2017 IEEE 37th International Conference on Distributed Computing Systems

On the Design of a Blockchain Platform for Clinical Trial and Precision Medicine

Zonyin Shae
Department of Computer Science and Information Engineering
Asia University
Taichung, Taiwan
zshae1@gmail.com

Jeffrey J.P. Tsai
Department of Bioinformatics and Medical Engineering
Asia University
Taichung, Taiwan
jjptsai@gmail.com

Abstract—This paper proposes a blockchain platform architecture for clinical trial and precision medicine and discusses various design aspects and provides some insights in the technology requirements and challenges. We identify 4 new system architecture components that are required to be built on top of traditional blockchain and discuss their technology challenges in our blockchain platform: (a) a new blockchain based general distributed and parallel computing paradigm component to devise and study parallel computing methodology for big data analytics, (b) blockchain application data management component for data integrity, big data integration, and integrating disparity of medical related data, (c) verifiable anonymous identity management component for identity privacy for both person and Internet of Things (IoT) devices and secure data access to make possible of the patient centric medicine, and (d) trust data sharing management component to enable a trust medical data ecosystem for collaborative research.

Keywords—*blockchain; clinical trial; precision medicine; data integrity; data integration; identity privacy; big data analytics; distributed and parallel computing; patient centric medicine; IoT;*

I. INTRODUCTION

This paper described briefly the scope, approaches, challenges, and system design concept of a blockchain platform for the medical domain, particularly for clinical trial and precision medicine. This project consisting of more than 30 researchers is currently undergoing at the Asia University, Taiwan. Various technology challenges and possible research areas and approaches will be highlighted. The methodologies and approaches to build clinical trial and precision medicine applications on top of the proposed blockchain platform will be also described. Our goals are to assist researchers in addressing clinical and precision medical issues and assisting in medical decision-making research. Exploring and integrating various data sets of disease, drug and clinical practice can reduce healthcare costs and provide better preventive, curative and post-treatment care in healthcare.

Blockchain [10, 12, 32-40] is a distributed and parallel computing mechanism with a distributed ledge to provide trust transaction across the Internet where trust is established through mass peer to peer distributed collaboration and smart contract code rather than through a central powerful institution for trustful transaction settlement. Once a transaction has been recorded in the blockchain distributed ledger, it is not changeable and not deniable. Smart contract [12] can be used to store various value digital assets into blockchain to claim the ownership of the asset. The asset is managed by the smart contract which is executed automatically by the program code. The smart contract code defines the rules and conditions to manage and trigger the action of the asset ownership.

Although Blockchain is much more investigated within the financial sector, it is gradually edging into other industries such as healthcare. For examples, HealthNautica [13] collaborates with blcokchain company Factom [14] leveraging bitcoin blockchain technology to securely store medical records and provide efficient audit trails. Blockchain company Gem [15] collaborates with Capital One to reduce long process time in healthcare insurance claim process. DNA.Bits [16] is planning to store genetic and medical data into blockchain as a personal "DNA wallet". The idea is to allow the genetic information and medical data accessible to the targeted researchers while maintain data security and privacy. BlockVerify [17] uses blockchain to fight counterfeit drugs via securely attaching a unique verification tag on drug packages which can be scratched off to verify the drug legitimacy against with blockchain.

There are also blockchain use cases in the area of distributed and parallel computing, for examples, FoldingCoin [18] from Stanford University, and GridCoin [20] from UC Berkeley. FoldingCoin created a "Proof of Fold" and GridCoin created a "proof of Research" concept to verify contributed computational power of each participated blockchain node. Participants contribute their computing cycles to medical research on the blockchain network instead of a "Proof of Work" or "Proof of Stake" tasks on a traditional blockchain. Berkeley built a special blockchain to run GridCoin, and Stanford run its FoldingCoin on Folding@home [19] platform which run on top of the existed bitcoin network with over hundreds of thousand participant bitcoin nodes with combined computing power around 80-100 petaFLOPS for the protein folding computing study. Both

This paper is supported from Ministry of Science and Technology, R.O.C, under the grant MOST105- 2632-E-468-002

1063-6927/17 $31.00 © 2017 IEEE
DOI 10.1109/ICDCS.2017.61

FoldingCoin and GridCoin are based on the similar blockchain based grid computing paradigm. We will investigate a more general distributed and parallel big data analytics mechanisms via blockchain network.

There are currently a hands full of blockchain networks with various protocols, distributed ledgers, and smart contracts. For examples, Bitcoin [21], Ethereum [22], Hyperledger [23], NASDAQ Linq [24], Ripple [25], Chain.com [26], etc. All these blockchain networks provide a similar horizontal technology with trust transactions in a virtual world that can be applied across various industries. However, blockchain is only a single critical system component, in order to make the entire system functional, we need to connect and integrate the blockchain trust mechanism with physical world applications and systems. Such that it is compelled to have a blockchain platform to allow users to build their own blockchain applications and systems comprehensively.

This paper discusses briefly the various aspects and provides some insights on the conceptual design of a blockchain platform which addresses the technology requirements of big data integration, data integrity, data storage, identity privacy, data access security, trust data sharing and collaboration, IoT integration, big data analytics, and general distributed and paralleling computing. Each application domain is expected to have their own unique situations and requirements, we will investigate the design of the blockchain platform starting from the medical domain, particularly the clinical trial and precision medicine. The healthcare industry looks for ways to manage massive amounts of disparity medical records in providing better health solutions, reduce insurance claim process time, and protect the data against cybercrime. Leveraging blockchain technology to build a secure data sharing and trust collaboration ecosystem is an appealing approach.

It would be very helpful in providing high quality personal healthcare if we can securely capture and store the full history of an individual's health and every doctor's visit, etc, on blockchain in which every data event is time-stamped and can't be tampered with. It is even better that each individual medical data can be in the control of individual who can access their records across healthcare providers and decide which physicians can see which records. With the diversity of captured medical related data is increasing and so with the increasing of concern of privacy and data access control security. In the premise of safeguarding patients' privacy, the technology of medical data capture for analysis is currently facing bottlenecks. Block chain technology, providing security, privacy, distributed and parallel computing and backup characteristics, may be used as effective data capture, storage and analysis mechanism to develop more efficient data computing model.

This paper will discuss Anonymous Identities Authentication [27] mechanisms based on Zero Knowledge Technology [28] to address the identity privacy issue. The purpose of the study was to hide the identity of the patient on the medical blockchain, but the legitimacy of the patient's identity can be systematically verified. The mechanisms can be applied in the IoT system to hide the IoT device identity, but can verify the legitimacy of the identity of the device. We can set different permissions for patient data to determine who can access the patient's data. For the IoT environment, the IoT device can be set to allow which applications can access the device sensor data.

In the field of scientific research, evidence suggests that trust is reduced because data is often manipulated. We will explore and develop an automated blockchain mechanism to prevent "hidden switching" of clinical trials, allow researchers to quickly verify the data integrity of results reported in medical journals. We will also explore in depth the use of blockchain smart contracts to make medical data more secure and private and facilitate the team data sharing and collaboration mechanisms. When data is trusted and protected, collaboration takes off. The value of sharable and secure electronic health records (EHRs) is easily apparent. Based on IBM healthcare for blockchain study report [48], according to the Premier healthcare alliance, sharing data across organizations could save hospitals USD 93 billion over five years in the U.S. alone.

This paper will discuss techniques for data sharing and exchange on blockchain. Various nodes on the blockchain can be grouped into groups. Only the nodes in the authorized group can access the user data through the user's authority setting. Moreover, discuss the mechanism to enable the exchange of information between different groups (such as EHR need to be exchanged between different groups).

This study plans to combine the blockchain technology and big data technology, integrating the clinical data and health information records from both hospitals of Chinese Medicine University Hospital and Asia University Hospital at Taiwan, and the Taiwan National Health Insurance Database Medical data. Blcokchain technology for the storage and management of medical data will make medical data more secure and private. The resulted trust in the system will be helpful in building a medical data sharing ecosystem; and the development of big data analysis technology will provide people with more accurate prevention of disease.

Our contributions in this paper are: (1) proposes a blockchain platform, built on top of the traditional blockchain network for leveraging its major components to achieve trust transaction properties for clinical trial and precision medicine and discusses various design aspects and provides some insights in the technology requirements and challenges. (2) proposes a blockchain platform architecture and identify 4 new system components and discuss their technology challenges: (a) a new blockchain based general distributed and parallel computing paradigm component to devise and study parallel computing big data analytics, (b) blockchain application data management component for data integrity, big data integration, and integrating disparity of medical related data, (c) verifiable anonymous identity management component for identity privacy including IoT and secure data access to make possible of the patient centric medicine, and (d) trust data sharing management component to enable a trust medical data ecosystem for collaborative research. (3) discussed the methodologies and approaches of clinical trial and precision medicine as two blockchain platform use cases.

The organization of the rest of this paper is following: the blockchain platform architecture will be described in section 2, and its customization for precision medicine will be discussed in section 3, in which the disparity data integration, data analysis acceleration, and general parallel computing via blockchain will be also discussed. Blockchain platform for clinical trial will be provided in section 4, in which, data integrity mechanisms and data sharing topic will be discussed. The identity privacy and verifiable anonymous identity including IoT will be addressed in section 5, and finally, a brief summary will be provided in section 6.

II. Blockchain Platform Architecture

The system architecture of a blockchain platform is shown in Figure 1. Our blockchain platform will be built on top of the traditional blockchain network for leveraging its major components to achieve trust transaction properties. We identify 4 system components in our platform: (a) a new blockchain based general distributed and parallel computing paradigm, (b) blockchain application data management, (c) verifiable anonymous identity management and secure data access, (d) trust data sharing management.

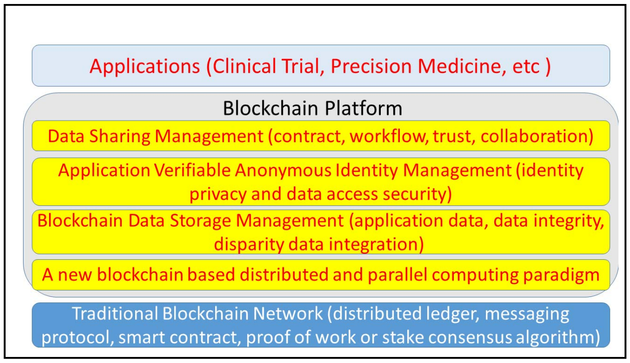
Applications (Clinical Trial, Precision Medicine, etc)
Blockchain Platform
Data Sharing Management (contract, workflow, trust, collaboration)
Application Verifiable Anonymous Identity Management (identity privacy and data access security)
Blockchain Data Storage Management (application data, data integrity, disparity data integration)
A new blockchain based distributed and parallel computing paradigm
Traditional Blockchain Network (distributed ledger, messaging protocol, smart contract, proof of work or stake consensus algorithm)

Figure 1. Blockchain Platform Architecture

The new blockchain parallel computing component will explore a new blockchain based general distributed and parallel computing paradigm component to devise and study parallel computing methodology for big data analytics.

Recently Hadoop computing [29], Grid computing [30], and Cloud computing [31] are among the most popular parallel computing paradigms. Hadoop computing paradigm is a centralized architecture. Each node requires high performance CPU and memory. The framework also requires a very high communication bandwidth between each computing node pair for intermediated data communication during the computing circle. Grid computing paradigm is a distributed and parallel architecture. Basically every node can ask to join the grid computing network and contribute its unused computing power for the aggregated computing. Cloud computing paradigm is a centralized computing architecture, in which the computing resources can be virtualized into virtual machines for parallel but individual computing model. The cloud computing resources featuring with the elasticity property. There is hybrid computing paradigm [41] in combing the cloud elasticity property into the grid computing using super computer grid.

The FoldingCoin and GridCoin belongs to the blockchain based grid computing paradigm. Both approaches make use of only the large aggregated computing power of the blockchain nodes. However, they did not leverage on the large aggregated communication bandwidth available in the blockchain network. Since there is no built in communication tools among each of the divided sub-tasks, the task partition model in this parallel computing paradigm can be limited.

It is often necessary to know the distribution function for statistical calculation in deriving the inference. If the distribution function is unknown, the distribution of the samples can be generated using permutation. If the number of the sample is large, random sample permutation is a very time consuming task. For example, the independent sample t-test is a commonly used statistical method to test whether the expected values of the two population characteristics are equal. When we want to compare whether there is a difference in the mean of the two groups, we can use the t-test for analysis. In the independent sample t-test, it is a comparison of whether the mean values obtained by the dependent variable of the different attributes (different groups) within the independent variable are significantly different. This distribution can be generated through a random reordering process. We will investigate the mechanism to leverage blockchain for generating the random sample permutation for big data sets.

We do believe that a new blcokchain based distributed and parallel computing paradigm can exist via exploring the leverage of both huge aggregated computing power and communication bandwidth of a blockchain network. This new blockchain distributed and parallel computing paradigm will leverage the blockchain parallel communication protocol and should be able to effectively support a general parallel computing tasks for big data analysis. We will explore this idea.

The blockchain application data management component is used to provide mechanism to achieve peer verifiable data integrity, as well as provide the mechanism for the disparity data integration. We make use of blockchain distributed ledger and smart contract to provide data integrity. Some insights along with use case will be discussed in the later section of this paper.

The verifiable anonymous identity management and secure access component will address the issue of identity privacy and data security for both person as well as the IoT devices. Although the user identity in tradition Blockchain is an anonymous hash value of user's public key. However, it was reported about 60% of the user Identity in the traditional blockchain, their real identities had been identified via the big data analysis across various data sets available in the Internet. On one hand, it is required for the users to maintain its own identity anonymous, on the other hand, in some applications, for examples, banking, it is absolutely required that user's real identity need to be verified. These are two contradict requirements, and this component will address this issues.

The trust data sharing management component is to provide a trustful data sharing environment. We will make use of blockchain smart contract to enforce the secure data sharing and its workflow. Once the trust can be accomplished, the data collaboration will come. This is a path to build a data sharing ecosystem and enable the large scale collaborative big data analysis.

III. Blockchain Platform for Precision Medicine

A. *A Precision Medicine Case Study*

Within the developed countries of the world, cerebrovascular diseases are among the top ten causes of death, and the main cause of the survivors' disability. For nearly two decades in Taiwan, cerebrovascular diseases are among the top ten causes of death. Acute and chronic cerebrovascular disease medical care has become a major burden of social society [42-47]. Although the understood relevant risk factors, especially the prevention and treatment of hypertension, as well as the progress of medical treatment of cerebral stroke, cerebrovascular disease mortality rate decreased only slightly. The high mortality and disability rate still remains as difficult issues of public health.

The Taiwan insurance coverage rate is almost 100%, and the project covers hospitalization, emergency, and out-patient. This database can faithfully record the patient's medical treatment process, including diagnosis, disposal, drugs and so on. However, the Taiwan health insurance database alone is not sufficient to have an in depth analysis of the stoke chronic illness since the stroke is a fairly diverse disease.

In order to have an in depth research into the stroke illness and its prevention and management, we need to integrate more data sets. Not only to take into account the past history of epidemiology of disease factors, but also began to consider the patient's mental state, living habits, income, environmental factors, regional and other more extensive information. Some large hospitals try to build a personalized system for rehabilitation planning, and to study the relationship between environmental factors and rehabilitation effects, such as the rehabilitation process of listening to music [49].

It would be helpful to investigate the risk factors of stroke at the genetic level, for examples, genetic risk factors, stroke prediction algorithm based on genomic data, and explore the effects of genetic factors on stroke efficacy influence after treatment. Moreover, to investigate the use of protein drugs and miRNA [50] drugs to assist the rehabilitation of stroke after electrotherapy.

Genetic factors, such as gene expression, SNP [51], and miRNA, are the most popular genome research projects. In general, the study of stroke need to integrate a large disparity sets of big data such as macrocosmic epidemiology and microscopic genomics. The study of stroke prediction, stroke treatment and rehabilitation after stroke has been towards the personalization of stroke treatment, and the establishment of personalized medical care.

B. *Blockchain Platform for Precision Medicine*

Integrating clinical stroke data sets is one of the major challenges in the development of stroke prevention and investigating the relationship between new drugs for stroke treatment and cancers and other diseases. We will develop a medical blockchain to store the Stroke Clinic Medical Data Library data set from Chinese Medical University Hospital (CMUH) Taiwan and the Taiwan Health Insurance Database data set, to investigate particularly these 2 technical challenges: (a) blockchain application data management component for integrating disparity of medical related data, (b) a new blockchain based general distributed and parallel computing paradigm component to devise and study parallel computing methodology for big data analytics,

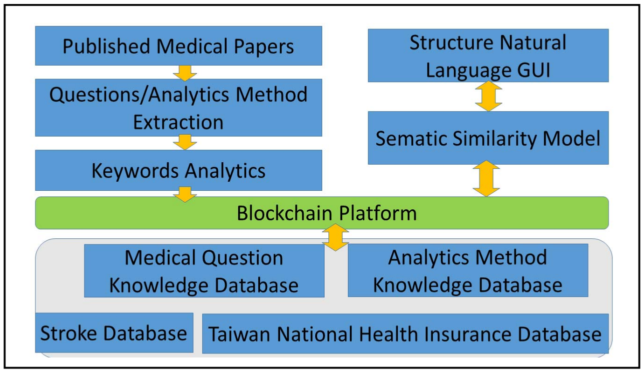
Published Medical Papers
Questions/Analytics Method Extraction
Keywords Analytics
Structure Natural Language GUI
Sematic Similarity Model
Blockchain Platform
Medical Question Knowledge Database
Analytics Method Knowledge Database
Stroke Database
Taiwan National Health Insurance Database

Figure 2. Blockchain Platform for Precision Medicine

This is to establish a precision medicine application use case with blockchain platform to provide more accurate prevention, treatment and care for stroke illness. The system architecture is shown in Figure 2.

NCBI PubMed [52] is a valuable database of biomedical research, currently contains about 24 million articles. Initially, we use the NCBI PubMed Biomedical Literature Library as a source of literature, apply semantic computation and text exploration techniques, analyze semantic similarity in the literature, and then use the implicit semantic model to group analysis to generate health knowledge base. Two health knowledge data bases will be generated via this literature analytics process: one is the medical question database and the other is analytics method knowledge database. Medical question database records the medical question under investigation and the analytics knowledge database records the results and the methodology/approach/analytics tools used to achieve the results. These two databases will be also under the integration and management of blockchain to ensure its secure data access. We will develop a user interface using structural natural language query, and apply semantic similarity model to analyze semantic similarity between the structural natural language query and meta data created for the problem knowledge data base and analysis method knowledge database to obtain accurate answers and analytical methods.

Such that in this use case, blockchain will manage and integrate 4 data sets: two are form the medical practice (Stroke Clinic Medical Data Library data set from CMUH and the Taiwan Health Insurance Database data set) and two are from the literature analytics (medical question database and analytics knowledge database). Note that these 4 datasets all have its own different data structure relationship, data access security policy, read/write throughput, and real time/off line processing requirements and properties. These are the interesting variables that we would investigate and explore the mechanisms how blockchain platform effectively store and manage data sets.

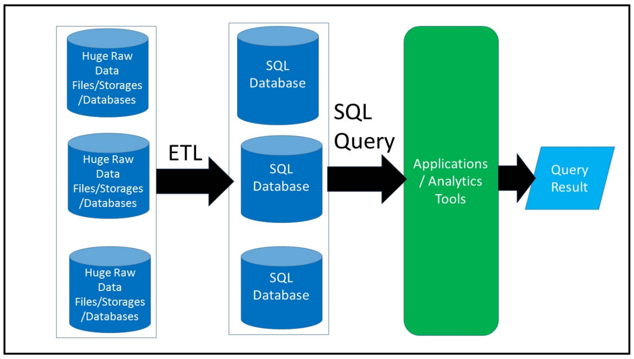
Huge Raw Data Files/Storages /Databases
Huge Raw Data Files/Storages /Databases
Huge Raw Data Files/Storages /Databases
ETL
SQL Database
SQL Database
SQL Database
SQL Query
Applications / Analytics Tools
Query Result

Figure3. Traditional Medical Data Analytics Model

C. *Integration Complexity of Disparity Medical Big Data Sets*

Usually we qualify a big data set with 4V properties: volume, variety, veracity (refer to the trustworthy of data), and velocity. Medical data is definitely having all of these 4V properties. Moreover, the value (the 5th V) of medical data is self-explained since it can save life. Big data (with the amount of data, trustworthy of data, frequency of data, data complexity and data structure) presents challenges to the traditional database system and computing system which cannot effectively store, calculate, process, and analysis to transform and interpret the data into knowledge.

The Taiwan national health insurance data structure of health care database, is a structured data format. However, the hospital treatment records consist of structured information, semi-structured electronic medical records (EMR) and unstructured (nuclear resonance imaging and computer tomography) data format. The integration of Taiwan national health insurance health-care databases with hospital records is very important to provide a full scope analysis of personal disease prevention and health treatment.

With the development of medical technology and the rapid adoption of personal healthcare related wearable IoT devices, the diversity of medical related data is increasing along with the increasing concern of privacy. The computing and storage technology to capture and integrate these disparity of medical related data with safeguarding patients' privacy and data security for efficient analysis is facing bottlenecks.

During the analysis for a given research medical question, large amount of disparity medical data need to be sampled filtering and aggregated before presenting to the analysis tools. Ideally researchers can write their own analytics program codes using arbitrary programming languages with arbitrary underline data sets of arbitrary data structures. Moreover, it would be more effective if the analytics code can be written to capable run in parallel computing. However, this will take a lot of cross domain knowledge training. This is especially true in the bio-medical area. As a result, open sources or commercial available analytics tools which are not executing in parallel are used in most of the studies. Since most of the most analysis tools (e.g. SAS) need a SQL like structure database as default data inputs, in general, there is a need to transform the filtered data into the SQL like data structure regardless of the underline raw data sources, see Figure 3.

Traditionally, this will need to create an individual data ETL (extraction, transfer, and load) for each SQL database for each individual medical research question. Most of the cases, this is formidable efforts with extremely expensive cost due to its big data set nature and to meet the security compliant requirements of medical data sets.

We will investigate a virtual mapping data analytics model to practically remove the ETL operation, see Figure 4. For each medical research question study, we provide virtual SQL database in which only the schema is logically defined per researcher's requested specification. There is no real data has been copied and stored there. The original medical raw data will be stored at its original location to fulfill HIPPA [53] requirements. The virtual SQL data base will store meta mapping to link the logical schema to the physical medical data. Such that researchers can modify the schema any time and the virtual SQL can be available immediately after schema modifications. The analytics tools will not tell any difference whether it is running on a virtual SQL data base or on a real one. So the analytics application codes can run as is without any modification or re-writing. This is an important and very useful feature since researchers usually need to modify the schema so many times during their study process that cause huge pain point for IT team. Moreover, the SQL queries can now be executed in parallel when it has been deployed in the Hadoop environment. For example, Hive which has already supports virtual mapping SQL for HBase.

We will investigate the mechanism to integrate Hadoop infrastructure into blockchain platform to provide data privacy and security in the virtual mapping data analytics model.

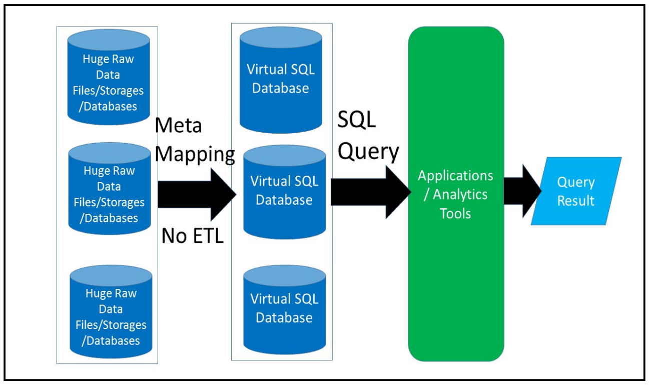
Huge Raw Data Files/Storages /Databases
Huge Raw Data Files/Storages /Databases
Huge Raw Data Files/Storages /Databases
Meta Mapping
No ETL
Virtual SQL Database
Virtual SQL Database
Virtual SQL Database
SQL Query
Applications / Analytics Tools
Query Result

Figure 4. Virtual Mapping Data Analytics Model

IV. Blockchain for Clinical Trial

A. Data Integrity Issue in Clinical Trial

In the field of scientific research, evidence suggests that trust is reduced because data are often manipulated. The pharmaceutical industry is eager to get a new drug market, and developers are also keen to test the positive results. Clinical trials not only need to monitor the effectiveness of new drugs, but also check whether there are any side effects, weigh the pros and cons of the two. Since 2007, US drug regulators have asked all clinical trials that test recruited subjects [2-3, 7-8] must be registered in the publicly accessible database ClinicalTrials.gov [9]. However, despite mandates for open access to the protocols and data captured in clinical trials, the issue of clinical trial data integrity remains [3-6]. According to COMPare[4], a recent project to monitor clinical trials, just nine in 67 trials it studied (13 percent) had reported results correctly. There are near 20000 registrations per year at ClinicalTrails.gov, COMPare project finding show only the tip of the iceberg of the data integrity problem.

When clinical trials are error-prone or manipulated, patient care is diminished and the consequences are serious for all. Blockchain techniques can increase trust and reduce suspicion of reporting data. Moreover, as clinical trials become more reliable and transparent, it will improve the accuracy of big data analyzes using clinical trials and save more lives. Carlisle [1] was first published in a clinical trial protocol using blockchain [10]. A recent study by Greg Irving and John Holden [2] showed that using the blockchain is a low-cost independent verification method for verifying the report data integrity of scientific research.

Blockchains that capture the lifetime history of clinical trials as they unfold could go a long way to restoring trust in science. Notably, the protocols of a trial, recorded and time-stamped as they are developed, could expose the practice of “outcome switching,” which increases the chance that the data being reported is random noise rather than a real result.

Transparency to all data could dissuade those who attempt to selectively report only good outcomes. Moreover, sometimes it is important to keep the clinical trial protocol secrete since it might contain research and commercial secrets. Blockchain could assure the trial data is recorded in realtime. The data integrity can then be verified after without exposing trial protocol secrets to competitors before the public release.

A blockchain recorded clinical trial data set is useful in the medical research. Since the test subjects in the clinical trial is very limited and most of the time do not have an appropriate population representation in the test set, the possible disease treatment and the side effects might not have been completely discovered in the trial. The trust trial data can then be integrated with the patient outcome data set after the drug has been approved. The integrated before and after data sets can be used to investigate the real and long term effect of the dug. Clinical trial records on blockchain would provide a good resource of trusted medical data set.

B. Blockchain for Peer Verifiable Clinical Trials

We need to explore and establish a mechanism to prevent the "hidden outcome switch" for clinical trials. We believe that blockchain techniques are a low-cost, stand-alone, and widely used tool for auditing and validating the reliability of scientific research. Allowing researchers quickly test the correct results reported in medical journals. Ultimately, the method can also be automated. Make it possible for the peer verifiable clinical trials.

We will explore in depth the use of blockchain smart contracts to explore the clinical trial process and the team's data sharing and collaboration mechanisms. This study will make full use of the smart contract to investigate and establish a blockchain platform that can work with teams to share data and collaborate on clinical trial workflows.

In order to be able to publicly share and collaborate on data, there must be a mechanism to record and enforce ownership of the data. If someone else later use data, they can either credit the data to the owner or the owner can explore monetization.

This will create a healthy data ecosystem that the whole community can benefit from. We will investigate various solutions to data ownership issues. When data is trusted and privacy is protected, shared and collaboration takes off.

To prevent pharmaceutical companies and researchers from fabricating data, Greg Irving, a researcher and medical doctor at the University of Cambridge, created a blockchain based system to document and independently validate those ongoing clinical trials. The study has been published by Greg Irvin and co-author John Holden [2] in the F1000 study. Greg Irving created a Proof of Concept (POC) with a research protocol on the ClinicalTrials.gov website to ensure that pharmaceutical companies or researchers will not deviate from the actual trial after initiating clinical trials and not modify previously approved Agreement to tie the test results.

The following is the method used by Greg Irving:

1. Prepare clinical trial raw file contain protocol and all prospective plan analysis files. Use a non-proprietary document format (such as an unformatted text file or LaTeX format).
2. Calculate the document's SHA256 hash value and convert it to a bitcoin key.
3. Import the key into a bitcoin wallet and create a transaction to its corresponding public address.

According to the same step, a public key can be generated for any unformatted text file containing a clinical trial protocol. If the newly generate public key matches the one in the blockchain, it not only proves the existence of the file with the timestamp, but also verifies that the document has not been altered in any way. Because in any case, the created SHA256 hash value will be different from the original, resulting in a different public key.

C. Blcockchain Platform for Clinical Trial

Data integrity and trust data sharing is two of the main challenges in the development of clinical trials. We will develop a blockchain platform to investigate (a) peer verifiable data integrity, and (b) data sharing and trust collaboration. We will collaborate with National Institutes of Health (NIH) USA and leverage its Integrated Biomedical Informatics System (IBIS) [11] for clinical trial data collection. We will integrate blockchain platform with IBIS to provide the peer verifiable data integrity and collaborative data sharing in the clinical trial. This is to establish a clinical trial application use case with blockchain platform. The system architecture is shown in Figure 5.

Greg Irving's blockchain POC for clinical trial only uses bitcoin blockchain distributed ledger to illustrate his approach of achieving data integrity. However, smart contracts are another key feature of the blockchain and are not currently used in clinical trials. A smart contract is a software program that executes programs in a blockchain. They can read other contracts, make decisions, and execute other contracts. We will explore the use of smart contracts to ensure the data integrity of clinical trials and to remove the possibility of human manipulation.

The clinical trial data can be linked and stored on the blockchain platform through the efficient and simplified automation contract design. Researchers of the future medical journals can also quickly store and verify the correctness of reports through smart contracts. We will study in depth the use of blockchain smart contracts to explore the data integrity of the clinical trial process and the team's data sharing and trust collaboration mechanisms.

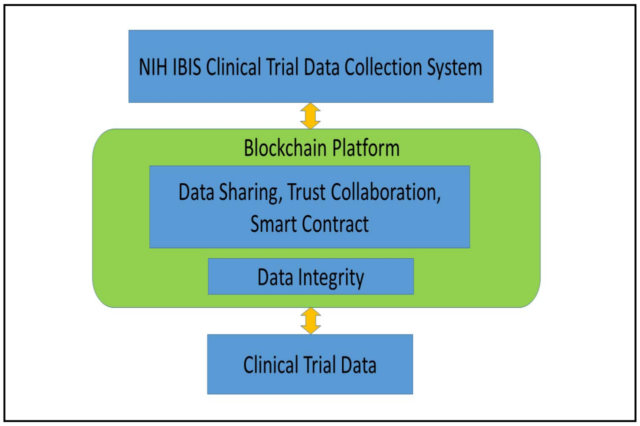
NIH IBIS Clinical Trial Data Collection System
Blockchain Platform
Data Sharing, Trust Collaboration, Smart Contract
Data Integrity
Clinical Trial Data

Figure 5. Blockchain Platform for Clinical Trial

V. Verifibale Identity Privacy and Secure Data Access

A. Verifiable Anonymous Identity Privacy

The transaction data stored in blcokchain ledger is transparent, everyone can access the transaction data that exists on the node, thus causing all transaction data to be spread out under the sun. This will present a major privacy risk for the blockchain users. It was reported that even the identity of all blockchain users is encrypted, over 60% of users their real identities have been identified [54-56] resulting from big data analysis across other data from Internet. For some applications, e.g., banking applications, this is not acceptable. The transparency and privacy are two contradict requirements. We will investigate mechanism to enable blockchain to achieve both the user's identity anonymity and the confidentiality at the same time.

Common types of user identification authentication include pass-through authentication, biometric authentication, and identity verification based on public key cryptography algorithms, etc. Different types of devices or systems can choose the appropriate identity verification for their security requirements or computing power technology.

In the traditional authentication process, the user account will be transmitted to the server in clear text, but in a public network environment, anyone can eavesdrop on the server's external connection. Recently, with the growing popularity of smart phones and the IoT applications, the identity privacy and data safety is even more at risk.

Anonymity's identity verification technology [57] replaces user fixed identity with a dynamic identity identification code to ensure user anonymity. Zero-knowledge Proofs technology are also applied to anonymous authentication. The zero-knowledge proof was proposed by Goldwasser and other scholars [58] in 1985. It uses cryptographic techniques to verify that a judgment is correct without providing the validator with any useful information. Since no new information is provided in the zero-knowledge verification process, this protocol is resistant to re-sending attacks. In addition, since the validator does not have the credentials of the user, the attacker cannot assume the control of the server as a user.

We will investigate the integration of blockchain platform with Zero Knowledge Technology to achieve anonymous identity authentication for blockchain users. The purpose of the study was to hide the identity of the patient on the medical blockchain, but to verify the legitimacy of the patient's identity. In the case of IoT blockchain applications, it can be used to hide the IoT device identity, but can verify the legitimacy of the identity of the device.

B. Secure data access

Data access control is critical with special complex policies in the medical blockchain. For example, in order to allow patients' healthcare providers to communicate and collaborate with healthcare groups (e.g., physicians) in disease prevention and treatment, healthcare providers can provide physicians with patient's healthcare information. Because the patient's data are private information, the blockchain system access control architecture must be able to achieve the user's identity certification even the anonymous identity is used in blockchain. Moreover, the system access control policy will need to be flexible to allow user to create arbitrary data access control policy to decide who, when, and what can be seen for any given individual medical records. If authorized by the patient, other users can legally access the patient's data.

To achieve patient centric healthcare, patient should have the authority to authorize the healthcare providers to allow other persons to access their medical data based on the access control policy that patient created. The access control policy can be more flexible, no longer only allow or deny access, it can allow users to set the access period and only allows specific parts of information can be accessed. Patient can determine how the data is to be shared, can assign access permissions to allow who has permission to query; can know who had already access to which data items; and can change permissions at any given time.

Blockchain can safeguard the privacy and security of the personal medical data, and put the personal data into the control of each individual. The sharing data will happen after the trust is established. This in turn will enable the notion of collaborative consumption of personal medical data.

We will investigate the mechanism to allow patient centric access control to set different permissions for patient data. The mechanism will also enable the IoT device to set permission to allow applications access the device sensor data.

We will investigate techniques for data exchange on block chains. Different nodes on the block chain can be grouped into groups. Only the nodes in the authorized group can access the user data through the permission setting of the user, allowing the exchange of information between different groups (such as electronic medical records need to be exchanged between different groups).

VI. Summary

This paper described briefly the scope, approach, challenges, and system design concept of a blockchain platform for the medical domain, particularly for clinical trial and precision medicine. This project currently undergoing at the Asia University, Taiwan. Various technology challenges and possible research areas and approaches are discussed. The methodologies and approaches of clinical trial and precision medicine as two blockchain platform use cases are discussed. This blockchain platform will be built on top of traditional blockchain and consist of 4 new system components to explore new blockchain based distributed parallel computing paradigm for big data analytics; to investigate data integrity, disparity data integration mechanisms, to study verifiable anonymous identity privacy and secure data access for both person and IoT devices; and to research on the secure data sharing and trust collaboration mechanisms to enable a trust medical data ecosystem for collaborative research.

Acknowledgment

The authors thank all researchers of the Asia University, Taiwan participating in this project for their collaborations. Particularly, Chao-Neng (Charles) Wang, Chi-Shiang Chan, Phillip Sheu, and Pei-Chun Chang for their insight discussions.

References

[1] Benjamin Gregory Carlisle, "proof-of-prespecified-endpoints-in-medical-research-with-the-bitcoin-blockchain", Web blog post. The Grey Literature. 25 Aug 2014. http://www.bgcarlisle.com/blog/2014/08/25/proof-of-prespecified-endpoints-in-medical-research-with-the-bitcoin-blockchain/

[2] Greg Irving, John Holden, "How blockchain-timestamped protocols could improve the trustworthiness of medical science", F1000research.8114.2 https://f1000research.com/articles/5-222/v2

[3] WMA Declaration of Helsinki - Ethical Principles for Medical Research Involving Human Subjects. 2016. http://www.wma.net/en/30publications/10policies/b3/

[4] COMPare - Full results. 2016. http://compare-trials.org/results

[5] Slade E, Drysdale H, Goldacre B, *et al.*: Discrepancies Between Prespecified and Reported Outcomes. *Ann Intern Med.* 2016; 164(5): 374.

[6] Goldacre B: How to get all trials reported: audit, better data, and

individual accountability. *PLoS Med.* 2015;12(4): e1001821.

[7] Anand V, Scales DC, Parshuram CS, *et al.*: Registration and design alterations of clinical trials in critical care: a cross-sectional observational study. *Intensive Care Med.* 2014; 40(5): 700–22.

[8] Medline trend. 2016. http://dan.corlan.net/medline-trend.html

[9] The CArdiovasCulAr Diabetes & Ethanol (CASCADE) Trial. Tabular View ClinicalTrials.gov. 2016. https://clinicaltrials.gov/ct2/show/record/NCT00784433?term=NCT00784433&rank=1

[10] Kyle Croman, Christian Decker, Ittay Eyal, Adem Efe Gencer, Ari Juels, Ahmed Kosba, Andrew Miller, Prateek Saxena, Elaine Shi, Emin Gün Sirer, Dawn Song, Roger Wattenhofe, "On scaling decentralized blockchains", 016/2/26, International Conference on Financial Cryptography and Data Security, 106-125.

[11] Integrated Biomedical Informatics System (IBIS) NIH, USA https://ibis.nih.gov/

[12] Ahmed Kosba, Andrew Miller, Elaine Shi, Zikai Wen, Charalampos Papamanthou,, "Hawk: The Blockchain Model of Cryptography and Privacy-Preserving Smart Contracts", Security and Privacy (SP), 2016 IEEE Symposium on, 839-858

[13] Healthnautica https://www.healthnautica.com/comppages/index.asp

[14] Factom https://bitcointalk.org/index.php?topic=850070.0

[15] Gem https://gem.co/

[16] DNA.Bits http://socialm1.wixsite.com/dnabits

[17] BlockVerify http://www.blockverify.io/

[18] FoldingCoin http://foldingcoin.net/

[19] Folding@home https://folding.stanford.edu/home/

[20] GridCoin wiki http://wiki.gridcoin.us/Proof-of-Research

[21] BitCoin wiki https://en.wikipedia.org/wiki/Bitcoin

[22] Ethereum consortium https://www.ethereum.org/

[23] Hyperledger white paper, https://docs.google.com/document/d/1Z4M_qwILLRehPbVRUsJ3OF8Iir-gqS-ZYe7W-LE9gnE/edit#heading=h.m6iml6hqrnm2

[24] Pete Rizzo, Hands on with Linq, Nasdaq's Private Market Blockchain Project", http://www.coindesk.com/hands-on-with-linq-nasdaqs-private-markets-blockchain-project/

[25] Ripple https://ripple.com/

[26] Chain.com https://chain.com/

[27] E. Brickell, J. Camenisch, and L. Chen, "Direct Anonymous Attestation," Proceedings of the 11th ACM Conference on Computer and Communications Security, pp. 132–145, 2004.

[28] M. Blum, P. Feldman, and S. Micali, "Non-interactive Zero-knowledge and Its Applications," Proceedings of the 20th Annual ACM Symposium on Theory of Computing, pp. 103–112, 1988.

[29] Hortonworks Hadoop computing tutorial, "Hadoop Tutorial: getting started with HDP", http://hortonworks.com/hadoop-tutorial/hello-world-an-introduction-to-hadoop-hcatalog-hive-and-pig/

[30] Michael Creel, William Goffe, "Multi-core, Clusters, and Grid Computing: a Tutorial", Computational Economics, Vol.32, No 4, 353-382, January 2008.

[31] Peter Mell, Timothy Grance, "The NIST Definition of Cloud Computing", Sep. 2011, http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf

[32] Andreas Antonopoulos, "Mastering BitCoin", O'Relly, ISBN-10: 1449374042

[33] Paul Vigna, Michael J. Casey, "The Age of Cryptocurrency: How Bitcoin and the Blockchain Are Challenging the Global Economic Order", St. Martin's Press, ISBN:978-1-250-08155-1

[34] Vitalik Buterin, "Ethereum: Platform Review, Opportunities and Challenges for Private and Consortium Blockchains" https://www.scribd.com/doc/314477721/Ethereum-Platform-Review-Opportunities-and-Challenges-for-Private-and-Consortium-Blockchains

[35] Thomas Hardjono, Ned Smith, Alex Pentland, "Anonymous Identities for Permissioned Blockchains", http://connection.mit.edu/wp-content/uploads/sites/29/2014/12/Anonymous-Identities-for-Permissioned-Blockchains2.pdf

[36] Thomas Hardjono, Alex Pentland, "Verifiable Anonymous Identities and Access Control in Permissioned Blockchains" https://static1.squarespace.com/static/55f6b5e0e4b0974cf2b69410/t/5717e2350442622ecf2d8739/1461183029767/ChainAnchor-Identities-04172016.pdf

[37] Joseph Poon, Thaddeus Dryja, "The Bitcoin Lightning Network: Scalable Off-Chain Instant Payments", https://lightning.network/lightning-network-paper.pdf

[38] Fan Zhang, Ethan Cecchetti, Kyle Croman, "Town Crier: An Authenticated Data Feed for Smart Contracts", ACM CCS'16, https://eprint.iacr.org/2016/168.pdf

[39] Guy Zyskind, Oz Nathan, Alex Pentland, "Enigma: Decentralized Computation Platform with Guaranteed Privacy", arXiv (whitepaper) , https://arxiv.org/abs/1506.03471

[40] Santosh Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System", https://bitcoin.org/en/bitcoin-paper

[41] Moustafa AbdelBaky, Manish Parashar, Hyunjoo Kim, Kirk E. Jordan, Vipin Sachdeva, James Sexton, Hani Jamjoom, and Zon-Yin Shae, Gergina Pencheva, Reza Tavakoli, and Mary F. Wheeler, "Enabling High Performance Computing as a Service", IEEE Computer, Oct. 2012.

[42] Wei-Min Ho, Jr-Rung Lin, Hui-Hsuan Wang, Chia-Wei Liou, Ku-Chou Chang, Jiann-Der Lee, Tsung-Yi Peng, Jen-Tsung Yang, Yeu-Jhy Chang, Chien-Hung Chang, Tsong-Hai Lee (2016) Prediction of in-hospital stroke mortality in critical care unit. *Springer plus*., **5**, 1051.

[43] Chia-Lin Wu, Chun-Chieh Tsai, Chew-Teng Kor, Der-Cherng Tarng, Ie-Bin Lian, Tao-Hsiang Yang, Ping-Fang Chiu, Chia-Chu Chang (2016) Stroke and Risks of Development and Progression of Kidney Diseases and End-Stage Renal Disease: A Nationwide Population-Based Cohort Study. *PLoS One*, 11, e0158533.

[44] Alberto Spalice, Francesca Del Balzo, Laura Papetti, Anna Maria Zicari, Enrico Properzi, Francesca Occasi, Francesco Nicita, Marzia Duse Ital (2016) Stroke and migraine is there a possible comorbidity? *J Pediatr*. 42, 41.

[45] Yu-Chi Tung, Guann-Ming Chang (2016) The Relationships Among Regionalization, Processes, and Outcomes for Stroke Care: A Nationwide Population-based Study. *Medicine (Baltimore),* 95, e3327.

[46] Melgaard, Anders Gorst-Rasmussen, Lars Hvilsted Rasmussen, Gregory Y. H. Lip, Torben Bjerregaard Larsen (2016) Vascular Disease and Risk Stratification for Ischemic Stroke and All-Cause Death in Heart Failure Patients without Diagnosed Atrial Fibrillation: A Nationwide Cohort Study Line. *PLoS One*., 11, e0152269.

[47] Naoki Kaneko, Toshihiro Mashiko, Taihei Ohnishi, Makoto Ohta, Katsunari Namba, Eiju Watanabe, Kensuke Kawai (2016) Manufacture of patient-specific vascular replicas for endovascular simulation using fast, low-cost method. *Sci Rep.,* 6, 39168.

[48] Healthcare Rallies for Blockchains, IBM report, https://public.dhe.ibm.com/common/ssi/ecm/gb/en/gbe03790usen/GBE03790USEN.PDF

[49] Susan Chow, "Music Therapy for Stroke", http://www.news-medical.net/health/Music-Therapy-for-Stroke.aspx

[50] MicroRNA wiki, https://en.wikipedia.org/wiki/MicroRNA

[51] SNP wiki, https://en.wikipedia.org/wiki/Single-nucleotide_polymorphism

[52] NCBI https://www.ncbi.nlm.nih.gov/pubmed

[53] Health Insurance Portability and Accountability Act https://en.wikipedia.org/wiki/Health_Insurance_Portability_and_Accountability_Act

[54] Chaitanya Katikala , John Phillips , Andy Mai ," The Hacker's Code: Finding Bitcoin Thieves Through the Similarity and Status Claims Between Users", Stanford CS 224W Fall 2013 - Team 39C - S 224W Final Report, http://snap.stanford.edu/class/cs224w-2013/projects2013/cs224w-039-final.pdf

[55] Reid, Fergal, and Martin Harrigan. "An analysis of anonymity in the bitcoin system." Security and Privacy in Social Networks. Springer New York, 2013. 197-223.

[56] Androulaki, Elli, et al. "Evaluating User Privacy in Bitcoin." IACR Cryptology ePrint Archive 2012 (2012): 596

[57] M. L. Das, "Two-Factor User Authentication in Wireless Sensor Networks," IEEE Transactions on Wireless Communications, Vol. 8, No. 3, pp. 1086–1090, 2009.

[58] S. Goldwasser, et al, "The Knowledge Complexity of Interactive Proof-Systems," Proceedings of the 17th Annual ACM Symposium on Theory of Computing, pp. 291–304, 1985.