Where is our data? A Blockchain-based Information Chain of Custody Model for Privacy Improvement

Wagner Silva
Federal University of Rio de Janeiro State (UNIRIO)
Oswaldo Cruz Foundation (Fiocruz)
Rio de Janeiro, Brazil
wagner.silva@uniriotec.br
wagner.silva@fiocruz.br

Ana Cristina Bicharra Garcia
Federal University of Rio de Janeiro State (UNIRIO)
Rio de Janeiro, Brazil
cristina.bicharra@uniriotec.br

2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD) | 978-1-7281-6597-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/CSCWD49262.2021.9437727

***Abstract*— The advancement of Information and Communication Technologies has brought numerous facilities and benefits to society. In this environment, surrounded by technologies, data, and personal information, have become an essential and coveted tool for many sectors. In this scenario, where a large amount of data has been collected, stored, and shared, privacy concerns arise, especially when dealing with sensitive data such as health data. The information owner generally has no control over his information, which can bring serious consequences such as increases in health insurance prices or put the individual in an uncomfortable situation with disclosing his physical or mental health. While privacy regulations, like the General Data Protection Regulation (GDPR), make it clear that the information owner must have full control and management over their data, disparities have been observed in most systems and platforms. Therefore, they are often not able to give consent or have control and management over their data. For the users to exercise their right to privacy and have sufficient control over their data, they must know everything that happens to them, where their data is, and where they have been. It is necessary that the entire life cycle, from generation to deletion of data, is managed by its owner. To this end, this article presents an Information Chain of Custody Model based on Blockchain technology, which allows from the traceability of information to the offer of tools that will enable the effective management of data, offering total control to its owner. The result showed that the prototype was very useful in the traceability of the information. With that it became clear the technical feasibility of this research.**

Keywords*— *privacy, blockchain, sensitive data, health data, traceability, information management, chain of custody

I. INTRODUCTION

Technology has dramatically changed the way people relate to each other and the services they offer to them. If someone asked us now if we know where our data are, the answer would probably be ”no”. In this technological world, data and information are present everywhere. The simple act of talking on a cell phone, sending an email, chatting in a chat application, and using any smart device is enough for our data to be collected, stored, and shared. This fast dissemination of information raises concerns about data privacy. It is necessary to identify everything that happens to our information for privacy is guaranteed [1], [2].

Privacy concerns have been heightened when we are talking about sensitive data, such as health data. During the recent COVID-19 pandemic, a lot of confidential data has been generated and shared between the medical community and institutions for research purposes [3]. Governments and private organizations have taken several initiatives to treat, mitigate and prevent the spread of the disease [4]. These initiatives include the development of technological solutions to control the pandemic through the capture, use, and sharing of personal information, which includes health data [5]. There are many cases of sharing and unauthorized use of personal data. An excellent example in the private sector comes from the partnership of two giant companies, Apple and Google. They worked together to develop technology that would allow communication between their mobile operating systems (Android and iOS). The idea is to allow infected people to send information about their health status anonymously [6]. An example of government action comes from the South Korean government, which has developed a system to track individuals suspected of being infected. Although the objective was good, revealing the places visited by the infected had negative economic impacts, such as the closing of some stores and restaurants [7].

Privacy regulations have been created and adjusted over the years to ensure the privacy of information. The recent General Data Protection Regulation (GDPR) [8] is a good example of regulation that address privacy issues. GDPR include standards beyond the entire data life cycle, from the source of information to disposal. Although this regulation exists, not all initiatives address their guidelines correctly, bring risks to privacy. This subject is so hot that in Brazil, the General Data Protection Law (LGPD), which has been expected to be launched in 2021, was brought forward to 2020. This regulation was prepared based on the GDPR, which makes them very similar [9]. GDPR, for example, points out that the owner of the information must have full control and management over his data. This is a fundamental characteristic of a privacy guarantee. This regulation says that systems and platforms must provide tools for users to choose what, with whom, and when they want to share, in addition, to know what is happening with their information; furthermore, everything that happens with their information must have their explicit consent [10].

Although regulations such as GDPR, it makes it clear that

the information owner must have control and management of their data [10], disparities have been observed [11], [12]. The user’s explicit consent is the primary legal basis for the processing of personal health data. However, it appears that it is difficult for users to be adequately informed about what happens

978-1-7281-6597-4/21/$31.00 ©2021 IEEE

to their data and, therefore, they are often not in a position to give consent or to have control over the data. Because many times the user does not know where their information is and what is happening to it. Once the information is shared, the information owner has no way of knowing where his data is being used.

Think about this problem. For the users to exercise their right to privacy and have significant control over their data, they must know everything that happens to them, where their data is, and where they have been. It is necessary that the entire life cycle, from generation to deletion of data, is managed by its owner. To this end, this article presents an Information Chain of Custody Model based on Blockchain technology, which allows from the traceability of information to the offer of tools that allow the effective management of information, offering total control to its owner, placing him as the main actor in this process.

After this introduction, the remainder of the paper is structured as follows. In Section 2, we provide an overview of the required acquirement. Section 3 shows our related work. As our solution in Section 4, we present a Blockchain-based Information Custody Chain Model for Privacy Improvement and a prototype for evaluation. Section 5 shows our prototype for the presented model. In Section 6, we make a discussion about the solution and, finally, in Section 7, we make our conclusions.

II. Background

A. What is Chain of Custody?

The chain of custody is a term used many times in the archival management of digital documents area. It can be understood as the environment in which the life cycle of the documents goes through. In other words, it defines who is responsible for applying archival principles and functions to documentation [13]. In this scenario, the custodian is seen as the owner of the document and responsible for ensuring integrity and management [14]. Still about digital documents, Flores (2016) [13] affirms that both the management and preservation of documents must be carried out through information systems, where it is possible to guarantee the security and management of documents.

When looking at personal data and information and making

a comparison, we can conclude that the concept of Chain of Custody can be correctly applied to this domain. When we talk about privacy guarantee, security, and data management and information, providing tools for the information owner to allow him to manage his information effectively, we are transforming him into a custodian of his information. No research was found using the Chain of Custody concept in information management. So this research appropriated this concept to apply it to information management, proposing na Information Chain of Custody Model based on Blockchain technology.

B. Privacy

With the emergence of the Internet and new perspectives for interaction and communication between individuals, people have faced an increasing number of decisions regarding their information privacy. Decisions ranging from aspects of setting visibility on social networks to the option to download a smartphone application based on access to confidential data that it requests [15].

Privacy is a concern that affects all people, online and off [16], but many people do not know what's privacy is. According to Westin, privacy is the right of every individual to determine who, when, and how his or her data will be collected, stored, used, and discharged [17]. Information privacy demands are strongly affected by the individuals' perception of consequences for sharing their data [18]. Although related, privacy and information security are different concepts. While security aims to deal with unauthorized personal data capture, privacy goes beyond involving protection against unwanted data usage and share that violate confidentiality and information integrity [19].

Privacy regulations have been created and adjusted over the years to guarantee the information privacy. The recent deployed European General Data Protection Regulation (GDPR) [8] aims to protect the privacy of personal data, and it was implemented in May 2018 in all countries of the European Union [8], [10]. This norm consists of a detailed and extensive document, which discusses important topics. Besides that, GDPR states that systems and platforms must provide tools for users to choose what and when they want to share, besides knowing what will happen with their information. It is essential to highlight that, although GDPR regulates the European Union members, its impacts affect the world because any company that needs to access any European citizen data must comply with the norm [8], [10].

C. Blockchain

Blockchain technology has attracted attention as the basis for transactions involving cryptocurrencies like Bitcoin. How- ever, its capabilities go far beyond that, allowing applications based on this technology to be vastly enhanced. Blockchain, like the Internet, is an open global infrastructure based on a distributed structure and consensus algorithms [20].

The blockchain is a network that works with very secure chained blocks that always carry content with a fingerprint. This content can be any information, from a financial transaction to a patient's medical record or complete medical history [21]. These blocks are dependent on each other and form a chain of blocks (hence the name: Blockchain). This possibility makes it this technology perfect for recording information that needs to be trusted. As illustrated in the figure 1, the great innovation is that the posterior block (except the first block, since there is none behind it) contains the fingerprint (hash) of the previous block plus its content and, with this two information, generate your fingerprint (hash).

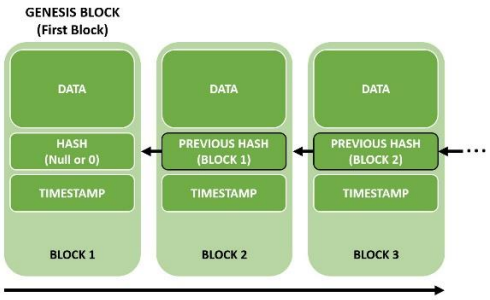

Fig. 1. Blockchain simple structure

This possibility ensures that the chain remains intact, and the data for each block cannot be changed or deleted. Otherwise, this would invalidate the chain after a certain point. Besides, it allows navigability between the blocks in the sequence in which they were created, functioning as a change history. [22].

The Blockchain's main characteristics are: Decentralized storage: Blockchain stores information transparently and delivers it to third parties with the owner's consent. Consent: the consensus algorithm controls the access, storage, and distribution of information on a network. Immutability: changing or altering data is impossible [23].

III. Related Works

Blockchain has been used to solve numerous problems in the health area; these problems range from medical research to treatments. Due to this research's nature, it is essential to investigate how this technology is being used in other research involving Blockchain and health data. Therefore, in this section, we review research related to health information sharing trends with a focus on control and data management based on blockchain technology.

An important application is in medical research. In this field, the user must provide consent for the use of the information. However, informed consent forms are the most common types of clinical fraud [24]. This situation includes changes and document forgery. To get around this situation, Benchoufi et al. [25], implemented a smart contract system based on blockchain that prevents doctors from using patient data until a key has been released at the end of the process, while also allowing revocation of consent.

Blockchain technology has significant power to offer data control to the patient. A particularly an interesting initiative in this direction is that proposed by Azaria et al. [26], which offers patients and doctors an immutable record of health data. In drug tracking, blockchain technology is another opportunity. It takes advantage of the immutability of the blockchain to develop tracking from the manufacturer to the patient. This solution allows healthcare providers to be in accord with the current safety standards pharmaceutical supply [27].

The research presented by Yue et al. [28] and Vora et al. [29] introduced a framework, based on blockchain, to allow the patient to control and share their data securely, without violating privacy.

Franciscon et al. presents a systematic literature review based on technologies relating to Blockchain architectures applied by governments to public services. In this review, the authors conclude that the solutions using Blockchains are very diversified, owing to efficient data treatment through the immutability and traceability [30].

IV. Blockchain-based Information Custody Chain Model

One of the pillars of GDPR is to offer total control to the information owner about his data. To enjoy this right, it is essential that, first, they know everything about their data, where they are, where they went, and what happened to them. For this, this article presents an Information Chain of Custody Model (figure 2) based on Blockchain technology, which allows, from the traceability of information to the offer of tools, the effective management of information, offering total control to its owner, placing him as the main actor in the privacy-related issues.

The proposed model was divided into three layers, each with its specific purposes. The Blockchain layer is the first and the most important. It provides all security in the storage of information. Besides, it allows traceability and information management mechanism. The second layer offers information management controls to users, and the third layer uses Digital Rights Management (DRM) technology to improvement the privacy of the data that has been accessed.

A. Blockchain Layer

The layer that receives the data to be stored is called the Blockchain Layer. It is the most important of the model for ensuring that the data is immutable and secure, while also allowing its traceability, transparency, and management.

Immutability and security is guaranteed because it is impossible to change data when the information is stored in a block of the blockchain [22]. Since implemented in a distributed architecture, this technology works with a consensus protocol, where the copies are always comparing. In case of violation in any chain, the consensus protocol quickly detects the change and updates the changed chain, bringing the original value before the change. Since there is no single point of attack, the information inserted in the blockchain is extremely safe and reliable.

The data and all access to them have been stored in the blocks. These blocks depend on each other through a cryptographic hash, forming a chain of blocks. This possibility makes it the perfect technology for recording information that needs trust and transparency. Since the blocks have been connected and each block represents data access, this allows efficient information management and traceability. Imagine that block 3 represents access to data by a specific hospital, block 4 would be access to data by a researcher from this hospital. Thus, the researcher did not obtain access to the original data, but through the hospital, all of this is registered in the Blockchain.

The left side of the figure represents the content of a selected block in the chain. In this case, the selected block is block 7. In each block, we have the block number, creation date and time, the hash of the selected block, the hash of the previous block (allowing the blocks to be linked to each other), personal data of the information owner, and, in this example, the stored health data. The information of who has accessed the data can be accessed through the option "Access data" (figure 3).

B. User Control Layer

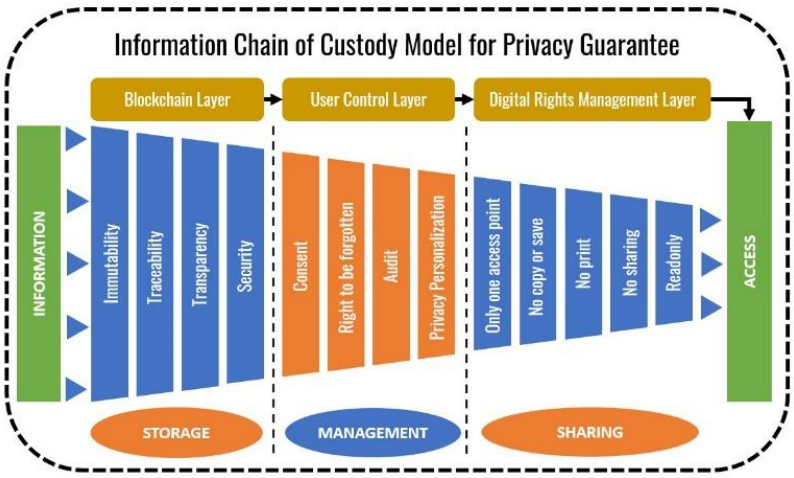

Fig. 2. Custody Chain Information Traceability Model

The User Control Layer is responsible for offering all the tools for managing their information. The management is done through the information contained in the blockchain (previous layer). Considering that it allows traceability and, consequently, the management of information, the user will be able to identify, in real-time, where his data is, who had access and other information, offering a complete audit. Besides, this layer offers privacy customization mechanisms, where the user can define, for example, what information will be shared and for how long time. The information owner may also revoke access to information already shared, guaranteeing his right to be forgotten. Finally, the user can consent or not access to their information.

All the above options are offered to ensure adherence to regulations such as GDPR, which has its main foundation, the defense that the user must always control their data.

C. *Digital Right Management Layer*

A challenge in this model is how to provide access to information, ensuring that its owner is still in control of it. We can make an analogy to an image shared on social networks, where once there, nothing guarantees that it was not or will be copied and shared. Because of that, it has been decided to use in this layer the concept of Digital Rights Management (DRM) [31].

Thus, this layer will determine several "rights" pre-established by the user in the previous layer to access the data. Some of the options allow or not copy, print, download, how many times they can be downloaded, or even offer a partial reading of the information. Summarizing, DRM restricts what we can do with the information according to what the user determines.

V. Prototype

The purpose of this prototype is to validate the proposal presented above. As this research has not been completed yet, some peculiarities in the model have not been implemented yet. To demonstrate technical feasibility, it was decided to implement the most critical layer of the project (Blockchain) first.

In the prototype, the information is stored in blocks of the Blockchain Layer, where each block represents the data, permissions, and accesses made to the information. Blockchain technology supports smart contracts, which allows us to automate and track specific state transitions (such as a change in viewing rights or creating a new record in the system). Through these smart contracts in a blockchain, we record all the information access. We have included a cryptographic hash of the blockchain registry to prevent a breach and allow the link between the blocks, thus ensuring data integrity and traceability. New records, associated with a specific user (owner of the information), can be added. Users (owner of the information) can authorize the sharing of records between interested parties.

In practice, the process is:

- The data is inserted into the Blockchain through a storage module, from there, this technology guarantees the integrity and immutability of the information.
- When data access occurs, the Blockchain generates a new block where the content to be accessed is stored, the link to the block containing the original information accessed and the access information such as who accessed it and the location where it has been done.
- From that point, everything is registered on the Blockchain, allowing the owner of the information to make a perfect management and tracking of his information.

Figure 3 shows a representation of the Blockchain Layer. On the right side, we can see a chain referring to the information previously stored. Block 1 stores the original data. The remaining blocks represent each data access. This access can be done directly in block 1 or indirectly, as is the case in blocks 4, 5, 6 and 7. This allows multi-level traceability.

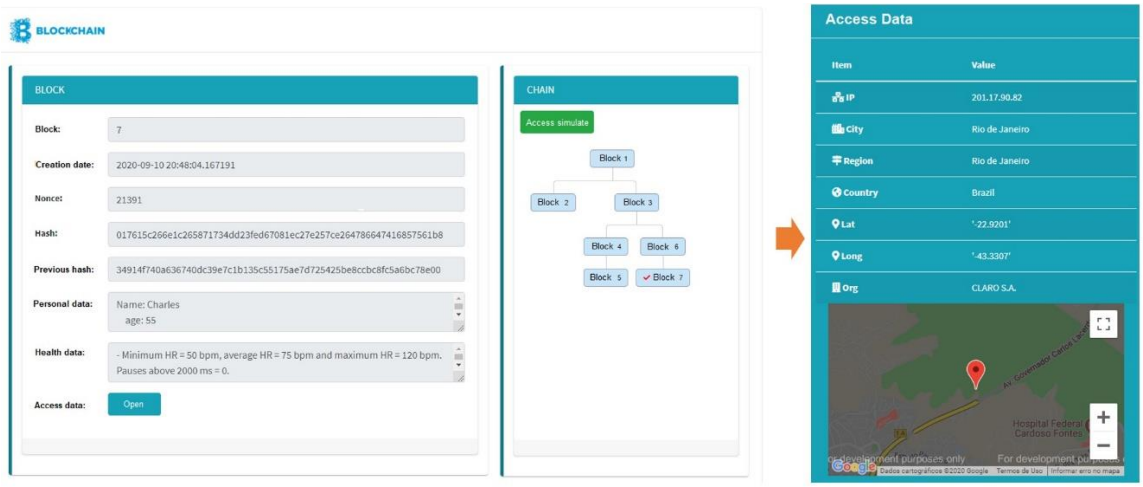

Fig. 3. Blockchain Layer

As each block in the Blockchain represents one access (except the first block), access information will be available when accessing each block individually.

The data access information is stored in the Blockchain too. Thus, the same security and immutability guaranteed for data are also guaranteed for data access information, where it is impossible to change to erase the tracks from which the data was accessed. The stored data access information can identify who accessed the data, from which dispositive the access occurs, the country, state, city, and location in terms of approximate Latitude and Longitude. So the information owner has full knowledge of where their data is, and what's happened with that. This opens many opportunities to offer greater control over the data to its owners. These controls will be implemented at the user control layer.

VI. Discussion

The analysis and evaluation of the prototype were done through access simulations. For that, the tool needed to be available on the internet so that it would be able to identify and store the information regarding who is accessing the data, their location, date, time, and other important information. The access was made from three different locations:

- In one institution: The objective was to verify that the prototype was able to identify where the access was being made from and which institution would be making that access to the data.
- Mobile device: the objective was to verify if the prototype would be able to locate that from its access.
- Personal computers: the objective was the same defined for access through mobile devices since personal computers do not identify their owner.

First, the health data of a fictitious user was entered into the Blockchain. After that, the first simulation was accessed by an institution. It was observed that this access allowed greater precision, identifying the institution's name that was accessing the data. Access by a mobile device and personal computers identified only the internet service provider and the access's approximate location. Figure 4 shows the result of the three accesses.

The result showed that the prototype was very useful in the traceability of the information. With that it became clear the technical feasibility of this research. As the access made by an institution gave us more precision in the captured information, it was decided to use more institutions in future evaluations.

VII. Conclusions

Ensuring personal information privacy is a significant challenge, but it is necessary, especially in sensitive health data. Although it is their right, the owner of the information generally has no control over his information. Therefore, serious consequences can occur due to the improper disclosure of his information, such as increases in health insurance prices or placing the individual in an uncomfortable situation with disclosing our physical or mental health.

This article presented an Information Chain of Custody Model based on Blockchain technology. This model allows from the traceability of information to the offer of tools that allow the effective management of information, offering total control to its owner, agreeing with regulations such as GDPR.

Some limitations were identified and have been addressed during the research. One of these is the latitude and longitude information. They returned the city information. This particular limitation is not a problem in institutions' case, as it is easy to identify an institution present in a city. The biggest problem was the access by a mobile device and personal computers.

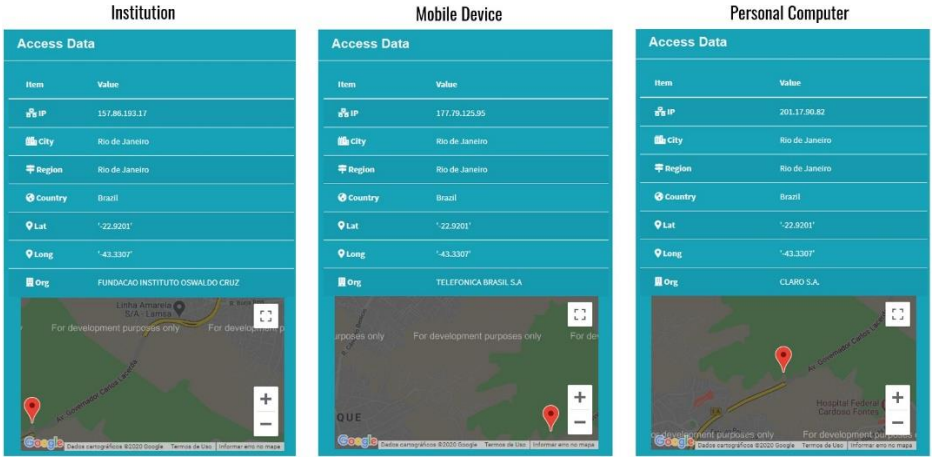

Fig. 4. Access Data Information

As a future work, which has already been started, is implementing the other layers of the model, corrections, and improvements such as the one mentioned in the previous paragraph and, after completion, an evaluation with more participants involving several institutions.

REFERENCES

[1] A. Westin, "Improving access and protecting privacy," Connecting Americans to Their Health Care, 2006.

[2] W. Silva et al., "Health information, human factors and privacy issues in mobile health applications," in 53rd HICSS, 2020.

[3] D. Wu et al., "The sars-cov-2 outbreak: what we know," Int. J. of Infectious Diseases, 2020.

[4] E. Luysterborg, "Privacy and data protection in the age of covid-19: Deloitte belgium: Covid-19," Apr 2020. [Online]. Available: https://www2.deloitte.com/be/en/pages/risk/articles/privacy- and-data-protection-in-the-age-of-covid-19.html

[5] D. S. W. Ting et al., "Digital technology and covid-19," Nature medicine, vol. 26, no. 4, pp. 459–461, 2020.

[6] R. Kitchin, "Using digital technologies to tackle the spread of the coronavirus: Panacea or folly," The Programmable City Working Paper 44. Available at: http://progcity . . . , Tech. Rep., 2020.

[7] S. Park et al., "Information technology–based tracing strategy in response to covid-19 in south korea—privacy controversies," Jama, 2020.

[8] P. Voigt et al., "The eu general data protection regulation (gdpr)," A Practical Guide, 1st Ed., Cham: Springer International Publishing, 2017

[9] P. P. Pinheiro, Proteção de Dados Pessoais: Comentários a lei 13.709/2018-LGPD. Saraiva Educação SA, 2020.

[10] G. D. P. Regulation, "Gdpr - general data protection regulation," URL: https://gdpr-info.eu/(visited on 25/06/2020), 2018.

[11] Y. Liu et al., "Analyzing facebook privacy settings: user expectations vs. reality," in 2011 ACM SIGCOMM, 2011, pp. 61–70.

[12] M. Netter, et al., "Privacy settings in online social networks–preferences, perception, and reality," in 2013 46th HICSS. IEEE, 2013, pp. 3219–3228.

[13] D. Flores, et al., "Cadeia de custo´dia para documentos arquiv´ısticos digitais," Acervo, vol. 29, no. 2, pp. 117–132, 2016.

[14] H. Jenkinson, A manual of archive administration including the problems of war archives and archive making. Clarendon Press, 1922, vol. 4.

[15] A. Acquisti et al., "Nudges for Privacy and Security: Understanding and Assisting Users' Choices Online," ACM Computing Surveys, vol. 50, no. 3, pp. 1–41, aug 2017. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3101309.3054926

[16] N. McDonald et al., "The politics of privacy theories: Moving from norms to vulnerabilities," in 2020 CHI, 2020, pp. 1–14.

[17] A. F. Westin, "Privacy and freedom," Washington and Lee Law Review, vol. 25, no. 1, p. 166, 1968.

[18] M. J. Dupuis et al., "Measuring the human factor in information security and privacy," in 49th HICSS, 2016, pp. 3676–3685.

[19] M. E. o. Whitman, Principles of information security. Cengage Learning, 2011.

[20] S. Underwood, "Blockchain beyond bitcoin," 2016.

[21] D. Drescher, Blockchain basics. Springer, 2017, vol. 276.

[22] M. Crosby et al., "Blockchain technology: Beyond bitcoin," Applied Innovation, vol. 2, no. 6-10, p. 71, 2016.

[23] H. M. Hussien, et al., "A systematic review for enabling of develop a blockchain technology in healthcare application: taxonomy, substantially analysis, motivations, challenges, recommendations and future direction," vol. 43, no. 10, p. 320, 2019.

[24] J. Barrett, "Fraud and misconduct in clinical research," Principles and Practice of Pharmaceutical Medicine, 1988.

[25] M. Benchoufi, et al., "Blockchain protocols in clinical trials: Transparency and traceability of consent," F1000Research, vol. 6, 2017.

[26] A. Azaria, et al., "Medrec: Using blockchain for medical data access and permission management," in 2016 2nd OBD. IEEE, 2016, pp. 25–30.

[27] V. Dhillon, D. Metcalf, and M. Hooper, "The hyperledger project," in Blockchain enabled applications. Springer, 2017, pp. 139–149.

[28] X. Yue, et al., "Healthcare data gateways: found healthcare intelligence on blockchain with novel privacy risk control," Journal of medical systems, vol. 40, no. 10, p. 218, 2016.

[29] J. Vora, et al., "Bheem: A blockchain-based framework for securing electronic health records," in 2018 IEEE - GC Wkshps. IEEE, 2018, pp. 1–6.

[30] E. A. Franciscon et al., "A systematic literature review of blockchain architectures applied to public services," in 2019 IEEE 23rd CSCWD. IEEE, 2019, pp. 33–38.

B. Rosenblatt et al., "Digital rights management," New York, 2002.