Proceedings of the 54th Hawaii International Conference on System Sciences | 2021

Data Sovereignty in Data Donation Cycles - Requirements and Enabling Technologies for the Data-driven Development of Health Applications

Markus Schinle
FZI Research Center for
Information Technology
schinle@fzi.de

Christina Erler
Karlsruhe Institute of
Technology
christina.erler@kit.edu

Wilhelm Stork
Karlsruhe Institute of
Technology
wilhelm.stork@kit.edu

Abstract

Personalized healthcare is expected to increase the efficiency and the effectiveness of health services using different kinds of algorithms on existing data. This approach is currently confronted with the lack of digital data and the desire for self-determined personal data handling. However, the issue of health data donation is on the political agenda of some governments. Within this work, a knowledge base will be created by reviewing existing approaches and technologies regarding this topic with the focus on chronic diseases. A list of requirements will be derived from which we conceptualize a data donation cycle to demonstrate the challenges and opportunities of health data sovereignty and its future possibilities concerning data-driven health application development. By linking the requirements to technological approaches, the baseline for future open ecosystems will be presented.

1. Introduction and Research Question

Personalized services will play a major role in future medical care. From today's perspective, this data-centered technological development faces interdisciplinary challenges in the form of ethical, legal, structural, and economic requirements, but also technological implications [1]. In addition to a politically motivated digital sovereignty of individual citizens, there are the interests of organizations to profitably incorporate data-hungry algorithms into products and services that demonstrably provide benefits for patients and medical professionals to their individual needs. Questions concerning the collection and use of digital health data have become the focus of public attention in recent years. This area of tension is being rebalanced by the current COVID-19 pandemic. In general, the use of personal data in medical applications got more relevant during the pandemic. So far, it has become visible how complex decisions are to weigh up the options regarding the use of personal data through government intervention or voluntary participation of citizens e.g. the use of the RKI Corona Data Donation App[1] to better understand the current COVID-19 pandemic in Germany. This illuminates the challenges in the context of the realization of data donation to support data-driven development of medical applications. The interdisciplinary nature of this topic led to discussions in several fields and ended at least in ethical, legal, structural, economic, medical and technological questions [2, 3, 4]. This paper will primarily address the last four perspectives as implications for self-determined medical data donation. Therewith, we want to find answers to the central research question, which requirements and technological prerequisites are necessary concerning a sustainable integration of health data in the development of medical applications. As chronic diseases are one of the main causes of rising healthcare costs and are seen as promising to benefit from digital medical applications, we focus on them [5].

2. Basics

As trust is one of the key factors regarding the sharing of personal data, trusted infrastructures are recently discussed as a base for data sharing in health (see subsection 4.1). Distributed Ledger Technology (DLT) emerged as a foundation for trusted infrastructures because of its decentralized and immutable way of linking transaction data [6]. One representative of DLT is Blockchain, which links the transaction histories (blocks) via linked lists to generate a ledger (chain) so that these are immutable [7]. For this purpose, each block references the previous block via the hash code of them. If a participant with fraudulent intentions changes a transaction from a previous block, the hashes of all following blocks change, and the change is recognized as invalid. Another concept in the context of service-oriented application development is the use of containerization technology. To overcome conflicting or missing dependencies and moving applications and corresponding from one system to another without

[1] https://corona-datenspende.de, accessed October 2020

URI: https://hdl.handle.net/10125/71098
978-0-9981331-4-0

losing executability, the bundling to so-called virtual containers is used to isolate software, data, libraries, and configuration files [8]. Using this technology allows the encapsulation of data and services by ensuring compatibility and resource-efficient virtualization. To create interoperability and integration of information and services, the interdisciplinary research area Semantic Web emerged [9]. For example, through semantic annotations using metadata on the Web, information retrieval can be improved. Semantic interoperability plays an essential role in the integration of health data, which is why modern medical data standards are supported by ontologies defined with Semantic Web technologies. Health information is one of the special types of personal data and is therefore particularly protected by law e.g. the General Data Protection Regulation (GDPR) in the EU. Therefore, it may only be collected, stored, used, and processed under strict conditions e.g. by patient consent or pseudo- or anonymization. Personal data are considered as pseudonymous when the data subject can no longer be identified without the use of additional data, whereas in the case of anonymous data, the personal reference is changed to such an extent that the data subject can no longer be identified [10]. When personal data is provided, the consent of the patient can be defined by stating a purpose for data processing. There are different models of patient consent management and approaches that allow e.g. the digital mapping of consents by using Semantic Web technologies, which semantically describes what can be done with the data, by whom and in what form [11].

3. Methodology

The following sections describe the research method and the scientific approach. As research method, we applied the Design Science Research Methodology (DSRM) provided by Peffers et al. (see subsection 3.1) [12]. For getting a broad knowledge base, we applied a structured literature review (see subsection 3.2).

3.1. Design Science Research

The DSRM aims to develop and evaluate new IT artifacts in the form of an agile process which iteratively combines the requirements of the theory with the practice [12]. The methodology proposed by Peffers et al. is mainly used for information systems research and consists of the following six process steps [12] :
(1) Problem identification and motivation; (2) Definition of objectives of solution; (3) Design and development of the solution artifact; (4) Demonstration of the solution artifact; (5) Evaluation of the effectiveness and efficiency; (6) Communication. The first four sections of this paper represent step 1 and 2 by explaining and motivating the research problem and defining the objectives of the solution artifact. Based on results of our conducted literature review, requirements and ideas for a possible solution to our research question are derived as IT artifacts. The findings, serve as the foundation for the 3. and 4. step presented as artifacts in the subsections 4.2. This paper serves as communication tool for the 6. step of the DSRM and concludes with the discussion of the results and presentation of connection points for possible future work linked to step 5.

3.2. Literature Review

In order to identify requirements and approaches for data donation in the context of chronic diseases, the following search string was defined for the structured literature review: *"(data OR information OR knowledge) AND (sharing OR donation) AND health AND (patient OR personal* OR sovereign OR disease) AND (mental OR diabetes OR cancer OR dementia OR oncology)"*. With the help of the search string the following scientific databases were searched to find publications, which contain these catchwords in their title, abstract or keywords: IEEE Xplore, ACM Digital Library, ScienceDirect and EBSCOhost. Through this search and filtering of duplicates, 740 potentially relevant publications could be found. Subsequently, a two-stage manual filtering was carried out by the authors. In the first step the title, abstract and keywords were examined in detail to identify relevance to our research problem. Potentially relevant publications were then flagged for a more detailed examination of the full texts, which was part of the second filtering step. After all 35 publications were marked as relevant. In addition, their forward and backward references were taken into account, which led to a total of 47 relevant publications. These results were cross-validated and the identified requirements were synthesized among fellow researchers.

4. Approach

The following section is intended to describe answers to the defined research question in the form of artifacts. First, approaches from the literature are considered in order to derive requirements for a possible solution. Based on the requirements, we describe a schematic approach for data-driven development of medical applications and suitable technologies for an ecosystem that supports it.

4.1. Approaches from Literature Review

The collection and use of digital health data is becoming increasingly relevant due to global developments in the digitization of the healthcare systems through personal or electronic health records (PHR or EHR), Internet of Things (IoT) sensors, artificial intelligence (AI) and other emerging technologies [13, 3]. Research on sharing medical data has been active for several years e.g. Rao et al. present a software tool called Collaborative Data Network (CDN) for clinical information sharing and querying using HL7 standard in 2010 [14, 15]. Nevertheless, there is currently still a lack of uniform and interoperable infrastructures to gather medical data for health research and product development that are available to the general public for participation due to the fact that many technical, structural, legal and ethical issues are unresolved [3].

4.1.1. Global Data Donation Efforts

MyHealthRecord[2] is the national health record system of Australia in which patient data is stored and made available to the citizens. It enables patients to share health information with doctors, hospitals and other health care providers. The use and participation in the system is activated by default for every Australian and must be actively contradicted. Data from MyHealthRecord is not yet available for research purposes, but it is planned. The government is currently waiting until robust processes and governance, security, privacy and technical arrangements are in place [3]. The FINDATA[3] authority is the central national authority for the management and release of health and social data of the finnish population, which was introduced by law that came into force in Finland on May 2019. To this end, it informs researchers about data availability, issues permits for secondary use and enables pseudonymized data processing using their provided secure environments. The access to the health and social data should be possible in Januar 2021 through the national eHealth infrastructure Kanta[4]. 23andMe[5] is a US biotechnology company which sell direct-to-consumer genetic tests based on saliva samples to identify the ancestry and the genetic predispositions to specific diseases and conditions. The business model also includes the voluntary possibility to release the test results combined with web behavior information and self-reported information to third parties (e.g. academic institutions or private companies) for research purposes, commercial applications, patents or operating licenses, which in return pay money to 23andMe. Due to this business model and the ethical implications, the company became the focus of media, politics, science and even the U.S. Food and Drug Administration (FDA) [16, 3]. Beside governmental and commercial data donation approaches there are also scientific ones and there seems to be a trend towards more transparency regarding the communication of research results to enable further knowledge, research and re-analysis. However, this also requires the active involvement of funders and journals, which are currently collecting research results [17]. Often, incomplete data are available or the re-use of this data is made difficult, which leads to financial effects, biases, less benefits for research and the care of patients [18]. To avoid these problems Chan et al. make three recommendations [18]: (1) Academic institutions and funders should be rewarded for fully sharing their research data; (2) Legislators, ethics committees, funders and journals should enforce policies for study registrations and the availability of full research data; (3) Standards are to be developed for the content of protocols, full study reports and data exchange practices. Dugas et al. present a community-driven multilingual platform for the exchange and discussion of data models in medical research and healthcare to improve and accelerate the design of medical data models by exchanging best practices and more standardized data models with semantic annotations [19]. To face the described problems on national level, the Medical Informatics Initiative (MII)[6] aims to establish an infrastructure for the integration of clinical data from patient care and medical research in Germany [20]. This funding is currently in the development and networking phase, in which each of the four consortia is setting up a data integration center in cooperation with the participating university hospital and partner institutions to develop IT solutions for specific use cases in different medical domains [21, 22, 23, 24]. These data integration centres should first demonstrate how data, information and knowledge from patient care, clinical and biomedical research can be linked across the borders of different locations. At present, these focus on university medical institutions and do not yet address medical practices, regional hospitals or patient care in inpatient and outpatient settings. A research project at European level provides the project MyHealthMyData (MHMD)[7]. The project aims to create the first open biomedical information network focusing on the link between organizations and individuals. It wants to encourage hospitals to start providing anonymized data for open research and citizens to make the most out of the available health data and become the owner of their data. They investigate Blockchain technology as the key element to connect the healthcare stakeholders and manage the data sharing process in a proactive, secure, open and decentralized manner [25, 26]. Therefore,

[2] https://www.myhealthrecord.gov.au/, accessed September 2020

[3] https://www.findata.fi/, accessed September 2020

[4] https://www.kanta.fi/, accessed September 2020

[5] https://www.23andme.com/, accessed September 2020

[6] https://www.medizininformatik-initiative.de, accessed July 2020

[7] http://www.myhealthmydata.eu/, accessed July 2020

they use a federated Blockchain network in which the authorized participants define the access control policies in terms of the Blockchain. An off-chain approach is used for health data storage to maintain GDPR compliance, since only references to the data via hash values are stored in the Blockchain and the sensitive health data does not leave its storage facilities [25, 26]. This off-chain approach is discussed and used in the healthcare context by several other identified publications [27, 28, 13, 29]. In the MHMD solution, the various participants in the network prove their identity through certificates issued by a certificate authority. After the data has been fed into the network by the data owner, a data consumer can request the data based on these certificates. Combined with user-defined consent settings, a data owner can manage and authorize data exchange and access to their own health data to get the opportunity to track, who has used the data, when and under which condition. The data can be found by the data consumers using metadata summarized in a central catalogue. Gordon et al. [29] discuss the challenges and limitations for facilitating the use of Blockchain technology in healthcare services. In particular, the topics of scalability, creation of incentives for the further development and maintenance of the interoperability solution, patient key management and patient engagement need to be addressed. To ensure privacy-by-design harmonization tools, encryption and de-identification technologies are combined with the MHMD solution [25, 26]. The overall objective of MHMD is to create a trustworthy information marketplace that offers added value to citizens, hospitals, research centres and businesses. Jaremko et al. address the trust issue regarding data transfer and processing as well as regarding the data consumer, by processing the data completely anonymously and only for the intended purposes also by envisaging Blockchain technologies in combination with containerization as well as validated methods for the de-identification of data [30].

4.1.2. Data Sharing for Personalized Medicine

Further challenges of embedding data sharing solutions in scientific, medical and patient settings are seen in ethical, legal and regulatory questions by Lawler et al. [4]. They consider in particular, data sharing for the combination of clinical and genomic data as an important aspect for personalized medicine and cancer research, but see genomic data as hard to anonymize. The study of Pickard examined how financial compensation affects the willingness to share health data [31]. Two thirds of the participants would be more motivated to share the information through a financial reward. The possibility of gaining more insight into current state of health of an individual and more self-determination as well as opportunities for participation in the treatment process could be a motivation, especially for people suffering from chronic diseases [32]. When managing their chronic disease, affected people are confronted with the continuous recording and management of their conditional data. Patient-centered self-management applications are particularly important for improving the diagnosis and treatment of chronic diseases because they enable remote data collection and monitoring [5]. Herewith, those applications should handle the sensitive data safe from scratch in order to guarantee general security and privacy. Accordingly, Al-Taee et al. identify security requirements needed to build up a solution for enhancing security and personal privacy in mobile health systems [33]. The main focus of the solution is the use of data protection mechanisms, cryptographic methods, access control and auditable procedures that give patients and healthcare professionals the right to control the disclosure of identifiable health data in this case. Further self-management systems (SMS) and assistive technologies with the aim of supporting patients with different chronic diseases like diabetes [5], serious mental illness [34] and dementia [35] could be determined. The continuous and frequent monitoring of longitudinal data in outpatient home environments, which provides the basis for analysis, is often not yet available to healthcare providers or research [36]. Ma et al. present an open source cloud platform for the exchange of health data and support for outpatient chronic disease care [37]. Especially, the aspects interoperability, security and privacy are identified as the key challenges to effectively exchange health data. Hu et al. introduced a hybrid cloud-based solution to share the data collected in this environment with healthcare provider via a private cloud as well as with third party health services and researchers via a public cloud to promote the treatment of chronic diseases [36]. The willingness to provide patient-generated data to research and central registers depends on whether it is possible to receive information and knowledge in return. In a study among cancer survivors, Smith et al. found that the participants were willing to exchange information on the long-term effects and treatment of cancer, but expressed the wish to receive further evidence based information from the central registries, which could be helpful for their self-management and shared care beyond the cancer [38, 39]. As SMS come along with the cognitive intake of large amounts of health information and communication with health care providers to empower informed decisions, both need visual tools to simplify and manage complex information [40]. Rajwan et al. provide a guideline for the design and implementation of such visualizations [40]. The study of O'Kane et al. found that there were discrepancies between the information needs of patients and clinicians [41]. Clinicians are mostly interested in the data itself and the needs of the patient. In contrast, patients want to get knowledge from the available data to better monitor and understand their health condition. Accordingly, the visualization must be adapted to these needs. In addition to the visualization, the entire data exchange and donation process must correspond to the data owners consent appropriate to data protection regulations which is why consent management is necessary. O'Kane et al. showed that consent decisions to share data change over the course of a lifetime, depending on the ever-changing health status and technological experience [42]. According to them, the possibility of modifying a given consent with regard to the patient's own current data protection sensitivity should be taken into account. The study of Grando et al. reviews patients and behavioral health providers opinions about consent preferences and the desired granularity when exchanging behavioral health data for care or research [43]. That study concludes that patients want granular consent control. However, from the provider's point of view, patients should be trained to make such decisions informed.

4.1.3. Shared Decision-Making

Currently, Odisho and Gore see shared decision-making between patient and healthcare providers based on data as blocked by the patients significant knowledge-barriers which impairs the ability to be an effective part of their own care process [44]. They propose a structured approach for patient-centered health information to help patients to obtain and better understand it. Jourquin et al. see inclusion and transfer of control to the patient himself as challenges for better, optimized therapy decisions [45]. These challenges face opportunities which are seen in the development and implementation of decision support systems (DSS) to enable personalized medical treatment [46, 47, 48]. Based on health-related data, decisions can be made using technological advancements and methods to support the process of these enormous amounts of data profitably. Wang et al. present a shared decision making framework which uses EHR data to provide a list of possible diabetes medication using a classification model, which should serve as the basis for the consultation between physician and patient [47]. Goletsis et al. present a decision support platform based on clustering techniques that identify patients with similar clinical characteristics and genetic predisposition [46]. They use their results to choose a personalized treatment plan based on the assumption that patients with a similar profile have a similar treatment plan. Semantic information, which is described using ontologies, serves as input for the clustering methods. Sqalli et al. present an conceptual AI-supported health coaching model and process that aims to support the patient in the long-term management of his chronic disease by producing personalized health plans using sensor data, machine learning and visualization tools [48]. Krithara et al. are building a big data system for decision makers which combines data from clinical notes, genomics, EHR and bibliography to identify patterns that help to adapt public health policies and individual diagnosis and treatment strategies for dementia and lung cancer [49]. Suinesiaputra et al. show how the sharing and open access to data are important for advancing the understanding of the heart functioning to find better ways to stop the progression and therapeutics concerning heart-related disease [50]. Within this context, Seth et al. use data from Body Sensor Networks (BSN) to model cardiac functioning for personalized medicine [51]. To build these kinds of DSS, Li et al. note that a large amount of data is needed from different healthcare facilities to train decision models. But according to Jaremko et al., due to the sensitivity of data, privacy rights must be given priority so that AI methods must be adapted to be functionally safe, understandable and avoid black boxing [52, 30].

4.1.4. Security and Privacy

Sajedi et al. review algorithms like cryptography, watermarking and steganography and their application in health information systems [53]. They recognize the trade-off between hiding and analyzing information and describe the challenges of security in healthcare systems. Addressing the privacy challenge, Li et al. present two distributed privacy-preserving ensemble approaches that can be used without disclosing sensitive patient-level data by first learning the data distribution from the facilities and then sharing and combining these local models into integrated data models [52]. Paddock et al. could prove that by means of homomorphic encrypted data can support real-time learning in personalized cancer medicine, as it allows analysis of the data while keeping it encrypted at all times [54]. According to them, homomorphic encryption does not neglect all privacy concerns, especially with respect to de-identification, since it is possible e.g. to identify individuals even from pooled samples of genomic data. Shahbaz et al. propose an anonymity algorithm for DICOM images and textual patient and communication information to ensure privacy and security of the data using a cloud-based system [55]. The anonymization is combined with an access policy so that it is applied according to the role of the user. The solution of a flexible, standards-based, privacy-enhanced user profile management approach based on an adaptive extended merkle structure from Sánchez-Guerrero et al. enable the patient to selectively disclose identity information and protect their privacy [10]. Those approaches can reduce the perception of privacy risks and highlight the benefits for individuals, which seems, besides accountability and conditionality, an important aspect to build up health data frameworks [56].

4.2. Requirements in Data Donation Cycles

The iterative development of medical applications enables a continuous integration of new data into the development process. These steps result in products, services or procedures, which in turn generate new data that can be used for continuous improvement of existing solutions or to enable the development of innovative approaches by combining different data sources. This is what we call data-driven development, which results in a cyclical approach that requires the consideration of the derived requirements based on the identified approaches outlined in Table 1. This shows the challenges, that must be considered from different points of view. Figure 1 outlines such a data donation cycle schematically. The data owner needs the possibility to request the data from the collecting and storing parties and to determine himself which data is passed on to third parties, who have access, and what may be done with it (S1,S1.1,S2). This requires a tool for visualization and administration of existing health data by the data owner itself, which is designed in a user-friendly way (T3). A research-compatible cross-institutional electronic health record could provide data adapted to the needs of health professionals and clinical researchers (M1) to best support decision making and enhance the patient's understanding of his or her health status (M3). In addition to the quality of this data (T5), the mapping and forwarding of the data in a internationally recognized standard play a role (T6.5). Semantic interoperability can be ensured through standardized descriptions of these resources and prevents lock-in effects (T6.3,T6.1,T6.2). Besides, it must be ensured that not only data but also processes, the identities and access rights linked to them are inter-organizational interoperable (S1.2). Technological approaches for this are seen in the application of DLT [25, 26, 27, 28, 13, 29]. Besides a possible legal fixation of this necessity for manufacturers, this additional effort must be balanced by compensation mechanisms for researchers and research projects (E2). The collected data itself can either be made available free of charge (philanthropic approach) (S3) or by providing services or monetary values in return (capitalistic approach) (E1). Anonymization can make entire health data histories available for research even after death (S3.1), or individual data sets which, in combination with others, provide a holistic view of a specific medical condition. Pseudonymization allows the continuous collection and monitoring of longitudinal data (S3.3) but requires the secure administration of the true identity by the data owner himself or by a data custodian. In this case, personal data are involved, which may make it legally necessary to specify a purpose concerning the data processing, i.e. by whom and for what the data may be used (S2.1). This consent should be digitally available and therefore in an automatically processable electronic format (see section 2). Here, trust and thus transparency must be guaranteed that the data is used exclusively for the purposes specified (S2.2). A further advantage of pseudonymization is the possibility to request further data for studies if the persons concerned have explicitly agreed to be contacted (S4.1,S4.2). The combination of health data, their metadata, and consent can be encapsulated by containerization to ensure integrity (T2.3) and traceability of the data processing steps. These data capsules can be kept either by the collecting agency, the data owner himself or a data custodian. The capsules themselves may have logic that implements transparent access control and validation mechanisms in combination with a trusted infrastructure. To release the data to other interested parties, a decentralized register can be created, making the data discoverable (S4.3,T4). The storage of the data itself should be decentralized as well (T7) but high availability must be guaranteed (T2.4). Information about the correctness of the data (T2.2) by the data originator as well as releases can also be mapped in the form of manipulation-proof access rights (T1.4) to trusted actors (T1.5) in a decentralized and distributed register. Authorized organizations (T1.5), with interest and competence in generating added value from the data, can contribute to the development processes by combining existing data sets while complying with data protection and security requirements (T2,T6,T6.4,T6.3). Using flexible (T11), expandable (T12), privacy and confidentiality preserving (T1, T2.5) data analytics tools, epidemiological predictions can be made in real-time (T8, T10) or new insights can be gained to support healthcare (T9). It must be kept transparent and persistent what the data capsules were used for and to what extent new findings can be traced back to them (S2.3,S4,T2.5). This information must be made available to the data owner (S3.2). The new findings are used as artifacts in the development of new treatment and care methods (M2) as well as medical products and

Structural	*S1* - **Sharing of health data:** The sharing of health data across institutional boundaries (e.g. between medical institutions, family members) [21, 25, 24, 22, 20, 3, 23, 50, 32, 43, 37, 52, 13, 42, 56, 10, 14, 15, 29, 36, 45]. *S1.1* - **Release of data to third parties:** The possibility of sharing the donated data with commercial user groups (e.g. university spin-offs, start-ups) in order to promote innovation e.g. in the field of analysis and artificial intelligence [30, 28]. *S1.2* - **Organizational interoperability:** Support of cross-system processes, identities and rights [21, 20, 29].
	S2 - **Data governance:** Patients gain access and control over their health data, allowing them to monitor the current and clinically correct data stored about them. The patient has the power to manage any exchange of data by defining the exact content, purpose and access rights (patients' right of informational self-determination) [21, 25, 26, 41, 24, 20, 3, 23, 28, 44, 32, 39, 33, 48, 29, 45]. *S2.1* - **Consent management:** Medical data may be processed only if the data owner has given his consent. The data owner has the right to revoke his consent at any time [21, 25, 26, 24, 22, 20, 3, 28, 50, 51, 43, 37, 17, 42, 56, 10, 38, 17, 29]. *S2.2* - **Trustworthiness:** As a data owner who makes his data available via a platform, trust must prevail. On the one hand, technologically in the data sharing and storing environment, but also in the other participants involved so that they do not do anything with the data that is contrary to the will of the data owner. In contrast, a data consumer must be able to trust that the data is trustworthy and correct. This trust must be noticeable [21, 26, 24, 3, 23, 28, 49, 32, 43, 37, 30, 4, 42, 54, 10, 29, 36]. *S2.3* - **Traceability of the data life cycle:** The data owner can view his health data and the history of all actions (e.g. access, use in research and clinical studies) on this data as well as the corresponding identities and the purpose behind that actions to track the entire data life cycle [25, 26, 20, 3, 23, 53, 27, 34, 45].
	S3 - **Data donation for research:** The possibility of selectively releasing one's own health data should exist so that data owners can find possible research projects and studies to support them with their data. Here, the data owner should be able to actively decide via informed consent whether to donate, what to donate and to whom [3, 21, 22, 24, 23, 50, 18, 51, 19, 43, 37, 30, 17, 4, 42, 38, 45]. *S3.1* - **Post mortem data donation:** Patients should have the opportunity to donate their medical data for research purposes after their death, similar to organ donation [57]. *S3.2* - **Sharing of research results:** Feedback of research results back to data owners and patients with similar diseases and symptoms [18, 30, 38, 36]. *S3.3* - **Monitoring of patient-generated longitudinal data:** Support and integration of mobile patient applications and IoT devices to collect, share and monitor patient-generated data, such as vital parameters, physical activities, nutritional intake and individual assessments and descriptions of the patient's own well-being [21, 41, 3, 5, 50, 32, 53, 37, 27, 33, 13, 48, 54, 56, 10, 38, 34, 29, 36, 45].
	S4 - **Transparency and reproducibility of research results:** The sharing of the corresponding research data (metadata, source code, variable definitions, de-identified data) with cost bearers of research projects, journals that publish the scientific publications and the public through accessible data repositories. Transparent documentation about the methodology and procedures used to collect the data to keep records about the whole process and possibilities to reuse the research data [17, 45]. *S4.1* - **Request for use in research:** For a specific research project, the health data of a patient can be made available to the requesting partner based on the patient consent [21, 24, 20, 3, 18, 30, 42, 14]. For data requests, the type and purpose of data processing by the data consumer must be specified [26, 23]. *S4.2* - **Secondary use of medical research data:** Provide medical research data for secondary use and analysis of the data by further researchers and studies [57, 24, 30, 17, 4, 42, 54, 34]. *S4.3* - **Prior registration of studies:** Registration of research studies including the indication of intended analyses in a public register before data collection [17].
Economical	*E1* - **Perceivable benefits of data donation:** The data donor receives compensation in the form of monetary or service benefits (e.g. via crypto token) by releasing his personal health data or research results to data consumers [28, 43, 30, 17, 4, 16].
	E2 - **Funding research data preparation and sharing:** Funding of researchers and research projects to compensate the additional effort needed to prepare and share research data for further use [45].
Technical	*T1* - **Data privacy:** Ensuring data privacy of patients and probands according to national regulations during the storage and transmission of data through privacy-by-design architecture [25, 57, 24, 21, 31, 26, 22, 3, 5, 23, 28, 49, 50, 18, 51, 32, 43, 53, 58, 37, 30, 45, 27, 4, 52, 33, 13, 42, 54, 56, 10, 14, 15, 55, 34, 29, 36]. Prevent data loss and data leakage [26, 28, 30]. *T1.1* - **Data deletion:** Deletion of the stored data by the data owner [30, 25, 5]. *T1.2* - **De-identification:** Possibility to remove direct identifying information from the medical data via anonymization or pseudonymization [21, 31, 26, 24, 22, 20, 23, 28, 50, 18, 51, 58, 30, 42, 54, 10, 15, 55, 29, 46, 45]. *T1.3* - **Privacy preserving analytics:** Support for privacy preserving analysis algorithms (e.g. privacy preserving smart contracts, secure multi-party computation, zero-knowledge, homomorphic encryption [25, 54], distributed computing, DataSHIELD, containerzation [30]) so that data does not have to leave its local storage facility, because the algorithms are brought to the data [21, 26, 24, 22, 30, 4, 52, 42, 54, 55]. The algorithms should be functionally safe, understandable and avoid black boxing [30]. *T1.4* - **Usage and access management:** Fine-grained definition, monitoring and control of access to health data by the data owner. Compliance should be controllable. Access and the purpose of use should be documented and authorised by the data owner for a specified period of time [21, 24, 22, 20, 5, 23, 43, 53, 27, 33, 13, 42, 54, 10, 14, 15, 55, 38, 29]. *T1.5* - **Users authentication and authorization:** Enable authentication of user (e.g. using a single sign-on approach, two-factor authentication) or user devices in order to be able to view the health data and to distribute, maintain and verify usage and access rights [26, 24, 22, 20, 23, 27, 33, 10, 15, 55, 29].
	T2 - **Security:** Apply the security-by-design approach to system design, implementation and operation [25, 21, 26, 24, 22, 20, 5, 23, 28, 50, 51, 32, 43, 53, 37, 30, 45, 27, 4, 33, 13, 42, 54, 56, 10, 14, 15, 55, 29, 36]. *T2.1* - **Data obfuscation:** Obfuscation of data via encryption, digital watermarks or steganography to ensure data security and confidentiality [5, 25, 57, 26, 28, 53, 37, 27, 33, 54, 10, 14, 29]. *T2.2* - **Data accuracy:** Guarantee the correctness of the data [32]. *T2.3* - **Data immutability:** Verification of the immutability of the data to build trust [25, 28, 30, 27, 13, 29, 26]. *T2.4* - **Availability:** The (permanent) availability of data for all authorized actors [53, 37, 43, 50, 33, 13, 55, 45]. *T2.5* - **Provability:** The linking of the data to a consent for use and proof of the existence of that link [26].

Technical	***T3*** **- Usability:** User-friendly and patient-centered tools for visualizing and managing existing health data and access rights. The entry hurdle for non-technical users should be low in order to increase acceptance for use [21, 22, 5, 44, 43, 48, 40, 14, 29, 46, 45].
	T4 **- Data identifier:** Provision of machine-processable description, data catalogue or data tags, which represents the aggregated metadata of the available data to enable the searches and data requests for data [26, 22, 20, 3, 23, 28, 54].
	T5 **- Data quality:** Measurement and documentation of data quality in its original state and along the data life cycle also in aggregated form [24, 22, 3, 5, 28, 45].
	T6 **- Sustainability:** Express data in technology-neutral and open formats so that data and its definitions can be efficiently migrated from one technology stack to another [21, 22, 23, 17]. ***T6.1*** **- Domain-driven data modeling:** Involvement of domain specialists (e.g. researchers, healthcare professionals) in the data modelling and the seperation between the technical implementation details from domain modeling [21, 24]. ***T6.2*** **- Avoidance of vendor lock-in:** The choice of standards, frameworks and system components so that vendor lock-ins are avoided, e.g. by preferring open source software, standards, models, frameworks and specifications [21, 24, 3, 22, 15]. ***T6.3*** **- Semantic interoperability:** Use internationally recognized, proven and open-technology terminologies and standards (e.g. HL7 FHIR, DICOM, IHE, openEHR, SNOMED CT) to enable communication between different systems [21, 24, 20, 23, 49, 50, 19, 37, 10, 14, 15, 46]. ***T6.4*** **- Data aggregation and harmonization:** Building a common understanding of information and data from different data silos by merging the different data types and building consistent data through computable semantic models (e.g. ontologies) [21, 46, 52, 14, 15, 29, 45, 24, 22, 49, 50, 19]. ***T6.5*** **- Structural interoperability:** Enable data flows between different systems and data silos [21, 22, 24, 20, 5, 23, 37, 29, 45].
	T7 **- Decentralization:** Decentralized storage of the data where they were generated. In case of necessary data transfer, the data will not leave the local data storage facilities without the consent of the data owner [21, 24, 23].
	T8 **- Predictive analytics:** Use of analysis and clustering methods for predicting or identifying patterns and connections between patients, their disease courses and findings to be able to make early decisions regarding diagnosis and treatments. [21, 49, 24, 22, 23, 50, 51, 37, 30, 48, 54, 46].
	T9 **- Data enrichment:** Analysis of aggregated data (e.g. using Natural Language Processing (NLP), ontological annotations or machine learning techniques) to extract meaningful knowledge in a structured form [21, 22, 23, 49, 51].
	T10 **- Scalability and performance:** Possibility for handling and processing a large amount of complex and constantly changing data [21, 24, 22, 5, 50, 51, 14, 36].
	T11 **- Flexibility:** A data analytics infrastructure should guarantee a high degree of flexibility in order to support different configuration and further processing options of the data according to the application (e.g. support of different file formats for export) [21, 10].
	T12 **- Extensibility:** The IT architecture should be modular and expandable, so that a multitude of use cases can be served [21, 24, 22, 5, 23].
Medical	***M1*** **- Clinical relevance:** The data that can be entered and made available though a platform or donation are adapted to the needs of healthcare professionals and clinical researchers. The information and services should be prepared in such a way that those groups of people are supported according to the clinical use case [21, 41, 3, 49, 44, 47].
	M2 **- Data-driven personalized medicine:** Improvement of patient care, especially therapy and diagnostics due to an increased amount of data. The aim is to design therapies as individually and precisely as possible for patients and to return possible findings to the patients [24, 22, 3, 28, 49, 50, 51, 39, 37, 30, 27, 35, 4, 48, 56, 38, 34, 46, 45].
	M3 **- Support collaborative decision making:** Provision of historical patient and study data, as well as the necessary background knowledge in a visually adapted form for patients, medical staff and clinical research to best support decision making and to strengthen a patient's understanding about their own condition. [21, 41, 3, 5, 44, 32, 27, 33, 48, 40, 46, 47].

Table 1: Identified Requirements for Data-driven Development of Medical Applications.

services (S1.1), which in turn generate new data whose effects must be analyzed.

5. Discussion and Future Work

Through a structured literature review, we identified requirements for self-determined health data management to embed them in a value-added data donation cycle. Currently, most approaches focus on enabling data-driven medical research for personalized medicine. The technological approaches outlined above could show a way to open data silos for data-driven medical application development to external actors such as companies. However, the approaches found are subject to national legislation and cultural influences about the willingness to make personal health data available. To address this interdisciplinary problem, we showed that not only technological but also structural, political, ethical, economic, and legal aspects must be resolved as well as compromises must be found for the development of a suitable open ecosystem. The derived requirements are based exclusively on the literature found by the review and depend strongly on the selected search string. Accordingly, further requirements can be found and are not limited to those mentioned here.

In the future, our research will continue to focus on the technological feasibility of the described data mobility mechanisms and their implications from different stakeholder perspectives. In particular, the self-determined control of decentralized health data flows and the privacy-preserving traceability of data analytics steps represent profound technological challenges. Besides, a prototype is to be developed

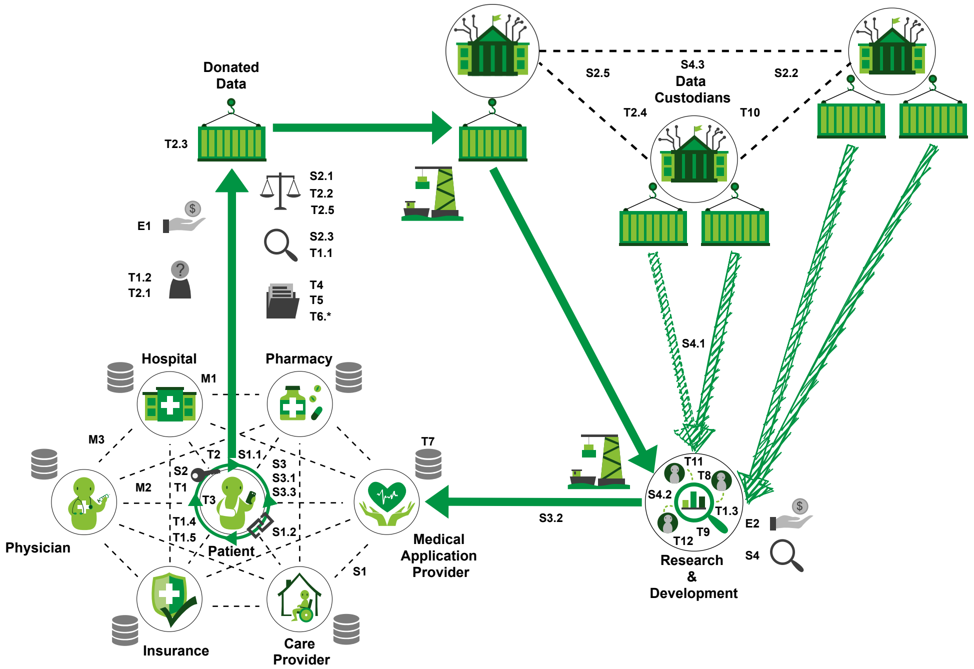

Figure 1: Schematic Data Donation Cycle.

based on a use case, which illustrates the cycle described above and makes it possible to verify it in real world scenarios.

References

[1] V. Estrada-Galiñanes and K. Wac, “Collecting, exploring and sharing personal data: Why, how and where,” *Data Science*, pp. 1–28, 2019.

[2] J. Krutzinna and L. Floridi, eds., *The Ethics of Medical Data Donation*. Philosophical Studies Series, 2019.

[3] V. Strotbaum et al., “Your data is gold – Data donation for better healthcare?,” *it - Information Technology*, vol. 61, 2019.

[4] M. Lawler and T. Maughan, “From Rosalind Franklin to Barack Obama: Data Sharing Challenges and Solutions in Genomics and Personalised Medicine,” *New Bioethics*, vol. 23, no. 1, pp. 64–73, 2017.

[5] A. Sunyaev and D. Chornyi, “Supporting chronic disease care quality: Design and implementation of a health service and its integration with electronic health records,” *JDIQ*, vol. 3, 2012.

[6] N. El Ioini and C. Pahl, “A Review of Distributed Ledger Technologies,” *OTM Confederated International Conferences*, pp. 277–288, 2018.

[7] Satoshi Nakamoto, “Bitcoin: A peer-to-peer electronic cash system,” 2008.

[8] D. Merkel, “Docker: lightweight linux containers for consistent development and deployment,” *Linux journal*, vol. 2014, no. 239, p. 2, 2014.

[9] P. Hitzler, *Semantic Web: Grundlagen*. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg, 2008.

[10] R. Sánchez-Guerrero et al., “Collaborative eHealth Meets Security: Privacy-Enhancing Patient Profile Management,” *IEEE J-BHI*, vol. 21, no. 6, 2017.

[11] H.J Pandit et al., “Creating a Vocabulary for Data Privacy,” in *OTM 2019 Conferences*, 2019.

[12] K. Peffers et al., “A Design Science Research Methodology for Information Systems Research,” *Journal of Management Information Systems*, 2007.

[13] M. A. Rahman et al., “An IoT and Blockchain-Based Multi-Sensory In-Home Quality of Life Framework for Cancer Patients,” in *IWCMC 2019*, pp. 2116–2121, 2019.

[14] P. Rao et al., “Towards Large-Scale Sharing of Electronic Health Records of Cancer Patients,” in *ACM SIGHIT*, 2010.

[15] P. Rao et al., “A Software Tool for Large-Scale Sharing and Querying of Clinical Documents Modeled Using HL7 Version 3 Standard,” in *ACM SIGHIT*, 2012.

[16] H.-C. Stoeklé, M.-F. Mamzer-Bruneel, G. Vogt, and C. Hervé, “23andme: a new two-sided data-banking market model,” *BMC medical ethics*, vol. 17, p. 19, 2016.

[17] Stephanie Wykstra, “Funder Data-Sharing Policies: Overview and Recommendations,” 2017.

[18] An-Wen Chan et al., “Increasing value and reducing waste: addressing inaccessible research,” *The Lancet*, vol. 383, no. 9913, 2014.

[19] M. Dugas et al., “Portal of medical data models: information infrastructure for medical research and healthcare,” *Database: The Journal of Biological Databases & Curation*, vol. 2016, 2016.

[20] S. C. Semler, F. Wissing, and R. Heyder, “German Medical Informatics Initiative,” *Methods Inf Med*, vol. 57, no. S 01, pp. e50–e56, 2018.

[21] B. Haarbrandt et al., “HiGHmed - An Open Platform Approach to Enhance Care and Research across Institutional Boundaries,” *Methods of information in medicine*, vol. 57, pp. e66–e81, 2018.

[22] H.-U. Prokosch et al., “MIRACUM: Medical Informatics in Research and Care in University Medicine,” *Methods of information in medicine*, vol. 57, pp. e82–e91, 2018.

[23] A. Winter et al., “Smart Medical Information Technology for Healthcare (SMITH),” *Methods of information in medicine*, vol. 57, pp. e92–e105, 2018.

[24] F. Prasser et al., “Data Integration for Future Medicine (DIFUTURE),” *Methods of information in medicine*, vol. 57, pp. e57–e65, 2018.

[25] A. Bayle et al., “When Blockchain Meets the Right to Be Forgotten: Technology versus Law in the Healthcare Industry,” pp. 788–792, 2018.

[26] M. Koscina et al., “Enabling Trust in Healthcare Data Exchange with a Federated Blockchain-Based Architecture,” in *IEEE/WIC/ACM*, pp. 231–237, 2019.

[27] K. Azbeg et al., “Blockchain and IoT for Security and Privacy: A Platform for Diabetes Self-management,” in *2018 4th Cloudtech*, 2018.

[28] X. Zheng et al., “Blockchain-based Personal Health Data Sharing System Using Cloud Storage,” in *2018 IEEE 20th Healthcom*, pp. 1–6, 2018.

[29] William J. Gordon and Christian Catalini, “Blockchain Technology for Healthcare: Facilitating the Transition to Patient-Driven Interoperability,” *Computational and Structural Biotechnology Journal*, vol. 16, 2018.

[30] J. Jaremko et al., “Canadian Association of Radiologists White Paper on Ethical and Legal Issues Related to Artificial Intelligence in Radiology,” *Canadian Association of Radiologists Journal*, no. 2, 2019.

[31] K. T. Pickard, “Exploring Markets of Data for Personal Health Information,” in *2014 IEEE ICDM*, 2014.

[32] D. Ose et al., “Persönliche Patientenakten im Internet. Ein narrativer Review zu Einstellungen, Erwartungen, Nutzung und Effekten,” *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen*, vol. 122, pp. 9–21, 2017.

[33] M. A. Al-Taee et al., “Mapping Security Requirements of Mobile Health Systems into Software Development Lifecycle,” in *2016 9th DeSE*, pp. 87–93, 2016.

[34] L. van Kleunen and S. Voida, “Challenges in Supporting Social Practices around Personal Data for Long-Term Mental Health Management,” in *UbiComp/ISWC ’19 Adjunct*, pp. 944–948, 2019.

[35] L. Koumakis et al., “Dementia Care Frameworks and Assistive Technologies for Their Implementation: A Review,” *IEEE RBME*, vol. 12, pp. 4–18, 2019.

[36] Yan Hu, Cong Peng, and Guohua Bai, “Sharing health data through hybrid cloud for self-management,” in *2015 IEEE ICMEW*, pp. 1–6, 2015.

[37] J. Ma, C. Peng, and Q. Chen, “Health Information Exchange for Home-Based Chronic Disease Self-Management – A Hybrid Cloud Approach,” in *2014 5th ICDH*, pp. 246–251, 2014.

[38] T. G. Smith et al., “Cancer survivor perspectives on sharing patient-generated health data with central cancer registries,” *Quality of Life Research*, 2019.

[39] J. Lewis and T. Siaw-Liaw, “Using information management systems and processes to support shared care for colorectal cancer survivors,” in *2017 IEEE ISTAS*, pp. 1–5, 2017.

[40] Y. G. Rajwan and G. R. Kim, “Medical Information Visualization Conceptual Model for Patient-Physician Health Communication,” in *IHI’10*, pp. 512–516, Association for Computing Machinery, 2010.

[41] A. A. O’Kane and H. Mentis, “Sharing Medical Data vs. Health Knowledge in Chronic Illness Care,” CHI EA ’12, Association for Computing Machinery, 2012.

[42] A. A. O’Kane, H. M. Mentis, and E. Thereska, “Non-Static Nature of Patient Consent: Shifting Privacy Perspectives in Health Information Sharing,” CSCW ’13, Association for Computing Machinery, 2013.

[43] M. A. Grando et al., “A Study to Elicit Behavioral Health Patients’ and Providers’ Opinions on Health Records Consent,” *Journal of Law, Medicine & Ethics*, 2017.

[44] Anobel Y. Odisho and John L. Gore, “Patient-centered approaches to creating understandable health information,” *Urologic Oncology: Seminars and Original Investigations*, vol. 35, no. 9, 2017.

[45] J. Jourquin et al., “Susan G. Komen Big Data for Breast Cancer Initiative: How Patient Advocacy Organizations Can Facilitate Using Big Data to Improve Patient Outcomes,” *JCO Precision Oncology*, vol. 3, 2019.

[46] Y. Goletsis et al., “Intelligent patient profiling for diagnosis, staging and treatment selection in colon cancer,” in *2008 8th IEEE BIBE*, pp. 1–6, 2008.

[47] Y. Wang et al., “A Shared Decision-Making System for Diabetes Medication Choice Utilizing Electronic Health Record Data,” *IEEE J-BHI*, 2017.

[48] M. T. Sqalli and D. Al-Thani, “AI-supported Health Coaching Model for Patients with Chronic Diseases,” in *2019 16th ISWCS*, pp. 452–456, 2019.

[49] A. Krithara et al., “iASiS: Towards Heterogeneous Big Data Analysis for Personalized Medicine,” in *IEEE CBMS2019*, pp. 106–111, 2019.

[50] A. Suinesiaputra et al., “Big Heart Data: Advancing Health Informatics Through Data Sharing in Cardiovascular Imaging,” *IEEE J-BHI*, 2015.

[51] D. Seth, N. Biswas, and D. Ghosh, “Big health data: Cardiac remodelling and functional interactions of big brain based implications in body sensor networks,” in *2017 7th CSNT*, pp. 339–344, 2017.

[52] Y. Li, C. Bai, and C. K. Reddy, “A distributed ensemble approach for mining healthcare data under privacy constraints,” *Information Sciences*, vol. 330, pp. 245–259, 2016.

[53] Hedieh Sajedi and Shabnam Rahbar Yaghobi, “Information hiding methods for E-Healthcare,” *Smart Health*, vol. 15, p. 100104, 2020.

[54] S. Paddock et al., “Proof-of-concept study: Homomorphically encrypted data can support real-time learning in personalized cancer medicine,” *BMC Medical Informatics & Decision Making*, 2019.

[55] S. Shahbaz, A. Mahmood, and Z. Anwar, “SOAD: Securing Oncology EMR by Anonymizing DICOM Images,” in *2013 11th FIT*, pp. 125–130, 2013.

[56] S. Patil et al., “Public preferences for electronic health data storage, access, and sharing - evidence from a pan-European survey,” *Journal of the American Medical Informatics Association*, vol. 23, no. 6, 2016.

[57] F. Holl et al., “Secondary Use of Research Data: A Review of Availability and Utilization of Public Use Files and Initial Steps for the Development of a Process to Facilitate Medical Research Data Donation,” in *2019 IEEE/ACS 16th AICCSA*, pp. 1–5, 2019.

[58] C. Lewis et al., “Prescription Medication Misuse Among American Indians in the Midwestern US,” *Journal of Health Disparities Research & Practice*, 2019.