Data Integration for Future Medicine (DIFUTURE): An Architectural and Methodological Overview

Oliver Kohlbacher, Ulrich Mansmann, Bernhard Bauer, Klaus Kuhn, Fabian Prasser

Angaben zur Veröffentlichung / Publication details:

Kohlbacher, Oliver, Ulrich Mansmann, Bernhard Bauer, Klaus Kuhn, and Fabian Prasser. 2018. "Data Integration for Future Medicine (DIFUTURE): An Architectural and Methodological Overview." Methods of Information in Medicine 57 (S01): e57–65. https://doi.org/10.3414/me17-02-0022.

Data Integration for Future Medicine (DIFUTURE)

An Architectural and Methodological Overview

Fabian Prasser[1]*; Oliver Kohlbacher[2,3]*; Ulrich Mansmann[4]*; Bernhard Bauer[5]*; Klaus A. Kuhn[1]*

[1]Institute of Medical Informatics, Statistics and Epidemiology, University Hospital rechts der Isar, Technical University of Munich, Munich, Germany;
[2]Department of Computer Science, Center for Bioinformatics and Quantitative Biology Center, Eberhard-Karls-Universität Tübingen, Tübingen, Germany;
[3]Max Planck Institute for Developmental Biology, Tübingen, Germany;
[4]Institute for Medical Information Processing, Biometry, and Epidemiology, Faculty of Medicine, Ludwig-Maximilians-University Munich, Munich, Germany;
[5]Department of Computer Science, University of Augsburg, Augsburg, Germany

Keywords
Health information systems, data warehousing, information dissemination, data sharing, privacy

Summary

Introduction: This article is part of the Focus Theme of Methods of Information in Medicine on the German Medical Informatics Initiative. Future medicine will be predictive, preventive, personalized, participatory and digital. Data and knowledge at comprehensive depth and breadth need to be available for research and at the point of care as a basis for targeted diagnosis and therapy. Data integration and data sharing will be essential to achieve these goals. For this purpose, the consortium Data Integration for Future Medicine (DIFUTURE) will establish Data Integration Centers (DICs) at university medical centers.

Objectives: The infrastructure envisioned by DIFUTURE will provide researchers with cross-site access to data and support physicians by innovative views on integrated data as well as by decision support components for personalized treatments. The aim of our use cases is to show that this accelerates innovation, improves health care processes and results in tangible benefits for our patients. To realize our vision, numerous challenges have to be addressed. The objective of this article is to describe our concepts and solutions on the technical and the organizational level with a specific focus on data integration and sharing.

Governance and Policies: Data sharing implies significant security and privacy challenges. Therefore, state-of-the-art data protection, modern IT security concepts and patient trust play a central role in our approach. We have established governance structures and policies safeguarding data use and sharing by technical and organizational measures providing highest levels of data protection. One of our central policies is that adequate methods of data sharing for each use case and project will be selected based on rigorous risk and threat analyses. Interdisciplinary groups have been installed in order to manage change.

Architectural Framework and Methodology: The DIFUTURE Data Integration Centers will implement a three-step approach to integrating, harmonizing and sharing structured, unstructured and omics data as well as images from clinical and research environments. First, data is imported and technically harmonized using common data and interface standards (including various IHE profiles, DICOM and HL7 FHIR). Second, data is preprocessed, transformed, harmonized and enriched within a staging and working environment. Third, data is imported into common analytics platforms and data models (including i2b2 and tranSMART) and made accessible in a form compliant with the interoperability requirements defined on the national level. Secure data access and sharing will be implemented with innovative combinations of privacy-enhancing technologies (safe data, safe settings, safe outputs) and methods of distributed computing.

Use Cases: From the perspective of health care and medical research, our approach is disease-oriented and use-case driven, i.e. following the needs of physicians and researchers and aiming at measurable benefits for our patients. We will work on early diagnosis, tailored therapies and therapy decision tools with focuses on neurology, oncology and further disease entities. Our early uses cases will serve as blueprints for the following ones, verifying that the infrastructure de-

Correspondence to:
Dr. Fabian Prasser
Institute of Medical Informatics, Statistics and Epidemiology
University Hospital rechts der Isar
Technical University of Munich
Ismaninger Straße 22
81675 Munich
Germany
E-mail: fabian.prasser@tum.de

* for the DIFUTURE Consortium

Methods Inf Med 2018; 57(Open 1): e57–e65
https://doi.org/10.3414/ME17-02-0022
received: December 1, 2017
accepted: April 17, 2018
Funding
The work of the DIFUTURE consortium during the conceptual phase was funded by the German Federal Ministry of Education and Research (BMBF) within the "Medical Informatics Funding Scheme" under reference numbers 01ZZ1603[A-D].

veloped by DIFUTURE is able to support a variety of application scenarios.
Discussion: Own previous work, the use of internationally successful open source systems and a state-of-the-art software architecture are cornerstones of our approach. In the conceptual phase of the initiative, we have already prototypically implemented and tested the most important components of our architecture.

1. Introduction

Future medicine will be predictive, preventive, personalized, participatory and digital [1]. Data and knowledge at comprehensive depth and breadth need to be available for research and at the point of care as a basis for targeted diagnosis and therapy. We aim at a learning health system, "where every clinical encounter contributes to research and research is being applied in real time to clinical care" [2]. Data integration and data sharing will be essential to achieve these goals. The consortium Data Integration for Future Medicine (DIFUTURE) is one of the four consortia selected by the German Ministry of Education and Research for funding during the development and networking phase of the Medical Informatics Initiative which started in January 2018. The consortium has been established by four core partners: Technical University of Munich and its University Medical Center Rechts der Isar, Ludwigs-Maximilians-University Munich and Munich University Medical Center, Eberhard-Karls-Universität Tübingen and Tübingen University Medical Center and Augsburg University.

Each participating university medical center will establish a Data Integration Center (DIC). Each DIC will form a central hub providing secure and reliable access to high-quality data integrated from healthcare and research sources. Interoperable interfaces will be offered to integrate and share data within the local sites, within the consortium and across consortia. We will adhere to and implement the FAIR data principles, which refer to data being findable, accessible, interoperable and reusable [3]. Our DICs will process heterogeneous data, comprising structured and unstructured data, clinical narratives, images and omics data. Data will be made available to researchers for cross-site (distributed) analyses and to physicians by providing innovative views on integrated data and decision support components for personalized treatments. Numerous challenges have to be addressed to realize our vision. In this article, we will describe our concepts and solutions with a specific emphasis on technical and architectural aspects.

Re-using healthcare data for research and sharing data across institutional boundaries immediately raises questions of interoperability, data quality and privacy. To address these challenges the DIFUTURE architecture is modular and extensible. On the data level, we will explicitly focus on structural, semantic and syntactic aspects of data collections, which will help us to avoid fundamental heterogeneity. We will use data and interface standards wherever possible and resolve heterogeneity by harmonization. On the application level, data processing is performed in a series of incremental steps utilizing several layers of services with clearly defined interactions. This approach will enable us to provide access to data in different forms suited for different usage scenarios in health care and research while being compliant with the interoperability requirements defined on the national level. The flexible working environments installed within the DICs will enable us to address quality aspects by checking for inconsistencies, implementing quality management processes and unambiguously tracking data provenance.

Data privacy and compliance to regulatory requirements are extremely relevant for building and maintaining patient trust. Therefore, innovative privacy-enhancing technologies and state-of-the-art IT security concepts play a central role in our approach. The modular nature of our architecture supports the separation of data, tasks and responsibilities into different organizational and technical units, which is fundamental to our data protection concept. Data will be shared using adequate methods selected for each application scenario based on rigorous risk and threat analyses. We will put a specific emphasis on distributed computing and remote analyses, where no patient-level data is actually shared [4].

From the perspective of health care and medical research, our approach is disease-oriented and use-case driven, i.e. following the needs of physicians and researchers. The aim is to show that our approach accelerates innovation, improves health care processes and results in measurable benefits for our patients. We will work on early diagnosis, tailored therapies and therapy decision tools with focuses on neurology, oncology and further disease entities. Our early uses cases will serve as blueprints, which show that the infrastructure developed by DIFUTURE is able to support a wide variety of (further) application scenarios.

2. Governance and Policies

Governance structures and strict policies for data use and sharing have been established by DIFUTURE based on our experiences and results from previous projects, such as the Leading Edge Cluster m4 (see also Sections 3 and 4). A general overview of the structure of the consortium is presented in ▶Figure 1.

The DIFUTURE Executive Board will be responsible for intra-consortium coordination and for coordination on the national level, i.e. with the national steering committee, its working groups and with other consortia. It will also supervise the development of the DIC infrastructure and guide the process of installing the DICs at all sites. The Executive Board will be advised by an International Scientific Advisory Board consisting of outstanding scientists who have agreed to share their expertise with the consortium [5]. The Executive Board will be overseen by a Supervisory Board consisting of the deans of the medical schools, the medical directors and business directors of the partners.

The supervisory board will ensure an alignment with the partners' plans for improving research and health care.

Each DIC will be directed by its own Executive Board comprising the Head of DIC, the CIO of the University Medical Center, the Director of the University Medical Center, the Dean of the School of Medicine and one User Representative speaking for the clinicians and clinical directors. Operations of the DIC will be coordinated by a DIC coordinator. A DIC's staff will be organized into three groups (coordination, informatics and analysis), each of which will be led by a team lead. The coordination group will focus on project management and compliance with legal and regulatory requirements as well as quality management. The informatics group will support the complete IT process from design to implementation and operations. The analysis group will consist of data scientists, bioinformaticians and biostatisticians, working on study design, data analyses and interpretation and also on quality management. Desk-to-desk intra-consortium working groups have been formed for important topics, such as data integration, security and privacy, high performance computing and bioinformatics. Renowned computer scientists, e.g. from TUM, have declared their willingness to advise these groups.

Other elements of our organizational structure will help us to manage change. Our interdisciplinary boards are a significant "horizontal" element of a matrix organization, consisting of participants from all professional groups involved in DIFUTURE, such as clinicians, informaticians and researchers. These groups have already been established and proven highly valuable for discussing problems and solutions, agreeing on standards, concepts, and other relevant decisions. Moreover, they are important for communicating success stories and for continuously educating the participants. These groups also compare outcomes to expectations and identify (and solve) potential problems early.

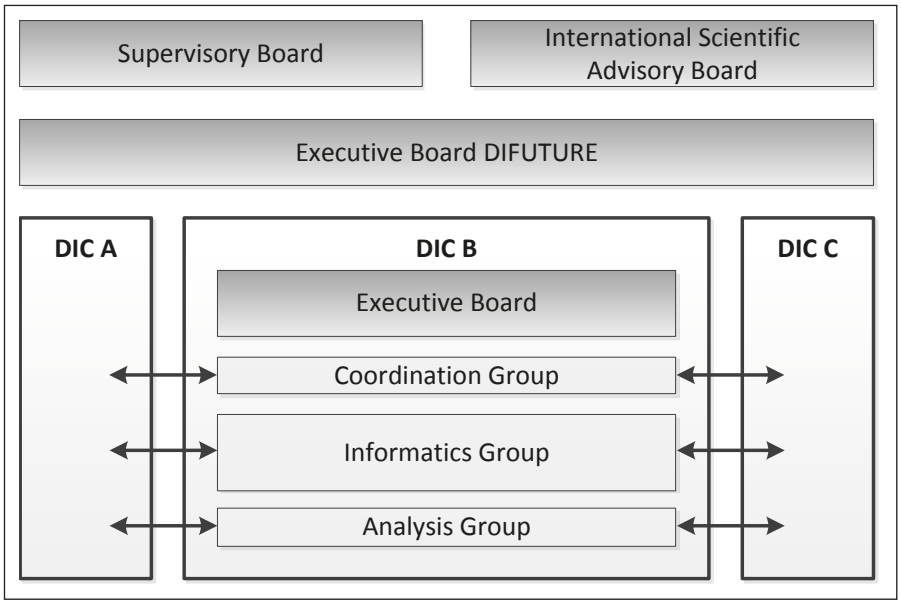

Figure 1 Organizational structure, management and collaboration.

DIFUTURE is fully aware of the data protection challenges of this initiative, with its unparalleled size and scope in Germany. Privacy, security and patient trust hence play a central role in our concept. We have specified strict policies for data use and sharing that need to be implemented and integrated with technical measures at each site to ensure a level of data protection that by far exceeds current standards. Our policies recommend that data processing in each DIC should be performed under the responsibility of the University Medical Center. Access to data will involve three layered organizational safeguards to ensure that ethical, legal and societal issues (ELSI) are considered adequately (see ▶Figure 2). These safeguards will be tightly integrated with the DIC's data protection model, which involves further organizational and technical measures. An example are secure methods of data sharing utilizing privacy-enhancing technologies, which will be described in more detail in the next section. The process of handling requests for data will be managed by the DIC's Data Transfer Unit (DTU), which will be formed by members of the DIC's coordination group.

Data will be imported into the DIC and be shared in strict accordance with data use agreements, applicable laws and informed consent. Requests will be reviewed by the local Data Access Committee (DAC), which will closely cooperate with the institutional ethical review board. The DIC's data protection specialists will review requests and analyze the risks of data sharing. In close cooperation with the DAC and the analysis and informatics groups of the DIC, appropriate methods for sharing the data will be recommended. Before any data is made accessible, applicants must sign a Data Use Agreement, where they agree in writing to adhere to the terms and conditions of usage (e.g. that the data are to be used solely for the requested and approved purpose and that the recipient must not

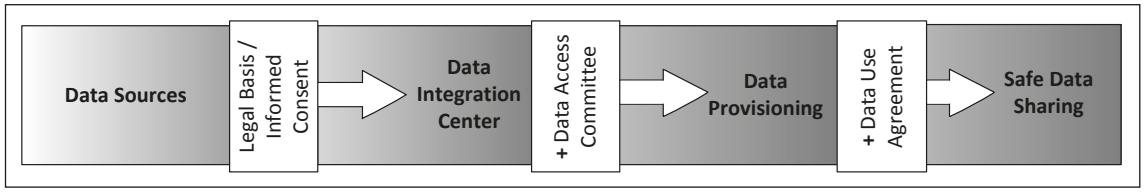

Figure 2 Principles for granting access to data.

try to identify or contact the data subjects). We are further intensely involved in the national activities on building a framework for data sharing.

3. Architectural Framework and Methodology

3.1 Basic Architecture of a DIFUTURE Data Integration Center

To address the challenges of interoperability, data quality and the need to support several different data consumers (e.g. from research and health care), each DIC will implement a three-step approach to data integration and sharing. In the first step, data is imported into the DIC aiming at technical harmonization. In the second step, data is enriched and harmonized within the DIC's data lake, which provides a central staging area and working environment. Already during this step, tools and processes for the harmonization of data can be shared between partners, including ETL workflows and registries of data structures, data items and terminologies. Further benefits of this strategy involve a pay-as-you-go approach to data integration, as the task of making data available to the DIC is clearly separated from the task of data preprocessing, harmonization, integration and utilization. In the third step, data will be made accessible through common data analytics platforms and with methods of secure data sharing. On the technical level, one of the pillars of our concept is to use internationally successful open source solutions, including i2b2 and tranSMART, and to combine and integrate them with commercial solutions on the basis of common data and interface standards, such as HL7, DICOM and IHE profiles [6].

An overview of the basic architecture of a DIC is shown in ▶Figure 3. In the following sections, we will describe the different steps along different axes, i.e. business architecture, applications architecture and data architecture, following common enterprise architecture modeling frameworks, such as TOGAF [7] and FEAF [8]. On the technical level, the DICs will be realized in a service-oriented manner, using lightweight RESTful interfaces for inter-system communication [9] and containerization [10, 11].

3.2 Data Import

On the business level, the data import services provide a blueprint for decoupling the clinical IT domain, which primarily focuses on data entry and retrieval, i.e. Online Transaction Processing (OLTP), from the DIC which primarily focuses on data integration, aggregation and analysis, i.e. on Online Analytical Processing (OLAP). From the applications perspective, data import is modeled as a pipeline consisting of several layers of services with clearly defined interactions, which process information in a series of incremental steps. On the first layer, connector services will be used to access source data using clearly specified and standards-based interfaces to clinical systems and research systems. These services implement an abstraction layer between the source systems and the DIC that technically harmonizes data into a common format which is suited for further processing in the pipeline and for pooling it in the DIC. Here, we will make extensive use of HL7 FHIR Resources [12]. Examples of further important data and interface standards on this layer are HL7, DICOM, CDA, CDISC ODM and IHE XDS. The import pipeline will be integrated with existing clinical interface engines available at the local sites. Further layers of the pipeline will perform cross-referencing and record-linkage by accessing the local clinical and research master patient index. Moreover, a pseudonymization service will remove all directly identifying information and identifiers and

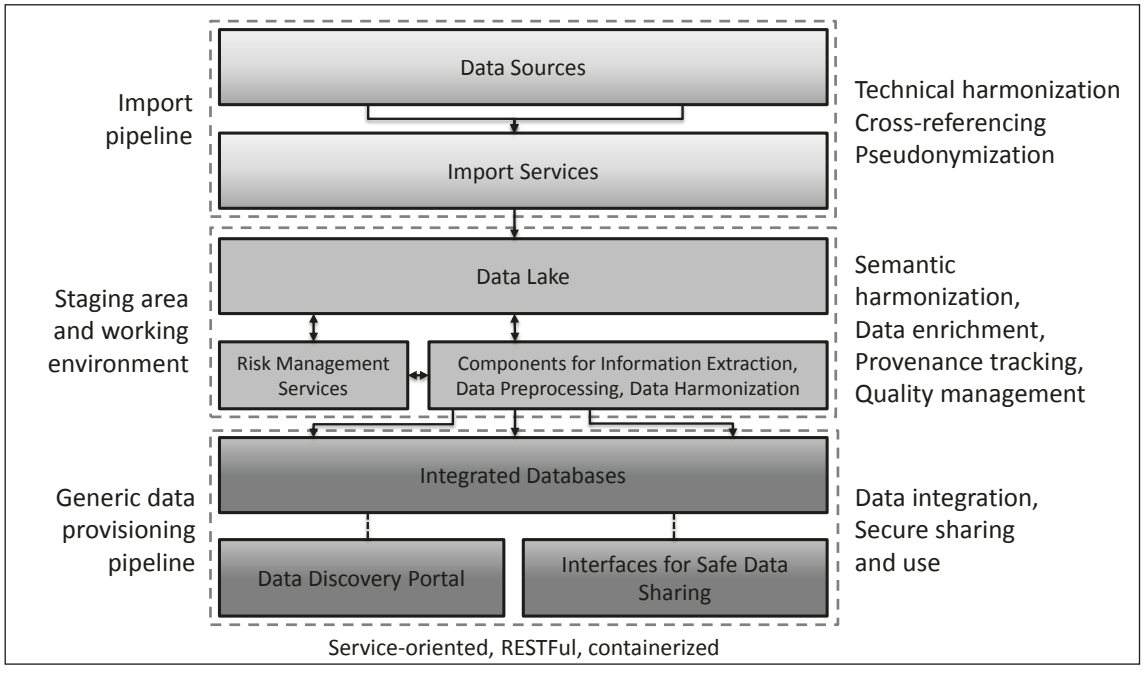

Figure 3 Basic Architecture of a DIFUTURE Data Integration Center.

insert pseudonyms that will be stable over time. The associated identifying information will be managed by the local Trust Center (see Section 3.4). Important interfaces on these layers are defined in IHE PDQ and PIX. Further services for data import will implement orchestration and monitoring (e.g. using OpenESB [13]) and a system registry. For high-volume data (e.g. images, omics) only metadata will be imported into the DIC. The actual payload data will be pulled on-demand if needed. Connectors to the PACSs will sanitize the metadata in DICOM images and they will de-identify images, e.g. using MIRC CTP [14]. Orthanc will be used to initiate image transfer via the DICOM WADO RESTful interfaces [15]. Patient identifying data will be removed from narratives using site- and context-specific information about the structure and content of documents.

3.3 Data Management

All incoming data will be pooled in the DIC's data lake, which is a staging area and working environment for data that is (almost) an exact copy of the data extracted from the source systems [16]. From here, data is further processed for downstream utilization by use of a clearly defined set of services. This architecture is aligned with strategic goals of the DIFUTURE concept. First, we emphasize that we aim to generally only load data into the data lake which is relevant to our internal, intra-consortium and trans-consortia use cases. This means that the data lakes will evolve in a gradual manner, which ensures that data harmonization remains manageable. Second, data provenance and data quality can be analyzed and documented from the beginning, i.e. before data has undergone significant transformations or has been aggregated for further purposes. Third, the DIFUTURE concept provides a blueprint for the technical properties of the environment. We hope to be able to avoid duplicate efforts by exchanging and re-using code and data processing workflows, which will be packaged into containers [10, 11]. Sharing will be implemented through common repositories and registries. Selected components for semantic harmonization will support federation to synchronize our data harmonization efforts.

From the applications architecture perspective, the working environment foreseen by DIFUTURE is layered and replicates data in various stages of its processing. Technically, the data lake will contain different systems for managing structured data, clinical narratives (notes, letters, reports), images and omics data. Examples of storage solutions that will be used include PostgreSQL databases, OpenBIS [17], a variant store and a PACS [15, 18]. The services provided in the working environment will consist of different solutions for processing the different types of data managed by the DIC. An overview is presented in ▶Figure 4.

For data transformation and harmonization, the medical data scientists working in each DIC will be provided with a typical software stack, including statistics software, such as R, scripting languages, such as Python and Bash and a variety of tools for querying and transforming data. Typical examples which are important in the biomedical context, are Pentaho Data Integration (PDI) [19], which is the standard ETL tool for tranSMART and Talend Open Studio [20], which is provided as an import tool for i2b2 by the Integrated Data Repository Toolkit [21]. Data processing workflows will be executed on elastic computing resources provided by a virtualization platform (OpenStack) to enable efficient and reproducible processing of large-scale data [22]. Complex workflows will be packaged into containers [10, 11]. High-Performance-Computing (HPC) resources will be provided for processing omics and image data. Standardized data analysis and quality control pipelines will be established based on existing solutions that use UNICORE/gUSE [23] and KNIME [24].

Data harmonization will utilize a metadata repository for maintaining definitions of data items, data structures (e.g. describing data modules or individual attributes and their metadata), mappings from data items and structures to URIs, reference ranges and validation functions. The data managed by this repository will be used for retrospective integration of semantically diverse data and also for the prospective harmonization of newly collected data. A solution following the ISO/IEC 11179 Metadata Repository Standard [25] will be provided by our partner Kairos [26]. Terminologies and controlled vocabularies will be synchronized with our text mining components which will be used to extract structured information from narratives [27].

Project archiving and the documentation of data provenance will be tightly integrated, covering source provenance and transformation provenance to enable validating and verifying the results of analyses [28]. We will base our efforts on the W3C PROV Model, which is the *de facto* standard representation for interoperable provenance information and which has also been adopted by HL7 FHIR. The process of documenting provenance is difficult to automate and we therefore plan to specifically design data processing workflows in such a way that provenance information is generated and persisted. In order to implement our provenance concept, assets which evolve over time, including data and data transformations, will be versioned. All project-specific data pools will be archived prior to sharing, following the guidelines of good scientific practice. An inventory of

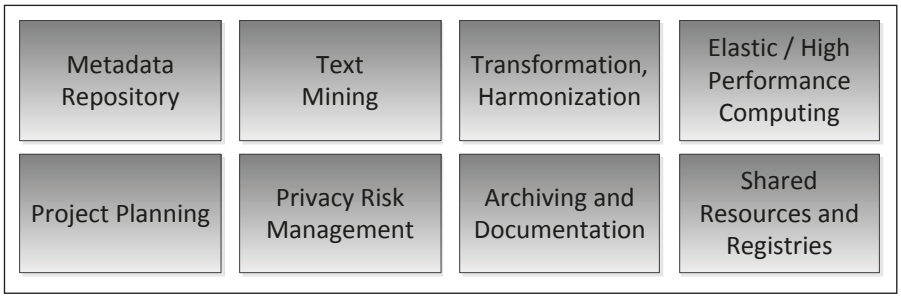

Figure 4 Most important services in the DIC's working environment.

archives will be maintained and projects will be registered with the national registry.

Using the methods supported by the DIC's data management environment, integrated databases will be created, which provide additional views on and organizations of data (e.g. organized along timelines) for research purposes. These different views and organizations are needed to support different types of downstream analyses. Target systems will typically be widespread and well-known analytics platforms, such i2b2 [29]. When a project has a genomic focus or requires integrated analyses of high-throughput data and clinical data, or comprehensive means of cohort comparison, we will provide instances of tranSMART [30], which can also be integrated with further platforms for omics data processing and knowledge management [31]. For projects involving visualization and analysis of genomic data, we will provide instances of cBioPortal [32]. For supporting additional statistical analyses, we will utilize RStudio [33]. On the data level, we will introduce a minimal common dataset with extensions specific to each use case. These efforts will be aligned with the national core dataset of the medical informatics initiative. Common terminologies, such as LOINC or SNOMED CT, will play a central role.

3.4 Data Protection and Sharing

Organizationally, the DICs of the participating sites will be integrated into the respective university medical centers and they will adhere to stringent data protection principles. Our architecture is designed from the ground up to be secure (security-by-design) and to guarantee that the privacy of patients and probands is respected (privacy-by-design). The underlying data protection concept implements the guideline by the TMF (Technologies, Methods and Infrastructure for Networked Medical Research e.V.), it considers relevant standards of the Federal Office for Information Security (BSI) and it is based on the concept implemented in the Leading Edge Cluster m4, which has been approved by three ethics committees (TUM, LMU, Bavarian Medical Association) and which has been reviewed by the data protection officer of Bavaria (BayLfD).

Already when data is imported into the DIC, patient identifying information will be separated from payload data and stored at a Trust Center, which is a separate organizational unit with its own personnel. As a consequence of this design, different types of data will be organizationally and technically separated from each other, implementing informational separation of powers. Structured and unstructured clinical data will be maintained in the DIC's data lake. High-volume data (images and omics) will remain stored in the source systems and only pulled temporarily and on demand for processing, typically directly into a data processing pipeline. Only the results of this processing will be stored in the data lake. The output of genomic analyses will be maintained in a separate variant store. For implementing our pseudonymization service, we will use the open source solution DIS (developed mainly out of the Leading Edge Cluster m4) which also covers data collection, biosample management and research identity management (concept see [34]).

Following the principle of informational self-determination, data will only be imported into the DIC and only be used or shared if permitted by data use agreements and laws or informed consent. At each site, a central component for managing consent and use agreement templates for health care and research data will be installed. A solution which we will evaluate for this purpose is gICS [35]. A reference to the status (e.g. accepted, declined, withdrawn, unknown) of associated consent policies must always be contained in the metadata of all data that is passed through the import pipeline. According checks are enforced by all services on this layer of our architecture.

The Trust Center will provide several services to the DIC. First, it will offer interfaces for retrieving permissions and restrictions expressed by informed consents, data use agreements and regulatory requirements, in a harmonized and structured form. For this purpose, semantics of policies from consent templates, agreements and regulations will be mapped to permission matrices expressed with ADA-M (Automatable Discover and Access Matrix) by the Global Alliance for Genomics and Health (GA4GH) [36]. Second, the Trust Centers will provide services for privacy-preserving cross-site record-linkage, which will for example be used for reconciling the identities of patients and probands. Our implementation will be fault-tolerant and based on Bloom filters [37]. Third, the Trust Center is also responsible for handling various processes, such as consent withdrawal or providing patients with access to their data.

In the research context, the DIFUTURE portal will provide different access tiers to efficiently and effectively support all relevant phases of research projects while keeping the complexity of our architecture low. In the *discovery phase*, researchers check whether sites have data which are relevant to their research question. Here, responses need to be provided quickly to facilitate data use. At the same time access to data of low detail or data dictionaries is typically sufficient. This phase will be supported by different layered access tiers providing views on the data with increasing granularity and interactivity. When users are confident that relevant data is available, a project plan is developed during the *design phase*. Decisions on feasible analyses and required sample sizes can be guided by investigation of single site data. In this intermediate phase, a higher overhead for the recipient is acceptable and it is not necessarily required that data from different sites is pooled. At the same time, a higher degree of flexibility regarding data processing and access to more fine-grained data is needed to develop an analysis plan. The actual *analysis* is performed in the last phase and potentially on an integrated data pool comprising data from multiple partners that is of sufficient sample size. The decision process of a site will be made transparent for applicants by providing information through the portal. For the DIC, the portal will provide functionalities to help organizing the processing of requests and for communicating with applicants and partners. The portal will be implemented based on Liferay, which is also often used in the biomedical context [38]. We plan to implement cross-site cohort selection based on SHRINE/i2b2 [39].

Before any data is shared, organizational safeguards will be applied (see Section 2) and the data will generally undergo a second pseudonymization procedure. Moreover, adequate data sharing mechanisms will be selected based on rigorous risk and threat analyses. There are various factors that can influence the decision on appropriate sharing mechanisms for performing a specified set of analyses. For example, cross-site record linkage will be used to determine the degree of population overlap for relevant data from the different sites. Moreover, re-identification risks of microdata (*safe data*) and macrodata (*safe outputs*) will be assessed and it will be determined to which degree privacy-enhancing technologies can be used to further protect the data. This will be implemented using ARX [40] and sdcMicro [41]. An important aspect will be the integration of these tools into data processing pipelines.

As is sketched in ▶Figure 5, various methods will further be used to create *safe settings* when sharing data. For example, each DIC will employ virtualized environments in which data analysis platforms are made available for remote access through secure network connections [4]. We will further apply the *platform as a service* (PaaS) concept known from cloud computing to support data sharing via remote analyses and distributed computing. Batch jobs, i.e. scripts containing analyses that can be performed remotely, will be reviewed, wrapped into containers and executed within the environment. Interfaces will be developed for accessing and performing operations on data from within these containers. Access to our infrastructure will be secured using an authentication and authorization infrastructure based on identity federation approaches, such as Shibboleth [42].

4. Use Cases

DIFUTURE will implement several disease-oriented use cases in a step-wise manner. All use cases address medical questions, typically comprising both health care and research aspects. The early use cases will serve as blueprints, which show that our infrastructure is able to support a wide variety of application scenarios. All use cases will utilize the data integration and data sharing capabilities provided by the DICs.

The first use case focuses on the development of an infrastructure for personalized optimal treatment of Multiple Sclerosis (MS). Data of relevance includes structured clinical and research data (including lab values, medication, various clinical scores), patient-reported outcomes, clinical narratives (notes, letters, reports), data extracted from MRI and OCT images as well as omics data. From cross-site harmonization of data and processes, we expect benefits for health care and research. Early steps will involve the standardization of imaging procedures, of documentation and of reporting, including the definition and introduction of structured forms. From the integration of retrolective data and additional data collection, we expect further benefits for health care and research by improved data availability and accessibility. For example, a data portal will be implemented to provide clinicians with a patient-centric view on available use-case specific information. Treatment decisions will be transferred back into the DIC for downstream analysis, but also to the EMR system for documentation. Improved data quality, availability and accessibility will be assessed using typical quality indicators, process analyses and user satisfaction surveys.

In the study context, the use case aims at a better understanding of the disease and on disease-modifying therapy of MS: markers and algorithms will be developed to predict the course of disease at onset and to allow early personalized treatment decisions. Treatment success is expected to be maximized and adverse effects to be minimized.

The process of data collection and analysis is sketched in ▶Figure 6. Known patients are already well-characterized, with retrolective data at the different sites comprising structured documentation and standardized letters. The methods described in the previous section will be used to implement further harmonization, the extraction of structured data from letters, curation, quality control and mapping to standards. Integrated views on harmonized retrospective datasets will form the training dataset for the development of treatment decision rules using secure access to local data and distributed approaches to generate functional hypotheses, algorithmic and prognostic rules. Additional harmonization and standardization of processes and data will prospectively result in a large longitudinal dataset collected across sites in a controlled fashion using the DIS system. This dataset will be used to validate the rules developed and to generate further hypotheses using distributed analyses and (virtual) common data pools.

Our use case on Parkinson's Disease aims at the development and intra-consortium deployment of an infrastructure supporting research on early diagnosis of the different subtypes of Parkinson's disease (PD). Similar to the MS use case, this infrastructure will support the development and validation of therapy decision support systems. Accordingly, both use cases will be rather similar on the level of the required

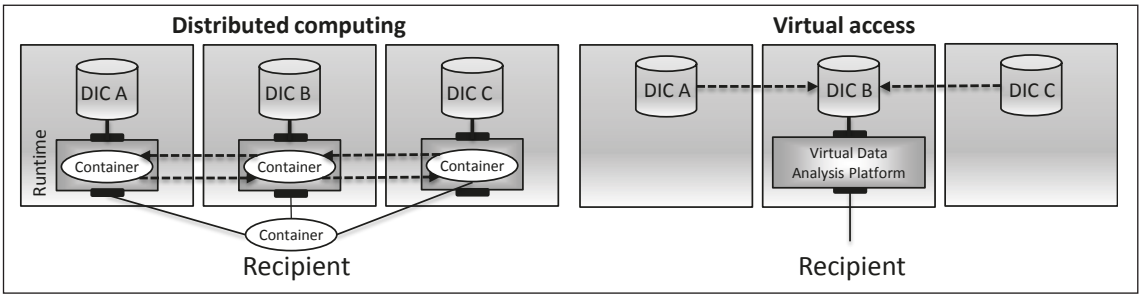

Figure 5
Example of secure methods of data sharing supported by DIFUTURE.

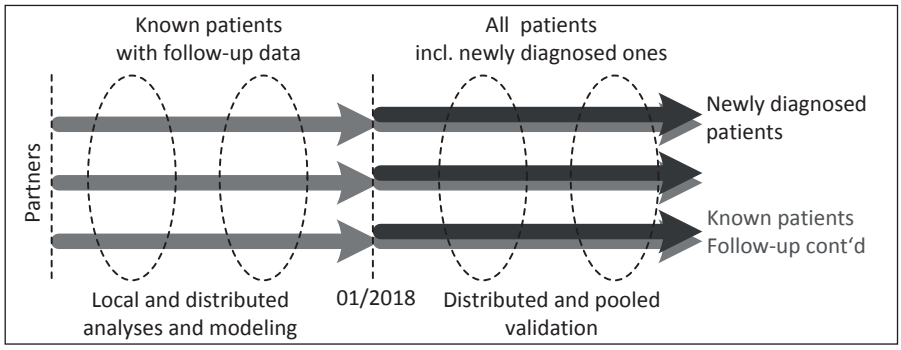

Figure 6 Overview of data analysis in the MS use case.

technical infrastructure. Analogously to the MS use case, tailored stratification strategies are expected to maximize treatment success and to minimize adverse effect. Further use cases will address stroke, cardiology and precision oncology. In order to focus our efforts and ensure early adoption by clinicians, we have selected specific clinical applications in oncology: indolent lymphoma and checkpoint inhibitor therapies in skin and lung cancer. DIFUTURE is also planning a cross-consortia use case and an international collaboration with ASCO CancerLinQ [26].

5. Discussion

In this article, we have presented an overview of the DIFUTURE concept with a specific emphasis on methodological, architectural and technical aspects. To realize our approach, we will employ a balance of internationally successful open source software, such as i2b2 and tranSMART and commercial solutions, e.g. by HIS, ERP, data warehousing and business intelligence product vendors. This will allow us to build – and to build upon existing – user communities, while creating an architecture that is advanced, flexible and easy to transfer to other sites.

Our approach follows other modern architectures, such as Vanderbilt's approach around its "derivatives" [43], the SHRINE/i2b2 solutions [39] and, in Europe, the integrated solution of Hôpital Européen Georges-Pompidou [44, 45]. The latter shows parallels, with its use of i2b2 and tranSMART, mirrored electronic patient records, its ETL suite and its large clinical data warehouse with layers including aggregated, anonymized and identified patient data. Our architecture for performing distributed analyses and for transferring algorithms between sites parallels the approach undertaken by the US network eMERGE for sharing "phenotyping algorithms" [46, 47]. Virtual data access environments have also been established in many domains, an important implementation is the Virtual Microdata Laboratory of the U.K. Office for National Statistics [48].

All three major partners have experiences with biomedical data warehousing and analytics platforms, including i2b2, tranSMART and cBioPortal. All partners have already developed initial versions of interfaces for connecting to their most important component systems. During the conceptual phase we have piloted common services, data transformation processes and data warehousing solutions with a specific focus on our audit use case. Moreover, we have started to work on our virtualization infrastructure by creating Docker images for analytics platforms (including tranSMART and i2b2) and piloted a consortium-wide registry for sharing Docker images. Moreover, we have implemented a prototype of a virtual data access environment [4]. We have further enhanced ARX, which is internationally recognized as a mature a tool for assessing privacy risks and protecting data from re-identification and which has been recommended by the European Medicines Agency for performing quantitative risk assessments when implementing its policy 0007 on clinical trial data sharing [49]. Moreover, all partners have made the strategic decision that identical business processes are to be implemented with the same application software and integrated solutions are to be preferred over solutions involving subsystems and interfaces. For example, the open source system DIS will be employed for research data collection, as a pseudonymization service and a research master patient index at all sites.

Due to space limitations and the technical focus of this article, there are several aspects which we could not discuss in detail, but for which concepts and solutions have been developed. Important examples are methods for sharing data with health insurance providers, details of our further cases and concepts for patient engagement, quality control, patient-centric portals and change management.

Acknowledgment

The authors gratefully acknowledge the invaluable contributions of the many members and supporters of the DIFUTURE team and specifically call out those who regularly attended our technical conference calls and our use case retreats as well as the leading clinicians of our primary use cases: A. Bayas, A. Berthele, R. Bild, R. Blaser, K. Bötzel, S. Endres, C. Gasperi, T. Gasser, B. Haslinger, B. Hemmer, G. Höglinger, R. Hohlfeld, M. Kerschensteiner, J. Kirschke, F. Kohlmayer, K. Kruber, M. Krumbholz, T. Kümpfel, M. Langermeier, H. Lautenbacher, C. v. d. Meijden, T. Meindl, M. Mühlau, T. Müller, M. Musick, S. Nahnsen, M. Naumann, N. Pfeifer, G. Pickert, J. Römhild, N. Rump, H. Spengler, F. Stahnke, M. Waltl, B. Wiestler, M. Wild, C. Zimmer.

References

1. Flores M, Glusman G, Brogaard K, Price ND, Hood L. P4 medicine: how systems medicine will transform the healthcare sector and society. Per Med 2013; 10(6): 565–576.
2. Dyke SO, Philippakis AA, Rambla De Argila J et al. Consent Codes: Upholding Standard Data Use Conditions. PLoS Genet 2016; 12(1): e1005772.
3. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016; 3: 160018.
4. Prasser F, Kohlmayer F, Spengler H, Kuhn KA. A scalable and pragmatic method for the safe sharing

of high-quality health data. IEEE J Biomed Health Inform 2018; 22(2): 611–622.
5. DIFUTURE – Scientific Advisory Board [cited 2017 Nov 27]. Available from: https://difuture.de/advisory-board/.
6. IHE IT Infrastructure Technical Framework. IHE International Inc; 2017 Jul 21 [cited 2017 Nov 30]. Available from: https://www.ihe.net/Technical_Frameworks/#IT.
7. 7® Version 9.1, an Open Group Standard. The Open Group [cited 2017 Oct 27]. Available from: http://www.opengroup.org/subjectareas/enterprise/7/.
8. Federal Enterprise Architecture Framework Version 2. The White House; 2013 Jan 29 [cited 2017 Nov 17]. Available from: https://obamawhitehouse.archives.gov/sites/default/files/omb/assets/egov_docs/fea_v2.pdf.
9. Fielding RT. Architectural styles and the design of network-based software architectures [dissertation]. Irvine: University of California; 2000.
10. Boettiger C. An introduction to Docker for reproducible research. ACM SIGOPS Operating Systems Review 2015; 49(1): 71–79.
11. Kurtzer GM, Sochat V, Bauer MW. Singularity: Scientific containers for mobility of compute. PLoS One 2017; 12(5): e0177459.
12. Bender D, Sartipi K. HL7 FHIR: An Agile and RESTful approach to healthcare information exchange. 26th IEEE International Symposium on Computer-Based Medical Systems; 2013. p. 326–331.
13. OpenESB – The Open Enterprise Service Bus [cited 2017 Nov 27]. Available from: http://www.open-esb.net/.
14. MIRC Clinical Trials Processor. Radiological Society of North America, Inc. [cited 2017 Nov 27]. Available from: http://mircwiki.rsna.org/index.php?title=MIRC_CTP.
15. Jodogne S, Bernard C, Devillers M, Lenaerts E, Coucke P. Orthanc – A lightweight, restful DICOM server for healthcare and medical research. 10th IEEE International Symposium on Biomedical Imaging; 2013. p. 190–193.
16. Stein B, Morrison A. The enterprise data lake: Better integration and deeper analytics. PwC Technology Forecast: Rethinking integration 2014; 1: 1–9.
17. Bauch A, Adamczyk I, Buczek P, Elmer FJ, Enimanev K, Glyzewski P, et al. openBIS: a flexible framework for managing and analyzing complex data in biology research. BMC Bioinformatics 2011; 12(1): 468.
18. Marcus DS, Olsen TR, Ramaratnam M, Buckner RL. The Extensible Neuroimaging Archive Toolkit: an informatics platform for managing, exploring, and sharing neuroimaging data. Neuroinformatics 2007; 5(1): 11–34.
19. Casters M, Bouman R, Van Dongen J. Pentaho Kettle solutions: building open source ETL solutions with Pentaho Data Integration. Indianapolis: John Wiley Publishing Incorporated; 2010.
20. Bowen J. Getting Started with Talend Open Studio for Data Integration. Birmingham: Packt Publishing Limited; 2012.
21. Bauer C, Ganslandt T, Baum B, Christoph J, Engel I, Löbe M, et al. The integrated data repository toolkit (IDRT): accelerating translational research infrastructures. J Clin Bioinforma 2015; 5(Suppl 1): S6.
22. de la Garza L, Veit J, Szolek A, Röttig M, Aiche S, Gesing S, et al. From the desktop to the grid: scalable bioinformatics via workflow conversion. BMC Bioinformatics 2016; 17(1): 127.
23. Streit A, Bala P, Beck-Ratzka A, Benedyczak K, Bergmann S, Breu R, et al. UNICORE 6 – recent and future advancements. Ann Telecommun 2010; 65(11–12): 757–762.
24. Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, et al. KNIME: The Konstanz information miner. In: Preisach C, Burkhardt H, Schmidt-Thieme L, Decker R, editors: Data Analysis, Machine Learning and Applications. Berlin: Springer; 2008. p. 319–326.
25. ISO/IEC 11179, Information Technology – Metadata registries (MDR). International Organization of Standardization (ISO) [cited 2017 Nov 30]. Available from: http://metadata-standards.org/11179/.
26. DIFUTURE – Partners [cited 2017 Nov 28]. Available from: https://difuture.de/partners/.
27. Averbis Information Discovery [cited 2017 Nov 27]. Available from: https://averbis.com/information-discovery/.
28. Ragan ED, Endert A, Sanyal J, Chen J. Characterizing Provenance in Visualization and Data Analysis: An Organizational Framework of Provenance Types and Purposes. IEEE Trans Vis Comput Graph 2016; 22(1): 31–40.
29. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc 2010; 17(2): 124–130.
30. 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.
31. Schumacher A, Rujan T, Hoefkens J. A collaborative approach to develop a multi-omics data analytics platform for translational research. Appl Transl Genom 2014; 3(4): 105–108.
32. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal 2013; 6(269): pl1.
33. Loraine AE, Blakley IC, Jagadeesan S, Harper J, Miller G, Firon N. Analysis and visualization of NA-Seq expression data using RStudio, Bioconductor, and Integrated Genome Browser. Methods Mol Biol 2015; 1284: 481–501.
34. Lautenschläger R, Kohlmayer F, Prasser F, Kuhn KA. A generic solution for web-based management of pseudonymized data. BMC Med Inform Decis Mak 2015; 15: 100.
35. Bialke M, Penndorf P, Wegner T et al. A workflow-driven approach to integrate generic software modules in a Trusted Third Party. J Transl Med 2015; 13: 176.
36. Automatable Discovery and Access Matrix. GA4GH [cited 2017 Nov 30]. Available from: https://www.ga4gh.org/ga4ghtoolkit/regulatoryandethics/.
37. Durham EA, Kantarcioglu M, Xue Y, Toth C, Kuzu M, Malin B. Composite Bloom Filters for Secure Record Linkage. IEEE Trans Knowl Data Eng 2014; 26(12): 2956–2968.
38. Schera F, Weiler G, Neri E, Kiefer S, Graf N. The p-medicine portal – a collaboration platform for research in personalised medicine. Ecancermedicalscience 2014; 8: 398.
39. McMurry AJ, Murphy SN, MacFadden D, Weber G, Simons WW, Orechia J, et al. SHRINE: enabling nationally scalable multi-site disease studies. PloS One 2013; 8(3): e55811.
40. Prasser F, Kohlmayer F, Lautenschläger R, Kuhn KA. ARX – A Comprehensive Tool for Anonymizing Biomedical Data. AMIA Annual Symposium; 2014. p. 984–993.
41. Templ M, Kowarik A, Meindl B. Statistical Disclosure Control for Micro-Data Using the R Package sdcMicro. J Stat Softw 2015; 67(4): 1–36.
42. Brandizi M, Melnichuk O, Bild R, Kohlmayer F, Rodriguez-Castro B, Spengler H, et al. Orchestrating differential data access for translational research: a pilot implementation. BMC Med Inform Decis Mak 2017; 17(1): 30.
43. Danciu I, Cowan JD, Basford M, Wang X, Saip A, Osgood S, et al. Secondary use of clinical data: the Vanderbilt approach. J Biomed Inform 2014; 52: 28–35.
44. Boussadi A, Caruba T, Zapletal E, Sabatier B, Durieux P, Degoulet P. A clinical data warehouse–based process for refining medication orders alerts. J Am Med Inform Assoc 2012; 19(5): 782–785.
45. Jannot AS, Zapletal E, Avillach P, Mamzer MF, Burgun A, Degoulet P. The Georges Pompidou University Hospital Clinical Data Warehouse: A 8-years follow-up experience. Int J Med Inform 2017; 102: 21–28.
46. Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, Peissig PL, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. J Am Med Inform Assoc 2016; 23(6): 1046–1052.
47. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, et al. The electronic medical records and genomics (eMERGE) network: past, present, and future. Genet Med 2013; 15(10): 761–771.
48. Ritchie F. Secure access to confidential microdata: four years of the Virtual Microdata Laboratory. The Labour Gazette 2008; 2(5): 29–34.
49. European Medicines Agency. EMA/90915/2016 (Version 1.3) – External guidance on the implementation of the European Medicines Agency Policy on Publication of Clinical Data for Medicinal Products for Human Use. 2017.