Please address the following aspects in your letter of intent

1. Binding letter of intent as advance notification or non-binding letter of intent

☒	Binding letter of intent (required as advance notification for proposals in 2019)
☐	Non-binding letter of intent (anticipated submission in 2020)
☐	Non-binding letter of intent (anticipated submission in 2021)

2. Formal details

- Planned name of the consortium
  National Research Data Infrastructure for Personal Health Data
- Acronym of the planned consortium
  NFDI4Health

- Applicant institution
  ZB MED Information Centre of Life Sciences, Gleueler Str. 60, 50931 Köln
  Head: Prof. Dr. Dietrich Rebholz-Schuhmann
- Spokesperson
  Prof. Dr. Juliane Fluck, fluck@zbmed.de, ZB MED and University of Bonn, Katzenburgweg 1a, 53115 Bonn

3. Objectives, work programme and research environment

- Research area of the proposed consortium (according to the DFG classification system)
  www.dfg.de/download/pdf/dfg_im_profil/gremien/fachkollegien/amtsperiode_2016_2019/fachsystematik_2016-2019_en_grafik.pdf
  22 (Medicine)

Concise summary of the planned consortium's main objectives and task areas

Epidemiological and clinical trial data have the advantage of being highly structured. Data collections in this field entail detailed phenotypes of study subjects and are generated according to a defined protocol. These data are most appropriate to investigate (1) the prognosis, aetiology and burden of diseases on individual and population level, (2) the influence of risk factors incl. nutritional factors and (3) the efficacy of preventive, diagnostic and therapeutic interventions.

Although individual studies are highly standardised and well-documented, they seldom fulfil FAIR data principles: (1) **Findability** is often hampered. For public health data, so far, internati-

onal attempts are limited to support standards for publication and to establish study/protocol registries. Clinical trials are registered in the WHO approved German Clinical Trials Registry hosted by DIMDI, but the collected data are not described in standard formats. (2) Modalities of data **access** are typically not reported in sufficient detail. (3) Different epidemiological databases are usually not **interoperable**. For instance, there is vast methodological heterogeneity in approaches to assess dietary intake and to investigate associations with health outcomes. (4) Data protection requirements restrict **reuse** of data because these highly sensitive data often cannot be anonymised and reuse is limited by the informed consent of study participants.

Based on these limitations, the key objectives of NFDI4Health are to create new opportunities for data analysis in the interest of improving population health, in particular: (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 accessing existing decentralised epidemiological/ clinical data infrastructures, (3) to enhance data sharing, record linkage, harmonised data quality assessments, federated analyses of personal health data in compliance with privacy regulations and ethics principles, (4) to enable the development and deployment of new, machine processable consent mechanisms and innovative data access services by operationalising the FAIR data principles, (5) to foster data sharing and cooperation between clinical 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.

The overarching aim of NFDI4Health is to create an infrastructure based on standards which thus enables harmonisation, is expandable and facilitates the retrieval and use of public health data, (dietary) exposure data as well as clinical trial data, facilitating structured combination and interoperability. The NFDI4Health consortium will address the following task areas (TA):

TA1 'Coordination' is responsible for the overall governance, dissemination/public relations and the coordination of use cases. Use cases will implement community-driven infrastructure requirements in the corresponding research data sources, e.g. existing epidemiological studies.

TA2 'Findability, Interoperability and Quality Standards of Health Data' (1) develops data management and publication policies for health data, (2) encourages data sharing, (3) informs about data access possibilities, (4) develops harmonised data quality standards incl. method standardisation and provision of method inventories, (5) establishes a joint Metadata Repository (MDR) to host the majority of data elements that are collected by the epidemiological, nutrition research and clinical trial communities. The MDR accounts for semantic enrichment of data to enable integration/interoperability and supports provenance concepts. All partners share their metadata and create, within allowances, interoperable data versions.

TA3 'NFDI4Health Services' handles all services provided by NFDI4Health. These include provision of a central registry and search services for health data, metadata repositories for health/medical/nutrition information models, software repositories for data analysis, standardised data quality assessments, certification services for software and registration of certified services (for automated access), (metadata) annotation services, publication services, (long-term) archiving services, data visualisation services, automatic consent services, central data access services and record/data linkage services.

TA4 'NFDI4Health Community' organises sustainable governance for outreach, interaction and dissemination to the community at large and for specific workshops addressing professional communities. New formats such as "crowdsourcing" to enhance user involvement and to ensure

continuous feedback were discussed in the framework of the first community workshop in Cologne (June 2019). The ongoing translation of community needs into new requirements will enable a timely advancement of NFDI4Health. TA4 establishes FAIR/reputation metrics for health data in coordination with national and international communities. Outreach and interaction with potential users, the public and other stakeholders and the organisation of standardisation processes together with the communities/various sub-disciplines are an important aim of this TA. Furthermore, a main focus is the development of training material, establishment of various training formats and establishment of lecture modules or graduate programmes.

TA5 'NFDI4Health in Context' organises all interactions with other NFDI consortia and coordinates common interfaces. This includes the overall coordination of cross-sectional topics identified across NFDIs. Outreach to funding bodies and political decision makers will be strived for to achieve a sustainable impact even beyond the funding period. Participation and networking in national and international communities will focus on (a) privacy and data protection of health data, (b) research data infrastructures, (c) method developments (d) distributed data analysis, (e) FAIR data sharing and changing the culture of data sharing, (f) training the next generation data scientist work force.

TA6 'Legal, Privacy & Data Access/Analysis in Concert' deals with conception, set-up and pilot implementation of (1) organisational structures and governance with regard to sensitive data and data protection laws (2) generic services and infrastructures with regard to data protection requirements as well as record/data linkage (3) common consent standards and data access procedures, initially for use cases. TA6 deals with the development of (4) criteria and automatic processes ensuring non-identifiability of study subjects in results of data analyses, (5) criteria to control privacy-relevant aspects for new software applications, (6) certification criteria for software artefacts, (7) a certification process for health data centres. It will work on (8) risk assessment of data sharing and the set-up of automatic control mechanisms for adherence to data protection requirements when sharing health data across sites/domains.

TA7 'Enable Future: Distributed Data Analysis and Computing' enables a better data access for distributed data analysis, for applying machine learning (ML) and artificial intelligence to personal health data. TA7 (1) evaluates existing and upcoming distributed data analysis infrastructures (e.g. Personal Health Train, DataSHIELD) with respect to data protection principles in use cases, (2) provides data quality analysis and metrics depending on different scenarios such as epidemiological studies, machine learning or simulation, (3) develops and validates software for distributed data analysis, (4) controls privacy of data derivates and data aggregation or trained ML models (→ TA6), (5) sets up a central access point for distributed data analysis with registered data centres and certified analysis engines, (6) generates test environments with virtual health data for open data access and virtual cohorts, (7) enables adaptation of data analysis according to future trends identified in TA5.

Brief description of the proposed use of existing infrastructures, tools and services that are essential in order to fulfil the planned consortium's objectives

NFDI4Health plans to integrate among others the following data infrastructures from (co-) applicants and participants of this consortium:

- population-based epidemiological studies (e.g., the German National Cohort (GNC), SHIP (University Medical Center Greifswald)) and nutrition-related epidemiological studies (e.g., EPIC-Potsdam/Heidelberg, Hamburg City Health, DONALD, LISAplus, GINIplus),

- the German Pharmacoepidemiological Research Database (GePaRD, BIPS),
- surveillance data collected by the Robert Koch Institute (RKI) on a routine basis,
- national health and nutrition monitoring data: RKI health surveys (DEGS, KIGGS), LIFE study (IMISE), nutrition surveys (NVSII, MRI),
- cancer registries, including the German Centre for Cancer Registry Data (ZfKD at RKI),
- administrative routine data like health insurance data hosted by DIMDI,
- the WHO approved German Clinical Trials Registry hosted by DIMDI,
- the KKS-Network of clinical trials,

NFDI4Health will be based on existing standards like:

- FAIR standards: DataCite metadata schema for register DOIs, ORCID, Research Data Alliance (RDA), GO FAIR,
- Terminology and communication standards: HL7 FHIR (Fast Healthcare Interoperability Resources), CDISC (Clinical Data Interchange Standards Consortium) standards, ICD-10 (ICD-11), LOINC, SNOMED CT,
- Common data models for electronic health records: HL7 FHIR, openEHR, OMOP OHDSI, CDISC ODM, ISO 13606:2019 - Electronic health record communication,
- International ISO standards released by relevant technical standardisation committees (e.g., ISO/TC 215 Health Informatics, ISO/TC 276 Biotechnology) and standards released by joint committees with the International Electrotechnical Commission (IEC) (e.g., ISO/IEC JTC 1 Information Technology),
- Diverse European standards, released by the relevant technical standardisation committees of the European Committee for Standardization (CEN) (e.g., CEN/TC 251 Health informatics).

NFDI4Health will exploit existing tools and services like:

- PUBLISSO Life Science Repository for publications, DOI services, data management planning tool RDMO4Life, terminology services,
- DataSHIELD, GO FAIR Personal Health Train,
- Meta databases of observational epidemiological studies (ENPADASI, InterConnect),
- EcoSoc Implementation Network,
- FAIRDOM SEEK Platform,
- i2b2/tranSMART tool for data exploring, querying and simple analysis of clinical trial and routine data,
- Leipzig Health Atlas (LHA), a web-based data sharing platform funded by the BMBF (i:DSem).

Interfaces to other proposed NFDI consortia: brief description of existing agreements for collaboration and/or plans for future collaboration

NFDI4Health and **NFDI4Medicine** address different medical and health-related data bodies but will work closely together in addressing publication policies and standards, metadata standards and common services. This is facilitated by the fact that different institutions and individuals are active in both consortia.

NFDI4Health will closely cooperate with all consortia in the medical/health domain, especially

with **NFDI4Medicine**, **NFDI4Neuroscience** and **GHGA**. Data protection challenges for making person-related data accessible are overarching themes where we will work closely together.

The **GHGA** consortium has agreed on a close partnership with the planned NFDI4Medicine and NFDI4Health consortia. Together, these three consortia provide perfectly complementary infrastructure components: bridging storage and management of public health and clinical trial data (NFDI4Health), healthcare data (NFDI4Medicine) and omics raw data (GHGA). By linking the data modalities addressed by these consortia (ideally, in a privacy-preserving manner), it will be possible to integrate previously disjoint data in an unprecedented manner. In this context, GHGA will provide large-volume data storage for omics raw data and expertise in the processing of this data. The data processing infrastructure in GHGA will yield interpretable data (e.g., genetic variants, gene expression quantification, epigenetic states), which can then be integrated and analysed with healthcare data and medical research data. GHGA will cooperate with NFDI4Health to facilitate linkage of omics data to information harvested in structured public health and clinical trial data. Ethical, legal and societal impacts are synergistic cross-sectional topics where all three consortia can contribute.

In this context, we coordinate our efforts together with **KonsortSWD** in addressing challenges that are related to making sensitive cohort data and survey data 1. re-usable and 2. interoperable. Partners of NFDI4Health aim to join the EcoSoc Implementation Network in order to cooperate directly under the auspices of the GO FAIR initiative.

NFDI4Neuroscience and NFDI4Health will share standardisation policies and processes. In the field of neuroscience, both consortia coordinate their activities with the aim to develop common (meta)data, quality and record linkage standards 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 explore opportunities for data linkage with environmental data of high relevance for health. In this respect, NFDI4Health is looking forward to a close collaboration with, e.g., **NFDI4Agri**, **NFDI4Earth**, **NFDI4BioDiversity** and **NFDI4NanoSafety**.

All overarching topics and standardisation with further Life Science consortia will be organised in close cooperation and concert with **NFDI4Life Umbrella**.

Furthermore, Frank Oliver Glöckner (spokesperson of **NFDI4BioDiversity**) already organised a meeting of all NFDI consortia in August to discuss further interfaces and cooperation regarding cross-cutting topics.

4. Cross-cutting topics

Please identify cross-cutting topics that are relevant for your consortium and that need to be designed and developed by several or all NFDI consortia.

For most cross-cutting topics, there are several levels at which these need to be addressed. In the context of **data privacy, data protection laws and data linkage** the cross-cutting activity needs to be addressed initially at the level of all health-related consortia and all consortia dealing with other highly sensitive, personalised data like KonsortSWD. The same holds true for relevant **data access services that concern personalised data**. On the next level, results can be exchanged and discussed with all NFDI consortia.

Overarching data standards and interoperability need to be addressed initially among all health-related consortia, next on the level of the life sciences in general and finally within the whole group of NFDI consortia. To achieve this, a close collaboration with relevant technical standardisation committees of recognised national, European and international organisations (ISO, IEC, CEN, DIN) as well as scientific standardisation initiatives from the different health and life science domains is crucial.

Similar stepwise approaches are necessary for cross-cutting topics such as

- search interfaces and services,
- terminology, lookup, annotation and curation services (based on metadata standards to describe the data semantics and context),
- standards for harmonising data formatting and description as a prerequisite for data comparison and integration,
- standards for distributed data analysis (interfaces/methods/opportunities to run software at the infrastructures),
- common requirements for NFDI for using machine learning and deep learning techniques,
- services for persistent identifiers,
- services for creation of standardised and tailored data management plans,
- digital long-term preservation,
- training/education for data creators, curators/stewards, data scientists and users,
- overarching FAIR/reputation metrics,
- cultural change within the community with respect to data structuring and sharing.

Please indicate which of these cross-cutting topics your consortium could contribute to and how.

- **Data privacy, data protection laws and data linkage**

  NFDI4Health will contribute to all areas where access to personalised data is required and this requirement needs specific solutions. Given the specific requirements with respect to data privacy, NFDI4Health can provide key knowledge and experience regarding human data protection and its implementation in study databases for the exchange with other NFDI initiatives. In this respect, NFDI4Health will benefit from a strong expertise provided by one co-applicant (University of Bremen) being a lawyer with focus on data protection and long-standing experience in ethical requirements of research studies involving human individuals. NFDI4Health strives for developing solutions to be applied on a broad basis to enable sharing of personalised data in Germany and to give advice for revising German law where necessary.

- **Training and education**

  NFDI4Health assigns a high relevance to specific training in good practice of health data collection, data access constraints and correct use and analysis of data. This does not only concern primary data collection but also the exploitation of existing databases and registries. Here, NFDI4Health can build on the strong expertise in “Good Practice in Secondary Data Analysis” of several co-applicants and participants, in particular the Otto von Guericke University Magdeburg. Moreover, together with the NFDI4Earth and NFDI4BioDiversity consortia and the city and state of Bremen, NFDI4Health is planning to

establish a graduate school for research data management and data science where also KonsortSWD and NFDI4Neuroscience expressed their interest to actively contribute to the development of the curriculum and the teaching modules.

– **Standardisation**

  Consortial members of NFDI4Health are already well established and active in relevant standardisation committees (e.g., ISO, IEC, CEN, DIN) and several standardisation initiatives, as well as meta initiatives that aim at a harmonisation of standardisation within the life sciences and beyond, such as the European network EU-STANDS4PM ("A European standardization framework for data integration and data-driven in silico models for personalized medicine") and the European COST action CHARME ("Harmonising standardisation strategies to increase efficiency and competitiveness of European life-science research"). These activities will be beneficial for many NFDI consortia, especially the ones related to health and life science.

– **Monitoring of data quality and federated data analysis infrastructures**

  To facilitate federated data analyses the quality of data has to be closely monitored according to predefined criteria. In this respect, NFDI4Health will build on the knowledge generated in the DFG-funded project "Standards and tools for data monitoring in complex epidemiological studies" where a broad community of German epidemiological research institutes was involved under the lead of the University Medical Center Greifswald. In addition, NFDI4Health also provides experience in the implementation of federated data analysis infrastructures. In this regard, further development of services and analytical tools will be of generic interest across 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 person-related sensitive data. In particular, publication processes that focus on metadata publication and method documentation will be worked out. In this respect, appropriate models and metadata sets will be developed, e.g., by extending the generic metadata standard of DataCite. Furthermore, requirements of long-term preservation for archiving metadata and data sets will be considered. These tools and models can be reused by other domains handling sensitive data, e.g., by KonsortSWD.