Joseph Annand

DSE6003

Assignment 1

For a business that analyzes, manages, and distributes personal health information (PHI) for informing business decisions, developing a thorough data privacy and security plan that satisfies legal and ethical standards is essential to protecting the company’s assets from attackers and shielding the company from lawsuits by the government or individuals whose PHI was compromised under the company’s watch. A data security plan must include detailed descriptions for how data is stored, accessed, analyzed, and distributed. These procedures and protocols must satisfy the requirements of the Common Rule and HIPAA, which includes de-identifying data and ensuring proper encryption of data for storage and transmission. Proper protocols for such activities are defined by HIPAA, including the Safe Harbor and the Expert Determination methods. These methods will be discussed in more detail about how they may be implemented in the company’s work. Ensuring data security also requires safety restrictions and other procedures when accessing sensitive data. Once data security policies are in place and practiced, maintain security by evaluating vulnerabilities and the risk associated with them as technology evolves and standards change. This document describes the many considerations and recommendations on how to de-identify, store, access, and transport sensitive data for a fictitious company that works with PHI.

De-identification of the data is the process of redacting or coarsening data to make it difficult to link an identified person to a record in a data release (Nissim, et. al., 2017, pp. 2-3). It is “designed to protect individual identity…while preserving some of the dataset’s utility for other purposes” (Garfinkel, 2015, p. 8). There are two methods defined by the HIPAA Privacy Rule for de-identifying PHI: the Safe Harbor method and the Expert Determination Method. The Safe Harbor method consists of removing eighteen identifiers, including the name, email address, birthdate, full ZIP code, telephone number, and full-face photographs (U.S. Department of Health and Human Services, 2012). While the Safe Harbor method is a straightforward, reliable way to ensure that data is de-identified according to the HIPAA Privacy Rule, depending on the objectives of the analysis of the data, removing all these identifiers might limit the utility of the data. In such cases, the Expert Determination method can be used instead. This method is defined as “a person with appropriate knowledge of and experience with generally accepted statistical and scientific principles for rendering information not individually identifiable” determines when “applying such principles and methods…the risk is very small that the information could be used, alone, or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information; and documents the methods and results of the analysis that justify such determination” (U.S. Department of Health and Human Services, 2012, p. 7). While there are no specific criteria defined by the HIPAA Privacy Rule for who an expert is or what specific qualifications they must possess, it is important that an expert should have “relevant professional experience and academic or other training…as well as actual experience…using health information de-identification methodologies” (U.S. Department of Health and Human Services, 2012, p. 10). All experts within the company will be granted such a title and authority by the Chief Information Officer. Experts may include, but are not limited to, data managers, data analysts, data scientists, engineers, and analytical scientists with sufficient qualifications as determined by the Chief Information Officer. Experts may employ a variety of techniques for de-identification that maximize the utility of the dataset while minimizing the risk of re-identification. The National Institute of Standards and Technology and The Department of Health and Human Services describe a list of acceptable methods for the de-identification of data containing PHI.

De-identification methods for PHI under the Expert Determination method can be grouped into a few categories: suppression, generalization, perturbation, and k-anonymization. A process called pseudonymization may also be employed, but this method has a few additional stipulations for it to meet legal standards. Suppression is the removal or elimination of certain features of the data prior to releasing it (U.S. Department of Health and Human Services, 2012). Which records and how many of the records to suppress will be dependent on the amount of risk of re-identification for the records within the data set. Entire features may be suppressed if “a substantial quantity of records is considered as too risky,” while individual records may be deleted if they are too risky to be shared or specific values from an individual value may be suppressed, retaining some values from the record for analysis (U.S. Department of Health and Human Services, 2012, p. 19). In regard to the removal of direct identifiers, values can be replaced with symbols such as “XXXX” or “\*\*\*\*” to suppress them, or values can be replaced with random values, and if the same identity appears more than once in the dataset, then the identifier receives unique random value for each appearance (Garfinkel, 2015, p. 15). Generalization is the process in which data is transformed into a “more abstract representation” (U.S. Department of Health and Human Services, 2012, p. 19). This may include converting specific values to a broader value like a range or group of values. Additionally, specific values may be replaced with equally specific, different values, like a random value within five units of the original value, in the risk mitigation approach called perturbation. Another approach to de-identification of PHI is k-anonymization (U.S. Department of Health and Human Services, 2012, p. 20). For a dataset to be k-anonymous, there must be at least “k” matching records for every combination of quasi-identifiers (Garfinkel, 2015). Quasi-identifiers, such as birthdate, sex, and ZIP code, may be used for linkage attacks, in which attackers connect quasi-identifiers from a company dataset to a publicly available dataset with direct identifiers still intact to identify individual records from the company dataset (Garfinkel, 2015). However, as mentioned before, removing values may damage the utility of the dataset; therefore, it is important to utilize the above approaches to mitigating risk while considering how the use of each approach will affect the value of the dataset during analysis. Additional strategies in de-identifying quasi-identifiers are to exchange quasi-identifiers between records or release only a sample of a de-identified dataset (Garfinkel, 2015, p. 20).

Additional techniques to protect individuals’ privacy include pseudonymization, hash functions, and differential privacy. The implementation of these techniques may not be as straightforward as those previously mentioned and should be used in consultation with the Chief Information Officer when deemed necessary to use on a particular dataset or collection of datasets. Pseudonymization is a “special kind of transformation in which the names and other information that directly identifies an individual are replaced by pseudonyms” (Garfinkel, 2015, p. 16). Extra caution must be taken with pseudonymization as tit may be easily reversed if a table linking the original identities to the pseudonyms is kept or if the substitution of original identities “is performed using an algorithm for which the parameters are known or can be discovered” (Garfinkel, 2015, p. 16). In these cases or any case in which pseudonymization can be easily reversed, the data is not anonymous under the Common Rule (Garfinkel, 2015). This issue can be avoided, however, with an appropriate data use agreement that forbids the code key to be shared” (Garfinkel, 2015, pp. 16-17). Data use agreements should be drafted and submitted to the Chief Information Officer for approval upon the use of pseudonymization for company datasets. The risk of linking specific individuals to the pseudonyms created for them remains with pseudonymization. A process in which a dataset similar to the original is created with some or entirely all newly generated data elements that do not map to actual individuals is called synthetic data generation (Garfinkel, 2015). Synthetic data generation can be performed using differential privacy (Garfinkel, 2015). Differential privacy is a set of techniques that prevent identity disclosure by adding “non-deterministic noise to the results of mathematical operations” (Garfinkel, 2015, p. 7). It applies only to online query systems in its most basic form, but in addition to synthetic data generation, it may also be used to produce machine-learning statistical classifiers (Garfinkel, 2015). Although differential privacy can be effective at evaluating and reducing the risk of re-identification, it comes at the cost of decreased data utility (Garfinkel, 2015). The implementation of differential privacy in data workflows should be done in consultation with the Chief Information Officer. The use of hash functions is also a common de-identification technique for PHI (U.S. Department of Health and Human Services, 2012); however, only hash functions approved by NIST should be employed (NIST website).

Access to sensitive data should be limited to those who are trained and qualified to handle such data and who must use the data to perform their duties (Federal Trade Commission, 2015). Managers may request approval from the data manager for direct reports to receive access to databases with sensitive information. Depending on qualifications and relevancy, the employee may be granted limited access to a database containing sensitive information, such as restrictions on the data the employee may query. Employees will create strong passwords that meet specific criteria to inhibit hackers with password-guessing tools and keep these passwords safe per recommendation from the Federal Trade Commission (2015). Employees will only have a specific number of login attempts before being locked out their account for security reasons (Federal Trade Commission, 2015). To sign on to the company network and access databases, employees will complete a two-factor authentication. According to the Cybersecurity and Information Security Agency (2020), one study showed that multi-factor authentication blocked 100 percent of automated bot attacks, 99 percent of bulk phishing attacks and 66 percent of targeted attacks on Google users’ accounts.

To ensure that only authorized users access sensitive data, security measures must be employed during storage and transmission of the data. The Federal Trade Commission recommends using strong cryptography to secure confidential information, including transport layer security encryption for transmitting data through the internet, data-at-rest encryption during storage, and cryptographic hashing, which may be able to verify file and message integrity. Storage encryption can be achieved with three different methods: full-storage encryption, volume and virtual disk encryption, and file/folder encryption (Scarfone, K., et. al., 2007). Decryption tools must be stored on a separate device or location from the encrypted data (U.S Department of Health and Human Services, 2013). To comply with the Federal Trade Commission’s recommendation of using accepted methods of encryption and security, for the transmission of sensitive data, transport layer security version 1.2 should be employed configured with cipher suites using NIST-approved schemes and algorithms (McKay & Cooper, 2019). Any computers connected to the company’s network must maintain updated firewalls and antivirus software to protect from malicious attacks over the internet. Creating a plan to keep data secure and confidential does not end with creating a data security plan; it is an ongoing endeavor as technology evolves and new standards and requirements for data security emerge. Therefore, company systems should be tested frequently for common vulnerabilities, like SQL injection attacks (Federal Trade Commission, 2015), and work with security researchers through co-ordinated security disclosure (OECD, 2021).

The goals of the company’s data privacy and security plan are simple: keep PHI safe and secure by complying with ethical and legal standards in order to protect individuals and the company. The actions necessary to accomplish these goals are not as straightforward and require specific knowledge and expertise to achieve. Fortunately, government agencies provide extensive guidance on the various methodologies for de-identification of PHI and leave company experts to decide which ones are best for specific datasets in minimizing privacy loss while maximizing data utility. Access to data must be limited to authorized users, which means restricting access to not only malicious attackers but also unauthorized employees within the company. With the help of all employees, unauthorized data access can be eliminated with good practices, such as strong passwords. Limiting unauthorized access also comes down to the use of federally approved storage and transmission encryption methods. Ultimately, technology is constantly changing and evolving and with it, the company will update its data security plan to uphold its legal and ethical obligations to keep individuals’ private data safe.

References

Cybersecurity & Information Security Agency (2020, October 2). *Capacity Enhancement Guide: Implementing Strong Authentication*. <https://www.cisa.gov/sites/default/files/publications/CISA_CEG_Implementing_Strong_Authentication_508.pdf>

Federal Trade Commission. (2015, June). *Start with Security: A Guide For Business*. <https://www.ftc.gov/system/files/documents/plain-language/pdf0205-startwithsecurity.pdf>

Garfinkel, S. (2015, October) *De-Identification of Personal Information*. U.S. Department of Commerce, National Institute of Standards and Technology. <https://nvlpubs.nist.gov/nistpubs/ir/2015/nist.ir.8053.pdf>

McKay, K. A. & Cooper, D. A. (2019, August). *Guidelines for the Selection, Configuration, and Use of Transport Layer Security (TLS) Implementations*. U.S. Department of Commerce, National Institutes of Standards and Technology. <https://doi.org/10.6028/NIST.SP.800-52r2>

Nissim, K., Steinke, T., Wood, A. Bun, M., Gaboardi, M., O’Brien D. R., Vadhan, S. (2017, March 3). Differential Privacy: A Primer for a Non-technical Audience. Vanderbilt Journal of Entertainment & Technology Law 21, no. 1 (2018): 209-275.

OECD (2021), “Encouraging vulnerability treatment: How policy makers can help address digital security vulnerabilities”, Directorate for Science, Technology and Innovation Policy Note, OECD, Paris, [www.oecd.org/digital/encouraging-vulnerability-treatment.pdf](http://www.oecd.org/digital/encouraging-vulnerability-treatment.pdf).

Scarfone, K., Souppaya, M., Sexton, M. (2007, November). *Guide to Storage Encryption technologies for End User Devices: Recommendations of the National Institute of Standards and Technology*. U.S. Department of Commerce, National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.SP.800-111>

U.S. Department of Health and Human Services (2012, November 26). *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule*. <https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf>

U.S Department of Health and Human Services (2013, July 26). *Guidance to Render Unsecured Protected Health Information Unusable, Unreadable, or Indecipherable to Unauthorized Individuals*. https://www.hhs.gov/hipaa/for-professionals/breach-notification/guidance/index.html