future internet

MDPI

Article

Healthchain: A Privacy Protection System for Medical Data Based on Blockchain

Baocheng Wang * and Zetao Li

School of Information Science and Technology, North China University of Technology, Beijing 100144, China; lzt953822835@gmail.com

* Correspondence: wbaocheng@ncut.edu.cn

Abstract: Recently, with the great development of e-health, more and more countries have made certain achievements in the field of electronic medical treatment. The digitization of medical equipment and the structuralization of electronic medical records are the general trends. While bringing convenience to people, the explosive growth of medical data will further promote the value of mining medical data. Obviously, finding out how to safely store such a large amount of data is a problem that urgently needs to be solved. Additionally, the particularity of medical data makes it necessarily subject to great privacy protection needs. This reinforces the importance of designing a safe solution to ensure data privacy. Many existing schemes are based on single-server architecture, which have some natural defects (such as single-point faults). Although blockchain can help solve such problems, there are still some deficiencies in privacy protection. To solve these problems, this paper designs a medical data privacy protection system, which integrates blockchain, group signature, and asymmetric encryption to realize reliable medical data sharing between medical institutions and protect the data privacy of patients. This paper proves theoretically that it meets our security and privacy requirements, and proves its practicability through system implementation.

Keywords: blockchain; e-health; encryption; group signature; privacy protection

Citation: Wang, B.; Li, Z. Healthchain: A Privacy Protection System for Medical Data Based on Blockchain. *Future Internet* **2021**, *13*, 247. https://doi.org/10.3390/fi13100247

Academic Editors: Peter Kieseberg and Thomas Moser

Received: 19 August 2021
Accepted: 22 September 2021
Published: 24 September 2021

1. Introduction

With the widespread application of information technology in the medical field, there is a surge in the amount of medical data being produced. Because there is a large amount of such data, which itself has diverse data structures, collecting and analyzing these data and exploring their potential value can effectively promote the progress of clinical medicine and drug research and development [1]. Alternatively, in order to record and track the patient's condition more efficiently, and to facilitate patient referrals, many medical institutions have jointly established shared databases to store data, such as electronic medical records.

However, while this increases the convenience of electronic medical treatment and helps with the analysis of medical data (which promotes the development of medical research), privacy disclosure problems exposed by medical data sharing are becoming more and more obvious. For example, in an analysis of thousands of online medical service systems around the world, Greenbone Networks found that more than 24 million patient data records could be easily accessed or downloaded from the network [2]. In addition, more than 1.19 billion medical images are transmitted on the network due to an improper configuration of the picture archiving and communication system (PACS). These data contain a large number of personal privacy information, such as name, date of birth, attending physician, diagnostic information, etc. For some particular diseases (such as AIDS), the disclosure of such information could seriously damage personal reputations. Worse, attackers will be able to analyze and use this information to commit fraud (e.g., using an ID card number to impersonate a patient's identity), extortion, or to tamper with this information. Therefore, the interests of many institutions and individuals are in conflict

with such technologies because of privacy concerns. Obviously, these problems have become a major obstacle to the comprehensive popularization of electronic medical systems.

Therefore, the design of a medical data-sharing solution which is able to protect patient privacy as much as possible in the life cycle of medical data, and in the future for data analysis, is particularly important.

It should be noted that due to the variety of types and sources of medical data, this paper mainly explores the protection of various clinical data and derived data generated by patients in the process of treatment in the hospital, such as personal identity information, medical institution information, electronic medical record data, and medical insurance information. In addition, in the whole life cycle of medical data (data acquisition, data storage, data sharing, and data analysis), this paper mainly solves the problem of the privacy protection of data in the storage and sharing stages.

This paper shows how we can use blockchain and mainstream cryptography technology to provide a relatively safe and efficient solution for medical data sharing between medical institutions within a certain range. It is worth mentioning that, because there is no absolutely secure encryption method, we try to consider the issue of privacy protection using another way of thinking, that is, by dividing the data and using different methods to hide it. In so doing, the sensitivity of the data is greatly reduced. Only authorized personnel can obtain the complete data. Otherwise, even if a certain part of the data is cracked, the attacker can only obtain the incomplete data, and the privacy can still be better protected. This is a major innovation outlined in this article.

The rest of this article is organized as follows: in Section 2, we review and analyze the existing medical data privacy protection solutions, and outline our contributions; in Section 3, we identify the security requirements of the system based on a literature review; in Section 4, we briefly describe the cryptographic primitives used in this article; in Section 5, we outline the system design; and in Sections 6 and 7, we prove its feasibility through security analysis and system implementation. Section 8 is the conclusion of this article.

2. Related Work

Today, with increasingly advanced technology, healthcare, as a huge industry closely related to everyone, is highly valued by governments and enterprises all over the world. Companies interested in healthcare not only include traditional medical companies, but also technology giants, such as Microsoft and Google. Especially in the current global background of the raging COVID-19 pandemic, people are increasingly placing high hopes in e-health, which has seemingly unlimited potential for development [3]. In response to a global pandemic like COVID-19, the source of medical data will no longer be limited to medical institutions, but will rely on mobile devices, including wearable devices [4]. The health data collected independently by the user is used as a direct digital data source to ensure the reliability, integrity, security, attribution, and auditability of the data. It helps to improve the overall assessment and monitoring of patients as individuals [5]. In addition, in order to better integrate and make full use of these large amounts of fragmented data, machine learning and artificial intelligence algorithms are used to build AI models that can predict COVID-19 symptoms, understand how it spreads, and speed up the use of medical data research and treatment [6]. These solutions have encountered and solved the problem of data source credibility, the problem of data sharing caused by mutual distrust between enterprises, and the problem of how to integrate and utilize a large amount of fragmented data. At the same time, they also encountered another common problem, that is, the problem of data privacy protection. It can be seen that in the future digital medical field, the issue of data privacy will be a major and inevitable issue.

Many solutions have been proposed for the privacy protection of medical data. In the data acquisition stage, traditional privacy protection schemes are mostly based on anonymous technology. The fundamental idea is to hide the relationship between individuals and data. However, if we are only deleting personal attribute information, it is still possible

for individuals to be identified [7], which is not conducive to data sharing in the scenario of electronic medical records. Thus, k-anonymity [8] was proposed, followed by a series of improvements [9–11]. However, such schemes often rely too much on assumptions of the attackers' background knowledge, and cannot provide rigorous and effective proof to users regarding their level of privacy protection. Therefore, differential privacy technology [12] has been introduced into this field. For the data storage stage, the existing solutions are mostly based on cloud platforms for data storage. For example, in 2014, Thiranant et al. [13] proposed a design scheme for a data privacy security framework for electronic health systems which are based on web services. This scheme limits the access rights of different types of users and encrypts the files before they are uploaded to the cloud in order to protect data privacy. Similarly, Ilokah et al. [14] proposed a framework based on ABE, which allows the secure outsourcing of computationally intensive data decryption processes to cloud servers. This reduces the time required for decryption on the user's side, and reduces the computing power required for users to access the data. Many encryption schemes (such as [15–17]) and audit schemes [18–20] have been proposed to solve data confidentiality and integrity problems in the data storage process, but how to select the appropriate auditors is a problem that still needs to be solved. At the same time, this approach relies heavily on the reliability of service providers and applications. If the third party is malicious, the effectiveness of these encryption schemes and audit schemes is compromised. Because of the natural limitations of the single-server architecture, such as the trust problem of the third party and the issue of single-point failure, such schemes cannot fully guarantee the integrity and confidentiality of data, and this therefore leads to the destruction of data privacy.

With the characteristics of decentralization, traceability, and non-tampering, the potential of blockchain technology in the field of electronic medicine is gradually being discovered. Therefore, many schemes based on this technology have emerged (such as [21–23]). However, these schemes still have some limitations. For example, in the electronic medical scene, the anonymity of blockchain nodes is not conducive to the traceability of transactions and related personnel—and yet if the identity is public, it will damage privacy. In addition, unlike Bitcoin [24], the particularity of the electronic medical scene means that the transaction itself is a sensitive piece of information, which is not conducive to the privacy protection of data because all transactions on the blockchain are open and transparent. In recent years, there have been many papers focusing on the use of blockchain technology to solve the problems encountered in the efficient transmission of medical data, with its need for privacy protection, such as [25–27].

In order to facilitate medical data sharing and protect the privacy of patients, this paper integrates blockchain with group signature. Taking medical institutions as group members and consensus nodes, after receiving transactions, the information is signed, and then packaged and uploaded to the blockchain, so that it cannot be tampered with. At the same time, due to the characteristics of group signature, only group managers can identify specific signers, which ensures traceability while protecting privacy. In addition, in order to further provide data confidentiality, this paper also introduces symmetric encryption and asymmetric encryption, and combines asymmetric encryption with group signature. We construct a consortium blockchain and analyze its security to prove the feasibility of the scheme.

3. Requirement Analysis

The requirement model of the solution proposed in this paper is shown in Figure 1.

By analyzing the model, we can conclude the following functional requirements and non-functional requirements:

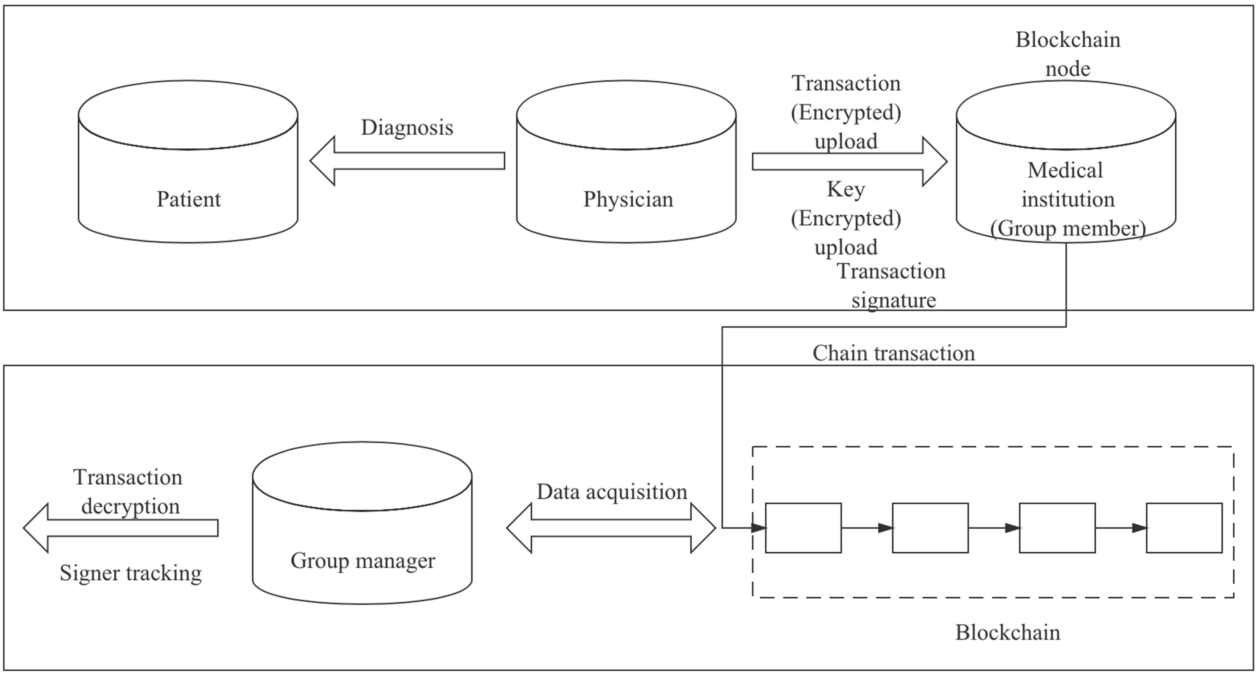

Figure 1. Requirement model.

3.1. Functional Requirements

Data encryption: After the doctor enters the patient information and diagnosis information into the system, the system should quickly and relatively securely encrypt each piece of data, and assign a unique serial number to enable quick retrieval without decrypting the data.

Data upload: After the data is encrypted, the physician should be able to upload the encrypted data to the relevant department of the medical institution where they are located through the system.

Data signature: After the relevant department receives the data, the data should be signed to incorporate the organizational information.

Chain transaction: All medical institutions should also act as blockchain consensus nodes to collect transactions and upload them to the blockchain after reaching a consensus.

Data acquisition: Only personnel authorized by the system should be able to quickly retrieve and obtain data from the blockchain using the serial number.

Data decryption: The authorized person should be the only person who can decrypt and obtain the complete data and be at the core of data sharing among medical institutions. (There can be more than one authorized person, but it should be as few as possible).

Signer tracking: The authorized person should be the only person who can identify the signer by their signature.

3.2. Non-Functional Requirements

Anonymity: The system should fully protect the privacy of relevant personnel, including, but not limited to, the privacy of patients, such as medical institution information, because such information may still be used to analyze the true identity of patients.

Traceability: In order to avoid malicious acts, the system should be able to trace the signers of relevant affairs when necessary while protecting the privacy of personnel.

Data integrity: The system should ensure that the data is difficult to tamper with and provide efficient data integrity verification.

Data confidentiality: Transactions in the electronic medical scene cannot be explicitly stored in the blockchain, so the system should be able to reliably encrypt data.

Data privacy: In addition to protecting personal privacy by anonymous means, sensitive information within the transaction should also be protected. The protection here means that if the data are cracked, the data do not leak, or, if they do, they leak as little sensitive information as possible.

Resistance of birthday collision: The system should prevent two blocks from being generated simultaneously.

Resistance of interception: The system should be able to detect the interception and modification of a transaction and invalidate it.

Resistance of various attacks: The system should be able to resist various traditional attacks, such as a DDoS attack, or a modification attack.

4. Cryptographic Primitives

In this section, we briefly introduce the basic principles of the cryptography techniques used in our proposed scheme.

4.1. Symmetric Encryption

Symmetric encryption is an encryption algorithm where the same key needs to be used for encryption and decryption. Because of its fast encryption and decryption speed, symmetric encryption is often used when a large amount of data needs to be encrypted. In our scheme, we use an AES algorithm [28] to encrypt medical data. As a substitute for a DES algorithm [29], this type of algorithm has been deeply analyzed and widely used all over the world. For different application scenarios, the key length is also different. There are three common schemes for an AES algorithm, namely AES-128, AES-192, and AES-256. This section provides an overview of AES-128.

As a block cipher, this AES algorithm groups pieces of plaintext together, with 128 bits (i.e., 16 bytes) in each group. If a group has less than 128 bits, it will be automatically supplemented. Each group of data is encrypted until the complete plaintext is encrypted.

The encryption process of this AES algorithm includes ten rounds, each involving four operations:

SubBytes: use a substitution-box to replace the group byte-to-byte.

ShiftRows: a simple displacement.

MixColumns: an alternative to using the arithmetic properties on the field GF (2^8).

AddRoundKey: bitwise XOR is performed on the current group and part of the extension key.

The encryption key of each round is obtained by the expansion of the initial key. In addition, the 16-byte plaintext, ciphertext, and key of each round are represented by a 4×4 matrix. The approximate encryption process is shown in Figure 2. The specific details of each operation in the encryption process are not repeated here. Interested readers can learn about it in the relevant literature [28].

The decryption process of the AES algorithm is the inverse operation of encryption process. Since each operation is reversible, decryption in the reverse order can recover the plaintext. However, the need for the receiver of the message to obtain the secret in order to decrypt the ciphertext is an urgent problem. This leads users to incorporate asymmetric encryption.

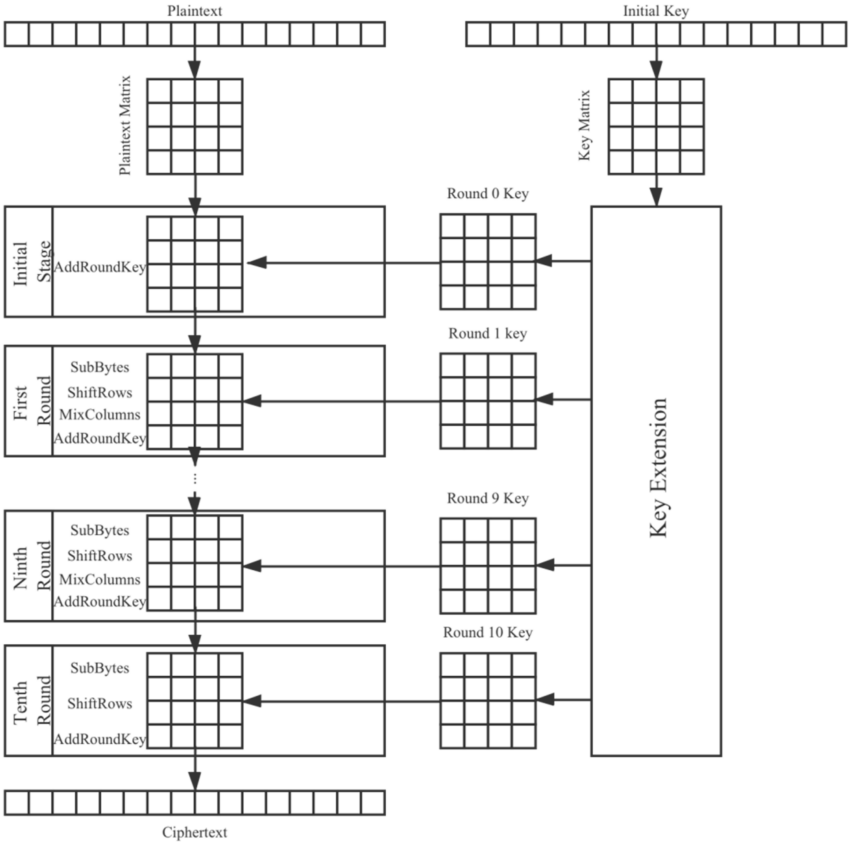

Figure 2. Encryption process of the AES algorithm.

4.2. Asymmetric Encryption

Asymmetric encryption uses two different keys for encryption and decryption. Compared with symmetric encryption, asymmetric encryption does not have the problem of secure key transmission, and has better security. However, the speed of encryption and decryption is reduced accordingly. Therefore, asymmetric encryption is usually used to encrypt the key generated by a symmetric encryption algorithm. The flow of an asymmetric encryption algorithm is shown in Figure 3. In the scheme proposed in this paper, we use an RSA algorithm [30] for the encryption of the symmetric encryption key. The security of the RSA algorithm is based on the factorization of large integers. It is the most widely used asymmetric encryption algorithm. It was jointly proposed in 1978 by three scholars of MIT: Ron Rivest, Adi Shamir, and Leonard Adleman.

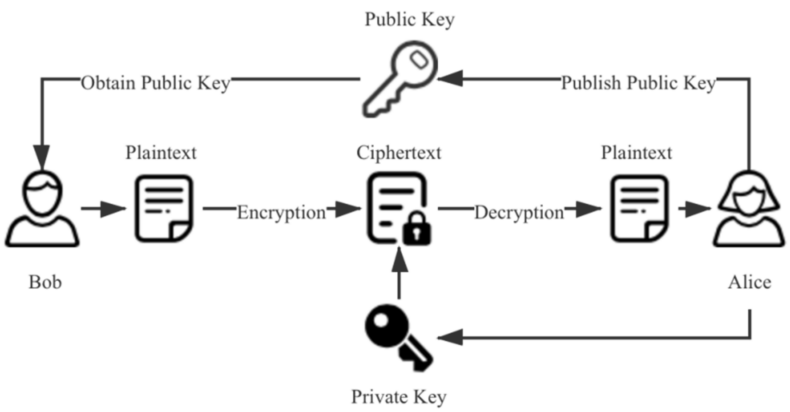

Figure 3. Asymmetric encryption algorithm flow.

The encryption and decryption process of an RSA algorithm includes the following steps:

1. Randomly select two different prime numbers p and q. According to the computational power of the current computer, these two numbers should comprise at least 200 bits before they can be considered safe enough in practice.
2. Calculate the common modulus n:

$$n = p * q \tag{1}$$

3. Calculate the Euler function for the common modulus n:

$$\varphi(n) = (p - 1)(q - 1) \tag{2}$$

4. Randomly select an integer e, which must meet two conditions: e and $\varphi(n)$ are coprime, and $1 < e < \varphi(n)$. Take (e, n) as public key P.
5. Calculate:

$$d = e^{-1} \bmod (p - 1)(q - 1) \tag{3}$$

and take (d, n) as the private key S and keep it private.
6. The encryption process from plaintext M to ciphertext C is:

$$C = M^e \bmod n \tag{4}$$

7. The decryption process is:

$$M = C^d \bmod n \tag{5}$$

The algorithm flow is shown in Figure 4.

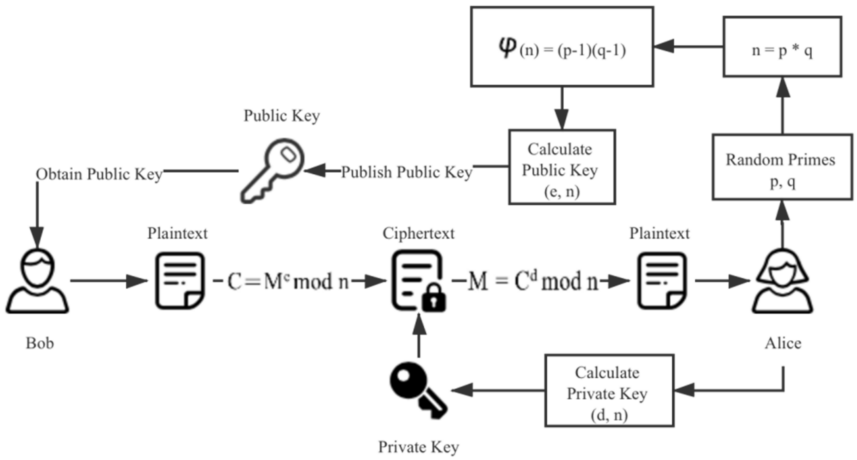

Figure 4. RSA algorithm flow.

4.3. Group Signature

Group signature [31] is a relatively new signature concept proposed by Chaum and van Heyst in 1991. This technology enables any member of a group to sign messages on behalf of the group. Like other digital signatures, the signature can be publicly verified. However, people other than the group manager cannot know which group member signed it, so the privacy of the group members is effectively protected. In this paper, we choose the scheme proposed by [32] because the size of its group public key and group signature are constant, which means that it does not increase with the increase of group members, which helps to reduce the system overhead. The group signature scheme is mainly composed of the following six algorithms:

1. Setup (λ): The initialization algorithm acts as a randomization algorithm with security parameters λ as the input, the generated key sk is used to register group members, the

key ok is used to identify the signer, and the group public key gpk is used to generate and verify group signatures. The key sk and ok are held by the group manager in secret, and the group public key is public to all.

2. Enroll (i, sk): The registration algorithm takes key sk as the input to provide user i with its private signature key gsk[i].
3. Gsig (gpk, gsk[i], M): The signature algorithm uses the key gsk[i] of user i and the group public key gpk to generate a signature σ on message M.
4. GVerify (gpk, M, σ): The verification algorithm is a deterministic algorithm, which takes the group public key gpk and the group signature σ on message M as inputs. It verifies whether the signature was generated by a group member.
5. Open (ok, M, σ): As a deterministic algorithm, the tracking algorithm uses the key ok to cancel the signature σ's anonymity. That is, it can track which group member the signature σ comes from.
6. Rev: When necessary, the group manager can disclose the private key component Xi · Zi of the group members to the blacklist through the revocation algorithm to revoke the permissions of the group members.

The algorithm flow is shown in Figure 5.

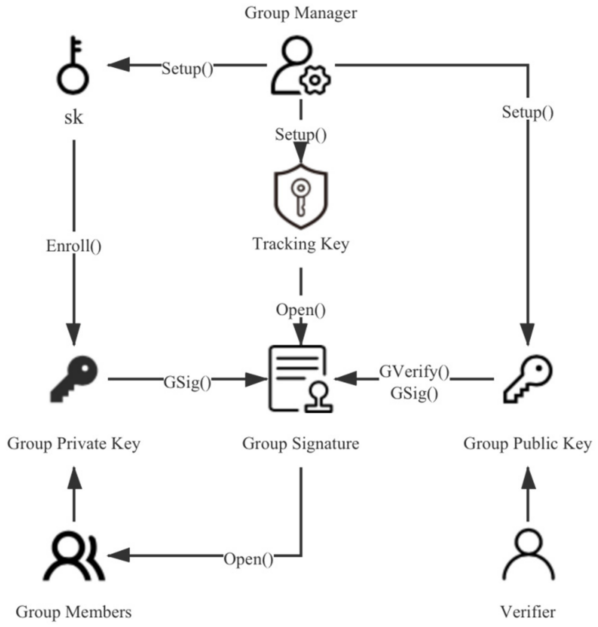

Figure 5. Group signature algorithm flow.

5. System Design

Below we present a series of pseudo code to better describe the system design (Algorithms 1–3).

Algorithm 1 Transaction encryption on HealthChain

```
serial_number = Function <Hash> ([name, ID])
cipher_text = Function <AesEnc> ([M, k])
encrypted_key = Function <RsaEnc> ([k, public_key])
WHILE True (DO)
    IF (cipher_text && encrypted_key) IS NOT NULL THEN
        transaction = Function <Upload> ([serial_number, cipher_text, encrypted_key])
    ELSE
        Function <Sleep> ([t])
    END IF
END WHILE
```

Algorithm 2 Transaction processing on HealthChain

gpk, *sk*, *ok* = **Function** <*Setup*> ([λ])
gsk[i] = **Function** <*Enroll*> ([*i*, *sk*])
WHILE True **(DO)**
    cipher_text = **Function** <*GetCiphertext*> ([*transaction*])
    IF *cipher_text* **IS NOT** NULL **THEN**
        σ = **Function** <*Gsig*> ([*gpk*, *gsk[i]*, *cipher_text*])
        Procedure *Consensus*
    ELSE
        Function <*Sleep*> ([*t*])
    END IF
END WHILE

Algorithm 3 Transaction decryption on HealthChain

transaction = **Function** <*Retrieval*> ([*serial_number*])
aes_key = **Function** <*RsaDec*> ([*encrypted_key*, *private_key*])
plain_text = **Function** <*AesDec*> ([*cipher_text*, *aes_key*])
IF Function <*GVerify*> ([*gpk*, *cipher_text*, σ]) == True **THEN**
    signer_id = **Function** <*Open*> ([*ok*, *cipher_text*, σ])
END IF

Now we introduce our system design in detail.

5.1. System Transaction

In the system proposed in this paper, the transaction content is mainly composed of a serial number, a ciphertext, an encrypted symmetric encryption key, and a group signature. Specifically, it can be expressed as msg = serial_ number||cipher_ text||encrypted_ key||group_ signature. The serial number is the hash value calculated by linking the patient's name and ID number, so it has the characteristics of being difficult to crack and easy to verify. From the perspective of privacy protection, a piece of medical data can be divided into people, time, place, and events. We hide identity information through serial numbers. In addition, the location information (i.e., medical institution) is hidden in the group signature, and only the group manager can know the information. Finally, the visit time, diagnosis content, and other information are encrypted and saved. By anonymizing and splitting a complete piece of medical data, the attacker cannot obtain more sensitive information, even if cracking the ciphertext, which effectively protects the privacy of patients. The transaction content structure is shown in Figure 6.

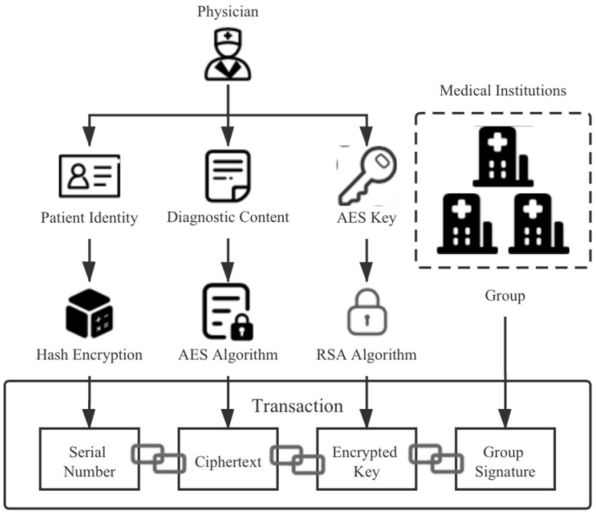

Figure 6. Transaction structure.

5.2. Module Design

The system proposed in this paper mainly includes five modules: transaction encryption, key encryption, transaction signature, chain transaction, and transaction retrieval.

5.2.1. Transaction Encryption

When a diagnostician completes the diagnosis or treatment of a patient, they need to enter the transaction upload page through the system. The physician needs to enter the patient's name and identity card number first. The two will then be linked to create the hash value, which will serve as the patient's unique identity number (i.e., serial number). They will then enter the time of treatment and medical information.

Before each transaction is published, the diagnostician gets an AES key to encrypt the transaction. In addition, in order to improve the security without increasing the additional system burden, the system generates different AES keys ('one key at a time') by generating random numbers every time.

5.2.2. Key Encryption

A number of medical institutions constitute a 'group'. The group manager needs to maintain an additional key pair, in which the public key is public to all for the encryption of the AES key, and the private key is stored by the group manager. After the physician encrypts the transaction, the asymmetrically encrypted public key is used to encrypt the AES key. Compared with the encryption of the AES key by the group members, this method does not have the problem of the secure transmission of unencrypted AES keys, which reduces the system overhead and completely avoids the need for the group members to see the transaction's content. Finally, the diagnostician sends the serial number, ciphertext, and encrypted AES keys to the corresponding group members (i.e., the medical institution). In order to realize the function of a 'one-time key' with a method similar to that of the transaction encryption, but without increasing the burden of the system, the group manager can randomly generate the private key and generate the public key. However, the 'one-day key' method is adopted. At the same time, the group manager maintains a table corresponding to the hash value of a public key and the corresponding private key one by one, which is convenient to retrieve and use when decrypting the AES key.

5.2.3. Transaction Signature

The main feature of group signature technology is that people other than the group manager can only verify whether the signature comes from the group, but not which group member specifically is signing. Therefore, based on this nature, we believe that the ciphertext does not need to include medical institution information, but only the medical time, medical information, and so on. The information of medical institutions can be directly reflected in the group signature. The advantage of doing this is that the attacker cannot obtain the medical institution information either by obtaining the signature or by cracking the ciphertext. This effectively protects the privacy of the medical institutions. At the same time, since the ciphertext only contains treatment time and treatment information without patient identity information, the sensitivity of affairs is also significantly reduced. The group members signed the hash value of the ciphertext, and collected the serial number, signature, ciphertext, and the encrypted AES key together. So far, a transaction collection was completed.

5.2.4. Chain Transaction

After a medical institution packs a certain number of transactions into the block, it will broadcast messages through a simple PBFT (Practical Byzantine Fault Tolerance) consensus algorithm [33]. After being accepted by other nodes, the medical institution will upload the block to the blockchain. The broadcasting process is shown in Figure 7. Obviously, since the consensus of the algorithm is reached through the passing of the message, its

efficiency is affected by the number of nodes. Therefore, in order to ensure the efficiency of the consensus, the number of nodes should not exceed 100.

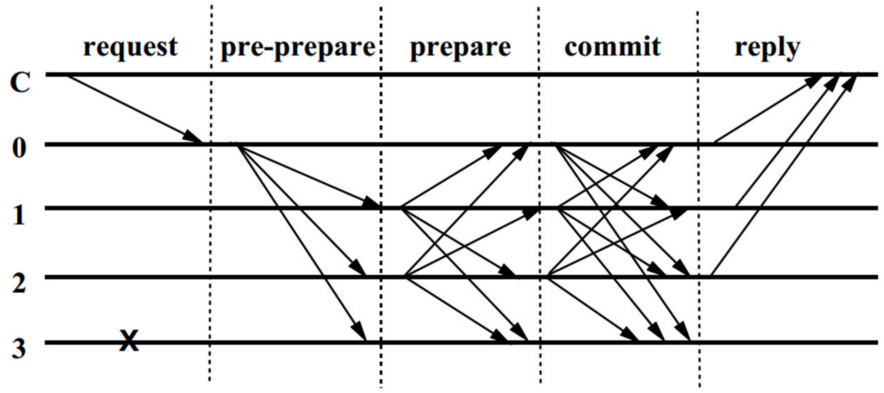

Figure 7. PBFT consensus algorithm broadcast process.

Regarding the choice of the consensus algorithm, we selected the appropriate algorithm by comparing the mainstream consensus algorithms across six aspects. The first is the management of nodes. As a public institution and the joint custodian of medical data, any medical institution in the blockchain network should first obtain system permission to join and exit. Next, because medical data has strong timeliness in many cases, which leads to a large number of transactions per unit time, there are high requirements for transaction delay and throughput. In addition, the cost of system construction and maintenance must also be considered, so the unnecessary waste of resources (such as computing power competition represented by PoW) is unacceptable. With good node management capabilities, the system's requirements for security will be slightly reduced. Moreover, the relatively small number of medical institutions also makes it unnecessary for the system to have strong scalability. The comparison of mainstream consensus algorithms in the above six aspects is shown in Table 1. Among them, DPoS refers to delegated proof of stake. From this, it can be seen intuitively that PBFT is the most suitable consensus algorithm on the whole. At the same time, the practical application represented by Fabric has fully proven its practicability.

Table 1. Comparison of mainstream consensus algorithms.

Characteristic	PoW	PoS	DPoS	PBFT
Nodes Management	No Permission	No Permission	No Permission	Permission Required
Transaction Delay	High	Low	Low	Very Low
Throughput	Low	High	High	High
Energy Saving	No	Yes	Yes	Yes
Security Boundary	1/2 Malicious Computing Power	1/2 Malicious Stakes	1/2 Malicious Stakes	1/3 Malicious Nodes
Scalability	Good	Good	Good	Poor
Typical Application	Bitcoin	Peercoin	BitShares	Fabric

5.2.5. Transaction Retrieval

In some cases, transactions need to be traced back. Similar to the mainstream blockchain system, the system proposed in this paper can query the serial number to carry out efficient data retrieval.

After finding the target transaction, the group manager must first decrypt the AES key by using the asymmetric encryption private key, then decrypt the ciphertext by using

the AES key, and finally decrypt the group signature by using the tracing key to obtain the identity of the signer (i.e., the information of the medical institution). Therefore, in theory, only the group manager can obtain complete transaction information, and if the attacker does not obtain all the keys, he can only obtain some data that do not have analytical value. This greatly protects the privacy of the data.

The overall design of the system is shown in Figure 8.

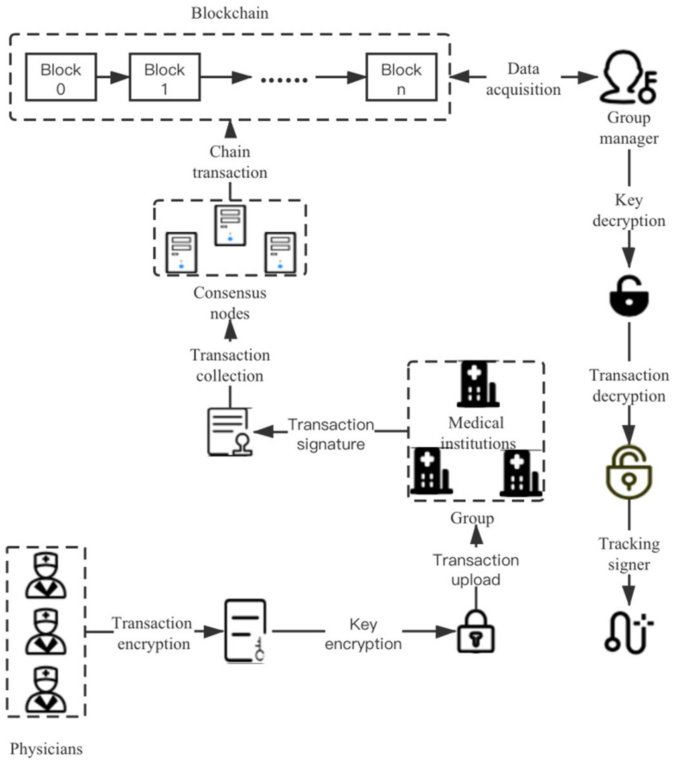

Figure 8. Overall system design.

6. Security Analysis

In this section, we explain how our scheme meets the security requirements mentioned above:

Anonymity: The hash value based on the patient's name and ID number is used as the unique identity of the patient, which is easy to verify, is irreversible, and which realizes the anonymity of the patient in the blockchain. At the same time, the information of medical institutions is separated from the data, which is directly reflected in the digital signature. Due to the characteristics of the group signature, only the group manager can obtain the information regarding the medical institutions from the signature. This not only protects the information of medical institutions, but also further prevents attackers from analyzing patient identity information from such information.

Traceability: The natural traceability of blockchain and the existence of a group manager in the group signature technology enables the system to trace back to specific medical data and their signers when necessary.

Data integrity: Blockchain technology itself has the characteristic of being tamper-proof, and each block is connected in turn. With the hash value of the block, data integrity verification can be easily carried out.

Data confidentiality: As the mainstream symmetric encryption algorithm, the speed and security of the AES algorithm are widely recognized, while key encryption is carried out using an asymmetric encryption algorithm. It will provide good data confidentiality for the system.

Data privacy: By anonymizing the identity and separating the information of medical institutions, only authorized personnel (such as the group manager) can obtain complete

medical information, and when the transaction data is leaked, attackers cannot obtain too much valuable information (only the time and content of medical treatment in the transaction data, while there is no information about the patient and the medical institution). This greatly protects the privacy of the data.

Resistance of birthday collision: Since the system is a consortium blockchain based on the PBFT consensus algorithm, chain transaction and block publishing rely mainly on its primary node, so it avoids the 'soft-fork' of blockchain and can resist birthday collisions.

Resistance of interception: Because each transaction is signed by the group members, and the attacker cannot modify the transaction under the premise of ensuring the validity of the signature, the system can prevent transaction interception and tampering.

Resistance of various attacks:

DDoS attack: Since the upload of transactions and blocks must be completed by trusted nodes, and because the size of the block and the number of transactions can be limited, the system can effectively resist DDoS attack.

Modification attack: Due to the characteristics of group signature and blockchain, any modification of transactions by attackers will be found and rejected.

User impersonation attack: In the consortium blockchain, only verified trusted nodes can enter the blockchain network, so the user impersonation attack is largely alleviated.

7. System Implementation

We implemented a simple consortium blockchain simulation system based on Hyperledger. The system consists of three organizations. Each organization is run by a machine. The system is configured as Ubuntu 16.04 (64-bit), with Intel(R)Core(TM) i7 6700 CPU 3.40 GHZ and 3 GB RAM. There are 10 nodes representing 10 medical institutions, and these nodes also sign transactions as group members. The startup method of the Hyperledger node adopts the virtual container method. The consensus algorithm uses the PBFT algorithm, and another terminal runs as a group manager. The main functions mentioned in the module design are realized in a simple form. In order to facilitate implementation, the environment configuration and some parameters are uniformly given a random value in the same range.

We analyzed the system from two aspects: transaction throughput and scalability. Unlike PoW (proof-of-work), the PBFT consensus algorithm does not need to wait for six blocks for transaction confirmation, and its transaction confirmation delay is only affected by network conditions, so this aspect was not analyzed in great detail.

Transaction throughput is an important index to measure the practical value of the system. The principle of the PBFT consensus algorithm and its wide application in consortium blockchain have fully proved that its throughput when the number of nodes is small is much higher than that of traditional consensus algorithms, such as PoW and PoS (proof-of-stake). At the same time, there is no computing power competition, which avoids the waste of resources and is fairer to each node. In the implementation process, we conducted multiple rounds of transaction throughput tests on a blockchain network composed of 30 nodes, and randomly selected 15 rounds of experimental results. The experimental results are shown in Figure 9.

As can be seen from the figure, since the system adopts the easily implemented PBFT consensus algorithm, there is a certain gap between its throughput and the ideal throughput of the PBFT consensus algorithm, but it is still much higher than traditional consensus algorithms, such as PoW. Therefore, the system can basically meet the throughput requirements of real scenes.

Scalability is a congenital deficiency of the PBFT consensus algorithm, which is caused by the principle of consensus (i.e., message passing). We tested the transaction throughput with the number of system nodes set at 30, 50, 80, and 120, respectively, conducting 15 rounds of tests in each case, and finally calculating the average throughput. The experimental results are shown in Figure 10.

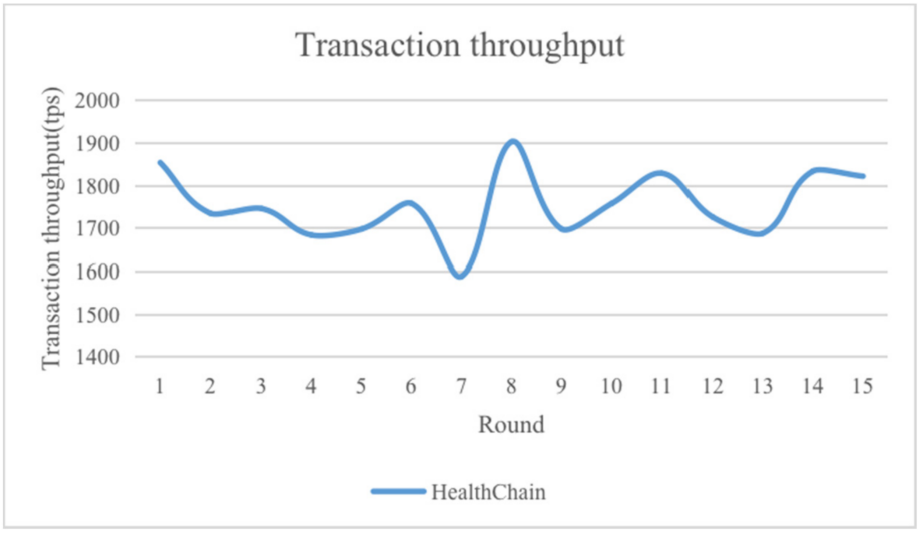

Figure 9. Throughput.

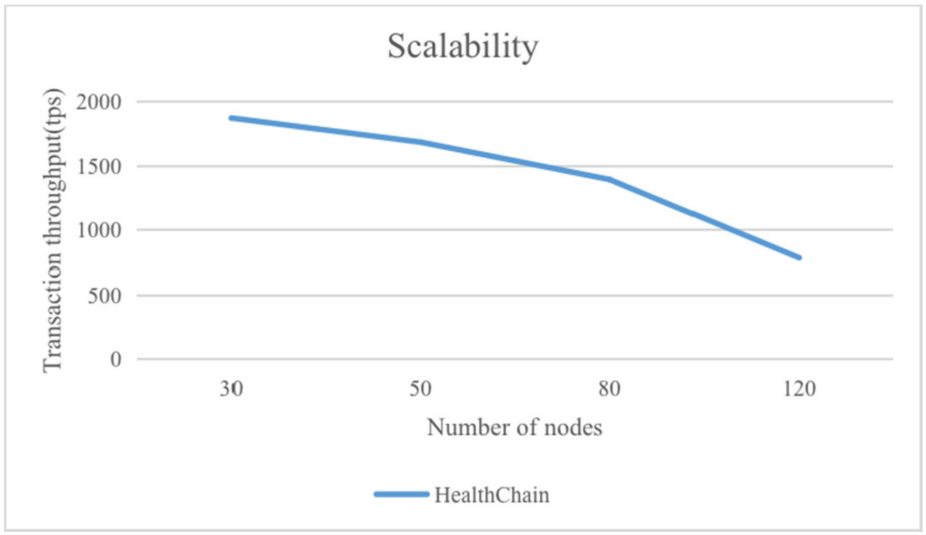

Figure 10. Scalability.

As can be seen from the figure, the throughput of the system will gradually decrease with the increase in the number of nodes. When the number of nodes exceeds 100, the throughput will decrease significantly. This is also basically consistent with the actual situation of the PBFT consensus algorithm. Considering that the nodes of the consortium blockchain constructed in this paper are medical institutions, 100 nodes can basically cover the major hospitals in a small and medium-sized city. Therefore, in practical applications, the number of nodes in the consortium blockchain will not be too large, and the throughput can be guaranteed.

To sum up, the throughput of the system proposed in this paper can fully meet the needs of the actual scenario. However, it should be noted that the system is more suitable for small and medium-sized application scenarios.

8. Discussion

In this paper, we discuss the design and implementation of a medical data privacy protection system based on blockchain and group signature, which can protect the privacy of patients and medical institutions without sacrificing efficiency. In contrast to the existing schemes, we split the transaction information in order to hide it in different ways,

considering that any encryption scheme cannot be absolutely secure. If data leakage is inevitable, the transaction must disclose as little sensitive information as possible. Our method is applicable to scenarios, such as electronic medical records and medical big data. Many existing blockchain platforms also make the implementation of the system relatively simple, greatly reducing the development and operation costs, and are more conducive to popularization. Separate streams of medical data are encrypted with different symmetric keys. Since the encrypted key is also stored in the transaction, there is no additional burden on the system. In addition, there are many options for encryption algorithms, such as the Paillier encryption algorithm [34]. Compared with the RSA algorithm, this algorithm is lighter and less computational overhead. At the same time, its public and private keys are also compatible with the RSA algorithm. Therefore, future research will mainly focus on selecting more secure and efficient data encryption algorithms and more appropriate blockchain consensus algorithms, or will appropriately increase the number of group managers in order to better explore the potential of the system.

Author Contributions: Conceptualization, Z.L.; methodology, Z.L.; software, Z.L.; validation, Z.L.; resources, B.W.; writing—original draft preparation, Z.L.; writing—review and editing, B.W. All authors have read and agreed to the published version of the manuscript.

Funding: Project NO.110052971921/021 Supported by "The Fundamental Research Funds for Beijing Universities".

Data Availability Statement: Not Applicable, the study does not report any data.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Obermeyer, Z.; Emanuel, E.J. Predicting the Future Big Data, Machine Learning, and Clinical Medicine. *N. Engl. J. Med.* **2016**, *375*, 1216–1219. [CrossRef] [PubMed]
2. A Review of Cyber Security Incidents in 2019 (International). Available online: https://www.freebuf.com/articles/network/226830.html (accessed on 10 February 2020).
3. Han, J.H.; Lee, J.Y. Digital Healthcare Industry and Technology Trends. In Proceedings of the 2021 IEEE International Conference on Big Data and Smart Computing (BigComp), Bangkok, Thailand, 17–20 January 2021; pp. 375–377. [CrossRef]
4. Korzun, D.G. *Internet of Things Meets Mobile Health Systems in Smart Spaces: An Overview*; Springer International Publishing: Berlin/Heidelberg, Germany, 2017.
5. D'Antrassi, P.; Prenassi, M.; Rossi, L.; Ferrucci, R.; Barbieri, S.; Priori, A.; Marceglia, S. Personally Collected Health Data for Precision Medicine and Longitudinal Research. *Front. Med.* **2019**, *6*, 125. [CrossRef] [PubMed]
6. Aich, S.; Sinai, N.K.; Kumar, S.; Ali, M.; Choi, Y.R.; Joo, M.-I.; Kim, H.-C. Protecting Personal Healthcare Record Using Blockchain & Federated Learning Technologies. In Proceedings of the 2021 23rd International Conference on Advanced Communication Technology (ICACT), online, 7–10 February 2021; pp. 109–112. [CrossRef]
7. Xiong, P.; Zhu, T.; Wang, X. Differential privacy Protection and application. *J. Comput. Sci.* **2014**, *37*, 101–122.
8. Sweeney, L. k-Anonymity: A Model for Protecting Privacy. *Int. J. Uncertainly Fuzziness Knowl. Based Syst.* **2002**, *10*, 557–570. [CrossRef]
9. Machanavajjhala, A.; Kifer, D.; Gehrke, J.; Venkitasubramaniam, M. L-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data* **2007**, *1*, 3. [CrossRef]
10. Li, N.; Li, T.; Venkatasubramanian, S. t-Closeness: Privacy beyond k-Anonymity and l-Diversity. In Proceedings of the ICDE 2007 IEEE 23rd International Conference on Data Engineering, Istanbul, Turkey, 11–15 April 2007.
11. Song, F.; Ma, T.; Tian, Y.; Al-Rodhaan, M. A New Method of Privacy Protection: Random k-Anonymous. *IEEE Access* **2019**, *7*, 75434–75445. [CrossRef]
12. Dwork, C. Differential Privacy. In *Proceedings of the 33rd International Conference on Automata, Languages and Programming-Volume Part II*; Springer: Berlin/Heidelberg, Germany, 2006.
13. Thiranant, N.; Sain, M.; Lee, H.J. A design of security framework for data privacy in e-health system using web service. In Proceedings of the International Conference on Advanced Communication Technology, PyeongChang, Korea, 16–19 February 2014.
14. Ilokah, M.; Eklund, J.M. A Secure Privacy Preserving Cloud-based Framework for Sharing Electronic Health Data. In Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) in conjunction with the 43rd Annual Conference of the Canadian Medical and Biological Engineering Society, Montréal, QC, Canada, 20–24 July 2020.
15. Narayan, S.; Martin, G.; Safavi-Naini, R. Privacy preserving EHR system using attribute-based infrastructure. In Proceedings of the Acm Cloud Computing Security Workshop, DBLP, Chicago, IL, USA, 8 October 2010; p. 47.

16. Choe, J.; Yoo, S.K. Web-based secure access from multiple patient repositories. *Int. J. Med Inform.* **2008**, *77*, 242–248. [CrossRef] [PubMed]
17. Yang, Y.; Zheng, X.; Liu, X.; Chang, V. Cross-domain dynamic anonymous authenticated group key management with symptom-matching for e-health social system. *Future Gener. Comput. Syst.* **2017**, *84*, S0167739X1730554X. [CrossRef]
18. Wang, C.; Wang, Q.; Ren, K.; Lou, W. Privacy-Preserving Public Auditing for Data Storage Security in Cloud Computing. In Proceedings of the 2010 Proceedings IEEE Infocom, San Diego, CA, USA, 14–19 March 2010.
19. Shang, T.; Zhang, F.; Chen, X.; Liu, J.; Lu, X. *Identity-Based Dynamic Data Auditing for Big Data Storage*; IEEE: Piscataway, NJ, USA, 2019.
20. Gope, P.; Amin, R. A Novel Reference Security Model with the Situation Based Access Policy for Accessing EPHR Data. *J. Med Syst.* **2016**, *40*, 242. [CrossRef]
21. Hossein, K.M.; Esmaeili, M.E.; Dargahi, T.; Khonsari, A. Blockchain-Based Privacy-Preserving Healthcare Architecture. In Proceedings of the 2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE), Edmonton, AB, Canada, 5–8 May 2019.
22. Alshalali, T.; Mbale, K.; Josyula, D. Security and Privacy of Electronic Health Records Sharing Using Hyperledger Fabric. In Proceedings of the 2018 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 13–15 December 2018.
23. Xu, J.; Xue, K.; Li, S.; Tian, H.; Hong, J.; Hong, P.; Yu, N. Healthchain: A Blockchain-Based Privacy Preserving Scheme for Large-Scale Health Data. *IEEE Internet Things J.* **2019**, *6*, 8770–8781. [CrossRef]
24. Nakamoto, S. Bitcoin: A Peer-to-Peer Electronic Cash System. Technical Report. Available online: http://bitcoin.org/bitcoin.pdf (accessed on 21 October 2019).
25. Abdellatif, A.A.; Samara, L.; Mohamed, A.; Erbad, A. MEdge-Chain: Leveraging Edge Computing and Blockchain for Efficient Medical Data Exchange. *IEEE Internet Things J.* **2021**, 1. [CrossRef]
26. Aileni, R.M.; Suciu, G. IoMT: A blockchain perspective. In *Decentralised Internet of Things*; Springer: Cham, Switzerland, 2020; pp. 199–215.
27. Zhang, H.; Li, G.; Zhang, Y.; Gai, K.; Qiu, M. Blockchain-Based Privacy-Preserving Medical Data Sharing Scheme Using Federated Learning. In Proceedings of the International Conference on Knowledge Science, Engineering and Management, Tokyo, Japan, 14–16 August 2021; Springer: Cham, Switzerland, 2021.
28. Daemen, J.; Rijmen, V. *The Design of Rijndael: AES-The Advanced Encryption Standard*; Springer: Berlin/Heidelberg, Germany, 2002.
29. Standards, N.B.O. Data Encryption Standard; Federal Information Processing Standards Publications; 1977. Available online: https://csrc.nist.gov/CSRC/media/Publications/fips/46/archive/1977-01-15/documents/NBS.FIPS.46.pdf (accessed on 23 September 2021).
30. Rivest, R.L.; Shamir, A.; Adleman, L. A method for obtaining digital signatures and public-key cryptosystems. *Commun. ACM* **1978**, *21*, 120–126. [CrossRef]
31. Chaum, D.; van Heyst, E. Group Signatures. In *Workshop on Advances in Cryptology-Eurocrypt*; Springer: Berlin/Heidelberg, Germany, 1991.
32. Ho, T.-H.; Yen, L.-H.; Tseng, C.-C. Simple-Yet-Efficient Construction and Revocation of Group Signatures. *Int. J. Found. Comput. Sci.* **2015**, *26*, 611–624. [CrossRef]
33. Miguel, C.; Barbara, L. Practical byzantine fault tolerance and proactive recovery. *ACM Trans. Comput. Syst.* **2002**, *20*, 398–461.
34. Paillier, P. Public-key cryptosystems based on composite degree residuosity classes. *Adv. Cryptol. Leurocrypt* **2004**, *1592*, 223–238.