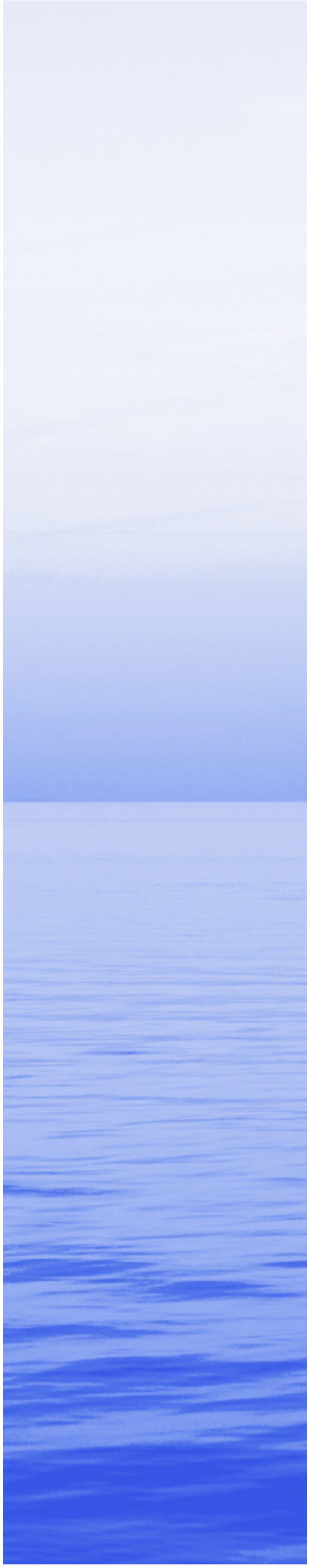
Clearwater

Whitepaper

The Privacy and Security Implications of Big Health Data Initiatives

Table of Contents

Introduction



In the 21st Century, personal data has become the new "gold rush" for the health care industry. The term "big data" is used to describe large amounts of a variety of data that is quickly available for analysis.

As you read this article, massive amounts of data are being analyzed, transferred, and collected by healthcare organizations to improve clinical care outcomes, reduce costs, explore new medical discoveries, and in some cases, monetize the data. Some research suggests that medical care is under-analyzed, which results in ineffective treatment, waste, and medical errors.[1] Perhaps in response to similar studies, big data initiatives are being used to make changes in care quality, develop research protocols to advance medical care, compare the effectiveness of different treatment interventions, and monitor medication, device, and treatment interventions and outcomes.[2]

To leverage big data, large health systems, universities and academic medical centers, and health plans have implemented separate data warehouses, sometimes referred to as "data lakes" for unrefined data, to receive and store vast amounts of data. Common reasons to create data warehouses including the sharing of information for research and public health reporting purposes. In addition to these purposes, data stored in warehouses may be used for other purposes, including data analysis to provide business insights and decision making, problem solving, and business cost savings. The types and sizes of projects, and the types and sizes of data sets vary, creating potential legal, regulatory, and ethical risk to both the healthcare organization and data warehouse providers.

Organizations may miss regulatory and ethical issues at the beginning of a data warehouse project often due to the complexity of regulatory requirements and technical processes. As a result, organizations are at risk for unintentionally implementing only basic controls that do not comprehensively protect and secure the data. This can create a situation in which risks to the data and the organization are exploited by malicious cyberattacks, including ransomware, or other types of unlawful disclosures. Big data means big risks to patient privacy, the organization, and the security of the data.

[1] Institute of Medicine. Committee on Quality of Health Care in America, the National Academies. To Err is Human: Building a Safer Health System (eds. Kohn LT, Corrigan JM, & Donaldson MS, National Academies Press, Washington, D.C., 2000)

[2] del Hoyo-Barbolla E, Lees D. The use of data warehouses in the healthcare sector. Health Informatics Journal. 2002;8(1):43-46. doi:10.1177/146045820200800108

Big data privacy and security risks

Understanding privacy risks

Data warehouses often store different types of data. For example, a warehouse may receive protected health information (PHI) but also receive information that is not subject to HIPAA Rules, such as personally identifying information, which may be subject to other data privacy rules enacted and implemented by states or other countries. When considering risks and requirements to data, both organizations and warehouses (if separate) should address risk-related issues from a data governance perspective: consider the type(s) of data collected, what rules apply to data sets, and understand how to limit the collection and access of data to the minimum necessary amount of information needed to complete a project. Additionally, to comply with the varying requirements of state and international data privacy laws, organizations should be aware of requirements to limit the use of the data to the purpose for which the data was collected. For example, the European Union General Data Protection Regulation (GDPR)[3] requires data controllers to limit the use of the data to the purpose for which the data was collected unless an exception applies.[4]

Compliance with data privacy laws often includes the requirement for transparency about how the organization will use or disclose the data, which usually takes the form of a privacy notice. Organizations operating in multiple jurisdictions may need to consider addressing notice requirements that comply with several laws or regulations. For example, organizations storing data in warehouses for future use for research and other noble purposes will need to consider whether the notice must inform individuals about research activities.[5]

Organizations maintaining data warehouses should also be aware of privacy rights that data subjects may exercise. Privacy rights often include the ability for the data subject to request access to or copies of the information, request corrections to the information the organization maintains, request information about 3rd parties that received the information, and in some cases, request that the organization delete or erase the data. Meeting the varied requirements described in state and international data privacy laws is complex and requires strict attention to detail. If the warehoused data is not copies of information owned and maintained in other locations, data warehouses should consider implementing processes and procedures to receive

[3] https://gdpr-info.eu/

[4] https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/principles/purpose-limitation/

[5] https://privacyruleandresearch.nih.gov/clin_research.asp

and respond to such requests. Failing to do so could lead to regulatory or legal complaints that result in monetary penalties[6] and reputational damage.

It is not uncommon for warehouses to collect and maintain more data than needed for an end project, creating additional risk and underscoring the importance of thoughtfully considering the lifecycle of the data. Consider the following situation:

Researchers love data. The more data that a research team can acquire or access, the happier they are because it is the data that ultimately serves as the foundation for research questions and studies. Data warehouses are considered ideal for allowing internal and outside research teams access to robust data because warehouses may receive and aggregate data from several sources. Warehouses may contain de-identified data, giving researchers the ability to access and use the de-identified data without obtaining regulatory approvals for some data privacy laws (e.g., HIPAA Privacy Rule).

Privacy questions to consider include:

- What type of data is needed for the project?
- Why is the data needed?
- Who will (or may) have access to the data?
- Are there legal or contractual requirements that prevent the collection or sharing of this data?
- Can we fulfill project goals by using data that is less sensitive and/or does not create unnecessary risk to people or the organization?
- Does any of the data require the provision of certain rights or special protections? How are individuals informed of how data about them will be used and disclosed?
- What is the minimum necessary amount of data that can be used or disclosed? Once identified, how will the organization validate compliance?
- If protected health information is required, what is the permitted purpose for both the use and the disclosure?
- Is there a valid business associate agreement in place, if required?
- Has the organization considered all relevant legal, regulatory, and ethical risks (ex. federal and state law requirements, current enforcement landscape, best practices, reputational risk, etc.)

[6] https://www.hhs.gov/hipaa/for-professionals/compliance-enforcement/examples/how-ocr-enforces-the-hipaa-privacy-and-security-rules/index.html

Demystifying De-identification

Throughout the healthcare ecosystem, there is confusion about what “de-identified” data really means. HIPAA provides specifications and requirements for de-identified data, but even when HIPAA does not apply, identifiable data can present risks. It is important that organizations understand when and how data may be properly de-identified, which means having robust policies, procedures, and training in place.

Ultimately, proper de-identification comes down to the type of data and the method being used. For PHI, this includes using either the “safe harbor” or the “expert determination” method.[7] To meet the “safe harbor” methodology, which is used most frequently, all “direct” identifiers must be removed prior to disclosure.[8] For example, an organization may not consider an IP address or a date of service as examples of identifiers, yet those identifiers are specifically included in the HIPAA de-identification standard. Other types of information, such as email addresses, certain images, and license plate numbers, are also specifically included in the standard, yet are often found in data believed to be properly de-identified.

Organizations may also choose to rely on a person with appropriate knowledge and experience with generally accepted statistical and scientific principles to properly de-identify data. The “expert determination” method is most often used by larger entities and for larger projects because it requires using experts, who may be costly, and the process may take time. Due to some of these possible burdens, the “safe harbor” methodology is often the default standard for de-identification, which raises other risks, because it must be performed methodically and requires the removal (sometimes manually) of a number of identifiers.

De-identification processes and questions should be addressed during the project start but also reviewed throughout the project to confirm the processes are being followed. This means that all people involved in the data lifecycle have some understanding of the requirements and expectations. For example, since de-identification can be performed manually or through automation, individuals interacting with the data must understand how de-identification is taking place and whether they have a specific role in the process. In other words, everyone involved should clearly understand what “de-identification” truly means, when it is required (usually due to the purpose of the data sharing and/or contractual requirements) and how procedures and practices are periodically audited and reviewed.

[7] https://privacyruleandresearch.nih.gov research_repositories.asp

[8] 45 CFR § 164.514(b)

Understanding Security Risks

It is common for 3rd parties to access, manipulate, transmit, and store data. As such, organizations should understand how the data is collected, accessed, protected, secured, transmitted, and destroyed to prevent future inappropriate use or disclosures.

Healthcare is all too familiar with the harmful outcomes associated with data privacy and security issues. The industry is a leader[9] in the number of breaches and cybersecurity attacks, exposing large numbers of data records and data subjects to risk.

Phishing schemes, ransomware and malware attacks, social engineering for credential theft, and insider threats are unfortunately common, and only becoming more prevalent. While these threats may be top-of-mind for healthcare privacy and security professionals, it is easy to forget about data sources outside of the electronic health record that may contain sensitive or protected data, such as data warehouses. Even when data has been disclosed for valid purposes, the receiving or processing entities, such as warehouses, must consider unique and appropriate administrative, technical, and physical security controls. Disclosing organizations may want or need to support these efforts because the warehouse or third party may lack sufficient resources or expertise to understand risks and requirements related to the data.

These types of repositories are perfect targets for attacks and frequently experience incidents because they often may not have the same regulatory and legal obligations that HIPAA covered entities and business associates have, and they may not have robust controls in place. This regulatory "gray space" is why it's critical to include privacy and security offices in the initial project discussions and throughout the project lifecycle. As mentioned above, discussions should begin by identifying the types of data needed and how that data will be used and disclosed by the entity that currently owns or maintains the data. Even if the data is de-identified (or anonymized, if permitted) before the disclosure, risk remains.

Questions that organizations, whether the disclosing or receiving entity, should consider:

- What administrative, physical, and technical controls are in place to protect and secure the data? Have those controls been implemented and documented appropriately?

[9] https://www.healthcaredive.com/news/healthcare-breach-costs/628344/

- If the organization is maintaining a variety of data, some of which requires certain protections, and other types that do not, what mechanisms are in place to appropriately secure the data and ensure data is not inappropriately comingled?
- What controls exist for transmission, storage, access, and destruction of the data?
- If de-identified data will be disclosed, which entity is responsible for proper de-identification? If the data is not protected by HIPAA, is the "anonymized" data truly anonymized? Can the data be re-identified?
- Who is responsible if there is an incident involving the data? Are multiple organizations responsible for different aspects of the incident assessment process?
- If the use or disclosure of the data involves a third party, is the organization periodically validating security practices of that entity?

While these questions will help drive decisions about the project and relevant controls, they may also be used to consider the structure of contractual agreements with vendors and business associates to ensure they understand requirements and expectations for privacy and security protections as the data is further used and disclosed downstream.

Who Is Responsible?

The question of who is ultimately responsible for protecting and security data in a warehouse is an important legal question, and it doesn't always have a clear answer.[10]

Considering the risks and requirements involving the data, including incidents affecting the data, usually depends on the origin of the data and data sharing requirements. The reality is that the data owner will usually have some liability and responsibility, even if the incident occurred downstream or outside of the entity.

When establishing relationships with partners, business associates, and vendors, organizations not only need to comply with relevant legal and contractual requirements, but they should consider roles and responsibilities. This includes

[10] https://www.hhs.gov/hipaa/for-professionals/special-topics/health-information-technology/cloud-computing/index.html

setting clear expectations and requirements for security and privacy controls, subsequent uses and disclosures, incident response and reporting, and other aspects related to the privacy and security of the data. These conversations are not always easy, but it is critical that these discussions happen early in the process. Waiting to have this conversation about expectations or requirements at the time of an incident adds unnecessary risk and tension to the already fraught situation. Organizations concerned about a third-party's practices or specific types of data sharing initiatives may want to consider whether there may be other ways to use or store data.Organizations, such as universities that are focused on research projects, often experience unique relationship challenges when data is transmitted throughout the organization, including in some cases, data repositories owned and maintained by the organization. In these situations, diverse data sets may be accessed by many people, such as clinicians, providers, faculty, and internal and external researchers, all of whom may have different relationships to the data and related processes. For example, these individuals may have access to several systems transmitting and receiving data and may be serving in various roles when interacting with the data. The more people with access to that data, the more risk exists for an organization, especially if those individuals haven't received appropriate training about when and for what purpose data can be used.

When an organization experiences an incident, the data type, jurisdiction, and contractual requirements will guide the assessment and decision-making. Organization must look at the type of data involved and whether it has certain reporting or notification requirements, by contract, rule, or law. Some organizations may have regulatory requirements, meaning the organization reports to regulatory agency(ies) and/or notifies affected individuals, and others contractual requirements, meaning the organization reports "upstream," to the entity that owns or maintains the data.

To comply with any applicable federal (such as HIPAA), state, and/or international laws, an organization may need to make incident assessments within a required time period,[11] report to specified regulators and/or the institutional review board, notify affected individuals, and take remediation steps. Each aspect of the incident assessment process is complex, and unfortunately, there may not be an easy answer or solution.

[11] https://www.hhs.gov/hipaa/for-professionals/compliance-enforcement/agreements/presence/index.html

This is another reason why it is vital that all key personnel are involved in the project lifecycle. Even within smaller organizations, there are often other departments, such as IT/security, privacy, compliance, legal, procurement, clinical, and other staff, who can and should be provided with opportunities to support the process. Consider working with subject matter experts—professionals and legal counsel—to supplement and support data warehouse efforts and relationships with third parties, so that appropriate and well-considered controls are in place.

Conclusion

Healthcare organizations and associated data warehouses should thoughtfully consider whether policies, procedures, and training are in place to protect and secure the data throughout the lifecycle. While there are clear benefits to data sharing initiatives, risks to the individual, to the organization, and to others should be considered. Organizations and key personnel should review those risks against business needs not only when the project starts and ends, but throughout the project. Relationships between healthcare organizations, such as between providers or health plans and associated data warehouses can be complicated, but when the right people, processes, and technology are in place, organizations on both sides of the relationship/data lifecycle can feel confident in advancing important data sharing initiatives.

Clearwater is the leading provider of Enterprise Cyber Risk Management and HIPAA compliance software and consulting services for the healthcare industry. Our solutions enable organizations to gain enterprise-wide visibility into cybersecurity risks and more effectively prioritize and manage them, ensuring compliance with industry regulations. Clearwater's suite of IRM|Pro® software products and consulting services help healthcare organizations to avoid preventable breaches, protect patients and their data, and meet OCR's expectations, while optimizing cybersecurity investments. More than 400 healthcare organizations, including 64 of the nation's largest health systems and a large universe of business associates that serve the industry, trust Clearwater to meet their information security needs.

Have questions regarding this white paper? Engage with Clearwater for a specific discussion about your organization's approach.

- ClearwaterSecurity.com/Contact