Leveraging Standards Based Ontological Concepts in Distributed Ledgers: A Healthcare Smart Contract Example

Mengyi Li*, Lirong Xia† and Oshani Seneviratne‡
Rensselaer Polytechnic Institute
Troy, NY 12180, USA
Email: *lim18@rpi.edu, †xial@cs.rpi.edu, ‡senevo@rpi.edu

***Abstract*—Sharing clinical data using Distributed Ledger Technologies (DLT) is increasing in momentum. Fast Healthcare Interoperability Resources (FHIR) standard provides a standard based shared vocabulary and widely accepted mechanisms enabling healthcare providers to share patient data across institutions. Many DLT based solutions are capitalizing on these standards to enable trustworthy electronic health record sharing among institutions. In this paper, we present our preliminary work on capturing the semantics of the FHIR standard in smart contracts. We also discuss the appropriate data to mine from transaction logs in decentralized ledgers to find any anomalies and information misuses by leveraging these standards-based ontological concepts.**

I. Introduction

Information is becoming more readily accessible across a wide range of institutions, both in centralized and decentralized applications. There are numerous problems associated with the increasing accessibility to information. For example, it is hard to measure the trustworthiness of entities involved in the information sharing ecosystem that can span many different systems and thousands, if not tens or hundreds of thousands, of different users. In centralized applications, data and the access/transaction logs can be accessed relatively quickly to be analyzed using machine learning algorithms. However, in decentralized data sharing applications, there are many challenges in determining aggregate statistics to elicit data usage patterns, as well as delays in processing transactions due to the need for achieving consensus, challenges in processing power (especially if proof of work is utilized), and other numerous scalability issues due to the distributed nature of the application. Nevertheless, DLT is a better choice to share medical data transfers and usages. It is based on robust security protocols, has no central point of failure which has plagued many health information systems such as the cases reported in [1] and [2], provides more privacy-friendly decentralized data ownership, and more importantly, it provides an immutable distributed ledger for all the transactions that guarantees the integrity of the data contained.

Making predictions about a user's intent upon data access needs to be implemented to prevent data misuse. To make decisions at scale on whether someone's intention for data sharing is good or not, we need a standard mechanism of analyzing transaction log data. In decentralized applications (DApps), such standard terminologies are favored as interoperability between various systems is a highly desired feature. Standardization ensures that we can utilize an algorithm for validation of appropriate and intended use. Since raw data can vary in formats and structures that are hard to use for analysis, standards that are grounded in well accepted ontological concepts can act as the *lingua franca* and can be used for effective feature selection in machine learning algorithms, as well as smart rule-based access control and data usage tracking.

In this short paper, we discuss an example from the healthcare domain that motivates the use of standards-based ontological concepts in DApps. We demonstrate a smart contract that uses terminologies from a well-adopted standard, the relevance of these terminologies for features used in machine learning models, and our proposal for detecting data misuses both retroactively and proactively.

II. Use Case

To make sure that our data resource is grounded in a real-world use case, we are using electronic health record exchange between various entities, such as attending physicians, referring physicians, pharmacists and insurance companies.

Healthcare systems must present shared data to institutions and practitioners in a structured and readable format; thus standards are required. Adherence to standards can help reduce risks and costs, thus support semantic interoperability, which plays a vital role in the healthcare system. Therefore, our specific focus is on data standardization of healthcare based smart contracts, and choosing a good healthcare standard remains to be an essential cornerstone.

A. Healthcare Standards for Information Exchange

We have explored the healthcare data standard FHIR [3], which is a standard for exchanging healthcare information electronically. We illustrate how a subset of the data records are represented in the transaction logs generated from a DApp that implements a health data sharing smart contract that uses terms from the FHIR standard. According to Bender et al. [3], the FHIR standard should be easier and faster to implement than the previous Health Level Seven (HL7) Version 2 & Version 3, since there is a drastic reduction in the number

of Common Message Element Types, which makes resource representation and transport simple but maintain the design. FHIR is built from a set of modular components and has a strong focus on the implementation. In other words, it makes the later implementations easier and faster. Additionally, FHIR defines a simple framework for extending and adapting the existing resources, thus allowing variability caused by the diverse healthcare process.

B. Information Misuse Problems

In the healthcare domain, there are many instances of secondary non-healthcare related usage of sensitive patient data. Examples include: using health data for marketing purposes [4], denying health insurance, or raising insurance premiums based on what is in a patient's health record [5], and even unauthorized research on patient data that the patients did not consent to when the data was collected [6]. There is also some news about how tech giants give their clients access to patients' records without their consent [7]. Thus, sensitive information, such as sexually transmitted diseases, can be divulged without a patient's consent. As can be imagined, in many of these cases, patients might not be even aware of such data usages, until after some harm is done. Once sensitive information is leaked, often the damage caused would be irreversible. Healthcare DApps can record the data flows between all the participants such as clinicians and researchers. Combined with machine learning techniques that continuously monitor and learn from the transactions on the ledger, we can detect and predict malicious transaction flows. So, there is a possibility to prevent these malicious transactions from executing, or in the worst case, determine who is accountable for the data breach thanks to the immutable ledger.

III. Related Work

Kuo, Kim, and Ohno-Machado have analyzed blockchain technology's benefits over other distributed healthcare/biomedical database systems [8]. In their analysis and prototype system, they illustrate how patients can manage their healthcare records, thus allowing decentralized management of the healthcare system. Records are source-verifiable, which largely reduces the risk of fraudulent activities on the records. Also, since data is stored on a decentralized network, institutions are not able to take forced possession of those data records, and there is no chance of other patient records. Therefore, DLT based systems have undoubtedly demonstrated improvements to medical record management. Other systems such as MedRec [9] too enables patient data sharing and incentives for medical researchers to sustain the system using DLT.

FHIRchain [10] presents an Ethereum based solution designed to meet the Office of the National Coordinator (ONC) for Health Information Technology's requirements by enforcing the use of FHIR to share clinical data by validating whether the generated reference pointers to sensitive patient records that are stored off-chain follow the FHIR API standards. FHIRchain does not analyze the semantics of the useful fields used for annotating the data using the FHIR standard as is the aim of the work presented in this paper. Additionally, they have listed the inability to control clinical malpractice as a limitation and a potential future work.

IV. Motivating Scenario

In the world of information security, to present appropriate levels of access, access control systems exist for participants to grant or revoke access according to personally identifiable information. Usually, access control systems are used in computer security to regulate access to critical or valuable resources. However, applying it on DLT can elevate the system to a higher level–the distributed public ledger can show all the transactions of the right to access a resource among all participants publicly. Hence a user can know who currently has the right to access specific information, preventing participants deceive others into thinking they have access to the information while their access requests are denied.

A. Smart Contract for Access Control on Personally Identifiable Information

On Hyperledger Composer[1], we created a relaxed, smart contract that deploys a network to illustrate access control on personally identifiable information. This network can allow participants to grant or revoke access to their information to others, and also transfer data to others. Therefore, the misuse of data might happen when a patient A grants access to practitioner B, but practitioner B shows the records of patient A to another member who is in organization C. This increases the possibility of the misuse of data, because organization C might use the records for research purposes without patient A's consent.

B. Invocation of the Smart Contract

Fig. 1 shows the participant registry of practitioner B before `authorizeAccess` transaction. After the method has been called, the transaction logs of patient A granting access to practitioner B to A's records can be seen in Fig. 2. The participant registry of practitioner B after `authorizeAccess` transaction can be seen in Fig. 3.

Fig. 1. Participant registry of practitioner B before authorizeAccess transaction

Since practitioner B now has access to patient A's records, practitioner B can include the records in a transaction and send it to an organization C. Fig. 4 shows the transaction logs of practitioner B showing A's records to organization C (Only organization C can see the transaction logs). This may be a

[1]https://hyperledger.github.io/composer/latest

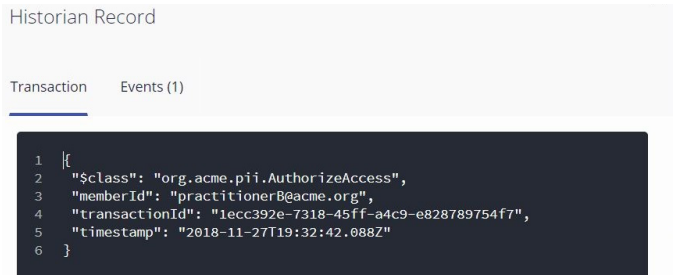

Fig. 2. Transaction logs indicating granting access to patient A's records to practitioner B

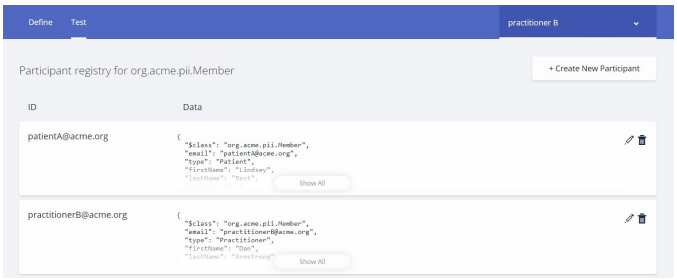

Fig. 3. Participant registry of practitioner B after authorizeAccess transaction

violation of the original intent of the data sharing agreement that patient A had with practitioner B.

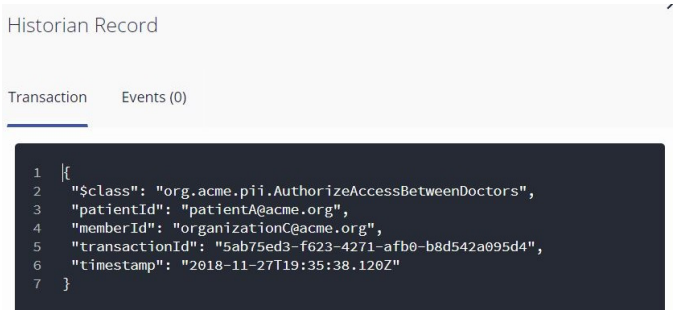

Fig. 4. Transaction logs of practitioner B showing A's records to organization C

Fig. 5 shows the rule of `authorizeAccess` transaction that allows member permission to view personal information of other members in smart contract, and Fig. 6 shows the rule of `authorizeAccessBetweenDoctors` transaction. Only the sender and the receiver can see the transaction logs that contain patient information.

```
rule ForeignRecordConditionalAccess {
    description: "Allow participants access to other people's records if granted"
    participant(p): "org.acme.pii.Member"
    operation: ALL
    resource(r): "org.acme.pii.Member"
    condition: (r.authorized && r.authorized.indexOf(p.getIdentifier()) > -1)
    action: ALLOW
}
```

Fig. 5. Permission rule to view personal information of other members

```
rule AuthorizeAccessBetweenDoctorsConditionalReadTransaction {
    description: "Request a communication"
    participant(p): "org.acme.pii.Member"
    operation: READ
    resource(r): "org.acme.pii.AuthorizeAccessBetweenDoctors"
    condition: (r.memberId === p.getIdentifier())
    action: ALLOW
}
```

Fig. 6. Permission rule that only allows sender and the receiver to see the transaction logs that contains patient information

C. The Need to Make the Smart Contract Smarter

As we can see from the smart contract description in the previous section, it allows one of the participants to grant or revoke access to another, and also allows one of the participants to send messages to another participant. Access control has a relaxed nature where the rights of participants to access the resources are expressed through access control policies, but we have no idea what the participants would do with the data afterward. Therefore, there could be some potential problems. After being granted access to some data, participants might use the data that they have access to in the interest of private corporations. For example, smart contracts can be misused by "bad actors" by raising insurance premiums, denying insurance, or sell to a company for marketing purpose without the user's agreement.

Therefore, a system that can make predictions about the intent of the participants requesting access control is needed to solve this problem. According to the past behaviors, if the method can recommend to the relevant institutions to make some changes about whether this person is a good or bad actor in the system, then there is a smaller chance of leaking the data to nefarious companies or agencies. Therefore, the smart contract needs these kinds of methods to solve problems like finding bad actors.

V. Proposal

A central challenge in mining patterns from medical records and transaction records is that they are stored in non-conformant or proprietary standards, which increases the difficulty of applying machine learning methods generically. We first need to process the data in a domain-specific manner, and then only can we apply the model to the data. However, if we can create models of standard transactions suitable for different types of requests, then there is no need to process the data in this arduous manner. Instead, when a user submits transactions, they will be doing it using a standard JSON form that conforms to the FHIR standard.

Therefore, a structured, smart contract using the FHIR standard mainly increases the ease of implementation and improves efficiency. For example, from the JSON template of the transaction log in FHIR[2], we can see features such as *requester*, *receiver*, etc. Those features can be closely examined for the use by machine learning algorithms for prediction of anomalous behavior.

[2]The JSON template for FHIR transaction logs is available at http://hl7.org/fhir/STU3.

The goal of this section is to analyze the FHIR standard to create a structured healthcare smart contract that can filter out apparent transactions that are not valid. We explore the combinations of features that are useful for creating the anomaly detection model.

A. Request Model for FHIR Based Smart Contracts

Using the smart contract, each user can submit a "request", which serves as the transaction on the ledger. Based on the FHIR standard, this request can have different types: `CommunicationRequest`, `MedicationRequest`, `ProcedureRequest`, and `ReferralRequest`.

We define the structure of different types of requests on the DLT using the FHIR standard. For example, for `CommunicationRequest`, there are fields like "identifier", "basedOn", etc. In the model file in Hyperledger Composer, we define a transaction `CommunicationRequest` with all these fields specified by the FHIR standard.

B. Fields Relevant for Anomaly Detection from Transactions in the Ledger

For each of the different types of requests identified above from the FHIR standard, we performed a qualitative analysis to determine if the corresponding fields are useful to the final decision as to whether a given transaction should be classified as an undesirable transaction or not. If they are not useful, we make it an optional field in the request resource type in the smart contract.

1) CommunicationRequest: This is a conveyance of information from one entity, a requester to another entity, a receiver. The requester and receivers may be patients, practitioners, related persons, organizations and devices. The information conveyed could be an alert, a reportable condition or other information. Possible misuse of data is that a practitioner is requesting the communication of patients' records to another practitioner/company without the patients' consent. The useful fields for these types of requests are illustrated in Table I.

2) MedicationRequest: This covers all orders for medications for a patient. For this type of request, bad actors may tend to request a medication that are not appropriate. Besides `identifier`, `basedOn`, `groupIdentifier`, `category`, `subject` (required), `requester` (required), `reasonCode`, `reasonReference`, there are other useful fields that can be used to predict the trustworthiness of the user as illustrated in Table II.

3) ProcedureRequest: This is a record of request for a procedure to be planned, proposed or performed. This type of request can be invalid if they are requesting a procedure that is not appropriate. Besides `identifier`, `basedOn`, `replaces`, `requisition` (same as `groupIdentifier`), `category`, `subject` (required), `requester` (required), `reasonCode`, `reasonReference`, there are other useful fields that can be used to predict the trustworthiness of a data sharing transaction as analyzed in Table III.

4) ReferralRequest: This is used to record and send details about a request for referral service or transfer of a patient to the care of another provider or provider organization. There is a chance that a partitioner is requesting to transfer patients' record to another partitioner/company without the patients' consent. Besides `identifier`, `basedOn`, `replaces`, `groupIdentifier`, `requester` (required), `subject` (required), `recipient`, `reasonCode`, `reasonReference`, there are other useful fields that were elicited in our analysis as illustrated in Table IV.

VI. Future Work

The next immediate goal in this project is to make use of the semantics of the FHIR standard in a smart contract to determine whether a given transaction is an anomaly or not and to classify behaviors of users with the goal of identifying information misuses. We will use statistical machine learning techniques for clustering and support vector machine methods for this purpose. These techniques have been proven successful in similar problems involving behavior based access control in scalable anomaly detection on TCP connections and HTTP requests [11]. One problem of our motivating scenario is the lack of ground truth, where the exception cases are human-annotated and very little is linked to actual misuse of data. To address this problem, Adler et al. simulated several attack variants and observables [11], which in our case will be the semantics of the FHIR standard in the smart contract. The methods that combines word embedding [12], K-means clustering [13] and continuous learning [14] are not novel. However, applying anomaly detection with K-means clustering and continuous learning to support real-time decision making on whether the actors involved in the transaction are bad results in novelty. Here is a simple outline of the possible algorithms that use the requests and fields identified through our analysis of the FHIR standard for health data sharing smart contracts:

A. K-means clustering

To analyze the semantics of the FHIR standard in a smart contract, we need to do the word embeddings first to capture the context of the word in semantic similarity to express the data. In our case, we need to do word embeddings on the fields that are useful to determine one's behavior profile using the transaction logs as explained in the previous section. We then plan to use the K-means clustering algorithm to cluster the semantics data from different transaction logs. The goal of clustering is to find out the centroids of each cluster, thus find the anomalies that are too far away from all centroids. For example, if each point on a clustering graph represents a member, the location of these members depends on the features extracted from the useful fields that we identified. A far away point from other clusters indicates that this member is an outlier which has abnormal behaviors based on the useful fields that we analyzed above. Therefore, we can conclude that this member is suspicious because his behavior profile is different from the normal ones.

Field	Description	Relevance for Machine Learning Algorithms
identifier	The unique ID of this request for reference purpose	It keeps track of the transaction and will be autogenerated as the transactionId.
basedOn	A plan or proposal that is fulfilled in whole or in part by this request	It records the reason that requester delivers such communicationRequest.
replaces	Records the requests replaced by this request	This allows analysis on the trace of the continuation of therapy through multiple requests. The field is not empty if the previous requests are replaced with these new requests. Thus, later the machine learning algorithm can look at the previously rejected requests to predict whether this is a similar problem that this request cannot be passed. This is useful for the algorithm to look back at this requester's previous requests, which can be used as training data for determining if this request is legitimate or not.
groupIdentifier	Common to all requests that were authorized simultaneously by a single author. According to FHIR, the author refers to the one who provides the prescription to the requester. This also represents the identifier of the requisition or prescription.	This allows the machine learning algorithm to look at the requisition with a shared identifier to see if it's spam.
category	The type of message to be sent such as alert, notification, reminder, and instruction.	This illustrates the purpose of the transaction.
medium	A channel that was used for this communication, such as email, fax.	This could be a feature to determine the behavior of the users. The behavior features can be used to determine the user's intent by the machine learning algorithm.
subject	The patient or group that is the focus of this communication request.	It is a piece of helpful information to check if the same patient's record has been sent multiple times.
recipient	The person, organization, clinical information system, device, group or care team which is the intended target of the communication.	This feature could be used for determining the intent of users. For example, if the user's record is sent to somewhere else like a marketing company or insurance company, then there is a larger possibility that the requester's request is not appropriate.
payload	The main content to be communicated to the recipients.	This feature is necessary but should not be used for the machine learning algorithm purpose because it might contain record data.
sender	The source of the communication. Where the record and communication content comes from is helpful for comparing with the requester.	If they are not the same individual, some investigation is needed.
requester	A required field that illustrates the individual who initiated the request and has responsibility for its activation.	Both requester.agent and requester.onBehalfOf are important information. If the transaction is later determined to be invalid, then the requester should be the one who is marked as a bad actor.
reasonCode	Describes why the request is being made.	It should be compared with the basedOn to see if they are consistent.
reasonReference	Indicates another resource whose existence justifies this request.	Optional but increases the credibility of the reason if provided.

TABLE I
USEFUL FIELDS FOR MACHINE LEARNING IN COMMUNICATIONREQUEST

Field	Description	Relevance for Machine Learning Algorithms
intent	A required field that can either be a proposal, plan or original order.	Whether the intent is a proposal, plan or an original order restrict the requester type. Therefore, this feature should be consistent with the according to requester type to validate this request.
recorder	The person who entered the order on behalf of another individual	For example, if we have identified multiple requests that are not valid by the same recorder, we need to be careful about this recorder. This feature should also be an important feature to input into the algorithm.
detectedIssue	Indicates an actual or potential clinical issue with or between one or more active or proposed clinical actions for a patient.	This is an optional field, but very straightforward in the way that it directly indicates possible problems of a patient. Usually, it is entered by a practitioner who is requesting medication for a patient.

TABLE II
USEFUL FIELDS FOR MACHINE LEARNING IN MEDICATIONREQUEST

B. Continuous Learning

Continuous learning methods redeploy the model with the new incoming FHIR based transaction logs. The smart contract can be augmented to save new training data (old transaction logs) as it receives new incoming transaction logs. When it has accumulated enough new data, the smart contract should test the model's accuracy against the machine learning model. If the accuracy is degrading over time, then the smart contract should call to redeploy the model using the new training data set.

Field	Description	Relevance for Machine Learning Algorithms
`intent`	A required field that determines whether the request is a proposal, plan, an original order or a reflex order.	It does not restrict the requester type as `MedicationRequest` does, but is still helpful since it reveals what the request is based on.
`code`	A required field that identifies a particular procedure, diagnostic investigation, or panel of investigations, that have been requested.	This is relevant to the subject (patient)'s record, thus should be considered as relevant feature because it identifies the requested procedure.

TABLE III
USEFUL FIELDS FOR MACHINE LEARNING IN PROCEDUREREQUEST

Field	Description	Relevance for Machine Learning Algorithms
`intent`	A required field that distinguishes the level of authorization/demand implicit in this request.	The codes for it are the proposal, plan, order, original-order, reflex-order, filler-order, instance-order, and option. Since it directly indicates the intent of the request, it should be considered when applying algorithms.
`type`	An indication of the type of referral.	The type of referral should be consistent with the patient's record. For example, if the patient's record does not include drug addiction, but the code for the type displays a patient referral for drug addiction rehabilitation, then the request is not valid. Also, if some practitioner/organization is specialized in a certain area that does not accord with `type`, then the request should not be passed. Therefore, the type is an important indicator of the validity of the request.
`serviceRequested`	The service that is requested to be provided to the patient.	The code for the `serviceRequested` should also be consistent with `type`. Also, if the same patient has been provided an abnormal times of the same service, it is considered an anomaly. This field is also a good feature.

TABLE IV
USEFUL FIELDS FOR MACHINE LEARNING IN REFERRALREQUEST

VII. CONCLUSION

In this paper, we have described how to leverage standards-based ontological concepts in distributed ledgers. We demonstrated the utility of this approach using the FHIR standard used in health data exchange for DApps and the applicability of several fields in machine learning models. By supporting FHIR on DLT based healthcare DApps, we enable interoperability in the healthcare system that allows each end of the transaction to understand the data more efficiently. By utilizing our analysis of the FHIR specific fields, we hope to successfully demonstrate the behaviors of bad actors sending requests for data access and usage. While future work is still needed to analyze the semantics on the larger set of data to predict such anomalies, the work presented in this paper is an initial step towards utilizing community accepted standards in smart contracts and utilizing those in machine learning models to make the smart contracts smarter.

ACKNOWLEDGEMENTS

This work is supported by the IBM-RPI Artificial Intelligence Research Collaboration (a member of the IBM AI Horizons Network). We thank our colleagues James A. Hendler and Geeth De Mel for their insight and expertise that greatly assisted the research, and also the reviewers for helpful comments and suggestions.

REFERENCES

[1] "Singapore personal data hack hits 1.5m, health authority says," *BBC News*, Jul 2018. [Online]. Available: https://www.bbc.com/news/world-asia-44900507

[2] "Hospital CEO forced to pay hackers in bitcoin now teaches others how to prepare for the worst," *CNBC News*, April 2018. [Online]. Available: https://www.cnbc.com/2018/04/06/hosptial-ceo-forced-to-pay-hackers-in-bitcoin-now-teaches-others.html

[3] D. Bender and K. Sartipi, "HL7 FHIR: An Agile and RESTful approach to healthcare information exchange," in *Computer-Based Medical Systems (CBMS), 2013 IEEE 26th International Symposium on*. IEEE, 2013, pp. 326–331.

[4] A. Tanner, *Our bodies, our data: How companies make billions selling our medical records*. Beacon Press, 2017.

[5] C. Safran, M. Bloomrosen, W. E. Hammond, S. Labkoff, S. Markel-Fox, P. C. Tang, and D. E. Detmer, "Toward a national framework for the secondary use of health data: an american medical informatics association white paper," *Journal of the American Medical Informatics Association*, vol. 14, no. 1, pp. 1–9, 2007.

[6] S. P. Mann, J. Savulescu, and B. J. Sahakian, "Facilitating the ethical use of health data for the benefit of society: electronic health records, consent and the duty of easy rescue," *Phil. Trans. R. Soc. A*, vol. 374, no. 2083, p. 20160130, 2016.

[7] A. Maxmen, "AI researchers embrace Bitcoin technology to share medical data," 2018.

[8] T.-T. Kuo, H.-E. Kim, and L. Ohno-Machado, "Blockchain distributed ledger technologies for biomedical and health care applications," *Journal of the American Medical Informatics Association*, vol. 24, no. 6, pp. 1211–1220, 2017.

[9] A. Azaria, A. Ekblaw, T. Vieira, and A. Lippman, "MedRec: Using Blockchain for Medical Data Access and Permission Management," 2016.

[10] P. Zhang, J. White, D. Schmidt, G. Lenz, and S. Rosebloom, "FHIR-Chain: Applying Blockchain to Securely and Scalably Share Clinical Data," 2018.

[11] A. Adler, "Using Machine Learning for Behavior-Based Access Control: Scalable Anomaly Detection on TCP Connections and HTTP Requests," 2013.

[12] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[13] S. P. Lloyd, "Least squares quantization in pcm," *Information Theory, IEEE Transactions*, vol. 28.2, pp. 129–137, 1982.

[14] Z. Chen and B. Liu, "Lifelong machine learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 10, no. 3, pp. 1–145, 2016.