4Health data holders to analyse the privacy risks of the data sets contributed to the infrastructure. For this purpose, we will suggest a structured risk and threat analysis methodology based on existing guidelines and quantitative measures derived from input data sets. The method will consider different types of privacy risks (e.g. singling out, linkage and inference) derived from legal and regulatory requirements as well as different attacker scenarios, covering the complete spectrum from (1) attackers with strong background knowledge targeting specific individuals to (2) attackers with little background knowledge trying to maximise re-identification probabilities. Tools will be provided to calculate these quantitative measures from data managed by NFDI4Health considering aspects such as the relationship between a data set and the population from which it was sampled as well as the stability, availability and reproducibility of data items that make individuals unique (**D6.7, M24**).

As a first step, we will create an overview of common approaches to measuring and reducing re-identification risks. In close cooperation with T6.1 we will analyse common processes of data sharing and relate the various steps to different privacy risk and threat models. The process of making data discoverable, specifying a data request application, assessing the risk of a data release, the data transfer itself and finally the risks along using the data in an external context will be analysed in detail. Here, we will reuse results from the well-established DZHK use and access blueprint[115] being operated at UM Göttingen.

We will then integrate selected tools, such as the ARX Data Anonymisation Tool[116] or sdcMicro[117], into the NFDI4Health infrastructure to enable LAPs to assess privacy risks in different contexts taking different aspects of data sharing into account. As an example, we will perform a structured risk and threat analysis for data sets from the NFDI4Health use cases resulting in a privacy impact assessment for the process of data sharing (**D6.10, M36**). After successful evaluation, customisation and integration into the NFDI4Health infrastructure these tools will be operated by the LAPs and ZB MED, respectively (see TA3).

T6.4 “Synthetic Health Data” (**Fraunhofer SCAI (lead), Charité/BIH**)

To enable users of the NFDI4Health infrastructure to implement data analyses in a privacy-preserving manner, e.g. using the PHT infrastructure, access to appropriate non-personal data sets is needed. Synthetic data generation, where sufficiently realistic but non-personal data are derived from an input data set, is an effective tool to support data sharing across the different organisations. Moreover, synthetic data allow for data mining and model development without having access to real-world data, e.g. due to legal constraints. This can help to determine the

feasibility of a planned project, and it can help to derive hypotheses and models that can later on be tested with real data.

In recent work[118], we have developed an AI-based approach called Variational Autoencoder Modular Bayesian Networks (VAMBN) that learns a generative model of longitudinal clinical data, which allows for generating realistic synthetic subject data. VAMBM takes into account typical features of clinical study or registry data, namely i) limited sample size; ii) many variables of different numerical scales and statistical properties; and iii) longitudinal data with many missing values.

Finally, using the established **concept of differential privacy**[119], which will also be covered by the risk assessment guidelines developed in T6.3, we have implemented and tested the possibility to train VAMBN models in a way that grants strict guarantees about the probability to derive personal data from the AI model.

The primary goal of TA6.4 is to provide a containerised software tool based on existing prototypical implementations of methods for synthetic data generation that can then be used by the partners in NFDI4Health to support data sharing (**D6.3, M12**). The software is envisioned to be compatible with the PHT infrastructure developed in TA3, e.g. by making sure that containerisation is compatible and that VAMBN models can be trained in a distributed manner as the same study is usually run at different sites in parallel. Moreover, the software will support the generation of appropriate quality control measures and figures, e.g. marginal distributions of variables and learned correlation structures. A close collaboration with TA3 (T3.4 and T3.7) will be required to reach these goals. In particular VAMBN will need access to the metadata repository developed in TA3 as well as to the computing facilities at each LAP. The software implementation will enable training of VAMBN models on a dedicated compute server at each of the LAP sites. Validation will be done with publicly available data sets, e.g. from the Parkinson's Progression Marker Initiative[120] and the ADNI TADPOLE challenge[121] (**D6.11, M48**). The tool will allow for training VAMBN models under differential privacy guarantees, which involves specifying a risk threshold: the stricter the threshold, the stronger the differences between synthetic data and training data, thus implementing a risk-utility trade-off. In cooperation with TA6.3 we will extend our tool to produce informative plots that allow the user to understand the trade-off between the degree of privacy protection and loss of accuracy of output data (**D6.8, M36**). Accordingly, the user will then be guided to choose a reasonable compromise.

T6.5 “Best practice of record/data linkage with regard to data privacy and data protection requirements” (**BIPS (lead), TMF (co-lead), DIfE, RKI, U Bremen, UM Göttingen, UM Greifswald**; participant: U Magdeburg)

Besides primary databases resulting from epidemiological or public health studies and clinical trials there is an at least partly unretrieved treasure trove containing valuable health-related information on an individual level in Germany, as for instance claims data stored by statutory health insurances, data recorded by pension funds, occupational history data stored by the Institute for Employment Research (IAB), and data from disease registries as e.g. (population-based) cancer registries. By combining different administrative data we can answer questions that require huge sample sizes. With such data we can generate evidence with a high level of external validity and high relevance for policy makers, also because they include information on hard-to-reach populations. However, since these social security and registry data are highly sensitive, rigid legislations protect them against any potential misuse. Current regulations make it challenging to reuse social security data for research purposes, particularly if done without obtaining individual consent. Asking millions of study subjects whose individual-level data are stored in secondary databases for informed consent may neither be feasible nor scientifically acceptable: asking for informed consent would lead to a highly selected sample due to non-responder bias which would in turn limit the validity and generalisability of research results. In Germany, social security data are protected by § 75 Social Code Book (SGB) X. This article allows the transfer of such data for research purposes without informed consent only upon application for a specified research project or research area and only for limited time periods, given that the legitimate rights for individual confidentiality of an affected party are not impaired or that the public interest in this research outweighs the right for confidentiality of an affected party by far.

Thus, alone the access to a single secondary data source is highly restricted and requires detailed justification and approval by the respective regulatory authority as well as by the data owner. **Moreover, linking secondary data sources** with each other or with primary databases is **extremely difficult** and time-consuming – hence often impossible – in Germany. Here we face two major challenges. First, linkage of personal data requires the **informed consent** of study subjects unless there exists a specific legal regulation. Second, the information entailed in the data sources to be linked often does not allow a deterministic linkage. Thus, record linkage is not only a legal problem but also a technical issue, since, unlike in many other countries, in Germany health-related data are stored without a unique identifier. This means for instance that an individual record from a cancer registry cannot be directly linked to the claims data record of

the same individual stored by a statutory health insurance. These data sources use different algorithms for generating and encrypting the personal information leading to different pseudonyms. However, depending on the applied procedure, record linkage approaches often entail the risk of introducing linkage errors[94]. Complementary to the **use case T5.3**, this measure will focus on the technical and legal framework that would be necessary to achieve a more research-friendly approach for record linkage across different data sources in Germany. T6.5 will pursue the following objectives. (1) **It will assess different linkage approaches for privacy preserving record linkage (PPRL)** that have been successfully implemented in Germany and other European countries, including probabilistic and deterministic linkage methods, depending on the research purpose, the structure and source of the data to be linked as well as the legal, organisational and technical environment, and particularly the EU GDPR privacy regulations. (2) **It will identify best practice approaches** (**D6.4, M12**) in consideration of their compliance with the GDPR while taking into account state-of-the-art scientific requirements. The experiences made in trying to link various health-related data sources in Germany in previous studies will be evaluated in close collaboration with the use case described in T5.3 (see also March et al.[93]). (3) Based on this evaluation, **solutions developed in the German legal and administrative context will be benchmarked** against the identified best practice linkage approaches to propose practical improvements that can be adapted to the situation Germany. (4) In close exchange with T5.3, respective **legislative demands will be formulated**.

Role of members and participants: UM Göttingen specialises in data access and analytics services and leads this TA. All members of TA6, in particular BIPS and TMF, will perform the processes described above with regard to legal support and cooperation with other TAs for all questions of privacy and data protection. Fraunhofer SCAI will take care of VAMBN software implementation, including quality measures for virtual subjects. ZB MED will provide central services (CAP). Furthermore, Charité/BIH adds profound knowledge in privacy-preserving record linkage, re-identification risk analyses and data anonymisation, U Bremen provides profound long-term experience in privacy questions and co-leads this TA.

Cooperation with other task areas: Privacy and data protection, as cross-sectional subjects, will be dealt with in close cooperation with all other TAs. TA6 has thus above all an advisory and supportive function since the legal conformity of the management of a data infrastructure is a prerequisite for NFDI4Health as a whole. This applies first of all to the cooperation with TA3 and

TA4 (T4.2, T4.5). In addition, legal support will be provided for TA5, since the respective legal framework conditions must also be considered in the development of different use cases.

Risks of implementation: (1) One potential risk is the significantly delayed access to project data via TA3, such that some technical solutions cannot be rolled out in time. To mitigate this risk we will test our demonstrator on publicly available data first (e.g. PPMI, TADPOLE). (2) To ensure that the computational resources within TA3 are sufficient for building and testing complex AI models, we will work closely together with TA3 to explain our needs (parallel computing infrastructure/GPU server). By running our software on publicly available data we will in any case be able to produce synthetic health data. (3) An additional risk may result from the fact that the concept of synthetic subject data will not be understood by project partners and will thus have no impact. To counteract this major problem we will offer training workshops to introduce project partners into the concept of VAMBN and explain the features of our software. (4) Finally, it may happen that no broad consent concept will be available which means that we will continue working with specific consent and that we use secure distributed computing and differential privacy approaches.

Table 10: Deliverables of Task Area 6

No.	Type*	Description (month; measure)
D6.1	INT	Detailed harmonised process descriptions for data sharing (BPMN and text) (M12; T6.1)
D6.2	DOC	“Legal map” of the applicable legal framework conditions for research data processing (M12; T6.2)
D6.3	SW	Containerised implementation of the VAMBN approach compatible with the PHT infrastructure (M12; T6.4)
D6.4	DOC	White paper on the status quo and the demands of record linkage in Germany (M12; T6.5)
D6.5	SVC	Prototypes for prioritised services, test bed (together with TA3 services and TA5 use cases) (M24; T6.1)
D6.6	DOC	Classification of different categories of health data based on risk assessment and legitimation basis for data processing (M24; T6.2)
D6.7	DOC	Guideline for re-identification risk assessment in NFDI4Health (M24; T6.3)
D6.8	SW	Implement visualisations and controls for specifying risk/utility trade-offs in a user-friendly manner (M36; T6.4)
D6.9	DOC	Report on using NFDI4Health CAP and harmonised NFDI4MED/MII ZARS service (M36; T6.1)
D6.10	SW	Integration of tools for privacy risk analyses into NFDI4Health and exemplary application to data from the use cases (M36; T6.3)
D6.11	DISS	Prototype of a synthetic cohort data set available (M48; T6.4)
D6.12	DOC	Report on the overall results concerning integration depth (M58; T6.1)

No.	Type*	Description (month; measure)
D6.13	DOC	Recommendations for a consistent legal framework balancing data protection and freedom of research as well as regulating rights in data not only from a data protection but also from an intellectual property perspective (M60; T6.2)
D6.14	INT	Common consent standards for further use of data (M60; T6.2)
D6.15	DOC	Catalogue of “suitable and specific measures” to safeguard privacy in accordance with data protection regulations (M60; T6.2)

* DOC: document, INT: interface definitions, APIs, plan designs, architectures, SVC: service, SW: software, EV: event, DISS: dissemination

Appendix

1 Bibliography and list of references

List of references

1 Arning U, Lindstädt B, Schmitz J. [PUBLISSO: an open access publication portal for life sciences]. GMS Med Bibl Inf 2016; 16(3): Doc15.

2 Poley C. [LIVIVO: new challenges for ZB MED’s search portal for life sciences]. GMS Med Bibl Inf 2016; 16(3): Doc21.

3 German Network for Bioinformatics Infrastructure. Available from: https://www.denbi.de/ (accessed 09 Oct 2019).

4 ELIXIR. Available from: https://elixir-europe.org (accessed 02 Oct 2019).

5 FAIRDOM. Available from: https://fair-dom.org (accessed 02 Oct 2019).

6 Wolstencroft K, Owen S, Krebs O, Nguyen Q, Stanford NJ, Golebiewski M et al. SEEK: A systems biology data and model management platform. BMC Syst Biol 2015; 9: 33.

7 International Organization for Standardization. ISO/TC 276 Biotechnology. Available from: https://www.iso.org/committee/4514241.html (accessed 10 Oct 2019).

8 EU-STANDS4PM. Available from: https://www.eu-stands4pm.eu (accessed 09 Oct 2019).

9 Liver Systems Medicine LiSyM. Available from: http://www.lisym.org (accessed 09 Oct 2019).

10 Hucka M, Nickerson DP, Bader GD, Bergmann FT, Cooper J, Demir E et al. Promoting coordinated development of community-based information standards for modeling in biology: The COMBINE initiative. Front Bioeng Biotechnol 2015; 3: 19.

11 Reseach Data Centre of the Robert Koch Institute. Available from: https://www.rki.de/fdz (accesses 01 Oct 2019).

12 Cologne Competence Center for Research Data Management C3RDM. Available from: http://fdm.uni-koeln.de (accessed 01 Oct 2019).

13 Winter A, Stäubert S, Ammon D, Aiche S, Beyan O, Bischoff V et al. Smart Medical Information Technology for Healthcare (SMITH). Methods Inf Med 2018; 57(S 01): e92-e105.

14 Leipzig Health Atlas. Available from: https://www.health-atlas.de (accessed 30 Sep 2019).

15 Meineke FA, Löbe M, Stäubert S. Introducing technical aspects of research data management in the Leipzig Health Atlas. Stud Health Technol Inform 2018; 247: 426-430.

16 International Organization for Standardization. Available from: https://www.iso.org/home.html (accessed 11 Oct 2019).

17 International Electrotechnical Commission. Available from: https://www.iec.ch (accessed 11 Oct 2019).

18 CEN/CENELEC. Available from: http://www.cencenelec.eu (accessed 11 Oct 2019).

19 DIN. Available from: https://www.din.de/en (accessed 11 Oct 2019).

20 HL7 Standards. Available from: https://www.hl7.org (accessed 11 Oct 2019).

21 COST action CHARME. Available from: https://www.cost-charme.eu (accessed 11 Oct 2019).

22 International Organization for Standardization. ISO/TC 215 Health informatics. Available from: https://www.iso.org/committee/54960.html (accessed 11 Oct 2019).

23 DIN Standards Committee Medicine. Available from: https://www.din.de/en/getting-involved/standards-committees/named (accessed 11 Oct 2019).

24 Joint Programming Initiative A Healthy Diet for a Healthy Life (JPI HDHL). Available from: https://www.healthydietforhealthylife.eu/ (accessed 01 Oct 2019).

25 FAIRsharing. A curated, informative and educational resource on data and metadata standards, inter-related to databases and data policies. Available from: https://fairsharing.org (accessed 02 Oct 2019).

26 FAIR4Health. Available from: https://www.fair4health.eu (accessed 11 Oct 2019).

27 DataSHIELD. Available from: http://www.datashield.ac.uk (accessed 01 Oct 2019).

28 Pinart M, Nimptsch K, Bouwman J, Dragsted LO, Yang C, De Cock N et al. Joint data analysis in nutritional epidemiology: Identification of observational studies and minimal requirements. J Nutr 2018; 148: 285-297.

29 Vitali F, Lombardo R, Rivero D, Mattivi F, Franceschi P, Bordoni A et al., on behaf of the ENPADASI consortium. ONS: An ontology for a standardized description of interventions and observational studies in nutrition. Genes Nutr 2018; 13: 12.

30 InterConnect. Global data for diabetes and obesity research. Available from: www.interconnect-diabetes.eu (accessed 02 Oct 2019).

31 MyNewGut. Available from: http://www.mynewgut.eu (accessed 01 Oct 2019).

32 STRATOS initiative. Available from: http://www.stratos-initiative.org (accessed 01 Oct 2019).

33 Standards and tools for data monitoring in complex epidemiological studies. Available from: https://gepris.dfg.de/gepris/projekt/315057723 (accessed 01 Oct 2019).

34 euCanSHare. The consortium. Available from: http://www.eucanshare.eu/consortium (accessed 01 Oct 2019).

35 DataCite Schema. Available from: https://schema.datacite.org (accessed 01 Oct 2019).

36 CDISC. Available from: https://www.cdisc.org (accessed 11 Oct 2019).

37 HL7 FHIR. Available from: www.hl7.org/fhir/ (accessed 11 Oct 2019).

38 PubMed. Available from: https://www.ncbi.nlm.nih.gov/pubmed (accessed 01 Oct 2019).

39 LIVIVO. Available from: https://www.livivo.de (accessed 01 Oct 2019).

40 FAIRDOM. Project Area Liaisons. Available from: https://fair-dom.org/communities/pals (accessed 02 Oct 2019).

41 International Organization for Standardization. ISO/WD 20691. Biotechnology — Requirements for data formatting and description in the life sciences for downstream data processing and integration workflows. In preparation in ISO/TC 276 Biotechnology. Available from: https://www.iso.org/standard/68848.html (accessed 09 Oct 2019).

42 EMBL-EBI. Available from: https://www.ebi.ac.uk (accessed 02 Oct 2019).

43 Wimalaratne SM, Juty N, Kunze J, Janée G, McMurry JA, Beard N et al. Uniform resolution of compact identifiers for biomedical data. Sci Data 2018; 5: 180029.

44 CDISC. ODM-XML. Available from: https://www.cdisc.org/standards/data-exchange/odm (accessed 02 Oct 2019).

45 CDISC. BRIDG. Available from: https://www.cdisc.org/standards/domain-information-module/bridg (accessed 02 Oct 2019).

46 Baum B, Bauer CR, Franke T, Kusch H, Parciak M, Rottmann T et al. Opinion paper: Data provenance challenges in biomedical research. it - Information Technology 2017; 59(4).

47 Parciak M, Bauer C, Bender T, Lodahl R, Schreiweis B, Tute E et al. Provenance solutions for medical research in heterogeneous IT-infrastructure: An implementation roadmap. Stud Health Technol Inform 2019; 264: 298-302.

48 Daumke P, Heitmann KU, Heckmann S, Martínez-Costa C, Schulz S. Clinical text mining on FHIR. Stud Health Technol Inform 2019; 264: 83-87.

49 Research Data Alliance. FAIR Data Maturity Model WG. https://www.rd-alliance.org/groups/fair-data-maturity-model-wg (accessed 11 Oct 2019).

50 FAIRDOM. SEEK. Available from: https://fair-dom.org/platform/seek (accessed 02 Oct 2019).

51 Virtual Liver Network. Available from: https://www.h-its.org/projects/virtual-liver (accessed 09 Oct 2019).

52 Henney A, Coaker H. The Virtual Liver Network: systems understanding from bench to bedside. Future Med Chem 2014; 6(16): 1735-1740.

53 ERA CoBioTech. Available from: https://www.cobiotech.eu (accessed 09 Oct 2019)

54 Bauer CR, Knopp C, Bender T, Kusch H, Sax U. Application of basic research data management with FAIRDOM/SEEK from a medical informatics perspective. In: Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (ed.) 63. Jahrestagung der GMDS: Osnabrück, 02.-06.09.2018. Düsseldorf: German Medical Science GMS Publishing House; 2018. Available from: https://www.egms.de/static/en/meetings/gmds2018/18gmds093.shtml (accessed 09 Oct 2019).

55 Haarbrandt B, Schreiweis B, Rey S, Sax U, Scheithauer S, Rienhoff O et al. HiGHmed - An open platform approach to enhance care and research across institutional boundaries. Methods Inf Med 2018; 57(S 01): e66-e81.

56 Leibniz Institute for Prevention Research and Epidemiology – BIPS. Instruments for health surveys in children and adolescence. Available from: https://www.bips-institut.de/en/pages/ifhs.html (accessed 02 Oct 2019).

57 eStandards. Available from: http://www.estandards-project.eu (accessed 09 Oct 2019).

58 euCanSHare. Available from: http://www.eucanshare.eu (accessed 09 Oct 2019).

59 FAIRDOM. Data management checklist. Available from: https://fair-dom.org/knowledgehub/data-management-checklist (accessed 09 Oct 2019).

60 Medical Informatics Initiative. Data Sharing Working Group – Uniform use and access policy key issues paper. Available from: https://www.medizininformatik-initiative.de/sites/default/files/inline-files/MII_03_Use_and_Access_Policy_Key_Issues_Paper_1-0.pdf (accessed 09 Oct 2019).

61 World Health Organization. International Clinical Trials Registry Platform (ICTRP). Available from: https://www.who.int/ictrp/unambiguous_identification/en (accessed 09 Oct 2019).

62 FAIRMetrics. Metrics. Available from: http://fairmetrics.org (accessed 09 Oct 2019).

63 Moreau L, Missier P (eds.). PROV-DM: The PROV data model. Available from: https://www.w3.org/TR/2013/REC-prov-dm-20130430/ (accessed 09 Oct 2019).

64 HL7 FHIR. Available from: https://hl7.org/implement/standards/fhir/provenance.html (accessed 09 Oct 2019).

65 ART-DECOR. Available from: www.art-decor.org (accessed 09 Oct 2019).

66 Simplifier. The FHIR registry. Available from: www.simplifier.net (accessed 09 Oct 2019).

67 Maelstrom Research. Available from: https://www.maelstrom-research.org/ (accessed 10 Oct 2019).

68 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(5): 1372-1378.

69 Integrative Data Semantics for Neurodegenerative research. Available from: https://www.idsn.info/en/idsn.html (accessed 10 Oct 2019).

70 Schmidt CO, Krabbe C, Schössow J, Albers M, Radke D, Henke J. Square² - A web application for data monitoring in epidemiological and clinical studies. Stud Health Technol Inform 2017; 235: 549-553.

71 Lodahl R, Bauer CR, Baum B, Bender T, Parciak M, Krawczak M et al. Enabling pedigree visualization and analysis in tranSMART. Stud Health Technol Inform 2018; 253: 75-79.

72 Bauer CR, Knecht C, Fretter C, Baum B, Jendrossek S, Rühlemann M et al. Interdisciplinary approach towards a systems medicine toolbox using the example of inflammatory diseases. Brief Bioinform 2017; 18(3): 479-487.

73 Herschel M, Diestelkämper R, Ben Lahmar H. A survey on provenance: What for? What form? What from? The VLDB Journal 2017; 26: 881.

74 Hoekstra R, Groth P. PROV-O-Viz. Understanding the role of activities in provenance. In: Ludäscher B, Plale B (eds). Provenance and annotation of data and processes. Lecture Notes in Computer Science. Cham: Springer International Publishing; 2015. p. 215-220.

75 Beyan O, Choudhury A, van Soest J, Zimmermann L, Stenzhorn H, Dumontier M et al. Distributed analytics on sensitive medical data: The Personal Health Train. Data Intelligence (Special Issue on Emerging FAIR Practices 1) 2019 (accepted for publication).

76 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(6): 1929-1944.

77 Budin-Ljøsne I, Burton P, Isaeva J, Gaye A, Turner A, Murtagh MJ et al. DataSHIELD: An ethically robust solution to multiple-site individual-level data analysis. Public Health Genomics 2015; 18(2): 87-96.

78 MII consortium SMITH. Available from: https://www.smith.care/ (accessed 10 Oct 2019).

79 MII consortium DIFUTURE. Available from: https://difuture.de (accessed 10 Oct 2019).

80 Association of the Scientific Medical Societies in Germany. Available from: https://www.awmf.org/en (assessed 12 Oct 2019).

81 FAIRDOMHub. Available from: https://fairdomhub.org (accessed 12 Oct 2019).

82 FAIRsharing. Collections. Available from: https://fairsharing.org/collections (accessed 12 Oct 2019)

83 Rat für Informationsinfrastrukturen (RfII). Digitale Kompetenzen - dringend gesucht! Empfehlungen zu Berufs- und Ausbildungsperspektiven für den Arbeitsmarkt Wissenschaft. Göttingen: RfII; 2019.

84 Fournier J. [For the qualified handling of research data. A report on the workshop “Wissenschaft im digitalen Wandel” (6 June 2017) at the University Mannheim. o-bib Das offene Bibliotheksjournal 2017; 4(3): 88-93.

85 Teperek M, Cruz MJ, Verbakel E, Böhmer J, Dunning A. Data stewardship addressing disciplinary data management needs. Int J Digital Curation 2018; 13: 141-149.

86 Gagliardi AR, Kothari A, Graham ID. Research agenda for integrated knowledge translation (IKT) in healthcare: What we know and do not yet know. J Epidemiol Community Health 2017; 71(2): 105-106.

87 Cloudy with a Chance of Pain. Available from: https://www.cloudywithachanceofpain.com (accessed 12 Oct 219).

88 Lakerveld J, van der Ploeg HP, Kroeze W, Ahrens W, Allais O, Andersen LF et al., on behalf of the DEDIPAC consortium. Towards the integration and development of a cross-European research network and infrastructure: The DEterminants of DIet and Physical ACtivity (DEDIPAC) Knowledge Hub. Int J Behav Nutr Phys Act 2014; 11: 143.

89 Arfè A, Scotti L, Varas-Lorenzo C, Nicotra F, Zambon A, Kollhorst B et al., on behalf of the SOS project consortium. Non-steroidal anti-inflammatory drugs and risk of heart failure in four European countries: Nested case-control study. BMJ 2016; 354: i4857.

90 Poluzzi E, Diemberger I, de Ridder M, Koci A, Clò M, Oteri A et al. Use of antihistamines and risk of ventricular tachyarrhythmia: A nested case-control study in five European countries from the ARITMO project. Eur J Clin Pharmacol 2017; 73: 1499-1510.

91 Herrmann WJ, Weikert C, Bergmann M, Boeing H, Katzke VA, Kaaks R et al. Assessing incident cardiovascular and metabolic diseases in epidemiological cohort studies in

Germany. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2018; 61: 420-431.

92 Neuhauser H, Diederichs C, Boeing H, Felix SB, Jünger C, Lorbeer R et al. Hypertension in Germany. Dtsch Arztebl Int 2016; 113: 809-815.

93 March S, Antoni M, Kieschke J, Kollhorst B, Maier B, Müller G et al. [Quo vadis data linkage in Germany? An initial inventory]. Gesundheitswesen 2018; 80: e20-e31.

94 Harron KL, Doidge JC, Knight HE, Gilbert RE, Goldstein H, Cromwell DA et al. A guide to evaluating linkage quality for the analysis of linked data. Int J Epidemiol 2017; 46: 1699-1710.

95 Schnell R, Bachteler T, Reiher J. Privacy-preserving record linkage using Bloom filters. BMC Med Inform Decis Mak 2009; 9: 41.

96 Schnell R, Borgs C. Secure privacy preserving record linkage of large databases by modified Bloom filter encodings. Int J Popul Data Sci 2017; 1: 013.

97 Contiero P, Tittarelli A, Tagliabue G, Maghini A, Fabiano S, Crosignani P et al. The EpiLink record linkage software: Presentation and results of linkage test on cancer registry files. Methods Inf Med 2005; 44: 66-71.

98 Ohlmeier C, Hoffmann F, Giersiepen K, Rothgang H, Mikolajczyk R, Appelrat HJ et al.. [Linkage of statutory health insurance data with those of a hospital information system: feasible, but also "useful"?]. Gesundheitswesen 2015; 77: e8-e14.

99 Ohlmeier C, Langner I, Garbe E, Riedel O. Validating mortality in the German Pharmacoepidemiological Research Database (GePaRD) against a mortality registry. Pharmacoepidemiol Drug Saf 2016; 25: 778-784.

100 Stallmann C, Ahrens W, Kaaks R, Pigeot I, Swart E, Jacobs S. [Individual linkage of primary data with secondary and registry data within large cohort studies - capabilities and procedural proposals]. Gesundheitswesen 2015; 77: e37-e42.

101 Schmidt CO, Reber K, Baumeister SE, Schminke U, Völzke H, Chenot JF. [Integration of primary and secondary data in the Study of Health in Pomerania and description of clinical outcomes using stroke as an example]. Gesundheitswesen 2015; 77: e20-e25.

102 Wahrendorf M, Marr A, Antoni M, Pesch B, Jöckel KH, Lunau T et al. Agreement of self-reported and administrative data on employment histories in a German cohort study: A sequence analysis. Eur J Popul 2019; 35: 329-346.

103 March S, Rauch A, Thomas D, Bender S, Swart E. Procedures according to data protection laws for coupling primary and secondary data in a cohort study: The lidA study. Gesundheitswesen 2012; 74: e122-e129.

104 March S. Individual data linkage of survey data with claims data in Germany - An overview based on a cohort study. Int J Environ Res Public Health 2017; 14: 1543.

105 Hebestreit A, Thumann B, Wolters M, Bucksch J, Huybrechts I, Inchley J et al. on behalf of the DEDIPAC consortium. Road map towards a harmonized pan-European surveillance of obesity-related lifestyle behaviours and their determinants in children and adolescents. Int J Public Health 2019; 64: 615-623.

106 Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, Schabath MB et al. Radiomics: The process and the challenges. Magn Res Imaging 2012; 30: 1234-1248.

107 Gillies RJ, Kinahan PE, Hricak H. Radiomics. Images are more than pictures, they are data. Radiology 2016; 278: 563-577.

108 Hosny A, Parmar C, Coroller TP, Grossmann P, Zeleznik R, Kumar A et al. Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. PLoS Med 2018; 15: e1002711.

109 Fave X, Zhang L, Yang J, Mackin D, Balter P, Gomez D et al. Delta-radiomics features for the prediction of patient outcomes in non-small cell lung cancer. Sci Rep 2017; 7: 588.

110 Schmutzler RK, Rhiem K, Breuer P, Wardelmann E, Lehnert M, Coburger S et al. Outcome of a structured surveillance programme in women with a familial predisposition for breast cancer. Eur J Cancer Prev 2006; 15: 483-489.

111 Deutsches Zentrum für Herz-Kreislauf-Forschung. Verfahrensbeschreibung und Datenschutzkonzept des Zentralen Datenmanagements des DZHK. Available from: https://dzhk.de/fileadmin/user_upload/Datenschutzkonzept_des_DZHK.pdf (accessed 11 Oct 2019).

112 HiGHmed. Important milestone reached: Initial HiGHmed privacy concept approved by TMF Data Protection working group. Available from: https://www.highmed.org/news/important-milestone-reached-initial-highmed-privacy-concept-approved-by-tmf-data-protection-working-group (accessed 11 Oct 2019).

113 Pommerening K, Drepper J, Helbing K, Ganslandt T. [Guideline for data protection in medical research projects]. Berlin: TMF; 2014.

114 Medical Informatics Initiative. [Report on the first phase]. Available from: https://www.medizininformatik-initiative.de/sites/default/files/2019-04/MII_Semler_DMEA_11-04-2019.pdf (accessed 11 Oct 2019).

115 Deutsches Zentrum für Herz-Kreislauf-Forschung. Use and access. Available from: https://dzhk.de/en/research/clinical-research/use-and-access (accessed 11 Oct 2019).

116 Prasser F, Kohlmayer F. Putting statistical disclosure control into practice: The ARX Data Anonymization Tool. In: Gkoulalas-Divanis A, Loukides G (eds.). Medical data privacy handbook. Cham: Springer International Publishing; 2015. p. 111-148.

117 Templ M, Kowarik A, Meindl B. Statistical disclosure control for micro-data using the R package sdcMicro. J Stat Softw 2015; 67(1): 1-36.

118 Gootjes-Dreesbach L, Sood M, Sahay A, Hofmann-Apitius M, Fröhlich H. Variational Autoencoder Modular Bayesian Networks (VAMBN) for simulation of heterogeneous clinical study data. 2019. Available from: https://www.biorxiv.org/content/10.1101/760744v1 (accessed 11 Oct 2019).

119 Dwork C, McSherry F, Nissim K, Smith A. Calibrating noise to sensitivity in private data analysis. In: Halevi S, Rabin T (eds.). Theory of Ccryptography. Berlin Heidelberg: Springer; 2006. p. 265-284.

120 Parkinson's Progression Marker Initiative. Available from: http://www.ppmi-info.org (accessed 11 Oct 2019).

121 Alzheimer's Disease Prediction Of Longitudinal Evolution (TADPOLE) Challenge. Available from: https://tadpole.grand-challenge.org (accessed 11 Oct 2019).

122 Perez-Riverol Y, Ternent T, Koch M, Barsnes H, Vrousgou O, Jupp S et al. OLS client and OLS dialog: Open source tools to annotate public omics datasets. Proteomics 2017; 17(19).

123 Gonçalves RS, O'Connor MJ, Martínez-Romero M, Egyedi AL, Willrett D, Graybeal J et al. The CEDAR Workbench: An ontology-assisted environment for authoring metadata that describe scientific experiments. 2019. Available from: https://arxiv.org/abs/1905.06480 (accessed 09 Oct 2019).

124 Wolstencroft K, Owen S, Horridge M, Krebs O, Mueller W, Snoep JL et al. RightField: embedding ontology annotation in spreadsheets. Bioinformatics 2011; 27(14): 2021-2022.

125 Gleim LC, Karim MdR, Zimmermann L, Kohlbacher O, Stenzhorn H, Decker S et al. Enabling ad-hoc reuse of private data repositories through schema extraction. J Biomed Semantics 2019 (accepted for publication).

126 Scheufele E, Aronzon D, Coopersmith R, McDuffie MT, Kapoor M, Uhrich CA et al. tranSMART: An open source knowledge management and high content data analytics platform. AMIA Jt Summits Transl Sci Proc 2014; 2014: 96-101.

127 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.