6-18-2022

Decision model to design a blockchain-based system for storing sensitive health data

Christina Erler
FZI Research Center for Information Technology, erler@fzi.de

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

Michael Dietrich
FZI Research Center for Information Technology, dietrich@fzi.de

Wilhelm Stork
ITIV, wilhelm.stork@kit.edu

DECISION MODEL TO DESIGN A BLOCKCHAIN-BASED SYSTEM FOR STORING SENSITIVE HEALTH DATA

Research Paper

Christina Erler, FZI Research Center for Information Technology, Karlsruhe, Germany, erler@fzi.de

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

Michael Dietrich, FZI Research Center for Information Technology, Karlsruhe, Germany, dietrich@fzi.de

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

Abstract

The storage and sharing of sensitive health data in Blockchain-based systems implicates data protection issues that must be addressed when designing such systems. Those issues can be traced back to the properties of decentralized systems. A blessing but also a curse in the context of health data is the transparency of the Blockchain, because it allows the stored data to be viewed by all participants of the network. In addition, the property of immutability is in contrast to the possibility to delete the personal data upon request according to the European General Data Protection Regulation (GDPR). Accordingly, approaches to tackle these issues have recently been discussed in research and industry, e.g. by storing sensitive data encrypted On-Chain or Off-Chain on own servers connected to a Blockchain. These approaches deal with how the confidentiality and integrity of stored data can be guaranteed and how data can be deleted. By reviewing the proposed approaches, we develop a taxonomy to summarize their specific technical characteristics and create a decision model that will allow the selection of a suitable approach for the design of future Blockchain-based systems for the storage of sensitive health data. Afterwards, we demonstrate the utility of the decision model based on a use case for storing test results from a digital dementia screening application. The paper concludes with a discussion of the results and suggestions for future research.

Keywords: Blockchain, Distributed Ledger Technology, Sensitive Health Data, Decision Model, Data Management, Medical Data Storage, Taxonomy.

1 Introduction & Basics

Due to global developments in the digitization of the healthcare sector, Electronic Health Records (EHRs) are seen as possibility to create holistic health profiles by aggregating the patient and health data collected by various healthcare stakeholders in order to improve medical delivery through enhanced cross-sector collaboration, reduced healthcare costs and avoidance of duplicate examinations (Häyrinen et al., 2008). Currently, some European countries, including Germany, lack of implementation of government-regulated EHR systems, although they have had the introduction of such systems on their political agenda for several years (Pohlmann et al., 2020). Accordingly, the existing EHRs are often siloed systems maintained by a specific healthcare provider with no connection to each other (Xu et al., 2019). The reasons for this are seen in challenges regarding the creation of structural and semantic interoperability based on the procedures of the analog paper world and the plurality of actors and electronic systems, as well as the lack of clear political regulations and incentive structures (Pohlmann

et al., 2020). In addition to government-funded efforts, companies with private-sector interests pursued the goal of establishing centralized solutions and thus building data monopolies (Beinke et al., 2019). In order to avoid the associated disadvantages such as a lack of self-determination, loss of control for the patient as well as trust, privacy and security concerns, Distributed Ledger Technology (DLT) is recently discussed as a trusted and interoperable infrastructure to enable and improve interorganizational exchange of highly sensitive patient data in healthcare (see section 3.4). The best known representative of DLT is the Blockchain, which chains up the transaction histories (blocks) via linked lists to generate an immutable ledger (chain) (Nakamoto, 2008). In Estonia, the Blockchain technology is already being used to secure healthcare data integrity (e-Estonia, 2020). By using that decentralized technology, no central intermediary needs to be trusted, but only the underlying technology with the applied cryptographic procedures, which is characterized by tamper-resistance, transparency and fail-safety (Xu, Weber and Staples, 2019). Using Blockchains directly to store or exchange sensitive health data is not common because of privacy constraints. Sensitive health data requires confidentiality, i.e., it must be ensured that it can only be viewed by authorized persons (Schinle, Erler and Stork, 2021). However, due to the transparency of the Blockchain, all data can be viewed by all participants of a private network or by anyone in case of a public network (Xu, Weber and Staples, 2019). Personal health data is a special kind of sensitive data, as strict guidelines, such as the GDPR, apply to it. This grants individuals the right to be forgotten. Deleting data from the Blockchain is contradictory to its immutability feature. The Blockchain's immutability may be advantageous for the detection of health data manipulation. Medical institutions could, e.g. try to cover up medical malpractice by forging EHRs in order to prevent liability claims (Cao et al., 2019). Recently, a variety of approaches which address these problems have been discussed and developed in research and industry. In those approaches, different design patterns related to Blockchain implementations are applied (Xu, Weber and Staples, 2019). For example, by encrypting the On-Chain data or storing the data Off-Chain on own servers and linking the data by storing a pointer to the data together with their hash value on a Blockchain. To capture the design decisions in the development of such Blockchain-based systems, an analysis of current approaches in research on the application of Blockchain technology in healthcare is provided (see section 3.4). To summarize the results and findings of that analysis, we have developed a taxonomy that captures the technical characteristics of those approaches in terms of health data storage and privacy mechanisms (see section 4). The decision model that maps the entire decision-making process with regard to the technical design decisions is then derived with the help of that taxonomy (see section 5). Afterwards, we evaluate our proposed decision model by applying it to a concrete use case (see section 6) and conclude with a discussion of the results and a presentation of the research and practical implications that arise from it (see section 7).

2 Research Challenge & Question

The Gartner Hype Cycle for Blockchain Business 2019 predicts that it will take more than 10 years before Blockchain technology is adapted in healthcare (Gartner Newsroom Press Release, 2019). Therefore, Blockchain in this field is still in an experimental phase, with researchers and developers experimenting with the technology to take advantage of the aforementioned benefits. However, according to the Gartner Hype Cycle analysis for Blockchain 2021, most projects are stuck trying to align the use cases with the technology (Litan, 2021). This points towards a research gap in Blockchain-based applications between meaningful healthcare use cases and their technical application. To address this research gap, we want to build up a decision model that could help developers and researchers of health data system as guideline to design Blockchain-based systems. With respect to the selected field of application in the healthcare sector, there is no comprehensive decision model that can provide support regarding those decisions. As a basis for such a decision model, it is first necessary to identify which design decisions have to be made when conceptualizing and implementing a Blockchain-based system for storing sensitive health data. Accordingly, we want to investigate the following two research questions within the scope of this work:

RQ 1 - What design decisions should be made when conceptualizing and implementing a Blockchain-based system for the healthcare sector?

RQ 2 - How to build up a comprehensive decision model to support developers and researchers when confronted with the identified design decisions?

3 Methodology

The methodological procedure, which forms the basis for answering the previously mentioned research questions of this paper, is presented in the following section.

3.1 Design Science Research

As research method we follow the Design Science Research Methodology (DSRM) contributed by Peffers et al. (2007) to iteratively develop and evaluate the decision model in form of an IT artifact. The DSRM of Peffers et al. (2007) is mainly used for information systems research and seeks in a repetitive process as combination of theory and practice for IT artifacts that should be generalizable. It is divided into six process steps: (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. Due to the desire to develop a comprehensive decision model that is universally applicable to the health sector, this method is suitable for our research objectives and was applied accordingly. In accordance with these steps, our research begins with the identification and motivation of a practically relevant problem, which was represented by the first two sections. The second step of DSRM rests upon the identified approaches and ideas that were uncovered by the conducted literature review, whose main findings have been summarized by us in section 3.4 by an overview and the developed taxonomy. The knowledge gained thereby, serves as the foundation for the design and development of our solution artifact, namely the decision model that is demonstrated in section 5. We outline the evaluation of the decision model based on a use case for storing test results from a psychometric dementia screening application in section 6. In summary, this paper serves as the communication step of the DSRM.

3.2 Literature Review

In order to identify the scientifically used design decisions and design patterns in Blockchain-based health data systems, a structured literature review was conducted using the following search string: *"(blockchain OR distributed ledger) AND (sensitiv* OR personal OR priva* OR confidential*) AND (data sharing OR data storage OR data exchange OR off-chain OR on-chain) AND health*"*. The defined search string was then used to search the four scientific databases ACM Digital Library, IEEE Xplore, EBSCOhost and ScienceDirect to cover a wide range of journals and conferences in the field of computer science and information systems. This search resulted in a total of 257 publications. In an initial step, the duplicates, news articles and publications not available in English or German were removed from the list of relevant publications. Here, the focus was on peer-reviewed papers in order to ensure a high quality of the articles found. Subsequently, the remaining 211 publications were analysed for relevant keywords, abstract and title, limiting the number to 53 relevant papers. The identified literature was then used for a more detailed examination of the full texts. In this detailed review, we excluded articles that did not relate to the design of a Blockchain-based health data system or associated design decisions. The result of the entire review is a total of 18 relevant articles.

3.3 Taxonomy Development

The relevant articles from literature review were then classified by a taxonomy that was developed with an iterative method provided by Nickerson et al. (2013). This method was chosen because it relates specifically to taxonomy development in information systems. The steps required for this are as follows: *(1) Determine meta-characteristics, which are used to derive characteristics of the taxonomy; (2) Determine ending conditions*; *(3) Choose an approach*: *a.) Empirical-to-Conceptual:* Examine objects for common characteristics and group them accordingly. Characteristics are grouped into dimensions. *b.) Conceptual-to-Empirical:* Conceptualize characteristics and corresponding dimensions of objects.

Use them for object examination and creation of the taxonomy; *(4) If the ending conditions are not met go back to step 3*.

3.4 Approaches from Literature Review

In the identified literature, different systems and approaches are used to exchange, store and manage sensitive health data. The majority of systems store EHRs. In addition, the approach of Thwin and Vasupongayya (2018) manages Personal Health Records (PHRs). Unlike EHRs, PHRs are not managed by medical providers but by the owners themselves and may contain data from different sources, e.g. EHRs from different healthcare provider (Thwin and Vasupongayya, 2018). Some systems also store medical measurements of sensors, e.g. blood sugar measurements (Hawig et al., 2019). These sensors are connected with a smartphone app that allows users to share their measurements with physicians for treatment. Stored data is especially useful for research (Theodouli et al., 2018, Wang et al., 2019, Zheng et al., 2018), e.g. to improve treatment and diagnosis of diseases (Zhang et al., 2018, Wang et al., 2019, Zhang and Lin, 2018, Zhou, Li and Zhao, 2019) or to supply training data for machine learning algorithms (Hanley and Tewari, 2018, Chang et al., 2018). Furthermore, users may also be able to manage their own data (Zaghloul, Li and Ren, 2019, Azaria et al., 2016). A couple of the identified solutions enforce the use of existing clinical data standards (e.g. HL7 FHIR) to facilitate data sharing through Blockchain-based systems (Theodouli et al., 2018, Azaria et al., 2016, Zhang et al., 2018). Apart from the type of stored data, the approaches from literature differ by technical characteristics which are discussed in more detail in section 4. Hawig et al. (2019), e.g. implement two different systems to compare an On-Chain and Off-Chain approach.

4 Taxonomy

For the classification of these approaches, we developed a taxonomy by using the method described in section 3.3. We were particularly interested in the characteristics of the data storage and data protection mechanisms, which we therefore chose to be our meta-characteristics. The development process was considered to be finished if all approaches had been examined and at least one approach was classified under each characteristic. By applying an empirical-to-conceptual approach as well as an conceptual-to-empirical approach, the following dimensions and corresponding characteristics have been identified:

- *Storage location*: On-Chain, Off-Chain, Hybrid
- *Blockchain type*: Public, Private, Consortium
- *Off-Chain storage*: Existing databases, Central database, Peer-to-Peer (P2P) network
- *Security measures*: *Encryption* (Symmetric (sym.), Asymmetric (asym.), Hybrid (hybr.)), *Access control* (Access Control Lists (ACL), Token-based (Token), Attribute-based (Attribute)), *De-identification*

The assignment of approaches to identified characteristics is shown in table 1 and the characteristics itself are discussed in more detail below.

Reference	**Storage location**			**Blockchain type**			**Off-Chain storage**			**Security measures**		
	On-Chain	Off-Chain	Hybrid	Public	Private	Consortium	Decentralized	Centralized	Distributed	Encryption	Access Control	De-identification
(Li et al., 2018)			x	x				x		hybr.		x
(Zhang and Lin, 2018)	x				x	x				asym.		
(Hawig et al., 2019), App.1	x			x						sym.		x
(Hawig et al., 2019), App.2		x		x					x	sym.		x

(Zheng et al., 2018)		x		x						sym.		x
(Zhou, Li, and Zhao, 2019)		x		x		x			x	hybr.	ACL	
(Thwin and Vasupongayya, 2018)		x			x			x		asym.	ACL	x
(Liu et al., 2018)		x				x		x		asym.	ACL	x
(Theodouli et al., 2018)		x				x		x			ACL	x
(Zaghloul et al., 2019)		x		x				x		hybr.	Attribute	
(Azaria et al., 2016)		x			x		x				ACL	
(Zhang et al., 2018)		x			x		x			asym.	Token	
(Xiao et al., 2018)		x			x		x			sym.	ACL	x
(Chang et al., 2018)		x		x		x	x				ACL	x
(Wang et al., 2019)		x				x		x		asym.	ACL	
(Nguyen et al., 2019)		x			x				x	asym.	ACL	
(Hanley and Tewari, 2018)		x			x		x					x
(Daraghmi et al.,2019)		x				x	x			hybr.	ACL	
(Dagher et al., 2018)		x				x	x			hybr.	ACL	

Table 1. A taxonomy of the identified approaches.

4.1 Storage Location

First, the identified approaches can be classified by their storage location of sensitive data into three categories: *On-Chain, Off-Chain* and *Hybrid*.

On-Chain approaches store sensitive data directly on a Blockchain. The amount of data stored on a Blockchain is often limited by maximum transaction and block sizes, e.g. Bitcoin has a limit of 40 bytes per transaction (Xu, Weber and Staples, 2019) and one megabyte per block (Daraghmi et al., 2019). The block size could be increased (Daraghmi et al., 2019), but that would also lead to longer replication times (Xu, Weber and Staples, 2019). On-Chain data profits from the Blockchain's immutability and decentralization, which means data is protected from manipulation and loss, but cannot be deleted (Li et al., 2018). An issue, when storing sensitive data that requires confidentiality and is not intended to be visible for every participant, is the Blockchain's transparency (Xu, Weber and Staples, 2019). The identified approaches therefore use encryption to ensure that only participants with a secret key can access stored data (Zhang and Lin, 2018, Hawig et al., 2019). This requires additional key exchange and management outside the Blockchain (Xu, Weber and Staples, 2019). As a rule of thumb, data smaller than its hash value should be stored On-Chain and larger data Off-Chain (Xu, Weber and Staples, 2019). Furthermore, data that needs to be modified or deleted should not be stored On-Chain.

The majority of approaches store sensitive data **Off-Chain** and only metadata On-Chain (Hawig et al., 2019, Zheng et al., 2018, Zhou, Li, and Zhao, 2019, Thwin and Vasupongayya, 2018, Liu et al., 2018, Theodouli et al., 2018, Zaghloul et al., 2019, Azaria et al., 2016, Zhang et al., 2018, Xiao et al., 2018, Chang et al., 2018, Wang et al., 2019, Nguyen et al., 2019, Hanley and Tewari, 2018, Daraghmi et al., 2019, Dagher et al., 2018). This metadata includes a reference to the storage location as well as a hash of stored data to recognize manipulation of Off-Chain data (Xu, Weber and Staples, 2019). By storing only metadata On-Chain the Blockchain's scalability is not impacted when storing large amounts of data. However, only the metadata profits from Blockchain's immutability and decentralization. While manipulation of Off-Chain data can be recognized with its hash value, it is not possible to prevent it. Therefore, additional security measures are necessary to protect Off-Chain data from unauthorized access (Xu, Weber and Staples, 2019). In summary, Off-Chain storage is especially suitable for storing large amounts of data or data that has to be changed or deleted in the future. In the identified literature this is the most common approach because health data is collected over the whole lifespan and has to be deleted on request due to the GDPR.

Instead of storing sensitive data On-Chain or Off-Chain, a **hybrid** approach can be used (Li et al., 2018). In a hybrid approach, sensitive data is either stored On-Chain or Off-Chain depending on its requirements which leads to increased flexibility and better scalability in comparison to an On-Chain approach. To decide whether to store data On-Chain or Off-Chain, a decision criterion is necessary. In the case of Li et al. (2018) multimedia files, e.g. images, are stored Off-Chain and text files On-Chain. Data could also be distinguished by its size. If neither an On-Chain nor Off-Chain approach are sufficient, e.g. if text files should be immutable and preserved forever, but also large image files have to be stored (Li et al., 2018), a hybrid approach should be considered.

4.2 Blockchain Type

In the examined literature, different Blockchain types are used which can be classified by the type of network management and participant's permissions into: *public, private* and *consortium Blockchains*.

Public Blockchains, e.g. Bitcoin, are decentralized, i.e., not controlled by a central instance. They have open networks where everybody can join and verify transactions without permission (Dagher et al., 2018). As network participants don't trust each other, incentive mechanisms are used to ensure correct operation of the Blockchain (Jin et al., 2019). Public Blockchains often suffer from a limited transaction processing rate and size of stored data (Xu, Weber and Staples, 2019). A public Blockchain is used, e.g. to enable users to control their own data and to minimize dependencies to healthcare institutions (Zaghloul et al., 2019). If decentralization is important, public Blockchains are particularly suitable.

Private Blockchains are permissioned as well as managed and operated by a single organization (Dagher et al., 2018), allowing a high configuration flexibility. The responsible organization decides who is able to create new blocks, authenticates participants and controls Blockchain access by assigning permissions to network participants (Xu, Weber and Staples, 2019). Because participants are verified and authorized to generate new blocks, more efficient consensus algorithms like Practical Byzantine Fault Tolerance (PBFT) can be used (Xu, Weber and Staples, 2019). In the examined literature, a private Blockchain is, e.g. used by Hanley and Tewari (2018) to build a platform for anonymized machine learning data. In this case, the government acts as single provider and central control instance.

The third type of Blockchains are **consortium** Blockchains (Dagher et al., 2018), which are similar to private Blockchains but are operated jointly by multiple organizations. The creation and validation of blocks is done by pre-authorized network nodes (Xu, Weber and Staples, 2019). Consortium Blockchains are, e.g. used in the system of Wang et al. (2019) for the exchange of health data between several healthcare providers. For this purpose, each healthcare provider provides a leader node to verify transactions and blocks. Daraghmi et al. (2019) propose an incentive mechanism integrated in Proof of Authority (PoA) consensus algorithm to decide which provider nodes are responsible for validation and adding new nodes or the creation of blocks based on the quality of the provider EHRs.

These three types of Blockchains can be combined by either **hooking into a popular Blockchain** at transaction level or using multiple private Blockchains (Xu, Weber and Staples, 2019). In the first approach, the hash value of the used Blockchain is periodically stored on a popular Blockchain to profit from its trust and security (Xu, Weber and Staples, 2019). Med-PPPHIS (Zhou, Li, and Zhao, 2019) and DeepLinQ (Chang et al., 2018) use this approach to recognize manipulation of their private or consortium Blockchain. To improve scalability, transactions of a single Blockchain can be split and distributed onto **multiple private Blockchains** that are connected by a common Blockchain (Xu, Weber and Staples, 2019). For example, Zhang and Lin (Zhang and Lin, 2018) connect multiple hospitals, while each hospital maintains an own private Blockchain with its own data and a joint consortium Blockchain as an index of the data of the private chains.

4.3 Off-Chain Storage

The usage of an Off-Chain approach requires the selection of an appropriate storage location. In the examined approaches, data is stored either *decentralized, centralized* or *distributed.*

A **decentralized** data store is mainly used for EHRs as they are created and managed by individual healthcare providers within their own infrastructure (Xiao et al., 2018). In this case, the Blockchain is used to connect the existing data stores of the individual healthcare providers. Stored data can then be located in the individual provider databases by firstly querying the location of the requested data from the Blockchain (Xiao et al., 2018, Azaria et al., 2016, Zhang et al., 2018, Xu et al., 2021). By using existing databases, no new single point of failure is created and no additional storage costs arise (Azaria et al., 2016). However, global queries on all provider databases are difficult due to differences in hardware and software (Jin et al., 2019, Wüst and Gervais, 2018).

As alternative to decentralized storage, a single **centralized** data store can be used to store health data of several providers. A central data store enables an easy setup, data management and access management, but is a single point of failure (Hanley and Tewari, 2018, Zhou, Li, and Zhao, 2019). Centralized storage is provided by a single party, e.g. the government or cloud providers (Hanley and Tewari, 2018). They can access stored data and therefore have to be trustworthy (Wang et al., 2019). PHRs, which may contain data from different healthcare providers and are managed by the owner, are stored in a central cloud data store (Thwin and Vasupongayya, 2018). However, EHRs can also be stored centrally to achieve independence from the record generators, e.g. to build a government-controlled research platform that provides anonymized copies of EHRs (Hanley and Tewari, 2018).

For **distributed** Off-Chain storage, P2P networks are used. In P2P networks, data is stored in the local storage of the individual network participants. Data can then be distributed between them without a central server (Xu, Weber and Staples, 2019). Advantages are no single point of failure, high storage throughput and short reading times, which is suitable for the exchange of large amounts of data (Nguyen et al., 2019, Hawig et al., 2019). Distributed storages tend to have a higher setup and management complexity (Hawig et al., 2019). Examples for P2P protocols are the InterPlanetary File System (IPFS) and Dat. IPFS allows the retrieval of data via its hash and is used by Hawig et al. (2019) to store medical sensor data with the goal of creating an independent system as well as by Nguyen et al. (2019) to store EHRs. In order to control network access and to enforce deletion of stored data, a private IPFS network has to be used (Hawig et al., 2019). Instead of operating a separate P2P network, it is also possible to store data directly in the local storage of the Blockchain nodes, e.g by splitting data into shards and distributing them between the nodes (Zhou, Li, and Zhao, 2019).

4.4 Security Measures

For the protection of sensitive health data, several security measures are proposed in the examined paperss including *encryption***,** *access control* and *de-identification*.

Encryption: Encryption may be used to ensure the confidentiality of data. The data can be encrypted either asymmetrically with a public key (PK) of the recipient or symmetrically with a key that is shared between the persons who have access to the data (Xu, Weber and Staples, 2019, Bouras et al., 2020). In order to share asymmetrically encrypted data with other persons, it must be decrypted and then re-encrypted with the recipient's public key. In the case of Nguyen et al. (2019), this re-encryption is performed on user request by a central server. Hybrid methods, e.g. Elliptic Curve Integrated Encryption Scheme (ECIES), use asymmetric methods to exchange symmetric keys (Zaghloul et al., 2019, Li et al., 2018, Hawig et al., 2019 Zhou, Li and Zhao, 2019). More advanced encryption methods in literature are: Ciphertext-Policy Attribute-Based Encryption (CP-ABE) that allows the integration of attribute-based access policies into ciphers (Zaghloul et al., 2019), Proxy Re-Encryption (PRE), which allows delegation of decryption rights to third parties without exposing the plain data (Thwin and Vasupongayya, 2018, Wang et al., 2019, Zhou, Li, and Zhao, 2019) and Searchable Encryption (SE), which allows ciphers to include keywords that can be searched without decryption (Wang et al., Zhang and Lin, 2018). Three of the investigated approaches (Theodouli et al., 2018, Hanley and Tewari, 2018, Chang et al., 2018) provide medical data as training data for machine learning and therefore deliberately dispense with encryption, since machine learning algorithms cannot work on encrypted data (Hanley and Tewari, 2018). In summary, encryption is useful to ensure confidentiality of sensitive data if no computations on the data are necessary.

Access Control: Access control is essential to protect sensitive Off-Chain data from unauthorized access (Zhang et al., 2018, Xu et al., 2021). For this purpose, Blockchain can be used to store access permissions in either a smart contract (Nguyen et al., 2019, Liu et al., 2018, Theodouli et al., 2018, Azaria et al., 2016, Zhang et al., 2018, Chang et al., 2018, Xu et al., 2021) or directly in transactions (Xiao et al., 2018, Wang et al., 2019). Thwin and Vasupongayya (2018) do not use a Blockchain, but a server to manage permissions. The first approach to access control is to store permissions in an Access Control List (ACL) that contains all persons who may access specific data, e.g. an EHR (Azaria et al., 2016). Each EHR has an own ACL. When accessing Off-Chain data, a signed request is then sent to a central server that acts as an oracle and invokes the necessary smart contract functions in order to verify whether the requester is authorized or not. This is the case, if the corresponding permissions for the requester exist on the Blockchain. Users are identified with their Blockchain address and the access request signature guarantees that the requester is the owner of the address. The Blockchain address can be mapped to a real world identity with the help of an additional smart contract that, e.g. stores an associated government-issued ID number, such as a social security number (Azaria et al., 2016). The name of the person (Chang et al., 2018) or the place of residence (Nguyen et al., 2019) can also be stored for this purpose. To revoke permissions, the corresponding smart contract variable can be updated (Azaria et al., 2016). Furthermore, ACLs can be combined with PRE by additionally storing re-encryption keys for authorized persons that allow re-encryption of the encrypted Off-Chain data (Wang et al., 2019, Thwin and Vasupongayya, 2018, Zhou, Li,, and Zhao, 2019, Daraghmi et al., 2019, Dagher et al., 2018). In addition to assigning access permissions based on identities, it is also possible to define access policies based on attributes by using Attribute-based access control. To do this, participants must first visit a certified registrar, which checks their attributes and stores them in a smart contract. Users can then set access policies for their data in another smart contract. These access policies specify which attributes are required for data access. In the case of an access, the smart contract verifies whether the accessor has the required attributes according to the policy. If this is the case, the access is approved and a key-issuer creates a key based on the attributes. Off-Chain data is encrypted with CP-ABE and can be decrypted with that key (Zaghloul et al., 2019). This approach still requires an oracle for the issuing of keys and the verification of attributes during registration. However, the definition of access policies allows for more flexibility, since the permissions do not have to be granted for each person individually, but e.g. for all doctors in a particular hospital in a single access policy. ACLs and Attribute-based access control can be used to restrict access to certain operations, e.g. to allow only read or write operations (Liu et al., 2018, Xiao et al., 2018). A problem with both is that a trusted oracle is required to invoke smart contract functions with the correct input data or to create the correct transactions on the Blockchain. FHIRChain (Zhang et al., 2018) uses Token-based access control that does not require an oracle. In order to grant a person access to the data, the user must first create an access token by signing the reference to the desired Off-Chain data and encrypting it with the person's PK. The signature ensures the authenticity of the token. The token is then stored in a smart contract. Authorized persons can invoke the smart contract to obtain their stored access token and decrypt it with their SK to obtain the storage location. The smart contract also provides an immutable log of token creations and usages (Zhang et al., 2018). However, revoking access is more difficult because tokens cannot be deleted from the Blockchain. In general, access may be limited to certain parts of the records for the purpose of data minimization (Liu et al., 2018, Theodouli et al.,2018, Zaghloul et al., 2019, Azaria et al., 2016, Zhang et al., 2018, Chang et al., 2018, Daraghmi et al., 2019), for example, by predefined SQL queries (Azaria et al., 2016) or temporary (Liu et al., 2018, Xiao et al., 2018, Theodouli et al.,2013, Zhang et al., 2018, Daraghmi et al., 2019), e.g. by using only a temporary URL (Xiao et al., 2018). A complete revocation of access after authorization is not possible, since a snapshot of the data can be taken (Xu et al., 2021). Blockchain-based access control provides an immutable transaction history of accesses and permission changes (Nguyen et al., 2019, Liu et al., 2018, Xiao et al., 2018, Theodouli et al., 2018, Zaghloul et al., 2019, Azaria et al., 2016, Zhang et al., 2018, Chang et al., 2018, Wang et al., 2019, Thwin and Vasupongayya, 2018, Zhou et al, 2019, Daraghmi et al., 2019). This is useful, for example, to find a person responsible in the event of data misuse (Thwin and Vasupongayya, 2018). As a result, access control is useful to prevent unauthorized access to Off-Chain data and to prevent data misuse.

De-identification: Anonymization allows personal health data to be shared or sold in compliance with the underlying regulations such as the GDPR by removing identifiers such as name, address and others (Zheng et al., 2018). This is used in several systems (Theodouli et al.,2018, Hanley and Tewari, 2018, Chang et al., 2018, Zheng et al., 2018), that allow users to donate their data to third parties e.g. for medical research. To reduce the risk of reversibility of the original data and the risk of linkability medical sensor data can be aggregated and obfuscated (Hawig et al., 2019). Often data is only pseudonymized because it is linked to a Blockchain address (Xiao et al., 2018, Theodouli et al., 2013, Hanley and Tewari, 2018, Chang et al., 2018, Li et al., 2018, Zheng et al., 2018). In conclusion, health data and personal data should be anonymized or at least pseudonymized before sharing it for research purposes.

5 Proposed Decision Model

Considering the aforementioned characteristics four questions arise for the design of a Blockchain-based system for storing sensitive health data. We try to answer these questions in the following and visualize central decisions in a decision model, which is shown in Fig.1. Note: Before applying our decision model, it should be checked whether a Blockchain-based solution is suitable for the specific use case. Existing work e.g. by Wüst and Gervias (2018) can be used for this purpose.

Which storage type should be used? On-Chain storage is particularly suitable for small amounts of data and for data that does not need to be changed or deleted in the future. For large amounts of data and data that has to be deleted or changed, it is recommended to use an Off-Chain storage. As a rule of thumb, data smaller than its hash value should be stored On-Chain and larger data Off-Chain (Xu, Weber and Staples, 2019). If neither an On-Chain nor Off-Chain storage is sufficient for the purpose and more flexibility is required, hybrid storage may also be considered.

What type of Blockchain should be used? The choice of Blockchain type depends primarily on whether the system is to be operated and controlled decentrally, centrally by a single healthcare provider or jointly by several healthcare providers. Public Blockchains are appropriate for decentrally operated systems that do not want to rely on any provider. For systems managed by a single healthcare provider or a governmental institution, a private Blockchain is suitable and for systems managed by several healthcare providers together, a consortium Blockchain is reasonable. However, further advantages and disadvantages of different Blockchain types must be considered. It may also be useful to use several Blockchains. To recognition of manipulations of private or consortium Blockchains by the responsible providers, their transactions can be hooked in a popular Blockchain. In addition, scalability problems of an On-Chain storage can be reduced by splitting data across several private Blockchains.

Which Off-Chain storage is suitable for storing data off the actual ledger? In the case of an Off-Chain storage, an appropriate storage type must be chosen. Data can be either stored decentralized, centralized or distributed. Decentralized storage should be used to store EHRs, as these are already stored in existing infrastructure of the individual healthcare providers (Xiao et al., 2018). If independence from record generating providers is desired, a centralized or distributed storage can be used. PHRs and medical measurements are stored by the system provider and not in already existing infrastructure. Hence, they should be stored centralized or a distributed in a P2P network depending on whether data should be stored in one location or multiple locations.

What data protection and data security measures are required? For the storage of sensitive data, additional measures should be taken to protect and secure it. Both On-Chain and Off-Chain data should be encrypted to provide confidentiality if no machine learning algorithms need to be run on the data. Furthermore, access control can be used to prevent unauthorized access to Off-Chain data and to log the access in order to find a person responsible in the event of data misuse. If a trusted oracle is available, ACLs or Attribute-based access control can be used. The latter allows more flexibility through access policies. If no trusted oracle is available, Token-based access control is appropriate. Personal data, especially personal health data, intended to be used for medical research, should be anonymized or at least pseudonymized for data protection purposes.

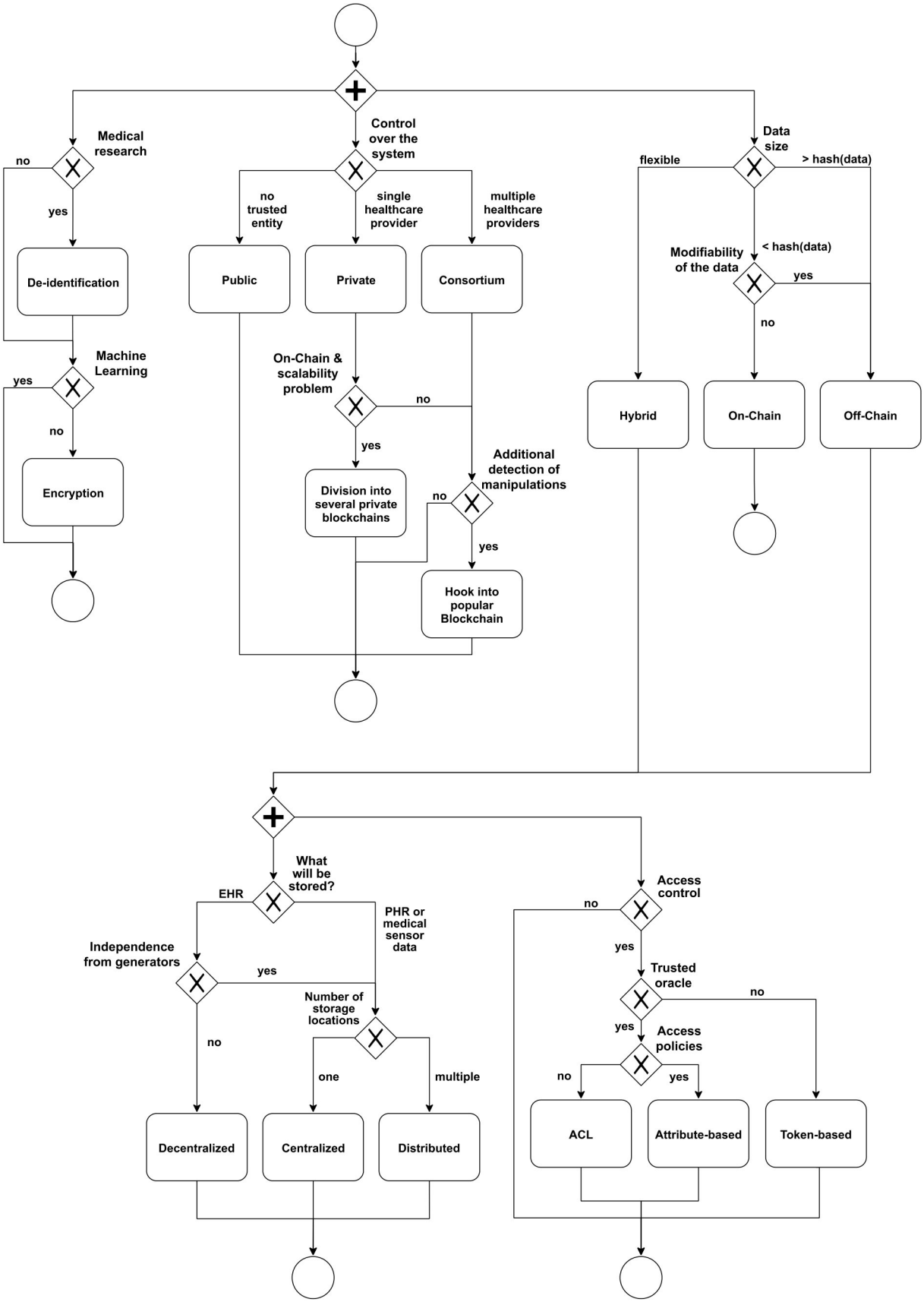

Figure 1. Decision model to determine a suitable approach for storing sensitive health data in a Blockchain-based system.

6 Evaluation

In this section, we evaluate the applicability of our proposed decision model by applying it to a concrete use case. Therefore, we first define our use case. Afterwards, we use the proposed decision model to determine the concrete technical characteristics of a Blockchain-based solution for this use case.

6.1 Use-Case: Storage of Dementia Screening Results

The prototype must store health data from DemPredict, a smartphone app that offers a dementia screening test (Schinle et al., 2018). In such a test, several tasks, e.g. connecting given numbers in ascending order, must be completed in one session. Each test session is associated with an user. The results of these sessions, including the scores achieved for each task, are provided in textual JavaScript Object Notation (JSON) format and must be stored by the prototype. For this, it must be able to retrieve and store data from the app provider on user request. The prototype must also display stored data to its users. Furthermore, it must be possible to delete data and to share data with third parties, e.g. for diagnosis, treatment or research purposes. Non-functional requirements, especially with regard to data protection, are also important. As the data may show signs of incipient dementia and sufferers fear of stigmatisation and social exclusion, they should not be disclosed without permission. In relation to the stored data, the prototype must guarantee confidentiality, integrity, scalability and high availability. Future extensibility for further health data from different providers is also desirable.

6.2 Application of the Proposed Decision Model

Before applying the proposed the decision model to the use case, it must be clarified whether a Blockchain is required. Based on the methodology proposed by Wüst and Gervais (2018), this is the case for the following reasons: A system state consisting of health data from different health data providers must be stored. Initially, DemPredict is the sole provider, but in the future additional health data from different providers will be added. These data providers as well as the users who own the data have write access in the system. Thus, there are multiple data writers. In order to gain the trust of users, in our case only the state can be considered as a trusted third party. However, since trust in the state is not always given and depends on the respective country, the state is not an option. Data writers such as providers and users are known, but not trusted, as a provider may be tempted to forge data in order to cover up medical malpractice. Since health data must be stored, we can apply our decision model from section 5 to determine the concrete characteristics of the prototype based on the previously mentioned requirements. In the following, the application of the decision model is discussed.

***Which storage type should be used?*:** Health data to be stored, such as DemPredict data, is larger than its hash value and should also be deletable by the user. Accordingly, Off-Chain storage is suitable.

***What type of Blockchain should be used?*:** The prototype should be controlled and operated by multiple organizations, the providers of the different health applications used by a patient such as DemPredict. Therefore, a consortium Blockchain is suitable. From the authors' point of view, the design decision was made that no additional detection of manipulations are required. On the one hand, to keep the system complexity low and on the other hand, we do not initially assume that several providers are conspiring.

***Which Off-Chain storage is suitable for storing data off the actual ledger?*:** Users should be able to modify their data, e.g. delete it, and share it with third parties. The prototype should be independent from individual providers to enable data modification independently of the provider. The stored data is comparable to a PHR, which can also contain data from health apps. Therefore, either centralized storage or distributed storage is appropriate. To achieve high availability, we choose a distributed storage in a P2P network that stores data in multiple locations.

***What data protection and data security measures are required?*:** Stored data can be shared with third parties, e.g. for research purposes. Accordingly, the data should be anonymized or at least pseudonymized. In order to be able to display data to users, we decide to store pseudonymized data because anonymized data cannot be linked to any user. We do not run any machine learning algorithms on the data and therefore the data is stored encrypted. Access control for Off- Chain data is also wanted.

The provider of our system can act as a trustworthy oracle. Since no complex access policies are required for our purpose, we will use an ACL.

7 Results & Limitations

In summary, the design of Blockchain-based systems is not trivial in the context of health data storage, as a large number of design decisions have to be made due to central properties of the Blockchain, such as transparency, immutability and limited scalability. Therefore, a decision model was required that facilitates the choice of an appropriate approach for storing sensitive health data. Through a structured literature review, we identified approaches and technical implementations for Blockchain-based health data storage in order to embed them in an iterative taxonomy development process. The resulting taxonomy helped to summarize the technical characteristics of the current healthcare Blockchain approaches and, in this context, to identify the points where design decisions must be made. Subsequently, the arisen taxonomy served as an auxiliary to build up the decision model. We applied our decision model to a real world scenario in context of dementia. This allowed us to evaluate and demonstrate the applicability of the decision model. Thereby, the choice of a suitable design fell on an Off-Chain storage, where the health data is stored encrypted and pseudonymized in a distributed P2P network and referenced by storing metadata together with a hash value on a consortium Blockchain. The advantages and disadvantages of this Blockchain-based design were further investigated by implementing a prototype, which is beyond the scope of this paper.

Of course, our research is not without methodological limitations. The developed decision model strongly depends on the literature found as well as on the selected search string. Access control goes hand in hand with identity management, especially in healthcare where correct authentication depends on secure and trusted entities (Bouras et al., 2020). Apart from the mapping of individual Blockchain addresses to real identities, the combination with decentralized approaches is not addressed in the publications found, although such approaches as Self-Sovereign Identity (SSI) and Decentralized Trusted Identity are currently being discussed in research (Bouras et al., 2020, Liu et al., 2020, Houtan et al., 2020). From the authors' point of view, giving patients the opportunity to fully control their identity while maintaining the confidence of identity data can add value in the development of patient-centered solutions in healthcare. Possible design patterns of SSI approaches are presented by Liu et al. (2020).

Further research should be undertaken to extend and transfer the decision model to other use cases and domains to enable generalizability. In addition, the applicability of the decision model is to be tested in the future as part of a research project in which the development of a Blockchain-based system in the healthcare sector is planned. Therefore, our research will continue to focus on figuring out the advantages and disadvantages that a decentralized infrastructure can bring in healthcare, especially to design patient-centered solutions.

References

Azaria, A., Ekblaw, A., Vieira, T. and Lippman, A. (2016). "MedRec: Using Blockchain for Medical Data Access and Permission Management". *2016 2nd International Conference on Open and Big Data (OBD),* 25-30.

Bayle, A., Koscina, M., Manset, D. and Perez-Kempner, O. (2018). "When Blockchain Meets the Right to Be Forgotten: Technology versus Law in the Healthcare Industry*"*. *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, 788-792.

Beinke, J. H., Fitte, C. and Teuteberg, F. (2019). *"*Towards a Stakeholder-Oriented Blockchain-Based Architecture for Electronic Health Records: Design Science Research Study*"*. *Journal of medical Internet research*, *21*(10), e13585.

Bouras, M.A., Lu, Q., Zhang, F., Wan, Y., Zhang, T. and Ning, H. (2020). "Distributed Ledger Technology for eHealth Identity Privacy: State of The Art and Future Perspective". *Sensors (Basel, Switzerland), 20*.

Cao, S., Zhang, G., Liu, P., Zhang, X. and Neri, F. (2019). "Cloud-assisted secure eHealth systems for tamper-proofing EHR via blockchain*"*. *Inf. Sci.,* 485, 427-440.

Casino, F., Dasaklis, T.K. and Patsakis, C. (2019). "A systematic literature review of blockchain-based applications: Current status, classification and open issues*"*. *Telematics Informatics, 36*, 55-81.

Chang, E.Y., Liao, S., Liu, C., Lin, W., Liao, P., Fu, W., Mei, C. and Chang, E.J. (2018). „DeepLinQ: Distributed Multi-Layer Ledgers for Privacy-Preserving Data Sharing". *2018 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, 173-178.

Dagher, G.G., Mohler, J., Milojkovic, M. and Marella, P.B. (2018). "Ancile: Privacy-Preserving Framework for Access Control and Interoperability of Electronic Health Records Using Blockchain Technology". *Sustainable Cities and Society, 39*, 283-297.

Daraghmi, E.Y., Daraghmi, Y. and Yuan, S. (2019). "MedChain: A Design of Blockchain-Based System for Medical Records Access and Permissions Management". *IEEE Access, 7*, 164595-164613.

e-Estonia. (2020), *E-Health Records — e-Estonia*. URL: https://e-estonia.com/solutions/healthcare/e-health-record/ (visited on October 30, 2021).

Fitte, C., Meier, P., Behne, A., Miftari, D. and Teuteberg, F. (2019). *Die elektronische Gesundheitsakte als Vernetzungsinstrument im Internet of Health.* INFORMATIK 2019: 50 Jahre Gesellschaft für Informatik – Informatik für Gesellschaft. Bonn: Gesellschaft für Informatik e.V, 111–124.

Gartner Newsroom Press Release. (2019). *Gartner 2019 Hype Cycle for Blockchain Business Shows Blockchain Will Have a Transformational Impact accross Industries in Five to 10 Years*. URL: https://www.gartner.com/en/newsroom/press-releases/2019-09-12-gartner-2019-hype-cycle-for-blockchain-business-shows (visited on March 15, 2022).

Hanley, M. and Tewari, H. (2018). "Managing Lifetime Healthcare Data on the Blockchain*". 2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, 246-251.

Hawig, D., Zhou, C., Fuhrhop, S., Fialho, A.S. and Ramachandran, N. (2019). "Designing a Distributed Ledger Technology System for Interoperable and General Data Protection Regulation–Compliant Health Data Exchange: A Use Case in Blood Glucose Data*". Journal of Medical Internet Research, 21*.

Häyrinen, K., Saranto, K. and Nykänen, P. (2008). "Definition, structure, content, use and impacts of electronic health records: a review of the research literature". *International journal of medical informatics*, *77*(5), 291-304.

Hepp, T. and Sharinghousen, M., Ehret, P., Schoenhals, A. and Gipp, B. (2018). "On-chain vs. off-chain storage for supply- and blockchain integration*". it - Information Technology*. 60.

Houtan, B., Hafid, A.S. and Makrakis, D. (2020). "A Survey on Blockchain-Based Self-Sovereign Patient Identity in Healthcare*". IEEE Access, 8*, 90478-90494.

Jin, H., Luo, Y., Li, P. and Mathew, J.P. (2019). "A Review of Secure and Privacy-Preserving Medical Data Sharing*". IEEE Access, 7*, 61656-61669.

Li, H., Zhu, L., Shen, M., Gao, F., Tao, X. and Liu, S. (2018). "Blockchain-Based Data Preservation System for Medical Data*". Journal of Medical Systems, 42*, 1-13.

Litan, A. (2021). *Hype Cycle for Blockchain 2021; More Action than Hype. Gartner. URL: https://blogs.gartner.com/avivah-litan/2021/07/14/hype-cycle-for-blockchain-2021-more-action-than-hype/* (visited on March 15, 2022).

Liu, J., Li, X.L., Ye, L., Zhang, H., Du, X. and Guizani, M. (2018). „BPDS: A Blockchain Based Privacy-Preserving Data Sharing for Electronic Medical Records*". 2018 IEEE Global Communications Conference (GLOBECOM)*, 1-6.

Nakamoto, S. (2008). *Bitcoin: A peer-to-peer electronic cash system*. URL: https://bitcoin.org/bitcoin.pdf (visited on November 15, 2021).

Nguyen, D.C., Pathirana, P.N., Ding, M. and Seneviratne, A.P. (2019). "Blockchain for Secure EHRs Sharing of Mobile Cloud Based E-Health Systems*". IEEE Access, 7*, 66792-66806.

Nickerson, R.C., Varshney, U., and Muntermann, J. (2013). "A method for taxonomy development and its application in information systems*". European Journal of Information Systems, 22*, 336-359.

Peffers, K., Tuunanen, T., Rothenberger, M.A. and Chatterjee, S. (2008). "A Design Science Research Methodology for Information Systems Research". *Journal of Management Information Systems, 24*, 45 - 77.

Pohlmann, S., Kunz, A., Ose, D., Winkler, E. C., Brandner, A., Poss-Doering, R., Szecsenyi, J. and Wensing, M. (2020). "Digitalizing Health Services by Implementing a Personal Electronic Health Record in Germany: Qualitative Analysis of Fundamental Prerequisites From the Perspective of Selected Experts*". Journal of medical Internet research*, *22*(1), e15102.

Schaar, P. (2014). *Anonymisieren und Pseudonymisieren als Möglichkeit der Forschung mit sensiblen, personenbezogenen Forschungsdaten*. Handbuch Ethik und Recht der Forschung am Menschen, Springer Berlin Heidelberg, 95–100.

Schinle, M., Wyszka, D., Schwarzler, F., Volz, K., Ruby, M., Sejdinovic, E. and Stork, W. (2018). "An Approach to digitalize Psychological Tests to support Diagnosis of Alzheimer's Disease in Ambulatory Care*". 2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, 1-6.

Schinle, M., Erler, C. and Stork, W. (2021). "Data Sovereignty in Data Donation Cycles - Requirements and Enabling Technologies for the Data-driven Development of Health Applications". *HICSS*.

Schneier, B. (1996). *Angewandte Kryptographie: Protokolle, Algorithmen und Sourcecode in C,* 1st Edition. ser. Reihe Informationssicherheit. Bonn: Addison-Wesley.

Theodouli, A., Arakliotis, S., Moschou, K., Votis, K. and Tzovaras, D. (2018). "On the Design of a Blockchain-Based System to Facilitate Healthcare Data Sharing*". 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, 1374-1379.

Thwin, T.T. and Vasupongayya, S. (2018). "Blockchain Based Secret-Data Sharing Model for Personal Health Record System". *2018 5th International Conference on Advanced Informatics: Concept Theory and Applications (ICAICTA)*, 196-201.

Wang, Y., Zhang, A., Zhang, P. and Wang, H. (2019). "Cloud-Assisted EHR Sharing With Security and Privacy Preservation via Consortium Blockchain*". IEEE Access, 7*, 136704-136719.

Wüst, K. and Gervais, A. (2018). "Do you Need a Blockchain?". *2018 Crypto Valley Conference on Blockchain Technology (CVCBT)*, 45-54.

Xiao, Z., Li, Z., Liu, Y., Feng, L., Zhang, W., Lertwuthikarn, T. and Goh, R. (2018). "EMRShare: A Cross-Organizational Medical Data Sharing and Management Framework Using Permissioned Blockchain*". 2018 IEEE 24th International Conference on Parallel and Distributed Systems (ICPADS),* 998-1003.

Xu, X., Weber, I. and Staples, M. (2019) *Architecture for Blockchain Applications*. 1st Edition. Cham: Springer International Publishing.

Yaga, D., Mell, P., Roby, N. and Scarfone, K. (2018). "Blockchain technology overview*"*. arXiv: Cryptography and Security.

Zaghloul, E., Li, T. and Ren, J. (2019). "Security and Privacy of Electronic Health Records: Decentralized and Hierarchical Data Sharing using Smart Contracts*". 2019 International Conference on Computing, Networking and Communications (ICNC)*, 375-379.

Zhang, A. and Lin, X. (2018). "Towards Secure and Privacy-Preserving Data Sharing in e-Health Systems via Consortium Blockchain*". Journal of Medical Systems, 42*, 1-18.

Zhang, P., White, J., Schmidt, D.C., Lenz, G. and Rosenbloom, S.T. (2018). "FHIRChain: Applying Blockchain to Securely and Scalably Share Clinical Data*". Computational and Structural Biotechnology Journal, 16*, 267 - 278.

Zheng, X., Mukkamala, R.R., Vatrapu, R. and Meré, J.B. (2018). "Blockchain-based Personal Health Data Sharing System Using Cloud Storage*". 2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom)*, 1-6.

Zhou, T., Li, X. and Zhao, H. (2019). "Med-PPPHIS: Blockchain-Based Personal Healthcare Information System for National Physique Monitoring and Scientific Exercise Guiding*". Journal of medical systems*, *43*(9), 305.