The Digital Age and the Insufficiency of HIPAA in Protecting Virtual Data

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Surabhi Ghatti

Spring semester 2022

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Signature Surabhi Date 07 May 2022
Surabhi Ghatti

Approved ______ Date 04 May 2022
Richard D. Jacques, Ph.D. Department of Engineering & Society

INTRODUCTION

Data guide the knowledge that society has, especially in medicine and healthcare. Healthcare is an ever-evolving field of continuous collaboration and innovation. This innovation and progress primarily stem from the increase in data needed to understand new phenomena. This increase in data has led to much computational innovation, coupled with exponential increases in computing power and processing capabilities. Some form of computation or program participates in every aspect of design and problem-solving. However, with such innovation comes increased risks of data exposure and breaches.

With such an increase in healthcare data, there was a need for standardized and controlled dissemination of patient information. As a result, The Health Insurance Portability and Accountability Act (HIPAA) of 1996 was established. The Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 and the HIPAA Omnibus Rule of 2013 cemented these frameworks that were instated to control and reduce any possible breaches of confidentiality and integrity of electronic personal information (Patil & Chakrabarti, 2021). However, with such a strict definition of protection, there are broad areas of information dissemination not explicitly covered in the bylaws. In such cases, there are severe breaches in patient privacy (Gostin & Nass, 2009).

In biomedical research, the use of patient data requires clearance through the Institutional Review Board (IRB), which includes a portion about HIPAA and its privacy rule. Though this protects patients' rights and proper handling in any medical cases, the ease with which this data is shareable is astonishing. It is up to the person with access to ensure trustworthy practices and abide by the restrictions endorsed by HIPAA. The Omnibus Rule, which was meant to cover any gaps in accessing stored data, was almost added a decade ago. Since its addition, the predominance of technology in every sector of the world, especially healthcare, has increased exponentially. This invisible network of data has expanded the possibilities for collaboration and advancement, while revealing more pathways for breaches and unlawful access. HIPAA clearly requires amendments that would help make data less susceptible to privacy breach risks, while also providing sufficient leeway for research studies (Patil & Chakrabarti, 2021). To determine how HIPAA can be reformed, this STS research will delve into the evolution of

HIPAA and investigate real-world examples in how this regulation has handled the electronic sharing of patient records. After determining areas of improvement, the discussion will look into data currently being protected and then propose a pathway for updating HIPAA to handle the evolution of Identifying areas that need changing is the first step to advocating for change and then eventual policy addition through legislation. Reform to HIPAA is essential for future advancements in medicine, while ensuring that there is controlled adaptability to the evolving technological world.

LITERATURE REVIEW

A Brief History of HIPAA

Before HIPAA was introduced, multiple federal regulations contained information protecting patient privacy; however, these rules were not consolidated and lacked the clarity needed to provide sufficient coverage. Recognizing the need for a set of uniform regulations, HIPAA was passed into law on August 21st, 1996, by the federal government to protect the confidentiality of medical records. In addition to covering privacy and security, HIPAA addresses five different sections, known as titles, when discussing confidential records. The one most commonly known by health-care professionals is Title II. This contains a provision that requires the U.S. Department of Health and Human Services (DHHS) to govern practices for electronic healthcare transactions, to protect the security and privacy of health data, and to help prevent health-care fraud and abuse. The DHHS is the designated body responsible for maintaining the safety of health records. These rules apply to what are known as covered entities, which by definition include health-care providers, health plans, and health-care clearinghouses (Moore & Frye, 2019).

Within the last two decades, important provisions to the initial law have been made, making it what we know today. In 2003, HIPAA-covered entities were required to comply with the HIPAA privacy rule- a regulation designed to “meet the pressing need for national standards to control the flow of sensitive health information and to establish real penalties for the misuse or improper disclosure of this information”(Choi et al., 2006). After this addition, in 2009, the HITECH Act, which supported the nationwide implementation of electronic health records (EHR) for integrated patient care and access, helped increase the prominence of HIPAA. As part of the federal government's 2009 economic stimulus package, HITECH also established payment incentives to help facilitate the transition from paper to electronic records (Harkins & Freed, 2018). With these EHRs, new concerns about patient privacy emerged. With this, the DHHS implemented the Omnibus Rule of 2013, which expanded the definition of the covered entities to include business associates in the hope that more parties would be responsible for maintaining the integrity of protected health information (PHI) (Moore & Frye, 2019). Examples of business associates include third-party administrators who process claims for health plans and managers who manage insurers' pharmacy networks and hospitals' consultants (Yaraghi & Gopal, 2018).

HIPAA in Action

A case study conducted by the Office of Civil Rights (OCR) in 2017 attempted to analyze the impact of the Omnibus Rule in reducing the frequency of medical data breaches. The data, gathered between October 2009 and August 2017, has shown a clear trend in this federal policy enhancing privacy protection efforts among business associates. However, a growing number of breach incidents are being seen in covered entities. There is no concrete evidence providing a reason for this growth. However, from a pure-definition standpoint, the primary mission for business associates is not to provide patient care. They have other specialized functions, which allow them to focus on providing sufficient training and building more-than-adequate IT infrastructure, unlike main healthcare institutions. This inference sheds light on the ease with which ransomware attacks are growing in sophistication (Yaraghi & Gopal, 2018).

Ransomware is a type of malware typically performed using a phishing email- an email with a normal looking attachment. The attachment, once opened, will encrypt critical files on the user's machine and once the encryption is complete, a display with instructions on the way to remit the ransom is shown, typically paid in bitcoins. With the rise of cryptocurrency in the early 2010's, it has become quite easy to conduct untraceable monetary transactions. Adding to the anonymity, ransomware-as-a-service (RaaS) has gained a strong footing. It has become one of the most popular forms of cyber extortion due to its notoriety and potential for high payout, now especially with the vast amount of data contained in the healthcare sector (Harkins & Freed, 2018).

Today, PHI data are more valuable on the black market than personal information from financial institutions. According to the Federal Bureau of Investigation (FBI) Cyber Division, PHI on the black market sells at a rate of $50 for each partial record as compared to the $1 for social security numbers or credit card numbers. Sadly, even with EHR being such valuable information, the healthcare sector is quite unprepared for the new cyberage (Harkins & Freed, 2018). In HIPAA, specifically dealing with the privacy policy, institutions must implement both equipment and administrative safeguards to help protect PHI and properly prepare in the event of a breach. This definition provides a range of freedom for each party in deciding what would be the best method to prevent potential breaches (Moore & Frye, 2019). A study conducted in 2015 by ABI research shows that medical identity theft and fraud are on the rise and healthcare providers are unable to cope with the data breaches leaking millions of personal records. Yet, compared to other industries, healthcare has the least amount of expenditure on cybersecurity. Cybersecurity spending, as projected by ABI Research back in 2017, only accounts for $10 billion dollars, less than 10% of the total spending on critical infrastructure security. In fact, another study conducted in 2015 by the Healthcare Information and Management Systems Society showed that 64% of individuals with some responsibility in information security had experienced a security incident within their healthcare organization. These professionals, so specialized in their clinical disciplines, are not as well trained in security awareness showing the gaps in the regulations enforced by HIPAA (Harkins & Freed, 2018).In fact, in an investigation of breaches reported by the US Department of Health and Human Services (HHS), 44% of the breaches were from accidental misses and 56% were malicious missuess. This almost 50-50 divide between accidental and malicious clearly highlights the need for both more thorough infrastructure and a more robust training program for designated specialists (Kafali et al., 2017).

HIPAA Definition Limitations

Furthermore, HIPAA doesn’t necessarily apply to the types of information shared; its definition and restrictions only extend to the covered entities and business associates as defined by the DHHS.

Consequently, healthcare data generated by non-covered entities are not protected by HIPAA- data ranging from posts made online to uploaded fitness tracker data. This concept was clearly brought to light in the report detailing the Facebook and Cambridge Analytica incident in which Facebook was approaching healthcare organizations to get deidentified patient data to link the data to individual Facebook users. It is important to recognize the complete invasion of privacy to individuals who don't even know that their PHI is being mined and associated with them on simple platforms they use for entertainment (Moore & Frye, 2020). This easy access to data poses unforeseen threats to millions of Americans, unaware that their personal medical data may be compromised.

Differential Privacy

Due to the importance of privacy and the necessity brought on by growing data mining and analyzing tools, there has been extensive research in preserving such privacy. In data mining, the more specific the information, the easier it is to collect and associate. For personal data, removing identifying attributes such as a name won't impact the data mining, but there is still the possibility of sensitive information leaking due to linking attacks based on public attributes of the data, known as quasi-identifiers. This susceptibility has led to research in privacy-preserving data mining (PPDM) and privacy-preserving data publishing (PPDP). PPDM allows data mining models to learn while controlling the disclosure of data about individuals. In PPDP, datasets are anonymized to allow for data disclosure without violating privacy (Clifton & Tassa, 2013).

One of the first formal PPDM models to gain momentum in the scientific community was k-anonymity, where k refers to the number of attributes for a given record in a database. The model requires that each released record be indistinguishable from at least *k-1* other records when considering the quasi-identifier attributes. Due to its arbitrary condition for randomizing, k-anonymity is great at preventing identity disclosure, but it is more susceptible to attribute disclosure. With enough given attributes, any hacker or mal-intention user can work backwards and find the individuals matching said attributes (Domingo-ferrer & Torra, n.d.). Given these shortcomings, there has been a lot of research into other variant models. These models use a concept of syntactic anonymity, in which they generalize the database entries until some syntactic condition, so that a malicious attack to link a quasi-identifier type to sensitive values is highly restricted. HIPAA's privacy rule currently uses syntactic models to describe how to implement privacy measures for data sharing and collection. However, these models seem as susceptible to attacks as k-anonymity (Clifton & Tassa, 2013).

Differential privacy, a rigorous notion of privacy based on adding noise to answers to queries on the data, has revolutionized the field of privacy-preserving data mining (PPDM). Unlike statistical queries, the added random noise adheres to a mathematical definition of database privacy. In this definition, an algorithm is differentially private if the addition or removal of a single individual from a given database alters the likelihood of output by an imperceptible amount. In differential privacy, the privacy is controlled by a parameter ε. The parameter quantifies the information leakage resulting from the addition or removal of a single participant; a smaller value of ε will result in more noise added to the query, i.e., adding more privacy (*Differential Privacy in Health Research: A Scoping Review | Journal of the American Medical Informatics Association | Oxford Academic*, n.d.). Moreover, this privacy model makes almost no assumptions about a potential attacker's background knowledge. There is widespread belief that differential privacy and related models are capable of withstanding attacks that syntactic anonymity is prone to, such as linkage, differing, and database reconstruction attacks (Clifton & Tassa, 2013; *Differential Privacy in Health Research: A Scoping Review | Journal of the American Medical Informatics Association | Oxford Academic*, n.d.).

There are fundamental differences between the concepts of PPDP and PPDM. In PPDP, the goal is to publish the data in anonymized manners without any assumption on how queries will be made, or another analysis will be done. For example, a hospital that wishes to release data about its patients in cases of public scrutiny but still wants to maintain anonymity is only concerned about publishing the data for any sort of analysis to be performed. In PPDM, which is what differential privacy targets, the query that will be done using the data must be known before applying the privacy-preserving princess. In a normal scenario, the data custodian, the person in charge of the data, maintains control and doesn't publish it. Instead, the data custodian provides queries that users can use to conduct the analysis they need to conduct on the data while also maintaining the privacy of the individuals in the database. The importance of knowing the query is so that the noise level to the global sensitivity and the value of the parameter ε can be calibrated (Clifton & Tassa, 2013).

Differential Privacy in Application

In 2018. The U.S Census Bureau announced that it would protect its publication of the 2018 End-to-End Census Tests (E2E) using differential privacy. When conducting internal research, the statistical disclosure limitation systems used in the 2000 and 2010 Census had serious vulnerabilities exposed by the database reconstruction theorem. The E2E was a dress rehearsal for the 2020 Census, designed to test the differential privacy publication system. With this new method, no real data gathered by the 2020 Census were released to the public, but the results of specific queries were (Abowd, 2018).

Since the last couple of years, the discussion on adopting differential privacy for healthcare data and research has grown in prominence. Australia, in particular, has faced several data breaches with regard to health data within the last two decades. A graph from South Australia Health reported data on children treated at hospital for respiratory infection, gastroenteritis, and whooping cough from 2005-2018. Unknown to the creators of the graph, it was linked to the source data, providing illegal access to names, dates of births, and test results. In another serious breach, data provided by individuals to the Red Cross Blood Service between 2010 and 2016 were stored in a file. This file was then transferred to an unsecure device and therein accessed by an unauthorized party, such devastating breaches led to the publishing of “Process for Publishing-Sensitive Unit Record-Level Public Data as Open Data", which provides detailed information on releasing datasets containing sensitive information. Furthermore, the Commonwealth Government Privacy Amendment (Re-identification Offense) Bill 2016 was also passed to make it a punishable offense to re-identify information disclosed by the commonwealth agency. Current strategies to ensure data privacy include de-identification, aggregation, authentication, authorization, and encryption. However, each of these is limited and has shown multiple shortcomings. Adopting the idea of differential privacy, Australia developed a system known as the COVID-19 Real-Time Information System for Preparedness and Epidemic Response (CRISPER). Differential privacy is useful for datasets with substantial amounts of data and variables. Furthermore, in addition to providing additional security, the primary datasets can remain with the data custodian, individuals responsible for the original data, so that data doesn't necessarily have to be released to be accessed (Dyda et al., 2021).

Differential Privacy: Not for All?

Even with differential privacy being so renowned, there are some limitations to the use of this algorithm. Differential privacy is less suitable for data where there is a small amount of data or when exact counts are needed (Dyda et al., 2021). Since its introduction in 2006 and the footing it has since gained, differential privacy has started to make other privacy models seem obsolete. However, there are often unstated assumptions and ignored conditions in which differential privacy will fail. In this way, syntactic models may not be as useless as they portray them to be (Clifton & Tassa, 2013).

The HIPAA "Safe Harbor" rules specify legally acceptable syntactic anonymization. This means that there are rules stating what sort of identifying information must be removed, giving the algorithm enough information on how to determine the value of *k* that will help set the level of anonymity that the rules are trying to achieve. However, with differential privacy, although the issue of identification is adhered to, it is difficult to quantify the value of ε because it doesn't reflect what might be traditionally defined as privacy. In fact, ε is a measure of the impact an individual has on the result, but not what is disclosed about an individual. Furthermore, differential privacy makes an assumption that the individual is independent, a statement that becomes problematic when dealing with relational learning, where values of an individual can influence what is learned about another. Syntactic models do not have such built-in assumptions (Clifton & Tassa, 2013).

Syntactic models are designed for privacy-preserving data publishing, while differential privacy is used for privacy-preserving data mining. One cannot replace the other. It is also important to recognize the tradeoff between privacy and utility. Syntactic models reduce privacy risk while preserving the utility of information communication by guaranteeing corrections on the analysis of the anonymized data. Differential privacy, due to its non-compact uncertainty, can have the quality of its results vary greatly.

The smaller ε, the greater the privacy; however, this greatly reduced the accuracy of the algorithm (Clifton & Tassa, 2013).

CONCLUSION

With easier access to technology and the automation of many processes, data sharing and mining have become quite vital. With the ease of accessing data that technology and software have brought, there has been an exponential rise in ways in which the data can be exploited, and users can be victimized. In the health-care sector collaboration is key for advancement. The introduction of HIPAA in 1996 provided a great foundation for protecting patient privacy in the U.S. With its 2003 Privacy Rule, 2009 HITECH act, and 2013 Omnibus Rule, HIPAA provided a clear definition on how the dissemination of patient information should be done and the infrastructure needed by responsible entities to ensure security. However, since its last major update in 2013, the way data are stored and accessed has evolved quite rapidly. There are significant gaps in the regulation's definitions, which has led to numerous privacy breaches, exposing millions of American lives. To begin approaching this gap in coverage, differential privacy, a PPDM technique, should be introduced into the bylaws of the privacy law. This new legislature, which discusses this state-of-the-art security algorithm, will supplement the syntactic model encryption design already included within the law. The combined use of these algorithms will help provide for a more robust infrastructure for patient privacy protection and will also help with the protection of data held by non-covered entities. HIPAA has revolutionized how healthcare is provided and research is conducted. It is vital that it continues to adapt to the ever-changing digital age to continue to provide concrete standards for privacy protection while also encouraging the advancement of the healthcare industry.

REFERENCES

Abowd, J. M. (2018). The U.S. Census Bureau Adopts Differential Privacy. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2867. https://doi.org/10.1145/3219819.3226070

Choi, Y. B., Capitan, K. E., Krause, J. S., & Streeper, M. M. (2006). Challenges Associated with Privacy in Health Care Industry: Implementation of HIPAA and the Security Rules. *Journal of Medical Systems*, *30*(1), 57–64. https://doi.org/10.1007/s10916-006-7405-0

Clifton, C., & Tassa, T. (2013). On syntactic anonymity and differential privacy. *2013 IEEE 29th International Conference on Data Engineering Workshops (ICDEW)*, 88–93. https://doi.org/10.1109/ICDEW.2013.6547433

Differential privacy in health research: A scoping review | Journal of the American Medical Informatics Association | Oxford Academic. (n.d.). Retrieved April 18, 2022, from https://academic.oup.com/jamia/article/28/10/2269/6333353?login=true

Domingo-ferrer, J., & Torra, V. ,. (n.d.). *A Critique of k-Anonymity and Some of Its Enhancements (Invited Paper)*.

Dyda, A., Purcell, M., Curtis, S., Field, E., Pillai, P., Ricardo, K., Weng, H., Moore, J. C., Hewett, M., Williams, G., & Lau, C. L. (2021). Differential privacy for public health data: An innovative tool to optimize information sharing while protecting data confidentiality. *Patterns*, *2*(12), 100366. https://doi.org/10.1016/j.patter.2021.100366

Gostin, L. O., & Nass, S. (2009). Reforming the HIPAA Privacy Rule: Safeguarding Privacy and Promoting Research. *JAMA*, *301*(13), 1373–1375. https://doi.org/10.1001/jama.2009.424

Harkins, M., & Freed, A. M. (2018). The Ransomware Assault on the Healthcare Sector. *Journal of Law & Cyber Warfare*, *6*(2), 148–164.

Kafali, O., Jones, J., Petruso, M., Williams, L., & Singh, M. P. (2017). How Good Is a Security Policy against Real Breaches? A HIPAA Case Study. *2017 IEEE/ACM 39th International Conference* *on Software Engineering (ICSE)*, 530–540. https://doi.org/10.1109/ICSE.2017.55

Moore, W., & Frye, S. (2019). Review of HIPAA, Part 1: History, Protected Health Information, and Privacy and Security Rules. *Journal of Nuclear Medicine Technology*, *47*(4), 269–272. https://doi.org/10.2967/jnmt.119.227819

Moore, W., & Frye, S. (2020). Review of HIPAA, Part 2: Limitations, Rights, Violations, and Role for the Imaging Technologist. *Journal of Nuclear Medicine Technology*, *48*(1), 17–23. https://doi.org/10.2967/jnmt.119.227827

Patil, A. P., & Chakrabarti, N. (2021). A review into the evolution of HIPAA in response to evolving technological environments. *Journal of Cybersecurity and Information Management*, *Volume 4*(Issue 2 : Special Issue-RIDAPPH), 5–15. https://doi.org/10.54216/JCIM.040201

Yaraghi, N., & Gopal, R. D. (2018). The Role of HIPAA Omnibus Rules in Reducing the Frequency of Medical Data Breaches: Insights From an Empirical Study. *The Milbank Quarterly*, *96*(1), 144–166. https://doi.org/10.1111/1468-0009.12314