A review on Privacy-Preserving Data Preprocessing

Mukesh Soni[1] *, YashKumar Barot[2] and S. Gomathi[3]

[1*]Smt. S. R. Patel Engineering College, Unjha , Gujarat, INDIA; soni.mukesh15@gmail.com

[2]Smt. S. R. Patel Engineering College, Unjha , Gujarat, INDIA; yashbarot6312@gmail.com

[3]Research scholar, Research & Development Centre, Bharathiar University, Coimbatore INDIA; gomathisrinivasan@gmail.com

Abstract

Health care information has great potential for improving the health care system and also providing fast and accurate outcomes for patients, predicting disease outbreaks, gaining valuable information for prediction in future, preventing such diseases, reducing healthcare costs, and improving overall health. In any case, deciding the genuine utilization of information while saving the patient's identity protection is an overwhelming task. Regardless of the amount of medical data it can help advance clinical science and it is essential to the accomplishment of all medicinal services associations, at the end information security is vital. To guarantee safe and solid information security and cloud-based conditions, It is critical to consider the constraints of existing arrangements and systems for the social insurance of information security and assurance. Here we talk about the security and privacy challenges of high-quality important data as it is used mainly by the healthcare structure and similar industry to examine how privacy and security issues occur when there is a large amount of healthcare information to protect from all possible threats. We will discuss ways that these can be addressed. The main focus will be on recently analyzed and optimized methods based on anonymity and encryption, and we will compare their strengths and limitations, and this chapter closes at last the privacy and security recommendations for best practices for privacy of preprocessing healthcare data.

Keywords: Privacy; information security; risk management; confidentiality; integrity; availability; HIPAA; anonymity; privacy appliance;

1. Introduction

As we know the health care delivery system adopts technology and advances in IT, large amounts of healthcare data are made available in digital form. Health care organizations commonly embrace data innovation to decrease costs and improve proficiency and quality in the healthcare system to make it fast and secure. Truth be told, with the help of the IT support the patient's data are becoming now digital and patient information is experiencing sensational and central changes in medical, working and plans of action and in the money related world mostly. Medical researchers are now trying to work on clinical data and patient's

Received: May 7, 2020 Revised: July 30, 2020 Accepted: September 1, 2020

DOI: 10.5281/zenodo.4014237 **Special issue on Research and Innovations in Data Protection and Privacy in Healthcare (RIDAPPH)**

medical records to discover valuable knowledge from the huge size of healthcare data that is not known in the health records of individual patients. Now by working on this clinical data will impact health care because the patient's data relies on a high degree of trust and confidence among the patient and physician. It is obvious that Doctors and medical teams need accurate and complete information about the patient and his/her illness in order to function constructively to help them. In any case, if patients don't believe the organization or the doctor to secure the secrecy of their human services data, and not maintain privacy, they can hold or ask the doctor not to record delicate data (California Healthcare Foundation, 1999) [4]. The company has clinical operational capacity for analysis and the valuable data is absolutely at risk of having incomplete and wrong information that can prevent the use of researchers for future research purposes. Transformation is a common phenomenon in the world of healthcare. These changes are fueled by the increase in aging and lifestyle changes. An increase in software and online devices like mobile, band new treatments, increasing main focus on the healthcare data care and value of data, and pre-tested medicine in contrast of all-inclusive clinical and medical decisions for healthcare - leading to providing valuable opportunities to support clinical decision-making, health care system upgrade for management of data and policy making of data, disease surveillance, resistance to adverse events, and improved treatment of multi-organ disease systems [1, 2]. Thus, a large-scale analysis of health care data has many advantages, promises and possibilities for health reform, but has many obstacles and challenges in the path. Indeed, concerns about the security of big health and Data protection data are increasing exponentially from year to year. In addition, public health care companies have discovered that an efficient, low-cost and innovative technologies-centric approach to the determination of privacy and security needs is not strong enough to protect the institution as well as its patients [3]. Inspired in this way, new techniques are required to prevent violations of confidential information as well as other types of security tragedies like trying to hack in order to obtain the highest possible usage of health records. Researchers would also discuss some of the related tasks as well as security of big health data and other new technologies risks and start concentrating mostly on major issue of data confidentiality in universal health care, by setting out the policies and regulations laid down by the Commission of the government and different agencies [5]. Use of such information is governed by the laws and regulations of That very many government healthcare services and health planners, payers, clearing and Research study related to risks such as the Health Insurance Portability Privacy Rules under Accountability Act of 1996 (HIPAA) [8, 9]. Data mining, notably when we have data from multiple sources, can cause real challenges. For instance, doctors and healthcare professionals are often obligated to obtain reports of each of this patient information from the survey to the community health database for a variety of purposes. All this may include patient 's mobile number, ethnicity, date of service, birth date, Postcode, sexuality, diagnostic codes, as well as doctor's Identifying number, doctor's Mailing address, total fees, social security no, and many more information. This information is being made publicly available to researchers and organizations. Since such compilations may not contain the patient's both first and last name, address, mobile number or social security card number, individuals meet the criteria as non-identified and unidentified. Disclosure of less risk to the privacy of the patient, therefore. However, by linking this data to other public datasets, procedures such as data analysis can link a person to specific diagnosis and treatment. Sweeney (1997) It describes how and when to recognize that kind info by combining certain criteria with the candidate with the help of the publicly available databases. Even with only the birthdate can be determined names and addresses of people who are up to 12%, if combine date of birth and gender 29%, date of birth and with the help of the zip code can find up to 69% accuracy. and with a local address and birthday 97%. Recent work has demonstrated the methods of identifying individuals by finding information from individuals using the internet (Malin and Sweeney, 2001) (Malin et al., 2003). By using Public hospital data for disease-related general illness. Data is collected mostly in the form of data that is used by the Internet to access public health information, health services and medical products, threatening privacy, but very few laws and regulations control Are using and disclosing this kind of data (Goldman and

Received: May 7, 2020 Revised: July 30, 2020 Accepted: September 1, 2020

DOI: 10.5281/zenodo.4014237 **Special issue on Research and Innovations in Data Protection and Privacy in Healthcare (RIDAPPH)**

Hudson, 2000). We will dig deep into issues in the management of privacy of healthcare data and health-related data security, by reviewing data on their core principles, Aspects and core values, and regulations imposed by the government. We also examine relevant literature on technical issues related to the protection of privacy.

2. PRIVACY AND SECURITY ANALYSIS OF HEALTHCARE INFORMATION

Many times, voluntary, odd, constructive, non-authentic or Surrounded by prescribed form and imperfection, as well as morality Ideas and legal barriers, health care of patient characteristics

data make them "confusing". Mainly Because they have evolved into one Direct patient care outcomes with benefits assessments They should be used for patient, research, or administrative purposes. Improper disclosure, data loss Integrity, or availability, can cause loss (Sios and Moore, 2002). Recently the Legislation and regulations such as HIPAA (Health Insurance Portability and Accountability Act) provide legal assistance to patients. patients' personal information such as health information and rights Establish responsibilities to protect and limit healthcare organizations.

In addition, previous research has not considered the Balance among privacy and availability of the data. There is a need to leverage multiple algorithms and a new privacy protection methodology that applies to a type of data mining method that balances privacy and data availability [10]

2.1 Healthcare Data Privacy

The word "privacy" has lots of meanings like The interest of the individual in defending his or her personally identifiable information as well as the respective duty of those individuals and entities who take part in the system for the uses of the data transaction of such data and information to accept those interests through fair information policies. Common meanings are to issue regulation of information to someone and to be free from Infringement or turmoil in one's private life. Privacy respects the Suitable usage of client information. The company could not sell its patient / user information to any third party without the user's prior permission. To obtain health care, one must disclose and share Quite private and often quite sensitive information. Any way we can influence the confidentiality of our medical information that we provide to our doctors and many others in the public health system. Even before people share personal information to our doctors, we 're no longer in charge of our privacy. Thus, the term "confidentiality" overlaps with the term "confidentiality". It is the right of the patient not to receive their information from other parties. To secure Information obtained from patients goes through unauthorized access and disclosure. Privacy is the ability to determine what and where a person's information is going.

2.2 Security in Healthcare Data

Although healthcare organizations create, store and transmit large amounts of confidential data to provide effective and appropriate care, still due to lack of support from technical and low levels of security they are vulnerable. Because of the Internet, information security has become highly important for our society. Because of the complicated structure of big data, the healthcare industry is most vulnerable to Infringements

Received: May 7, 2020 Revised: July 30, 2020 Accepted: September 1, 2020

DOI: 10.5281/zenodo.4014237 **Special issue on Research and Innovations in Data Protection and Privacy in Healthcare (RIDAPPH)**

of publicly disclosed data. In reality, attackers could use data mining processes and technologies to detect and disclose sensitive information to the public and thereby causing data breach and infringements of data. While this is a complicated process to implement the security measures that because of new methods of security, security controls are becoming cleverer. It is therefore important for organizations to incorporate solutions to health data security that protect critical assets and highly confidential data while satisfying health care mandates. Various technologies, such as encryption, firewalls, etc. These are mainly used to prevent the data from being compromised by vulnerabilities in systems in the company's database and technology. This concern in healthcare is relatively new, but nowadays information technology and its security are a well-established and well improved domain. There is a huge amount of knowledge available to protect health care information from possible threats. A general brief understanding of the security can be understood by:

1. Data protection element.
2. Principles of Security.
3. Threats, security issues, policy laws and information security.
4. Security of Information: organizational, Physical, Technological Security.

2.2.1 Data protection elements

Security is mainly achieved by maintaining its elements like: confidentiality, integrity, availability and accountability.

1. Confidentiality is one of the most important properties of Data protection elements for privacy of healthcare data and related information. We have to make sure that data will not be made available or disclosed to unauthorized persons or processes.

2. Integrity is also a property of the Data protection element mainly concerned with that data and information that will not be destroyed in an unauthorized manner.

3. Availability of the data is also the property of a Data protection element and information is most accessible and usable when it’s needed on demand by an authorized person at any time anywhere.

4. Accountability of data is the Capacity to review the behavior of all parties and methods interacting with the data and to evaluate when action needs to be taken.

2.2.2 Principles of Security

Let's be clear on what data will be collected or even why.

1. To be transparent to people they have to inform them and give them choices about the use of the medical data, organizations should make it easy to understand what data they collect, how it is used, and why. It's being transparent.
2. Haven't ever sold sensitive personal information to anyone. Institutions are using the information to make their product and services as helpful as feasible.

Received: May 7, 2020 Revised: July 30, 2020 Accepted: September 1, 2020

DOI: 10.5281/zenodo.4014237 **Special issue on Research and Innovations in Data Protection and Privacy in Healthcare (RIDAPPH)**

3. Make it possible for consumers to exercise control over their personal information. Whenever it comes to privacy, humans understand that such a size of data doesn't fit all of it. Users can choose the privacy that's right for them. And as technology advances, privacy settings and control also change, making sure that privacy has always been a personal decision that every user has rights.

4. Planning to build the strongest technologies for security. Trust for the confidentiality of customers involves defending the health information they truly respect. To keep all data, secure for users, constantly improving the technology and updating its structure. This basically involves enhancing built-in security innovations to identify and defend against advancing online threats ever since they approach users' personal information.

2.2.3 Security Threats, Security issues, Policy laws and Information security

There are many security threats possible to a computer system and electronic stored data which is produced either inside and outside of organizations like hospitals. Security threats include malicious codes such as viruses, Trojan horses and worms can be very dangerous to organizations if security is not maintained. Malicious code mostly takes advantage of security vulnerabilities in the Database and Network system; however, the situation also depends on organizational weaknesses like Failure to employ or use of hacked or cracked software and in the use of fake and dummy antivirus software. Malicious code can harm the system like starting to deny service attacks and deception Information theft and other data leaks. Attacks by malicious code such as Virus exposes the risk of "hackers" such as: External agencies are willing to harm external organizations and try to downplay network activities in general. Insider with Network Operations and Privileges Saying nothing about workers who are not really trained against their employer is so painful that they accidentally make mistakes.

2.2.4 Security of Information: organizational, Physical, Technological Security

Activities such as exposing the security threats and data loss risks can be managed by organizational, Security, physical, and technological security:

These three classifications of security measures are most important. The HIPAA security standard briefly describes "administrative security measures. "organizational roles, policy and procedure for managing the collection, 'Production, implementation and management of safety measures Protecting and preserving digital secure health information The workforce of the unit concerned in regards to its security Information"(Department of Health and Human Services, 2003, p. 261).Protection requires policy and procedure such as threat assessment and management, allocation of responsibilities for data protection develop such rules with procedures to provide security and access Information, admitting misconduct, for fight with threat and provide security incidents and Implementing a safety training and awareness program [41, 45].

2.3 Few Steps to Improve Privacy in Healthcare

Here are few Steps which can be to use for Improve Privacy in the Healthcare system and to provide Security mainly by Attempting to avoid security breaches such as infringements requires a systematic, preventative measures and the forward-looking approach, enough so security and privacy policies, risk management and

Received: May 7, 2020 Revised: July 30, 2020 Accepted: September 1, 2020

DOI: 10.5281/zenodo.4014237 **Special issue on Research and Innovations in Data Protection and Privacy in Healthcare (RIDAPPH)**

prevention techniques are in position as soon as a risk turns up. The privacy and security capacities of the organization are exponential, incremental and longstanding [11, 20].

We can Start with some of these following steps:

1. Construct the safety and protection practices of healthcare organizations on the industry-standard with mainly top-down approach.

2. Create and improve best integrity and confidentiality practices, including relevant legislation, privacy principles, data collection requirements, data analysis requirements, data preprocessing and data processing needs. Discuss protection and privacy specifications for collection, Using, maintain, reveal and disposal of data.

3. Incorporate a privacy and security vulnerability safety-based solution to help build a really well-targeted and efficient solution that maximizes the organization's restricted security and data protection plan and maximizes loss of data prevention.

4. Reduce the risk of vulnerabilities as well as other security threats with a comprehensive, in - depth plan that recognizes the entirety data protection threats.

5. Using the latest hardware-assisted security solutions to boost the accuracy and efficiency of technological security protocols.

6. Increase the value of one's security protocols via a developing performance way of ensuring that the security control system's hardware as well as software are suitable and live in collaboration to ensure maximum confidentiality in medical information.

2.4 Technologies in use

Here are some of the variety of ways to ensure the confidentiality, data security and privacy of large-scale medical data. Mostly used technique with help of technologies are as follow [22, 29],

1) Authentication

The term Authentication is mostly known for confirmation or validation which claims about the specific topic, which is Valid and real. It serves important functionality in any organization like: having access to systems, defending access to networks, protecting every users' identities as well as making sure the user is a real witch who is assuming like he or she is.

2) Encryption

The term Data Encryption is an effective tool to prevent unauthorized access to sensitive data. From data storage facilities to end-users (which also include mobile devices mainly used by doctors, practitioners and administration staff) and its solutions in the cloud service, it protects and manages data ownership over its life cycle. Encryption is useful in preventing breaches and data loss.

3) Masking Data

Received: May 7, 2020 Revised: July 30, 2020 Accepted: September 1, 2020

DOI: 10.5281/zenodo.4014237 **Special issue on Research and Innovations in Data Protection and Privacy in Healthcare (RIDAPPH)**

Here masking the data means replacing sensitive data with unknown values. This is not really an encryption technology, and because of this method original data value will not be returned from the original masked value. Here we use techniques to identify data sets and suppress personal identifiers like names, social security numbers and masking or semi-activators such as birth dates and postal codes. Therefore, Masking the data is one of the most effective and popular approaches for data anomaly [33].

4) Access control

When a user is authenticated there only after the user can submit the information into the system, but the user's access is still controlled by the access control policy, and by this effective, typically a powerful and flexible mechanism that allows users to rely on each doctor authorized by a patient or granted as well as trusted third party. Various methods have been proposed for this problem by many researchers to mainly deal with security and access control issues. Some of them are like; Role-based access control (RBAC) [34] and attribute-based access control (ABAC) [35, 36] which are mainly the most effective and powerful models for EHR. Although RBAC and ABAC have some certain limitations when they are used mainly for the medical related application and healthcare system. Paper [36] proposes an important and effective cloud-oriented storage efficient dynamic access control scheme cipher text based on the CP-ABE and a symmetric encryption algorithm (such as AES).

5) Auditing and Monitoring

Security surveillance gathers and investigates and some network events to capture intrusions. The audit and monitoring is frequently recording of the health care system's user activities, such as maintaining logs for each access and modifying the healthcare data. These are two alternative safety parameters to measure and ensure the safety of the healthcare system [38].

6) **Biometrics & Cryptographic Algorithms:**

Various Cryptographic techniques like Symmetric Key [53], Encryption Algorithm [54] and Biometrics techniques [55-56].

3.1 Methods for Privacy preserving in big data

Some of the traditional in use methods for preservation of privacy in big data like De-identification are briefly described here.

3.1.1 Anonymized-identification

Anonymized-identification is the traditional method of prohibiting the revelation of maintaining confidentiality of information in order to reject the information required by the first method of removing the patient's specific identifiers, second statistical method of patient verification that Insufficient identifiers are deleted [6]. However, a hacker may find more external and crucial information support to detect in big data. As a result, we can say that de-identification alone is not enough to protect the privacy of big data [7]. This is made possible by developing effective privacy-protection algorithms for de-identification to help reduce the risk of re-detection of patient's data. The concepts of k-anonymity [46–47], t -gallogens [46, 50] and L-variation [47, 49, 50] have been introduced to improve this old traditional method.

Received: May 7, 2020 Revised: July 30, 2020 Accepted: September 1, 2020

DOI: 10.5281/zenodo.4014237 **Special issue on Research and Innovations in Data Protection and Privacy in Healthcare (RIDAPPH)**

k-anonymity

In the k-anonymity method, when the higher the k value will be, the lower the probability of re-detection will be ensured. However, this leads to data distortion and therefore more and more information is lost due to K-anonymity. In addition, high anonymity makes the data non useful for recipients because some of the analyzes may be impossible or may yield biased and inaccurate results. In K-Anonymous, a semi-identifier containing data to identify individuals is used to link other publicly available data, with one identifier (such as disease) being sensitive.

L-variation

It is group-based anonymization used to protect the privacy of a data set by reducing the nuances of the important healthcare data representation. This particular model (typical, entropy, iterative) [46, 47, 48] is an updated version of k-anonymity, in which it uses methods including like generalization as well as suppression to reduce the nuances of particular data representation, so that at least k records can be map the separate records in data. The L-diversity model maintains some of the weakness in the k-anonymous model, in that the preserved identity is not equivalent to protecting sensitive feelings that are normalized or suppressed to the level of k-individuals. The main problem with this kind of method is that it depends on the sensitive kind of feature. If we want to make the data L-variant, we must include binary heuristic data, even if there are no different values for the sensitive attribute. It improves data security, but may cause problems during the data analysis. Consequently, the L-Diversity method can be vulnerable to distortion and parity attacks [48] and therefore does not prevent the disclosure of symptoms.

T-closeness

Another improvement of the anonymity of the L-Diversity group is T-intimacy. T -collections model (same / hierarchical distance) [47, 50] improve the L-variance model mainly by considering the values of a feature, taking into account the distribution of data values for that attribute. This technique is mainly important for when it accepts attribute disclosure, and its problem is most likely the size and variability increase, re-detection increases highly.

4. PRIVACY-BASED DATA PREPROCESSING

Here we have advanced the differential security-based method for data pre-preprocessing with a practical mechanism for clustering of distance-based. Adding Laplace noise to preprocessing data prevents the appearance of differential privacy data. We try to accept an adaptive mechanism to minimize the impact of the privacy budget parameter, adjust the distortion, and ensure the appropriate conditions for privacy protection and data availability [30, 39].

Data preprocessing

Data taken from the real-world is rarely clean and complete, especially in the health care sector. The data purge cleanup phase is related to the deal with missing values, errors and discrepancies in the healthcare dataset. Various techniques and methods have been developed to solve this problem. The choice of appropriate particular technique depends on various factors [5]. Proper management of incomplete data includes some of the steps described below

1) Treatment of missing value

Received: May 7, 2020 Revised: July 30, 2020 Accepted: September 1, 2020

DOI: 10.5281/zenodo.4014237 **Special issue on Research and Innovations in Data Protection and Privacy in Healthcare (RIDAPPH)**

Missing value (MV) is defined as the value of data that is not in the cell of the corresponding column of the dataset into the database. Because of this MV in the health care context may be human defaults, not applicable, not electronically recorded by sensors, patient not on a ventilator due to medical judgment, patient's condition is a specific variable, power failure, unrelated to database synchronization and others. Any statistical analysis or machine learning activities on MV data can give undesirable results or biased results, and mismanagement of missing data can lead to misleading conclusions and this can be dangerous for patients [6]. According to the survey [6] problems associated with MV can lead to loss of efficiency, difficulties in managing and analyzing data. Before applying any treatment, we must identify missing samples. Handling the MV can be used to ignore or disrupt missing data.

2) Releases missing values

The most common way to solve this problem is to ignore MVs. But this approach is not practical. If the number of missing values in the dataset is low, we must ensure that the analysis does not produce bias estimation bias for the remainder. Removal of MVs can be done in the following ways:

List-wise Removal (Row Removal): In this case a full case analysis is performed and eliminates all cases observed with more than one missing value. But this approach helps in a small number of missing cases. This works well and is rare when there are no MCAR models in the dataset.

Pair Removal: This method attempts to reduce errors in list-wise removal. If another feature is not used as a case when analyzing the data, the attribute with MV is removed. This not only strengthens the diagnostic power but creates other problems such as standard error output.

Exclusion of Symptoms: It is very rare and sometimes you can omit the symptom if the missing observations exceed 60% and the symptom is very low in the analysis. Values that are missing due to high values should be kept for a while.

- **Survey section**

Survey section	**Survey Items**	**Example of Questions**	**Source/Publication of Survey Item(s)**
Section A - Technology	Supervision of the Data privacy and Institutional Review with permission for each project.	How to maintain data privacy with use of the best technology available.	[1-7]
Section B -Data use and Ownership	Data use Violations and Penalties monitored over time and extended upto security knowledge and training for data processes	Maintain the Ownership of Data gathered from individuals and make them secure for pre-processing and provide a secure Data storing environment.	[8-20]
Section C -. Access control and policies.	Use good access control for violation of policy and	who have authority for each session and personal	[25-45]

Received: May 7, 2020 Revised: July 30, 2020 Accepted: September 1, 2020

DOI: 10.5281/zenodo.4014237 **Special issue on Research and Innovations in Data Protection and Privacy in Healthcare (RIDAPPH)**

	many applicable policies	question asked to patent?	

Comparisons

In Alberts C, Doroffe A. (2003). and Behlen, F.M., Johnson, S.B. (1999). shows that changes are fueled by the increase in aging and lifestyle. An increase in software and online devices like mobile, we get huge amount of healthcare data and value of data, and pre-tested medicine in contrast of all-inclusive clinical and medical decisions for healthcare - leading to providing valuable opportunities to support clinical decision-making, health care system upgrade for management of data and policy making of data, disease surveillance, resistance to adverse events, and improved treatment of multi-organ disease systems.

According to Lowrance, W. (2002). Meany, M.E. (2001). Murphy, S.N., Chueh, H.C. (2002). Oliveira, S.R.M., Zalane, O.R. (2003) and South Tyneside NHS Foundation Trust and Indiana Health Information Exchange. http://www.ihie.org/. Here are few Steps which can be used to Improve Privacy in the Healthcare system and to provide Security mainly by Attempting to avoid security breaches such as infringements requires a systematic, preventative measures and the forward-looking approach, enough so security and privacy policies.

Department of Health and Human Services (July 13, 2004) and Ferris, T.A., Garrison, G.M., Lowe, H.J. (2002). is providing a traditional method of prohibiting the revelation of maintaining confidentiality of information in order to reject the information required. However, a hacker may find more external and crucial information support to detect in big data. As a result, we can say that de-identification alone is not enough to protect the privacy of big data.

In this research made possible by developing effective privacy-protection algorithms for de-identification to help reduce the risk of re-detection of patient's data. The concepts of k-anonymity, t -glogens and L-variation have been introduced to improve this old traditional method. K‐anonymity and Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M. L‐diversity 2006. Li N, et al. and Samarati P. Sweeney L. t‐Closeness, Ton A, Saravanan M. Ericsson research.

5. CONCLUSION

An authoritative approach to managing the use and disclosure of personal health information is best for patients, individual researchers, healthcare organizations and society also. For those who do not follow good security and privacy practices, the risk is higher. Improper use or disclosure of future laws and regulations may increase data loss and data breach for harmful purposes. Despite the increasing emphasis on research, organizations should apply the same general policies to support the conduct of healthcare for research.

"Researchers do not have the right to review patient data. As described here, in addition to developing a strategy to reduce patient risk, researchers take general steps to ensure that users meet the requirements" (Burman, p. 33,2002). A recent publication recommended: "First, sensitive data, such as identifiers, names, addresses and social security no must be edited, changed or truncated from the original database, so that anyone who receives the data cannot compromise the privacy of the patents. Second, data mining algorithms also may be compromised from the database due to such knowledge being compromised by data privacy into

Received: May 7, 2020 Revised: July 30, 2020 Accepted: September 1, 2020
DOI: 10.5281/zenodo.4014237 **Special issue on Research and Innovations in Data Protection and Privacy in Healthcare (RIDAPPH)**

the database itself. The key to securing data is to make sure that private data remains private even after the data goes through the mining process.

Apart from this when we have such a good system but still, they are not sufficient for healthcare data mining. Mostly the original data is available as it is, there is still a risk to the privacy, Dignity and Accessibility of the data. Therefore, an effective privacy program relies on the implementation of strong security controls. Medical data miners must use several important safety techniques, as mentioned below:

1. compulsory supervision of the Data privacy and Institutional Review Board must be established with permission for each project.

2. violations and Penalties should really be monitored over time and extended to security knowledge and training. process.

3. Establish robust audit procedures.

4. Sanctions shall apply to violation of policy and many applicable policies.

5. Use good access control and authority for each session and question asked to patent.

6. Training is required for all theoretical researchers to strengthen their responsibilities.

7. If possible, delete useless common identifiers (such as age, names, addresses, data of birthday) of data contents or hide data from the user.

REFERENCES

1. Alberts C, Doroffe A. (2003). Managing Information Security Risks: The OCTAVE Approach. Boston, MA, Addison-Wesley.

2. Behlen, F.M., Johnson, S.B. (1999). "Multicenter Patient Records Research: Security Policies and Tools," JAm Med Inform Assoc. 6(6) 435-43.

3. Berman, J.J. (2002). "Confidentiality Issues for Medical Data Miners," Artif Intell Med. 26(1-2):25-36

4. California Healthcare Foundation (1999). Medical Privacy and Confidentiality Survey Summary and Overview, http://www.chcf.org/documents/ihealth/survey.pdf

5. Defense Advanced Research Project Agency (July 19, 2002). "Total Information Awareness Program (TIA) System Description Document (SDD)," Version 1.1.

6. Department of Health and Human Services (July 13, 2004). Protecting Personal Health Information in Research: Understanding the HIPAA Privacy Rule, (NIH Publication Number 03-5388), http://privacyruleandresearch.nih.gov/pr~02.asp

Received: May 7, 2020 Revised: July 30, 2020 Accepted: September 1, 2020

DOI: 10.5281/zenodo.4014237 **Special issue on Research and Innovations in Data Protection and Privacy in Healthcare (RIDAPPH)**

7. Ferris, T.A., Garrison, G.M., Lowe, H.J. (2002). "A Proposed Key Escrow System for Secure Patient Information Disclosure in Biomedical Research Databases," in Proc AMIA Symp. 245-9.

8. Goldman, J. and Hudson, Z. (2000). "Perspective Virtually Exposed: Privacy and E-Health," Health Affairs, 19(6), 140-8.

9. Goodwin, L.K. and Prather, J.C. (2002). "Protecting Patient Privacy in Clinical Data Mining," J Healthc Inf Manag, 16(4):62-7.

10. Islan, M.Z., and Brankovic, L., A. (2004). "Framework for Privacy Preserving Classification in Data Mining, School of Electrical Engineering and Computer Science," Australasian Computer Science Week.

11. Lowrance, W. (2002). "Learning from Experience: Privacy and the Secondary Use of Data in Health Research," The Nuffield Trust; www.mffield trust.0rg.uk

12. Meany, M.E. (2001). "Data Mining, Dataveillance, and Medical Information Privacy," in Privacy in Health Care. J, Humber, ed., Humana Press, pp. 145-164.

13. Murphy, S.N., Chueh, H.C. (2002). "A Security Architecture for Query Tools Used to Access Large Biomedical Databases," in Proc AMIA Symp. 552-6.

14. Oliveira, S.R.M., Zalane, O.R. (2003). "Protecting Sensitive Knowledge by Data Sanitization," in Proceedings of the Third ZEEE International Conference on Data Mining, Melbourne, Florida, USA, 613-616.

15. Burghard C. Big data and analytics key to accountable care success. Framingham: IDC Health Insights; 2012.

16. Fernandes L, O'Connor M, Weaver V. Big data, bigger outcomes. J AHIMA. 2012;83:38–42.

17. David Houlding, MSc, CISSP. Health Information at Risk: Successful Strategies for Healthcare Security and Privacy. Healthcare IT Program Of ce Intel Corporation, white paper. 2011.

18. South Tyneside NHS Foundation Trust. Harnessing analytics for strategic planning, operational decision making and end‐to‐end improvements in patient care. IBM Smarter Planet brief. 2013.

19. Indiana Health Information Exchange. http://www.ihie.org/. Accessed 24 Mar 2016.

20. Transforming healthcare through big data, strategies for leveraging big data in the healthcare industry. Institute for Health. 2013.

21. General Dynamics Health Solutions white paper UK. "Securing Big Health Data"©2015. http://gdhealth.com/globalassets/health‐solutions/documents/brochures/securing‐big‐health‐data_‐white‐paper_UK.pdf.

22. Zhang R, Liu L. Security models and requirements for healthcare application clouds. In: IEEE 3rd international confer‐ ence on cloud computing. 2010.

Received: May 7, 2020 Revised: July 30, 2020 Accepted: September 1, 2020

DOI: 10.5281/zenodo.4014237 **Special issue on Research and Innovations in Data Protection and Privacy in Healthcare (RIDAPPH)**

23. Linden H, Kalra D, Hasman A, Talmon J. Inter‐organization future proof HER systems—a review of the security and privacy related issues. Int J Med Inform. 2009;78:141–60.

24. Marchal S, Xiuyan J, State R, Engel T. "A big data architecture for large scale security monitoring", Big Data (BigData Congress), Anchorage, AK. 2014. p. 56–63.

25. Duygu ST, Ramazan T, Seref S. A survey on security and privacy issues in big data. In: The 10th international confer‐ ence for internet technology and secured transactions (ICITST‐2015).

26. Liu L, Lin J. Some special issues of network security monitoring on big data environments. Dependable, Autonomic and Secure Computing (DASC), Chengdu. 2013. p. 10–5.

27. Big Data security and privacy issues in healthcare—Harsh KupwadePatil, Ravi Seshadri. 2014.

28. Sectoral healthcare strategy 2012–2016‐Moroccan healthcare ministry.

29. Patil P, Raul R, Shroff R, Maurya M. Big data in healthcare. 2014.

30. Samrati P. Protecting respondents identities in microdata release. IEEE Trans Knowl Data Eng. 2001;13:1010–27.

31. Samarati P. Protecting respondent's privacy in microdata release. IEEE Trans Knowl Data Eng. 2001;13(6):1010–27.

32. Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M. L‐diversity: privacy beyond k‐anonymity. In: Proc. 22nd international conference data engineering (ICDE). 2006. p. 24.

33. Chawala S, Dwork C, Sheny FM, Smith A, Wee H. Towards privacy in public databases. In: Proceedings on second theory of cryptography conference. 2005.

34. Sweeney L. K‐anonymity: a model for protecting privacy. Int J Uncertain Fuzziness. 2002;10(5):557–70.

35. Meyerson A, Williams R. On the complexity of optimal k‐anonymity. In: Proc. of the ACM Symp. on principles of database systems. 2004.

36. Mehmood A, Natgunanathan I, Xiang Y, Hua G, Guo S. Protection of big data privacy. In: IEEE translations and con‐ tent mining are permitted for academic research. 2016.

37. Mohammadian E, Noferesti M, Jalili R. FAST: fast anonymization of big data streams. In: ACM proceedings of the 2014 international conference on big data science and computing, article 1. 2014.

38. Xu K, Yue H, Guo Y, Fang Y. Privacy‐preserving machine learning algorithms for big data systems. In: IEEE 35th international conference on distributed systems. 2015.

39. Wei L, Zhu H, Cao Z, Dong X, Jia W, Chen Y, Vasilakos AV. Security and privacy for storage and computation in cloud computing. Inf Sci. 2014;258:371–86.

Received: May 7, 2020 Revised: July 30, 2020 Accepted: September 1, 2020

DOI: 10.5281/zenodo.4014237 **Special issue on Research and Innovations in Data Protection and Privacy in Healthcare (RIDAPPH)**

40. Behlen FM, Johnson SB. Multicenter patient records research security policies and tools. J Am Med Inform Assoc 1999;6(6):435–43.

41. Berman JJ, Moore GW, Hutchins GM. Maintaining patient confidentiality in the public domain internet autopsy database. J Am Med Inform Assoc (JAMIA), Symp. Suppl. 1996;328–32.

42. Bouzelat H, Quantin C, Dusserre L. Extraction and anonymity protocol of medical file. Proc AMIA Annu Fall Symp 1996;323–27.

43. Department of Health and Human Services. 45 CFR (Code of Federal Regulations), Parts 160 through 164. Standards for Privacy of Individually Identifiable Health Information (Final Rule). Federal Register: vol. 65, number 250, 28 December 2000. p. 82461–510.

44. Department of Health and Human Services. 45 CFR (Code of Federal Regulations), 46. Protection of Human Subjects (Common Rule). Federal Register, vol. 56, 18 June 1991. p. 28003.

45. Duygu ST, Ramazan T, Seref S. A survey on security and privacy issues in big data. In: The 10th international confer‐ ence for internet technology and secured transactions (ICITST‐2015).

46. Li N, et al. t‐Closeness: privacy beyond k‐anonymity and L‐diversity. In: Data engineering (ICDE) IEEE 23rd interna‐ tional conference. 2007

47. Ton A, Saravanan M. Ericsson research. http://www.ericsson.com/research‐blog/data‐knowledge/big‐data‐privacy‐preservation/2015

48. Samarati P. Protecting respondent's privacy in microdata release. IEEE Trans Knowl Data Eng. 2001;13(6):1010–27.

49. Sweeney L. K‐anonymity: a model for protecting privacy. Int J Uncertain Fuzziness. 2002;10(5):557–70.

50. Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M. L‐diversity: privacy beyond k‐anonymity. In: Proc. 22nd international conference data engineering (ICDE). 2006. p. 24.

51. Samarati P, Sweeney L. Protecting privacy when disclosing information: k‐anonymity and its enforcement through generalization and suppression. Technical Report SRI‐CSL‐98‐04, SRI Computer Science Laboratory. 1998.

52. Samarati P. Protecting respondent's privacy in microdata release. IEEE Trans Knowledge Data Eng. 2001;13(6):1010–27.

53. Abhishek Anand, Abhishek Raj, Rashi Kohli, Dr. Vimal Bibhu: Proposed Symmetric Key Cryptography Algorithm for Data Security.In:1st International Conference on Innovation and Challenges in Cyber Security (ICECCS 2016).

54. Deeksha Priya Jha, Rashi Kohli, Archana Gupta: Proposed Encryption Algorithm for Data Security Using Matrix Properties.In: 1st International Conference on Innovation and Challenges in Cyber Security (ICICCS 2016).

Received: May 7, 2020 Revised: July 30, 2020 Accepted: September 1, 2020

DOI: 10.5281/zenodo.4014237 **Special issue on Research and Innovations in Data Protection and Privacy in Healthcare (RIDAPPH)**

55. Garima arora, P.Lakshmi Pavani, Rashi Kohli, Dr. Vimal Bibhu: Multimodal Biometrics For Improvised Security.In: 1st International Conference on Innovation and Challenges in Cyber Security (ICICCS 2016).

56. Preeti Chourasia, Rashi Kohli, Archal Garg:Biometrics Minutiae Detection and Feature Extraction. In: Lambert Academic Publishing.

Received: May 7, 2020 Revised: July 30, 2020 Accepted: September 1, 2020

DOI: 10.5281/zenodo.4014237 **Special issue on Research and Innovations in Data Protection and Privacy in Healthcare (RIDAPPH)**