MIRACUM: Medical Informatics in Research and Care in University Medicine

A Large Data Sharing Network to Enhance Translational Research and Medical Care

Hans-Ulrich Prokosch[1]; Till Acker[2]; Johannes Bernarding[3]; Harald Binder[4,5]; Martin Boeker[5]; Melanie Boerries[6]; Philipp Daumke[7]; Thomas Ganslandt[1,8]; Jürgen Hesser[9]; Gunther Höning[10]; Michael Neumaier[11]; Kurt Marquardt[12]; Harald Renz[13]; Hermann-Josef Rothkötter[14]; Carmen Schade-Brittinger[15]; Paul Schmücker[16]; Jürgen Schüttler[17]; Martin Sedlmayr[1,18]; Hubert Serve[19]; Keywan Sohrabi[20]; Holger Storf[21]

[1]Chair of Medical Informatics, Department of Medical Informatics, Biometrics and Epidemiology, Friedrich-Alexander-University Erlangen-Nürnberg, Erlangen, Germany;
[2]Institute of Neuropathology, Justus-Liebig-University Giessen, Giessen, Germany;
[3]Chair of Medical Informatics, Institute for Biometry and Medical Informatics, Otto-von-Guericke-University Magdeburg, Magdeburg, Germany;
[4]Institute of Medical Biostatistics, Epidemiology and Informatics, University Medical Center of the Johannes Gutenberg University Mainz, Mainz, Germany;
[5]Institute of Medical Biometry and Statistics, Medical Faculty and Medical Center – University of Freiburg, Freiburg, Germany;
[6]Institute of Molecular Medicine and Cell Research and Comprehensive Cancer Center Freiburg (CCCF), University Medical Center, Faculty of Medicine, University of Freiburg; German Cancer Research Center (DKFZ), Heidelberg and German Cancer Consortium (DKTK) partner site Freiburg, Freiburg, Germany;
[7]Averbis GmbH, Freiburg, Germany;
[8]Department of Biomedical Informatics, University Medicine Mannheim, Ruprecht-Karls-University Heidelberg, Mannheim, Germany;
[9]Experimental Radiation Oncology Department, University Medical Center Mannheim, Central Institute for Scientific Computing (IWR), Central Institute for Computer Engineering (ZITI), Heidelberg University, Mannheim, Germany;
[10]Department of Information Technology, University Medical Center of the Johannes Gutenberg University Mainz, Mainz, Germany;
[11]Chair for Clinical Chemistry, Medical Faculty Mannheim of Heidelberg University, Mannheim, Germany;
[12]University Hospital of Giessen and Marburg, Giessen, Germany;
[13]Chair for Clinical Chemistry, Philipps University Marburg, Medical Director of the University Clinic Marburg, Marburg, Germany;
[14]Institute of Anatomy, Otto-von-Guericke-University Magdeburg, Dean of the Medical Faculty, Magdeburg, Germany;
[15]Chair of the Coordinating Centre for Clinical Trials, Philipps University Marburg, Marburg, Germany;
[16]University of Applied Sciences Mannheim, Institute for Medical Informatics, Mannheim, Germany;
[17]Department of Anesthesiology, University of Erlangen-Nürnberg, Dean of the Medical Faculty, Erlangen, Germany;
[18]Institute of Medical Informatics and Biometrics, Carl Gustav Carus Faculty of Medicine, Technische Universität Dresden, Dresden, Germany;
[19]Department of Hematology and Oncology, University Hospital Frankfurt, Goethe University, Frankfurt am Main, Germany;
[20]Faculty of Health Sciences, University of Applied Sciences – THM, Giessen, Germany;
[21]Medical Informatics Group, University Hospital Frankfurt, Goethe University, Frankfurt am Main, Germany

Correspondence to:
Prof. Hans-Ulrich Prokosch
Friedrich-Alexander-University Erlangen-Nürnberg
Department of Medical Informatics, Biometrics and Epidemiology
Wetterkreuz 13
91058 Erlangen
Germany
E-mail: ulli.prokosch@uk-erlangen.de

Summary
Introduction: This article is part of the Focus Theme of Methods of Information in Medicine on the German Medical Inform-

atics Initiative. Similar to other international data sharing networks (e.g. OHDSI, PCORnet, eMerge, RD-Connect) MIRACUM is a consortium of academic and hospital partners as well as one industrial partner in eight German cities which have joined forces to create interoperable data integration centres (DIC) and make data within those DIC available for innovative new IT solutions in patient care and medical research.

Objectives: Sharing data shall be supported by common interoperable tools and services, in order to leverage the power of such data for biomedical discovery and moving towards a learning health system. This paper aims at illustrating the major building blocks and concepts which MIRACUM will apply to achieve this goal.

Governance and Policies: Besides establishing an efficient governance structure within the MIRACUM consortium (based on the steering board, a central administrative office, the general MIRACUM assembly, six working groups and the international scientific advisory board), defining DIC governance rules and data sharing policies, as well as establishing (at each MIRACUM DIC site, but also for MIRACUM in total) use and access committees are major building blocks for the success of such an endeavor.

Architectural Framework and Methodology: The MIRACUM DIC architecture builds on a comprehensive ecosystem of reusable open source tools (MIRACOLIX), which are linkable and interoperable amongst each other, but also with the existing software environment of the MIRACUM hospitals. Efficient data protection measures, considering patient consent, data harmonization and a MIRACUM metadata repository as well as a common data model are major pillars of this framework. The methodological approach for shared data usage relies on a federated querying and analysis concept.

Use Cases: MIRACUM aims at proving the value of their DIC with three use cases: IT support for patient recruitment into clinical trials, the development and routine care implementation of a clinico-molecular predictive knowledge tool, and molecular-guided therapy recommendations in molecular tumor boards.

Results: Based on the MIRACUM DIC release in the nine months conceptual phase first large scale analysis for stroke and colorectal cancer cohorts have been pursued.

Discussion: Beyond all technological challenges successfully applying the MIRACUM tools for the enrichment of our knowledge about diagnostic and therapeutic concepts, thus supporting the concept of a Learning Health System will be crucial for the acceptance and sustainability in the medical community and the MIRACUM university hospitals.

1. Introduction

Reuse of clinical data is a fast-growing field recognized as essential to realize the potentials for high quality healthcare, improved healthcare management, reduced healthcare costs, population health management, and effective clinical research [1]. In this context "Big Data" and "Data-driven Medicine" are buzzwords often used, when institutions aim at connecting and leveraging various types of clinical data, images and omics data in order to characterize treatment pathways at scale [2] and to enhance diagnostic and therapeutic decision making. Typically such approaches require the cooperation of large numbers of institutions in so-called data sharing networks. Examples of such networks are the Observational Health Data Sciences and Informatics (OHDSI) collaboration [3], the PCORnet clinical data research networks (CDRNs) and patient-powered research networks (PPRNs) [4, 5] and the eMerge network [6]. European projects have for example focused on sharing aggregated data and metadata among rare disease researchers [7] and biobanks [8], creating a general distributed infrastructure for life-science information [9] or enabling data-intensive life science research in the Netherlands [10]. Most of such projects build their concepts around the FAIR guiding principles for scientific data management and stewardship [11]. In Germany this issue has recently been tackled with the announcement of the Medical Informatics Funding Scheme, aiming at networking data and improving health care [12]. MIRACUM (Medical Informatics for Research and Care in University Medicine [13]) is one of the four consortia funded by the BMBF Medical Informatics Initiative (BMBF MI-I) for the development and networking phase. It brings together eight German University Hospitals (Erlangen, Frankfurt, Freiburg, Giessen, Magdeburg, Mainz, Mannheim and Marburg) and Medical Faculties, two Universities of Applied Sciences (Giessen and Mannheim) and one industrial partner (Averbis GmbH, Freiburg). University Medicine Dresden and University Medicine Greifswald have currently applied to also become a MIRACUM partner in autumn 2018.

2. Objectives

All these partners have agreed to share data by employing data integration centers (DIC), to develop common interoperable tools and services, to use the power of such data collections and tools in innovative IT solutions, which shall enhance both patient-centered collaborative research as well as clinical care processes. Finally the partners intend strengthening biomedical informatics in research, teaching and training.

At the national level, MIRACUM actively participates in the National Steering Committee (NSC) with two senior coordinators, as well in three national working groups on interoperability, data sharing and consent that have been implemented by the NSC together with the BMBF MI-I supporting project (committee's offices).

This paper aims at illustrating the major building blocks and concepts which MIRACUM will apply to achieve this goal.

3. Governance and Policies

Governance and project organization within MIRACUM will be based on (1) the Steering Board, (2) six Working Groups, (3) a central coordinating office (located at the site of the MIRACUM coordinator at Friedrich-Alexander University Erlangen-Nürnberg) supported by local offices at each of the eight

MIRACUM sites, which will establish a DIC, (4) the MIRACUM General Assembly, and (5) an international scientific advisory board (▶Figure 1). The definition of an appropriate management structure (self-assessment and management) is crucial to the success of MIRACUM. It aims at efficient decision making, useful and satisfactory internal communication, and technical and administrative project control. The project is managed and controlled by the Steering Board (SB), in which every university/university hospital partner is represented by two named persons: the principal partner coordinator (PI) and a named deputy (Co-PI) who together not only provide excellent competencies in medical informatics and in medical research/care, but also have a high decision-making authority in their organization (e.g. faculty dean, medical director, chair of medical informatics, hospital chief information officer).

According to the three major goals of the funding scheme, we have further defined six working groups (WG):

- **WG1 "DIC Competence Centers"** will focus on the detailed specification of the DIC architecture and on the development and implementation of the DIC components, as well as their interfaces on a technical and organizational level,
- **WG2 "Data Sharing and Access, Consent and Quality Management"** will deal with all aspects and regulations concerning patient consent, data sharing and access to the data, quality management, IT security and data protection,
- **WG3 "Strengthening Medical Informatics"** will concentrate on measures for strengthening biomedical informatics within all MIRACUM partner universities (this includes e.g. the development of a joint master program "Biomedical Informatics and Medical Data Science", as well as the establishment of summer schools, staff training programs, medical data science continued education programs for clinicians and medical researchers),
- **WG4 "Alerting in Care – IT Support for Patient Recruitment"** will be responsible for implementing and evaluating the subproject of the first MIRACUM use case,
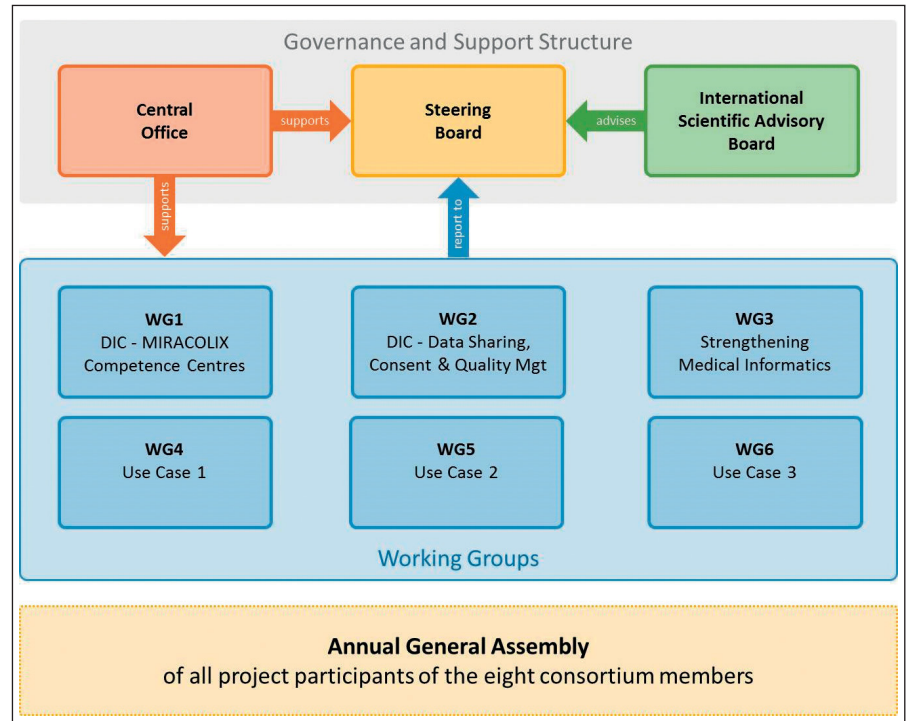

Figure 1   MIRACUM Governance Structure.

- **WG5 "From Data to Knowledge – Clinico-molecular Predictive Knowledge Tool"** will be responsible for implementing and evaluating the subproject of the second MIRACUM use case,
- **WG6 "From Knowledge to Action – Support for Molecular Tumor Boards"** will be responsible for implementing and evaluating the subproject of the third MIRACUM use case.

Each WG will define actions and monitor progress according to the specified goals outlined in its work plan. The working groups follow an agile, lean management process (SCRUM) and utilize the project management infrastructure (Confluence, JIRA, Chat) which is complemented by regular meetings and bi-monthly web conferences. The elected WG speakers will also ensure close collaboration between the working groups due to close interdependencies.

4. Architectural Framework and Methodology

To achieve a common use of research and patient care data within the consortium

and beyond, the MIRACUM DICs at the eight universities/university hospitals will comprise a modular set of components, which will be established at each local partner site, and further central components, which are established in order to support cross-institutional data sharing. Such components are not only technical IT-implementations, but (of equal importance) comprise a set of rules, policies, governance structures and data.

4.1 Data Governance

Reusing patient care data for research purposes, generating new knowledge and then transforming new knowledge into actionable support tools for patient care requires a close collaboration and integration with the local IT systems established at the MIRACUM universities/university hospitals. Thus, the DIC teams are closely linked to the routine IT departments. The scientific leaders (directors) of a DIC are members of the board of directors of the university hospitals, respectively the board of directors of the medical faculty and/or a lead person of the medical informatics/biometrics research group or the hospital's chief information officer (CIO). Further,

every MIRACUM partner has established a scientific Use & Access Committee (UAC), which is responsible for managing and evaluating all project proposals aiming at the use of DIC data either within the environment of the local MIRACUM site, but also for data sharing proposals within the MIRACUM consortium or even across all consortia. The work of this UAC will be guided by local Use & Access Policies (UAP), which have been defined at each of the MIRACUM partner sites and are closely aligned to the "Cornerstone Use & Access Policy" defined by the NSC Working Group on Data Sharing. Further, the day-to-day work of the MIRACUM partners' UAC is based on the definition of local bylaws and supported by an electronic project evaluation and management platform. Each MIRACUM partner will establish a trust center as a separate organizational entity, independent from the DIC, in order to provide ID management functionalities, such as pseudonymization and record linkage.

4.2 DIC Architecture

The technical components of a DIC are defined as parts of a modular architecture and may interact with each other and interchange data based on ETL processes, as well as standardized application programming interfaces (REST service interface). This architectural framework is built upon a **M**edical **I**nformatics **R**eus**A**ble e**C**osystem of **O**pen source **L**inkable and **I**nteroperable software tools (MIRACOLIX). For this ecosystem, we aim at reusing many open source software tools, which have proven their value in other international projects on data integration and data sharing (e.g. i2b2, tranSMART, the OMOP common data model and the OHDSI tools, XNAT, Samply.MDR, the gICS generic informed consent service, ARX). The fully released MIRACOLIX 4.0 based DIC architecture shall comprise the following technical components:

- primary data sources (mainly the EHR system and other clinical applications supporting the routine care processes, but also results of molecular/genomics high throughput analysis)
- ID-management tools (pseudonymization and privacy preserving record linkage)
- a data anonymization tool
- the MIRACUM metadata repository (M-MDR)
- data harmonization/data mapping tools
- a consent management system
- a natural language processing tool
- a hospital-wide trial-/project registry
- a project proposal management tool
- a set of ETL tools
- several data integration and data exploration repositories
- an IT infrastructure to share and easily deploy software pipelines for the analysis of omics data
- tools for data quality analysis, reporting and vizualisation
- modules for innovative user-friendly and efficient patient care process visualization
- connector component(s)
- a long-term research data archive
- a federated authentication system
- tools for development, deployment and monitoring of DIC IT components (e.g. a continuous integration test pipeline)
- a quality management system with a comprehensive set of standard operating procedures, describing the development, testing, deployment, maintenance, usage and revision of MIRACOLIX tools

4.3 MIRACOLIX Development

During the conceptual phase of nine months that preceded the launch of MIRACUM within the BMBF MI-I, we have implemented a basic architecture framework, which has served as a proof-of-concept architecture and was based on the MIRACOLIX 0.9 release. Each year in the development and networking phase, new MIRACOLIX updates will be released. Such new releases may constitute functional upgrades of already established architectural components, moving those to a higher level of maturity, but also the introduction of new components into the DIC architecture.

4.4 DIC Contents

In the current funding phase, DIC data will be mainly integrated from clinical care processes with the prospect of including research data e.g. from clinical trials in future applications. Thus, EHR data and data from various other clinical departmental systems shall be the major input sources for the data integration centers. The breadth of the data elements to be provided within the DIC will be extended incrementally in the four years of the development and networking phase and follow the BMBF MI-I roadmap and core dataset defined by the NSC working group interoperability. During the conceptual phase, the eight MIRACUM DIC have already been established and loaded with data according to the NSC basic core dataset modules person, demographics, encounters, diagnosis and procedures (this matches the official German claims data set for reimbursement of inpatient hospital stays), thus mainly comprising demographic patient information including age and gender, the hospital's provider data, the patient's diagnosis (ICD-10 Codes) and the procedures which have been performed for a patient (OPS 301 Codes). For most of the MIRACUM sites, those data are available for the years 2004–2017, for some only back until 2008/2009. In future releases detailed clinical and omics data elements will be included in the MIRACUM core data set depending (1) on the data required for the MIRACUM use cases and (2) on the interoperability recommendations defined by the NSC Working Group on Interoperability.

Adding new clinical data sources during the upcoming years will require the extension of our ETL processes, as well as the integration and mapping of the respective new data elements to the research data repositories and their data models (e.g. the generic i2b2 entity-attribute-value database structure or the OMOP CDM). The latter processes are described in the chapter on data harmonization. For the definition of data routes and interfaces between the different components of the MIRACUM DIC, we have strictly aligned our processes along the recommendations of the "Guidelines for data protection in medical research

projects – generic solutions of the TMF 2.0" [14] and experiences from the Cloud4Health project [15].

The processes to be supported in the eight MIRACUM DIC comprise the so-called Clinical Module, ID Management and the Research Module. Data export from the local EHR systems (and relevant departmental systems) can vary depending on previous local developments and experiences (e.g. Talend Open Studio based ETL processes, FHIR based provision of respective resources and even parsing of communication streams via a communication server) (1a–1c in ▶Figure 2).

Before transferring such data from the Clinical Module into the Research Module and loading it into the DIC research data repositories, several intermediate steps are pursued (e.g. checking for patient consent, data pseudonymization, data harmonization based on metadata definitions and natural language processing in case of narrative clinical texts). The implementation of the electronic consent module in MIRACUM will be based on the open source tool

gICS which was developed in the MOSAIC project [16].

Even though the harmonized DIC research data repository in ▶Figure 2 looks like one singular data store, in reality it will be a set of data integration and data exploration repositories, which shall be used for dedicated purposes depending on the clinical/research scenario and the types of data which shall be integrated (e.g. i2b2 [17, 18] and an OMOP DB [3] for clinical data, tranSMART for clinical and molecular/genomic data [19, 20, 21], XNAT for imaging data [22, 23, 24]).
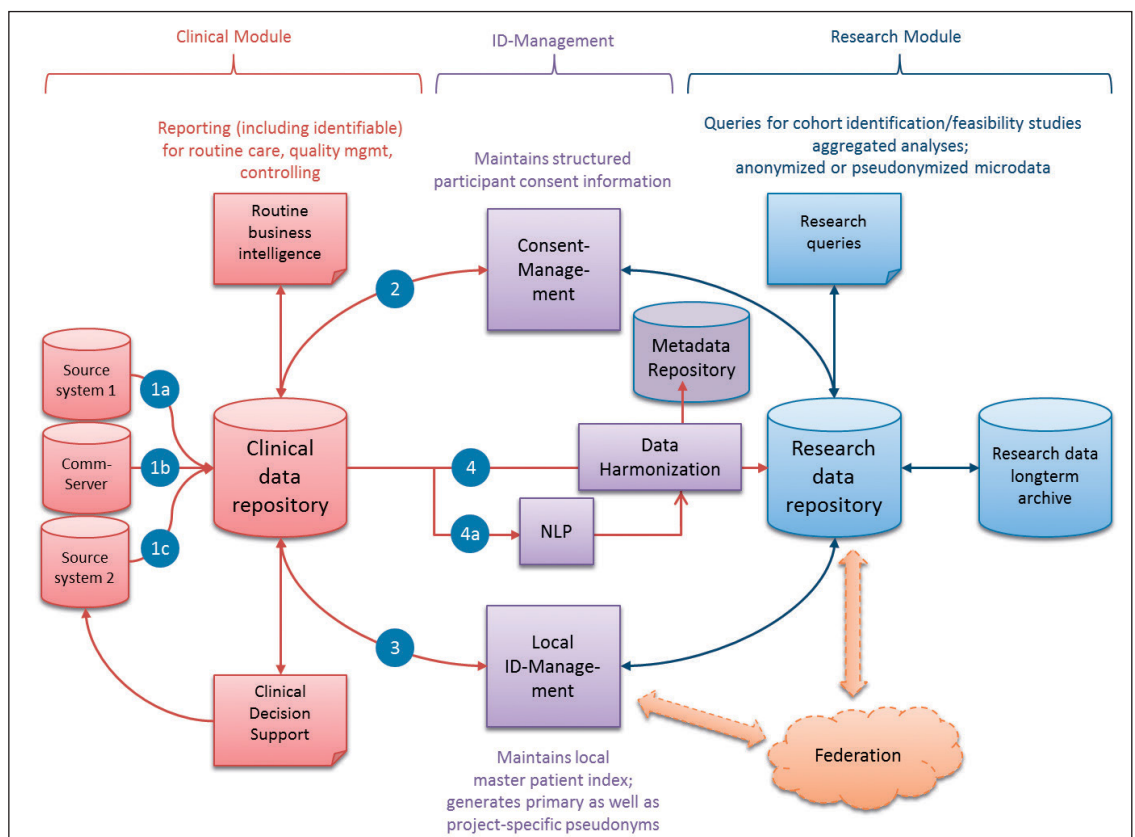
4.5 Data Harmonization

As stated by a multitude of researchers, managing and harmonizing very large amounts of data from different previously established sources is a significant challenge [25, 26, 27]. In the past, for multi-center research studies within the MIRACUM hospitals as well as in most research projects worldwide data heterogeneity was usually addressed on a project-by-project basis: first, the aims of a research project

were defined and then the researchers identified and extracted relevant data from their internal databases. The design of retrospective multi-center projects typically delineates variables of interest (VOIs), which may be different from the variables recorded within the project partners' local original assessment forms and measurement protocols [28]. Thus, finding suitable data for a data integration project and combining data sets wherever variables are comparable is a major challenge and causes repeated data harmonization efforts for every project. To reduce future harmonization efforts, MIRACUM applies a central metadata repository (M-MDR) and a standard data harmonization process (similar to the processes defined by Spjuth and colleagues [28]). For the definition of data elements' metadata an international standard has been agreed upon (ISO/IEC 11179) and the M-MDR has been built compatible to this standard [29] and has already been successfully applied in other research projects (Samply.MDR: e.g. [30, 31]).

Harmonization needs to be pursued on two different levels [28]. First, on the level

Figure 2
Local components and data flows from source systems into the DIC clinical/research data repositories as well as the integration of ID- and consent management platforms. 1. extraction of data from source systems into the clinical data repository (a: through direct database access, b: through HL7 data streams, c: through FHIR resources); 2. integration of consent management; 3. integration of ID management; 4. data harmonization and transfer of de-identified data into the research data repositories. 4a. optional natural language pipelines for narrative text annotations.

of metadata, 'vocabularies' or common data/information models (compare e.g. the PCORNET or the OHDSI OMOP common data models [32]) and second, on the level of the actual patient data themselves. In MIRACUM for the process of the first harmonization level all data elements (and their metadata description, including – in future versions – also provenance information and data quality categories) to be uploaded in the DIC of a MIRACUM site will be first defined within the MIRACUM metadata repository (M-MDR [33]). When describing the data required for a particular use case or research question, clinicians or medical researchers typically set up medical concepts without directly defining available and precisely described data elements. Thus, in an iterative process, moderated by methods researchers or data managers, and including the clinicians and medical researchers, it is first necessary to precisely describe such medical concepts (e.g. data type, validation rules, value lists, links with internationally standardized vocabularies), so that computer scientists can define respective database structures, thus creating a harmonized vocabulary (HV)

for the variables of interest (VOIs). This harmonized vocabulary in MIRACUM will not only be defined with respect to one specific research question or use case, but shall represent the incrementally extended core data terminology of the MIRACUM consortium (compare [25] for an exemplary approach). In a subsequent step, based on this MIRACUM core data terminology we shall define the MIRACUM Common Data Model (CDM). For MIRACULIX 0.9 we have used the OMOP V5 common data model as a starting point for our own work. We believe there is no need to reinvent a common data model from scratch again, and have decided to closely cooperate with the OHDSI project, which is maintaining and extending the OMOP CDM. Similarly, for the development of data harmonization tools, we will closely cooperate with the BBMRI ERIC Common Service IT project and the German Biobank Alliance [34, 35] to assure interoperability with similar large data (and biomaterial) sharing projects.

4.6 Data Sharing and Data Federation

The basic concept for data querying, data analysis and data exploration across all MIRACUM sites (and prospectively, also across different consortias) is data federation (which means that data sets are not combined in one big data store, but rather kept locally at the MIRACUM partner sites). In order to perform joint analyses in this federated environment we follow the example and experiences of the DataSHIELD (Data Aggregation Through Anonymous Summary-statistics from Harmonized Individual levEL Data-bases) concept, which has been proposed to facilitate the co-analysis of individual-level data from multiple studies without physically sharing the data [36] and has been successfully applied in the BioSHaRE Project [25]. This concept has also been taken up by the OHDSI community, which also retains data at the participant's site, simplifying patient and business privacy issues [3]. In the MIRACUM conceptual phase, data federation has been implemented for queries (e.g. in multi-center feasibility studies,
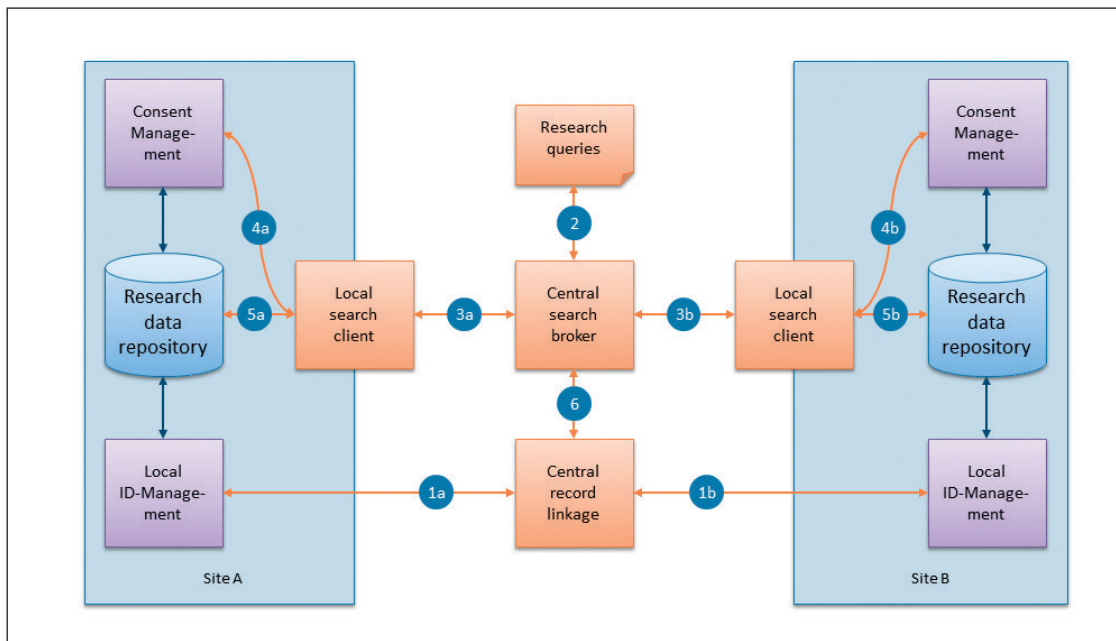

Figure 3   Federated components and data flows to support cross-site record linkage and queries. 1. privacy-preserving record linkage (subject to appropriate consent); 2. research queries are formulated and transferred to a central search broker; 3. the local search clients retrieve queries from the central search broker; 4. access to data is determined according to appropriate consent information; 5. the local research data repositories are queried and results reported back to the central search broker; 6. the central search broker accesses the central record linkage to merge duplicate records (subject to appropriate consent) and reports the aggregated results back to the data consumer.

compare ▶Figure 3) as well as for data analysis. In the development and networking phase federated approaches for machine learning will be added (compare use case 2).

5. Use Cases

The below described clinical/research use cases illustrate exemplarily how MIRACUM hospitals will benefit from the shared use of integrated data resources by applying those for different types of clinically integrated scenarios (e.g. recruiting patients being admitted in the hospital for a clinical trial, or supporting diagnostic and therapeutic decisions based on prediction models and/or molecular analysis results). Such applications will extend currently existing clinical systems (e.g. the EHR system) via integrated small applications (smart apps). In cases were such applications require writing data back to the clinical systems this would be based on standard interfaces, via FHIR resources and based on IHE profiles, to provide a high level of interoperability.

5.1 Alerting in Care – IT Support for Patient Recruitment

Clinical trials (CTs) are the gold standard for testing therapies or new diagnosis techniques that may improve clinical care. However, many trials fail in their objectives, because of the difficulty of meeting the necessary recruitment targets in an effective time and at a reasonable cost [37, 38, 39]. Based on previous research results from a joint project within five German university hospitals [40, 41, 42] as well as the European EHR4CR project [43] we will implement (and integrate into the local EHR system environments) and evaluate a comprehensive IT infrastructure to support and improve efficient patient recruitment processes. Within this use case, eligibility criteria of clinical trials running at all MIRACUM partner sites will be analyzed. Additionally, building on the data set already identified within EHR4CR [44], a core list of data elements, which are typical and most often used in trials shall be defined. Identifying such data items within

the respective local EHR systems, defining them in the M-MDR and integrating them into the MIRACUM data repositories will incrementally extend the size of the MIRACUM DIC core data sets. In parallel, we will also verify the completeness and quality of those data items (compare e.g. [42]) and establish feedback loops into clinical practice in order to increase such measures over the project years.

5.2 From Data to Knowledge – Clinico-molecular Predictive Knowledge Tool

An ever-increasing amount of data (e.g. clinical, longitudinal research data, omics) from patients are being created in our health care systems. Yet, to generate actionable knowledge, these data have to be jointly analyzed in order to identify patterns that are relevant for the treatment of patients. From such patterns, diagnostic and predictive models can be developed that must be transferred back to the patient care setting. Despite the advances in predictive modelling research in recent years, closing the loop into clinical routine – the dissemination and translation of predictive modelling research findings into healthcare delivery – is still challenging. As Khalilia and colleagues [45] have aptly described, in many cases the evaluation of the feasibility of predictive modelling marks the end of a project with no attempt to deploy the developed models into real practice. To unleash their full potential, researchers should aim at deploying and disseminating their algorithms and tools into day-to-day decision support. This is the challenge we have decided to tackle within the second MIRACUM use case. It is our aim to demonstrate for at least two major medical conditions (asthma/COPD and Neuro-Oncology) how to develop, train and evaluate predictive models (including the use of omics data), using machine learning approaches such as deep learning on federated data repositories, and how to implement them as decision support tools for treating physicians in routine care processes. Specifically, we will develop a deep learning approach that will identify patient subgroups from distributed data. For example, this will allow identifying endotypes in the

asthma/COPD application. Subsequent assignment of patients to these groups in routine care will inform personalized treatment. The said clinical projects explicitly emphasize the interdisciplinary approach that is critical for taking advantage of modern integrated medical data. For example, the topic of Neuro-Oncology includes major diagnostic and therapeutic disciplines i.e. Neurosurgery, Neurology, Radiation Oncology, Neuroradiology, Neuropathology and Laboratory Medicine for comprehensive patient management.

5.3 From Knowledge to Action – Support for Molecular Tumor Boards

In the last decade, the development of next-generation sequencing technologies has enabled in-depth genetic characterization of tumor samples. Large national and international consortia including The Cancer Genome Atlas (TCGA) Project and the International Cancer Genome Consortium (ICGC) have sequenced tumors from thousands of patients with over 100 different cancer entities. Databases like COSMIC (Catalogue Of Somatic Mutations In Cancer) harbor the accumulated data and represent the world's largest such repository. The data gained from these research projects has brought tremendous advances to our understanding of cancer biology and to the detection of relevant biomarkers. For many tumors, it is now possible to identify so-called "driver mutations" through in-depth genetic characterization that may be targeted by therapeutic interventions. Despite this progress, the very large, rapidly increasing number of genetic mutations pose an overwhelming diagnostic and clinical challenge in interpreting the importance of these variants for tumor patients. In this context, the annotation of gene variants is an important part of the bioinformatics analysis pipeline. The more accurately they can be characterized in terms of their pathogenicity, the better the classificators stratify patients for possible therapy options. Similarly, the need for studies that examine their relevance for tumor treatment and biology is very large. Moreover, it remains unclear how in-depth molecular characterization of tumors and subsequent

targeting of identified driver lesions can improve the outcome of cancer patients. To answer this important question several clinical trials have been initiated that test the implementation of Molecular Tumor Boards (MTBs) and measure the effectiveness of personalized treatment strategies on patient outcome.

In the MIRACUM conceptual phase, we have already performed an in-depth analysis of the clinician experiences and attitudes towards genome-guided therapy support [46] as well as analysis of activities, processes and IT solutions at all MIRACUM sites [47] to gain a comprehensive understanding of the requirements and the processes involved in MTBs across these institutions. Further, a comprehensive literature review was performed to learn from experiences of previous research towards the integration of pharmacogenomics testing and molecular-guided therapy decisions in clinical care environments [48]. To cope with the complexity of the generated tumor sequencing data all MIRACUM sites have commenced in efforts to implement multi-disciplinary MTBs. Towards harmonization of the currently heterogeneous organization of molecular tumor boards at the individual MIRACUM sites we have already identified common processes that can be significantly supported and improved by new IT solutions [47]. Those are (1) analysis of the sequencing data from several sources and platforms, (2) annotation of genetic variants for clinical interpretation, (3) presentation of the analysis results, (4) integration of the MTB into the clinical workflow with documentation in the EHR system and (5) archiving of data and analysis results. Thus, within our third use case we aim at establishing a generic framework supporting all steps from the analysis of omics data, their interpretation leading to a final therapy decision in the MTBs and its documentation in the EHR at all MIRACUM partner sites. This requires a close collaboration with the members of the interdisciplinary MTB. To support the interpretation of the complex and elaborate tumor analysis, MIRACUM patient visualization modules will be incorporated in the MTB platform for state-of-the-art presentation of total mutational burden and annotated mutations within a signal pathway of interest.

6. First Results

Internationally large scale data reuse and data sharing initiatives have already been initiated some years ago (e.g. [1, 3–10, 43]). Many of those have developed tools which have already been successfully applied in different research contexts. Researchers have also shown, that multiple approaches for data sharing networks can coexist and ETL processes as well as data repositories can still be used for varying networking approaches [49]. Thus, the MIRACUM partners have decided not to reinvent all those tools, but instead to apply as many successful concepts, architectures and tools as possible. This paradigm is manifested in our MIRACOLIX ecosystem (**M**edical **In**formatics **R**eus**A**ble e**C**o-system of **O**pen source **L**inkable and **I**nteroperable software tools) and has proven successful in the last year.

Despite the short time frame of the BMBF MI-I conceptual phase (nine months) MIRACUM has realized first achievements based on the data integration center infrastructures described above. DIC research repositories have been implemented at all eight MIRACUM university hospitals based on the i2b2 suite and an OMOP PostgreSQL database. ETL processes were modelled and implemented in Talend Open Studio. All components were provided to the MIRACUM partners for download as virtual machine images for VirtualBox as well as VMware in Owncloud together with extensive documentation. The actual data repository loading has been performed at each MIRACUM site locally based on the NSC core dataset definition (patient demographics, encounters, diagnosis and procedures). Such data were mostly available for the years 2004–2016 except two sites whose data reached only back to 2008/2009. Overall, the data of about 3,000,000 patients with 70,000,000 facts was made available for two different studies. The i2b2 installations were used for a federated feasibility querying prototype, which was applied for the identification of stroke and colorectal cancer cohorts.

Analysis packages for research questions concerning those two cohorts were created and tested at local sites (Erlangen and Freiburg) and then made available for distribution to the other sites. After approval of such analysis by the local use and access committees (UAC), the predefined analysis were retrieved at each side and executed on the respective local data repositories. Research questions analyzed in the DIC's stroke patients subcohort focused on acute treatment measures of acute ischemic stroke patients (and their development over time) between 2010 and 2016 [50]. The colorectal cancer cohort was analyzed to compare the distribution and pathway of therapeutic procedures within this patient subset. Similar to the approach taken by the OHDSI consortium [2] results were visualized as sunburst plots [51]. The latter analysis was even extended to include three additional hospitals of the HD4CR consortium [52] within a period of only two months.

7. Discussion

Based on the first successful studies above described, MIRACUM will continue to expand its DICs in the upcoming years following the needs of the three clinical/research scenarios described in the use cases. Defining efficient software development strategies (e.g. SRCUM) as well as development, unit testing, integration testing and deployment environments will be essential for releasing new DIC versions. Quality management, IT security, data protection, privacy by design (compare e.g. [53, 54]) will be major cornerstones for successful further development.

Despite all such technological challenges however, applying the MIRACUM tools for the enrichment of our knowledge about diagnostic and therapeutic concepts, thus supporting the concept of a Learning Health System [55] will be crucial for the acceptance and sustainability of our work in the medical community and the MIRACUM university hospitals. Therefore, additional large scale data analysis will be continuously developed and performed

(e.g. scenarios and research questions for psychiatric patients as well as for patients with rare diseases are already under development). As already illustrated in the conceptual phase, MIRACUM will also very actively contribute to the National Steering Committee working groups and be open for further cross-consortial analysis.

Sustaining the MIRACUM efforts will depend on two major factors: (1) The proven value of the DIC for clinical care as well as translational research and based on this, the continuation of stakeholder support in the board of directors of our university hospitals and medical faculties. (2) On future cooperation and alignment with similar large international projects. Our current cooperation with OHDSI and the world wide i2b2/tranSMART foundation and major partner's involvement in BBMRI-ERIC are important first steps on this way. Nevertheless, all such cooperations across borders are on risk, because in Germany we are still not able to rely on an internationally applied standard clinical nomenclature such as SNOMED CT. The positive result of current nation-wide licensing discussions are therefore very important for achieving international comparable research results within large scale cross-country networks.

Acknowledgment

We are grateful to the many members of the MIRACUM team who have actively participated in the design and first implementation of the described architecture and data integration centers.

References

1. Meystre SM, Lovis C, Bürkle T, Tognola G, Budrionis A, Lehmann CU. Clinical Data Reuse or Secondary Use: Current Status and Potential Future Progress. Yearb Med Inform 2017; 26(1): 38–52.

2. Hripcsak G, Ryan PB, Duke JD, Shah NH, Park RW, Huser V, Suchard MA, Schuemie MJ, DeFalco FJ, Perotte A, Banda JM, Reich CG, Schilling LM, Matheny ME, Meeker D, Pratt N, Madigan D. Characterizing treatment pathways at scale using the OHDSI network. Proc Natl Acad Sci U S A 2016; 113(27): 7329–7336.

3. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, Suchard MA, Park RW, Wong IC, Rijnbeek PR, van der Lei J, Pratt N, Norén GN, Li YC, Stang PE, Madigan D, Ryan PB. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform 2015; 216: 574–578.

4. Collins FS, Hudson KL, Briggs JP, Lauer MS. PCORnet: turning a dream into reality. J Am Med Inform Assoc 2014; 21(4): 576–577.

5. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. J Am Med Inform Assoc 2014; 21(4): 578–582.

6. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, Sanderson SC, Kannry J, Zinberg R, Basford MA, Brilliant M, Carey DJ, Chisholm RL, Chute CG, Connolly JJ, Crosslin D, Denny JC, Gallego CJ, Haines JL, Hakonarson H, Harley J, Jarvik GP, Kohane I, Kullo IJ, Larson EB, McCarty C, Ritchie MD, Roden DM, Smith ME, Böttinger EP, Williams MS; eMERGE Network. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genet Med 2013; 15(10): 761–771.

7. Gainotti S, Torreri P, Wang CM, Reihs R, Mueller H, Heslop E, Roos M, Badowska DM, de Paulis F, Kodra Y, Carta C, Martìn EL, Miller VR, Filocamo M, Mora M, Thompson M, Rubinstein Y, Posada de la Paz M, Monaco L, Lochmüller H, Taruscio D. The RD-Connect Registry & Biobank Finder: a tool for sharing aggregated data and metadata among rare disease researchers. Eur J Hum Genet 2018. Forthcoming. doi:10.1038/s41431–017–0085-z.

8. Biobanking and Biomolecular Resources Infrastructure –European Research Infrastructure Consortium [cited 2018 Mar 28]. Available from: http://www.bbmri-eric.eu/.

9. ELIXIR: A distributed infrastructure for life-science information [cited 2018 Mar 28]. Available from: https://www.elixir-europe.org/.

10. Eijssen L, Evelo C, Kok R, Mons B, Hooft R; other founding members of DTL Data (see Acknowledgements). The Dutch Techcentre for Life Sciences: Enabling data-intensive life science research in the Netherlands. Version 2. F1000Res 2015; 4: 33.

11. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJ, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PA, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016; 3: 160018.

12. Semler SC, Wissing F, Heyder R. German Medical Informatics Initiative: A national approach to integrating health data from patient care and medical research. Methods Inf Med 2018; 57(Open): e50–e56.

13. MIRACUM (Medical Informatics for Research and Care in University Medicine) [cited 2018 Mar 28]. Available from: http://www.miracum.org/.

14. Pommerening K, Drepper J, Helbing K, Ganslandt T. Leitfaden zum Datenschutz in medizinischen Forschungsprojekten. TMF Schriftenreihe. Berlin: Medizinische Wissenschaftliche Verlagsgesellschaft; 2014.

15. Christoph J, Griebel L, Leb I, Engel I, Köpcke F, Toddenroth D, Prokosch HU, Laufer J, Marquardt K, Sedlmayr M. Secure Secondary Use of Clinical Data with Cloud-based NLP Services. Towards a Highly Scalable Research Infrastructure. Methods Inf Med 2015; 54(3): 276–282.

16. Bialke M, Bahls T, Havemann C, Piegsa J, Weitmann K, Wegner T, Hoffmann W. MOSAIC – A Modular Approach to Data Management in Epidemiological Studies. Methods Inf Med 2015; 54(4): 364–371.

17. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, Kohane I: Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc 2010; 17(2): 124–130.

18. Kohane IS, Churchill SE, Murphy SN. A translational engine at the national scale: informatics for integrating biology and the bedside. J Am Med Inform Assoc 2012; 19(2): 181–185.

19. Scheufele E, Aronzon D, Coopersmith R, McDuffie MT, Kapoor M, Uhrich CA, Avitabile JE, Liu J, Housman D, Palchuk MB. tranSMART: An Open Source Knowledge Management and High Content Data Analytics Platform. AMIA Jt Summits Transl Sci Proc 2014; 2014: 96–101.

20. Christoph J, Knell C, Naschberger E, Stürzl M, Maier C, Prokosch HU, Sedlmayr M. Two years of tranSMART in a university hospital for translational research and education. Stud Health Technol Inform 2017; 236: 70–79.

21. Christoph J, Knell C, Bosserhoff A, Naschberger E, Stürzl M, Rübner M, Seuss H, Ruh M, Prokosch HU, Sedlmayr B. Usability and Suitability of the Omics-Integrating Analysis Platform tranSMART for Translational Research and Education. ACI 2017; 8(4): 1173–1183.

22. Marcus DS, Archie KA, Olsen TR, Ramaratnam M. The open-source neuroimaging research enterprise. J Digit Imaging 2007; 20 Suppl 1: 130–138.

23. Herrick R, Horton W, Olsen T, McKay M, Archie KA, Marcus DS. XNAT Central: Open sourcing imaging research data. NeuroImage 2016; 124, Part B: 1093–1096.

24. He S, Yong M, Matthews PM, Guo Y. tranSMART-XNAT Connector tranSMART-XNAT connector – image selection based on clinical phenotypes and genetic profiles. Bioinformatics 2016; btw714.

25. Doiron D, Burton P, Marcon Y, Gaye A, Wolffenbuttel BH, Perola M, Stolk RP, Foco L, Minelli C, Waldenberger M, Holle R, Kvaløy K, Hillege HL, Tassé AM, Ferretti V, Fortier I. Data harmonization and federated analysis of population-based studies: the BioSHaRE project. Emerg Themes Epidemiol 2013; 10(1): 12.

26. Fortier I, Burton PR, Robson PJ, Ferretti V, Little J, L'Heureux F, Deschenes M, Knoppers BM, Doiron D, Keers JC, Linksted P, Harris JR, Lachance G, Boileau C, Pedersen NL, Hamilton CM, Hveem K, Borugian MJ, Gallagher RP, McLaughlin J, Parker L, Potter JD, Gallacher J, Kaaks R, Liu B, Sprosen T, Vilain A, Atkinson SA, Rengifo A, Morton R, Metspalu A, Wichmann HE, Tremblay M, Chisholm RL, Garcia-Montero A, Hillege H, Litton JE,

Palmer LJ, Perola M, Wolffenbuttel BH, Peltonen L, Hudson TJ. Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies. Int J Epidemiol 2010; 39(5): 1383–1393.

27. Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, Hide W, Hill DP, Kania R, Schaeffer M, St Pierre S. Big data: the future of biocuration. Nature 2008; 455(7209): 47–50.

28. Spjuth O, Krestyaninova M, Hastings J, Shen HY, Heikkinen J, Waldenberger M, Langhammer A, Ladenvall C, Esko T, Persson MÅ, Heggland J, Dietrich J, Ose S, Gieger C, Ried JS, Peters A, Fortier I, de Geus EJ, Klovins J, Zaharenko L, Willemsen G, Hottenga JJ, Litton JE, Karvanen J, Boomsma DI, Groop L, Rung J, Palmgren J, Pedersen NL, McCarthy MI, van Duijn CM, Hveem K, Metspalu A, Ripatti S, Prokopenko I, Harris JR: Harmonising and linking biomedical and clinical data across disparate data archives to enable integrative cross-biobank research. Eur J Hum Genet 2016; 24(4): 521–528.

29. Kadioglu D, Weingardt P,Ückert F, Wagner T. Samply.MDR – Ein Open-Source-Metadaten-Repository. HEC 2016: Health – Exploring Complexity. Joint Conference of GMDS, DGEpi, IEA-EEF, EFMI. München, 28.08.-02.09.2016. Düsseldorf: German Medical Science GMS Publishing House; 2016. DocAbstr. 425.

30. Lablans M, Kadioglu D, Muscholl M,Ückert F. Exploiting Distributed, Heterogeneous and Sensitive Data Stocks while Maintaining the Owner's Data Sovereignty. Methods Inf Med 2015; 54(4): 346–352.

31. Storf H, Schaaf J, Kadioglu D, Göbel J, Wagner TOF, Ückert F. [Registries for rare diseases: OSSE – An open-source framework for technical implementation]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz. 2017 Mar 13. doi: 10.1007/s00103–017–2536–7.

32. Garza M, Del Fiol G, Tenenbaum J, Walden A, Zozus MN. Evaluating common data models for use with a longitudinal community registry. J Biomed Inform 2016; 64: 333–341.

33. MIRACUM Metadata Repository. M-MDR [cited 2018 Mar 28]. Available from: http://mdr.miracum.de/.

34. Mate S, Kadioglu D, Majeed RW, Stöhr MR, Folz M, Vormstein P, Storf H, Brucker DP, Keune D, Zerbe N, Hummel M, Senghas K, Prokosch HU, Lablans M. Proof-of-Concept Integration of Heterogeneous Biobank IT Infrastructures into a Hybrid Biobanking Network. Stud Health Technol Inform 2017; 243: 100–104.

35. Mate S, Vormstein P, Kadioglu D, Majeed RW, Lablans M, Prokosch HU, Storf H. On-The-Fly Query Translation Between i2b2 and Samply in the German Biobank Node (GBN) Prototypes. Stud Health Technol Inform 2017; 243: 42–46.

36. Gaye A, Marcon Y, Isaeva J, LaFlamme P, Turner A, Jones EM, Minion J, Boyd AW, Newby CJ, Nuotio ML, Wilson R, Butters O, Murtagh B, Demir I, Doiron D, Giepmans L, Wallace SE, Budin-Ljøsne I, Oliver Schmidt C, Boffetta P, Boniol M, Bota M, Carter KW, deKlerk N, Dibben C, Francis RW, Hiekkalinna T, Hveem K, Kvaløy K, Millar S, Perry IJ, Peters A, Phillips CM, Popham F, Raab G, Reischl E, Sheehan N, Waldenberger M, Perola M, van den Heuvel E, Macleod J, Knoppers BM, Stolk RP, Fortier I, Harris JR, Woffenbuttel BH, Murtagh MJ, Ferretti V, Burton PR. DataSHIELD: taking the analysis to the data, not the data to the analysis. Int J Epidemiol 2014; 43(6): 1929–1944.

37. Cuggia M, Besana P, Glasspool D. Comparing semi-automatic systems for recruitment of patients to clinical trials. Int J Med Inform 2011; 80(6): 371–388.

38. Dilts DM, Sandler AB. Invisible barriers to clinical trials: the impact of structural, infrastructural, and procedural barriers to opening oncology clinical trials. J Clin Oncol 2006; 24: 4545–4552.

39. Campbell MK, Snowdon C, Francis D, et al. Recruitment to randomised trials: strategies for trial enrollment and participation study. The STEPS study. Health Technol Assess 2007; 11: iii, ix–105.

40. Trinczek B, Köpcke F, Leusch T, Majeed RW, Schreiweis B, Wenk J, Bergh B, Ohmann C, Röhrig R, Prokosch HU, Dugas M. Design and multicentric implementation of a generic software architecture for patient recruitment systems re-using existing HIS tools and routine patient data. Appl Clin Inform 2014; 5(1): 264–283.

41. Schreiweis B, Trinczek B, Köpcke F, Leusch T, Majeed RW, Wenk J, Bergh B, Ohmann C, Röhrig R, Dugas M, Prokosch HU. Comparison of electronic health record system functionalities to support the patient recruitment process in clinical trials. Int J Med Inform 2014; 83(11): 860–868.

42. Köpcke F, Trinczek B, Majeed RW, Schreiweis B, Wenk J, Leusch T, Ganslandt T, Ohmann C, Bergh B, Röhrig R, Dugas M, Prokosch HU. Evaluation of data completeness in the electronic health record for the purpose of patient recruitment into clinical trials: a retrospective analysis of element presence. BMC Med Inform Decis Mak 2013; 13: 37.

43. De Moor G, Sundgren M, Kalra D, Schmidt A, Dugas M, Claerhout B, Karakoyun T, Ohmann C, Lastic PY, Ammour N, Kush R, Dupont D, Cuggia M, Daniel C, Thienpont G, Coorevits P. Using electronic health records for clinical research: the case of the EHR4CR project. J Biomed Inform 2015; 53: 162–173.

44. Doods J, Lafitte C, Ulliac-Sagnes N, Proeve J, Botteri F, Walls R, Sykes A, Dugas M, Fritz F. A European inventory of data elements for patient recruitment. Stud Health Technol Inform 2015; 210: 506–510.

45. Khalilia M, Choi M, Henderson A, Iyengar S, Braunstein M, Sun J. Clinical Predictive Modeling Development and Deployment through FHIR Web Services. AMIA Annu Symp Proc 2015; 2015: 717–726.

46. Hinderer M, Boeker M, Wagner SA, Binder H, Ückert F, Hülsemann JL, Neumaier M, Schade-Brittinger C, Acker T, Prokosch HU, Sedlmayr B. The experience of physicians in pharmacogenomic clinical decision support within eight German University Hospitals. Pharmacogenomics 2017; 18(8): 773–785.

47. Hinderer M, Börries M, Haller F, Wagner S, Sollfrank S, Acker T, Prokosch HU, Christoph J. Supporting Molecular Tumor Boards in Molecular-guided Decision-making – the Current Status of Five German University Hospitals. Stud Health Technol Inform 2017; 236: 48–54.

48. Hinderer M, Boeker M, Wagner SA, Lablans M, Newe S, Hülsemann JL, Neumaier M, Binder H, Renz H, Acker T, Prokosch HU, Sedlmayr M. Integrating clinical decision support systems for pharmacogenomic testing into clinical routine – a scoping review of designs of user-system interactions in recent system development. BMC Med Inform Decis Mak 2017; 17(1): 81.

49. Klann JG, Abend A, Raghavan VA, Mandl KD, Murphy SN. Data interchange using i2b2. J Am Med Inform Assoc 2016; 23(5): 909–915.

50. Haverkamp C, Gansland T, Horki P, Boeker M, Dörfler A, Schwab S, Berkefeld J, Pfeilschifter W, Niesen W-D, Egger K, Kaps M, Brockmann MA, Neumaier-Probst E, Szabo K, Skalej M, Bien S, Best C, Prokosch U, Urbach H. Regional differences in thrombectomy rates: secondary use of billing codes in the MIRACUM (Medical Informatics for Research and Care in University Medicine) Consortium. Clin Neuroradiol 2018; 28(2): 225–234.

51. Maier C, Lang L, Storf H, Vormstein P, Bieber R, Bernarding J, Herrmann T, Haverkamp C, Horki P, Laufer J, Berger F, Höning G, Fritsch HW, Schüttler J, Ganslandt T, Prokosch HU, Sedlmayr M. Towards implementation of OMOP in a German university hospital consortium. Appl Clin Inform 2018; 9(1): 54–61.

52. Available from: http://www.medizininformatik-initiative.de/de/konsortien/hd4cr-konzeptphase [cited 2018 Mar 22].

53. Prasser F, Kohlmayer F, Lautenschläger R, Kuhn KA. ARX – A Comprehensive Tool for Anonymizing Biomedical Data. AMIA Annu Symp Proc 2014; 2014: 984–993.

54. Prasser F, Kohlmayer F, Kuhn KA. Efficient and effective pruning strategies for health data de-identification. BMC Medical Informatics and Decision Making 2016; 16: 49.

55. Friedman C, Rubin J, Brown J, Buntin M, Corn M, Etheredge L, Gunter C, Musen M, Platt R, Stead W, Sullivan K, Van Houweling D. Toward a science of learning systems: a research agenda for the high-functioning Learning Health System. J Am Med Inform Assoc 2015; 22(1): 43–50.