A Semantically Rich Knowledge Graph to Automate HIPAA Regulations for Cloud Health IT Services

Dae-young Kim, Karuna P. Joshi
University of Maryland, Baltimore County, Baltimore, MD 21250 USA
{leroy.kim, karuna.joshi}@umbc.edu

Abstract—As healthcare organizations adopt cloud-based services to manage their patient data, compliance with the rules and policies of the Health Insurance Portability and Accountability Act (HIPAA) regulation becomes increasingly complex. At present, HIPAA rules are available only in large textual format and require significant human effort to implement in the Health IT systems. Moreover, every change in the regulation, like the recent relaxation in telehealth policy due to the COVID-19 pandemic, has to be manually implemented in the IT system. We have developed a semantically rich Knowledge graph, using Semantic Web technologies to represent HIPAA rules in a machine-processable format. This will significantly help in automatically reasoning of HIPAA policies. In this paper, we describe our design along with the results of our study of the current status of research on HIPAA ontology. We have validated our design against use cases defined by the US Department of Health and Human Services (HHS). This knowledge graph can be integrated with existing healthcare systems to provide automated compliance with HIPAA policies.

Index Terms—HIPAA, Compliance, Knowledge Graph (Ontology), Semantic Web, Policy Expressions

I. INTRODUCTION

Adherence with HIPAA is essential in all cases where patients’ private information is shared, such as clinical research, and when continuity of care is required. The current scenario of sharing patient data for managing the COVID-19 pandemic is no exception. Many organizations of various sizes provide testing for infection in several dispersed places, and many research institutions are investigating the characteristics, treatment, and prevention of COVID-19. When HIPAA came into force, some raised concerns about the efficiency of surveillance [1], [2]. However, it did not receive much attention till the outbreak of COVID-19, which brought drastic changes in the healthcare landscape. For example, after the outbreak, the need for non-face-to-face medical care (or telehealth) increased to prevent the nosocomial transmission by airborne virus-laden aerosols in healthcare facilities [3]. Also, contact tracing and social distancing became an indispensable part of disease control. All of these efforts access patient information and are, therefore, within the scope of HIPAA. However, it consumes substantial human resources and time to satisfy HIPAA requirements, which delayed their efficiency.

Through Coronavirus Aid, Relief, and Economic Security (CARES) Act [4], some of the HIPAA regulations applied to telehealth were relaxed, and telehealth coverage for Medicare was expanded. As a result, patients’ access to medical services has increased, and medical institutions can improve infection control, minimize overall risk to public health, and provide necessary medical services.

Nevertheless, the CARES Act only relaxed HIPAA’s regulations on telehealth. Simultaneously, more strains of the coronavirus and numerous patients without prognostic symptoms have been reported [5], [6]. Therefore, it is necessary to collect a large amount of information through information exchange between medical institutions to understand the virus thoroughly. However, due to HIPAA’s Security and Privacy Rule, a quick response to be in pace with the virus is being delayed [7]. Therefore, it is necessary to automate HIPAA policy compliance to facilitate data exchange efficiency and security in healthcare big data.

In short, organizational health data interoperability has become a crucial issue in the COVID-19 pandemic and further disease control in the future. Policy and rules automation would facilitate the compliance process to help organizations focus on protecting patients’ data, and organizations can efficiently access the big data for public health. We need novel approaches to represent and reason over the knowledge embedded in HIPAA rules and regulations to achieve this. Therefore, more work is required on the HIPAA knowledge graph that explicitly specifies real-world conceptualization [8].

A semantically rich, machine-processable knowledge graph (or ontology) that captures HIPAA regulation concepts will significantly help reduce the healthcare IT services overhead of complying with HIPAA. In addition to saving organizational resources dedicated to compliance adherence, it will also help proactively identify invalid private medical data sharing. We have developed a semantically rich Web Ontology Language (OWL) ontology to define HIPAA privacy and security rules. This ontology extends our previous HIPAA ontology [9] developed in 2016 by including the relationships between HIPAA rules and stakeholders and the hierarchies of provisions. In this paper, we described the updated knowledge graph in detail and validated the knowledge graph against the use cases defined by HHS.

This study has three main contributions. First, we surveyed the current state of the art of ontological research done to represent HIPAA. Second a semantically rich HIPAA knowledge graph that forms the underlying knowledgebase for compliance automation. Lastly, we have validated our knowledge graph against the HIPAA use cases defined by HHS and present the results in this paper.

The following sections II and III describe the background
and related work in this area. Section IV and V illustrate the knowledge graph and the validation results. Lastly, Section VI concludes by describing the future work and the overall conclusions of this research effort.

II. BACKGROUND

A. Health Insurance Portability and Accountability Act

HIPAA regulation protects the privacy of individually identifiable health information and ensures the lawful use and dissemination of the health information to promote high-quality health care [10]. The Health Information Technology for Economic and Clinical Health Act (HITECH) Act, authored by the Office for Civil Rights (OCR) of HHS, puts Business Associates (BAs) under HIPAA Security Rule coverage and the HITECH Act’s privacy and security provisions. Furthermore, a maximum penalty of $1.5 million for all violations of an identical provision is established [11].

From the consecutive announcement and statement between February and April in 2003, HHS emphasized HIPAA’s proper functions rather than its dysfunction, listing the many social benefits that can be achieved through it [12], [13]. HHS believed that by harnessing information technologies, HIPAA could ensure appropriate privacy safeguards and maintain a common-sense balance between protecting patients’ privacy and providing the best quality care. Also, HHS affirmed that the rules do not interfere with physicians’ ability to treat their patients and allow critical public health initiatives such as the surveillance of outbreaks of infectious diseases and the reporting of adverse drug events [13].

However, soon enough, discrepancies were reported between ideal outcomes and actual real-life environment results during the regulation application. Roberta B. Ness surveyed thirteen societies of epidemiology, and 1527 eligible professionals answered. 67.8% of respondents reported that HIPAA Privacy Rule made research more difficult at a considerable cost and time [14]. In the last question of the survey, the respondents also identified two main factors that impede health research:

- Substantial variability between institutions in their interpretation of Privacy Rule regulations.
- Confusion within government agencies about the demarcation between public health surveillance.

The adverse effect of HIPAA was quite visible. O’Herrin et al. reported that HIPAA increased the Institutional Review Board (IRB) requirements for medical record research. Notably, 77% of the applications which need full IRB approval were abandoned. The considerable workload is imposed on the Human Subject Committee (HSC) and researchers; therefore, investigators could not meet the regulatory requirements [15]. Also, Wolf and Charles proclaimed that HIPAA hindered patient accrual and increased mean personnel time spent recruiting [16].

Even after HIPAA was modified through the laws represented by The HITECH Act and the 2013 Omnibus Rule over more than ten years, the aforementioned negative aspects were hardly improved. Lenert and McSwain claimed that mixtures of state and HIPAA privacy regulations nowadays still inhibit effective Health Information Exchange (HIE) for patient care by pointing out antiquated rules, complexity, and conflict at the state and federal level [7].

B. Transition of Health IT Services to the Cloud

With the rise of cloud technology, many organizations began migrating their Electronic Health Record (EHR) systems to the cloud. The advantage of cloud EHR is that it can provide healthcare more efficient and effective at a lower cost [17]. National Institute of Standards and Technology (NIST) also focused on the massive cost savings with increased IT agility offered by cloud computing and proclaimed that the government and industry have to consider adopting cloud technology seriously. It is estimated that the global healthcare cloud computing market size will reach $51.9 billion by 2024, from $23.4 billion in 2019 [18].

However, we have to consider the multi-tenancy characteristics of cloud architectural designs, which most traditional in-house systems do not have [17]. In other words, intertwined interests between several BAs and the ambiguity of interpretation are significant obstacles to HIPAA compliance in cloud systems. To resolve this, organizations and cloud service providers must write a Business Associate Contract (BAC) and specify compliance requirements, service levels, and legal liability. However, though BAC can clarify the responsibilities, the expenditure spent on compliance became unreasonable because of its complexity, unlike HIPAA’s initial goal of reducing healthcare costs in data exchange between the organizations.

III. LITERATURE REVIEW AND RELATED WORK

A. HIPAA Ontology

This study presents a systematic literature review of HIPAA ontology/knowledge graph for regulation compliance in information systems. For the research’s strict rigor, we adopted the PRISMA statement for systematic reviews [19]. The primary data sources we referenced for the survey include MEDLINE, IEEE Xplore, ACM Digital Library, Springer Link, Science Direct, AIS Conferences Collections, Academic Search Ultimate, and Web of Science.

The literature review aimed to find out articles that focus on the ontological analysis of HIPAA. Therefore, the articles’ inclusion criteria were ontological analysis of HIPAA and application of HIPAA ontology written in English. This study follows a recall-focused retrieval strategy to retrieve as many relevant papers as possible [20]. The final result excluded articles when (1) the article analyzed HIPAA by the non-ontological method, (2) the article adopted HIPAA as a use case for customized ontology, and (3) the article whose method is too vague or adopted HIPAA superficial way. This allowed us to select five related studies out of 1082 candidate papers.
1) **Formal Ontology:** Karuna Joshi et al. demonstrated a semantically rich OWL HIPAA ontology, which gives us a hierarchical view. They investigated key stakeholders in HIPAA and what detailed rules are under the security and privacy rules imposed upon the stakeholders [9]. The critical idea lying at the heart of the ontology was to consolidate the cloud life cycle and HIPAA compliance to automate the acquisition and consumption of cloud-based services by defining healthcare domain-specific security and privacy measures.

Ramaprasad et al. conducted a meta-analysis and synthesis of HIPAA to formulating ill-structured problems by providing illustrative components and glossaries [21], [22]. They presented a comprehensible horizontal view ontology divided into five dimensions and revealed current research status done on combinations of components in each dimension. However, the research was more likely for a literature review than an ontological approach to regulatory compliance.

2) **Access Control:** Maithilee Joshi et al. proposed Attribute-Based Encryption (ABE) for cloud EHR using the "Organizational Knowledge Base" knowledge graph partially derived from the HIPAA ontology mentioned above [9], [23], [24]. The knowledge graph stores the roles and attributes of the medical organization’s different stakeholders and various relationships. Based on the user’s attributes specified in the knowledge graph, the Access Broker unit grants secure data access by semantic reasoning. This study made it possible to conduct much more flexible access control based on the characteristics of the data and the person who wants access, considering the subtle and sensitive nature of healthcare data.

**B. Semantic Web**

We have used semantic web technologies to develop the HIPAA ontology and reasoning with policy expression. Semantic web tools enable data to be annotated with machine-understandable meta-data, allowing the automation of their retrieval and their usage in the correct contexts. Semantic web technologies include languages such as Resource Description Framework (RDF) [25], and OWL [26] for defining knowledge graphs and the metadata, as well as tools for reasoning over these descriptions. Our most fundamental requirement is for a representation that supports interoperability at both the syntactic and semantic levels. OWL has well-defined semantics grounded in first-order logic and model theory, allowing programs to draw inferences with the assurance that the subsequent interpretation is sound. An important advantage for OWL over many other knowledge-representation systems is that it has well-defined subset profiles guaranteeing sound and complete reasoning with various levels of reasoning complexity and designed to work with popular implementation technologies.

**IV. HIPAA Knowledge Graph**

This section describes the knowledge graph that we have developed in OWL [26] to represent the Health Information, HIPAA Privacy Rules, HIPAA Security rules, and the main stakeholders. The HIPAA ontologies from previous research provide us abstract views of HIPAA in hierarchical and horizontal structures [9], [21], [22]. However, they do not include the dynamics between the provisions and stakeholders. The other studies that focused on access control based on a health IT ontology were not easy to generalize because their solution was made to fit into the specific environment [23], [24].

We enhanced the HIPAA knowledge graph from our previous work [9]. Our HIPAA knowledge graph incorporated relationships between HIPAA rules and stakeholders and refined hierarchies of provisions to make it more universal. We reviewed the HIPAA Privacy [27] and Security Rule [28] summary to make the ontology more semantically rich and add more details. For example, twelve new sub-classes are added under the "PublicInterestAndBenefitActivities" class, and object properties are introduced, such as "useAndDisclose," "provideService," and "enterInto." As a result, new relationships between the classes were defined, and sub-classes were added to the ontology’s leaf nodes. We used the Protégé software to build our knowledge graph [29].

Although several other studies have established HIPAA Ontology [30]–[33], the difference of our study is that our ontology mirrors HIPAA official regulation. Past studies may have referred to the HIPAA document, but they adopted a top-down approach to develop HIPAA ontology tailored to the final system they wanted to develop. Therefore, their HIPAA ontology has become weaker in expressing HIPAA's unique characteristics as it is developed according to the ready-developed framework or modified to fit into a system’s goal.

Figure 2 illustrates the high-level view of our proposed knowledge graph that includes more classes to represent

---

**TABLE I**

**SYSTEMATIC LITERATURE REVIEW RESULT**

<table>
<thead>
<tr>
<th>Index</th>
<th>Title</th>
<th>Author</th>
<th>Year</th>
<th>Journal/Conference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Attribute based encryption for secure access to cloud based EHR systems</td>
<td>Joshi, Maithilee and Joshi, Karuna and Finin, Tim</td>
<td>2018</td>
<td>CLOUD</td>
</tr>
<tr>
<td>2</td>
<td>The bright, light, and blind/blank spots in HIPAA research: An ontological analysis</td>
<td>Ramaprasad, Arkalgud and Syn, Thant and Win, Khin Than</td>
<td>2015</td>
<td>HICSS</td>
</tr>
<tr>
<td>3</td>
<td>Delegated authorization framework for EHR services using attribute based encryption</td>
<td>Joshi, Karuna Pande and Yesha, Yelena and Finin, Tim and others</td>
<td>2019</td>
<td>TSC</td>
</tr>
<tr>
<td>4</td>
<td>An ontology for a hipaa compliant cloud service</td>
<td>Ramaprasad, Arkalgud and Syn, Thant and Win, Khin</td>
<td>2015</td>
<td>ICACON</td>
</tr>
<tr>
<td>5</td>
<td>Ontological meta-analysis and synthesis of HIPAA</td>
<td>Ramaprasad, Arkalgud and Syn, Thant and Win, Khin</td>
<td>2014</td>
<td>PACIS</td>
</tr>
</tbody>
</table>
provisions in detail. ‘HealthInformation’ is the new class that represents health data covered by HIPAA. The ‘Identifier’ subclass is an enumerated class consists of 18 HIPAA identifier individuals. The ‘ProtectedHealthInformation’ sub-class is equivalent to “HealthInformation and (hasIdentifier some Identifier)” which means that if a piece of health information includes at least one individually identifiable information, it becomes Protected Health Information (PHI).

In this design, we have added many new sub-classes under the ‘HipaaPrivacyRule’ class to include the policy changes since our previous version. The privacy rule regarding access control comprises three parts – “Required Disclosures,” “Permitted Uses and Disclosures,” and “Authorized Uses and Disclosures.” Since the categories under each part do not share the exact requirements, many new sub-classes in the ‘HIPAAPrivacyRule’ class manifest these differences.

Figure 1 illustrates the object properties introduced in the refined ontology to provide an overview of the relationship between stakeholders and provisions. Object properties between ‘CoveredEntities,’ ‘BusinessAssociates,’ and ‘BusinessAssociateContract’ represent that BA provides service to Covered Entities (CEs), and they have liability entered into a BAC.
The other classes, ‘AuthorizedUseDisclosure,’ ‘Required-Disclosures,’ and ‘PermittedUsesAndDisclosure,’ represent the significant three parts of HIPAA privacy rule that apply to CEs. They have transitive characteristics, so sub-classes of the domain classes also can be a domain of the properties. For example, ‘Marketing’ sub-class under ‘AuthorizedUseDisclosure’ class also has ‘needsAuthorization’ relationship with ‘CoveredEntities’ class.

V. VALIDATION AND RESULTS

We have validated our knowledge graph’s design by applying it against the HIPAA use cases developed by HHS [34]. This section presents the results obtained by applying two use cases and reason over our design. We used the SPARQL Protocol and RDF Query Language (SPARQL) [35] to query our knowledge graph. We simulated a dataset for the validation, representing ten pharmacies and their BA law firms and law enforcement agencies with whom they share the data.

1) Pharmacy Chain Enters into Business Associate Agreement with Law Firm Covered Entity: This example covers the allegations that a law firm illegally disclosed PHI during an administrative proceeding on behalf of a pharmacy chain. There was no evidence of the allegation at the organizational level, but it turned out that they have not entered into a BAC.

The problem here is that the CE (the pharmacy chain) and the BA (the law firm) did not recognize the BAC's existence. The policy expression in SPARQL in figure 3 can prevent the problem. The BAC's absence may imply no PHI safeguard, procedures after a data breach, and other essential subjects to protect PHI. Therefore, compliance automation can guarantee if CE, BA, and possibly disclosed PHI is HIPAA compliant before considering organization policy.

Figure 3 also illustrates the result of the SPARQL BAC checking policy expression. To test the expression, we simulated ten pairs of pharmacy and law firm individuals. Each pair shares the same BAC which they have entered into, except two pairs. The Health Mart, King Spalding pair does not have BAC, and the Rite-aid, Schiff Hardin pair maintains different BACs. It was possible to filter out ineligible pairs who attempt to exchange PHI without proper BAC through the policy expression.

2) Pharmacy Chain Revises Process for Disclosures to Law Enforcement: The “Permitted Uses and Disclosures” may mislead the CEs and BAs by its name. In this example, chain pharmacies disclosed PHI to municipal law enforcement officials. It may seem that there is no problem on the surface because it is permitted disclosure. However, it becomes permitted disclosure only when the law enforcement requested data in writing or the circumstance satisfies the other five conditions specified in HIPAA.

To reflect these subtle differences, it is necessary to examine whether the conduct between organizations is valid through a semantically rich knowledge graph and clear policy expression. HIPAA compliance expression in figure 4 can prevent this situation.

Figure 4 illustrates the filtering result of the policy expression. The test set includes ten pharmacies and ten municipal law enforcement official. Every pharmacy has a written request from the law enforcement official except the Family Pharmacy, Bel Air Police Department pair, and Walgreens, University Park Police Department pair. As a result, the result filtered out those two pharmacies and law enforcement agencies.

VI. CONCLUSION

We have created a semantically rich knowledge graph to represent the knowledge in the official HHS HIPAA Privacy Rule, and Security Rule summary [27], [28]. We validated the ontology against official HHS HIPAA use cases. It also illustrated the feasibility of safe healthcare data exchange with real-time HIPAA compliance automation. Our knowledge
graph, which is available in the public domain, is flexible to incorporate future revisions of HIPAA with minimal effort.

Future Work

As part of our future work, we will have HIPAA experts also validate our knowledge graph. We will also expand this work by integrating the knowledge graph with the Attribute-based access control framework, which considers organizations’ policy and attributes of users and data they want to access together [23], [24]. This enhancement will help prevent illegal use and disclosure of health information by adding a layer of reasoning for the preemptive intervention of HIPAA compliance before the organization’s policy enforcement. Also, we will revise and validate our ontology by subsequent case studies in a real-life environment. Our final goal is to facilitate secure healthcare data exchange by compliance automation, leading to a healthy healthcare big data ecosystem and patient rights enhancement.

References


[18] Research and Markets, “Healthcare cloud computing market by product (emr/ehr, pubs, vna, phm, telehealth, rcm, crm, fraud management), service (saas, ias), deployment (private cloud, hybrid cloud), pricing (pay as you go), component (software) - global forecast to 2024,” 2019.


