INTERNATIONAL JOURNAL OF RESEARCH CULTURE SOCIETY ISSN: 2456-6683 Volume - 5, Issue - 11, Nov - 2021
Monthly Peer-Reviewed, Refereed, Indexed Journal Scientific Journal Impact Factor: 5.743
Received on : 15/11/2021 Accepted on :29/11/2021 Publication Date: 30/11/2021

Privacy Preserving in EHR Using Data Mining Local Suppression Methods

[1]UMA K., [2]M. HANUMANTHAPPA

[1]Research Scholar, Department of Computer Science and Applications, Bangalore University, Bengaluru, India
[2]Professor, Department of Computer Science and Applications, Bangalore University, Bengaluru, India
Email – [1]k.uma6568@gmail.com, [2]hanu6572@gmail.com

Abstract: *Abundant medical records generation causes storing the records electronically called Electronic Health Records. During the storage of health records, there is a risk of identification at various phases of data collection and processing. Data mining is one of the most promising areas in which it may be used to make a significant difference in healthcare. Health data has tremendous potential for improving patient results, predicting epidemic occurrences, gaining useful information, avoiding preventable diseases, reducing the cost of delivery of health care, and generally improving the quality of life. But it is a difficult task to decide on the permissible usage of data while protecting and securing it. Data mining plays a vital role in the process of preserving electronically stored healthcare data by applying aggregation, generalization, and reduction techniques. However, no matter how data mining is used efficiently in the medical field without ensuring security and privacy in the healthcare data. This paper explores the survey current security and privacy challenges in the healthcare industry, assesses how security and privacy issues arise in the context of healthcare data, and discusses possible solutions. This article provides an overview of several data security rules, followed by an analysis of India's data privacy status, Health Insurance Portability Accountability Act (HIPAA) rules for healthcare privacy, and data mining approach for data security and privacy using data suppression and generalization.*

Key Words: *Electronic Health Records, Data Security, Data Privacy, Data Mining, Anonymization, HIPAA.*

1. INTRODUCTION:

Nowadays, the healthcare industry generates voluminous data in everyday activities. The various types of data are patient health records, hospital records, and so on. In earlier days, hospitals and other healthcare sectors used to store the data in the format of the paper, file, and manual filing system. While the healthcare sector increase the productivity of data drastically, the traditional methods are not sufficient to hold the data. The healthcare providers motivated by the digital format of data storing methods. The healthcare industry of the federal government started to store the data digitally called Electronic Health Records (EHR) according to an article in AMA journal of ethics. Later, all healthcare industries significantly started to store in the form of EHRs. An electronic health record (EHR) is an electronic version of a patient's medical records that includes personal information, lab reports, diagnoses, prescriptions, and other information.

Healthcare organizations collects, maintains, analyses, and sharing massive amount of electronic data. Much of the data includes Protected Heath Information (PHI) and Personally Identifiable Information (PII) from several sources including transactional systems such as enrolments, medical claims, billing, and patient details. The extensive growth of Electronic Health Records Systems and digitization of healthcare is reducing costs and improving the quality of healthcare [1]. However, increasing the sharing of healthcare data requires more protection. Most healthcare information requires protection against unauthorized access, fraud and abuse. The healthcare industry is relatively unprepared when it comes to data security. Each healthcare organizations must take proactive, preventive measures to prevent fraud, breaching, and stolen medical data [2]. The data security is the biggest concern of the healthcare because the medical data are not perishable, which makes them particularly valuable. In India, privacy and security standards are linked together in order to provide security for data. Organization need to consider various aspects for adopting security measures, and satisfying the security requirements and technology is left to the individual organizations.

The most arduous area of research is data mining, which aims to locate valuable information and patterns in data. The basic goal of data mining is to find knowledge hiding in a large amount of data. Because of the continuous growth of data in every business, data mining is getting more common these days. Data mining is the best method for analysing and discovering usable information in a variety of industries, including financial data, retail data, biological data, scientific and engineering applications, intrusion detection, telecommunications, and healthcare. Among these, Healthcare industry is the vast area to study and apply data mining techniques to improve Healthcare services. In

particular, data mining draws data from various sources which creates different problem. For example, in healthcare, hospitals and physicians generally have to report the certain information for the several purposes. This often includes patient data such as demographic data, diagnosis data, treatment data, and physician data as well.

2. RELATED WORK:

Preserving the privacy and security of healthcare data using big data techniques is the latest study of researchers. A lot of work has been done on data security and protection by the researchers, academicians, hospitals, companies, and so on.

To provide the security and de-risking environment for the healthcare data, Tata Consultancy Services (TCS) IT business solution proposes a unique persona-based approach [3]. This method forces a secure repository to effectively assess, remediate, and monitor (ARM) data risks in the test environment can benefit healthcare companies protected the production environment. The authors suggested to healthcare organization for regular monitoring the physical, administrative, and technical security polices to safeguard Electronic Health Records. Authors used the ARM framework to prioritize, protect, and observe the susceptible data and claims. Finally, the IT solution team choses the persona-based approach accepted by the ARM framework, which ensures the development activities take place quickly and securely in the test environments. The proposed ARM framework is guarantees the development activities take place quickly and security in test environments by accepting the persona-based approach. The framework has many advantages, they are, improved regulatory compliances, risk reduction, integrity of application maintenance, and data masking consistence. The ARM framework is an extremely efficient method to arranging, persevering, monitoring susceptible data and applications.

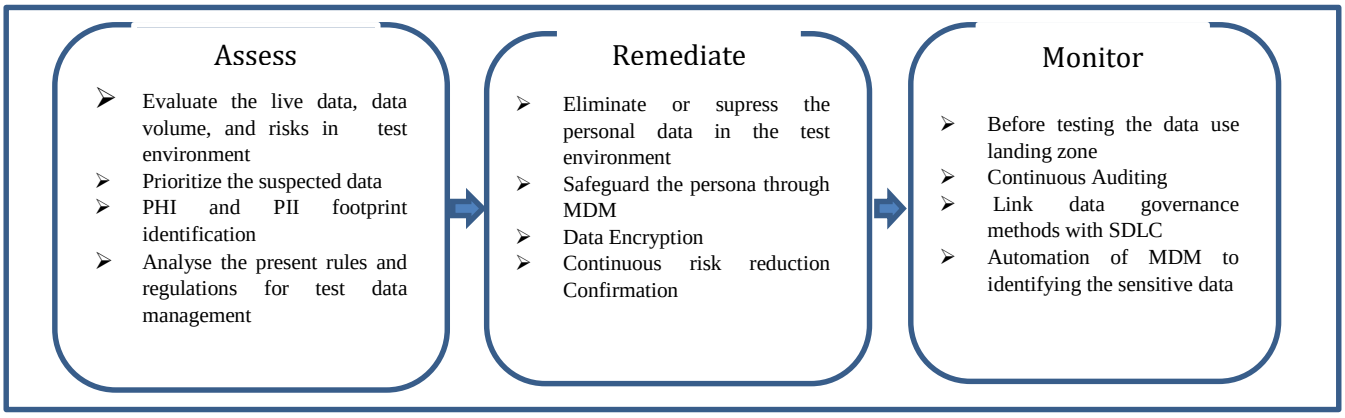


Fig.1. ARM framework.

The proposed ARM framework is guarantees the development activities take place quickly and security in test environments by accepting the persona-based approach. The framework has many advantages, they are, improved regulatory compliances, risk reduction, integrity of application maintenance, and data masking consistence. The ARM framework is an extremely efficient method to arranging, persevering, monitoring susceptible data and applications.

Data mining provides various privacy preserving data transmission methods in healthcare [4]. The researchers selected one of those methods to secure the healthcare data called anonymization. The authors presented a privacy-preserving framework that improves accuracy by relying on the history of security knowledge extraction approaches. The framework consists of three steps, first step is to anonymising the sensitive information. Secondly, creating privacy metadata to store history of privacy-preserving data transformation in standardized way. The last step includes an analysis of anonymised event logs using preserving metadata by conducting the privacy-preserving process mining. Using three publicly available healthcare event logs, the researchers evaluated the effects of various anonymization strategies on various process mining outcomes. The results of the experiment reveal that the influence of anonymization methods differs for different process mining algorithms and is dependent on the log's attributes.

Big data has changed the way of collecting, organizing, analyzing and managing the data. Big data technology is also successfully applied in healthcare industry. As other approaches for ensuring the data privacy and security big data also guarantees the preserving privacy and security [5]. The authors worked on addressing the attacks and threats to healthcare data by proposed method of combing existing methods. The approach consists a life cycle of big data security.

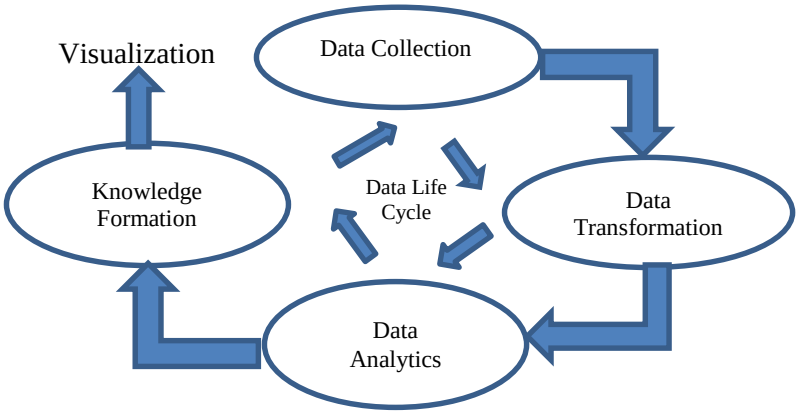


Fig.2. Big data life cycle

According to this life cycle, in every stage one should ensures the privacy and security in healthcare data. In the data collection phase, after collecting data from various sources must be protected from unauthorized access, duplication, modification, misuse, etc. During the data transformation phase also safeguard the data by filtering sensitive information by using security measures like anonymisation, data partitioning, etc. Later, in the data modelling phase, high level of security is needed since this phase uses the data mining algorithms. Only authorized staff allowed to access the data. In the last phase of data life cycle, sensitive information will be ready to take, so this phase requires security compliance. The researchers ensures the data security in every stage by applying various technologies such as data identification, encrypting, data manipulating, access control, observing and reviewing.

3. HEALTHCARE DATA SECURITY CHALLENGES:

In digital era, healthcare industry has a biggest challenge to secure the data from threats and security attacks. Electronic Health Records (EHRs) contains the patient personal information, lab reports, diagnosis details, and so on.

- **Electronic Health Records (EHRs) and Health Information Exchanges (HIEs)** - The Health Information Technology for Economic and Clinical Health (HITECH) Act was passed by US Congress. It encourages healthcare practitioners to use electronic health records (EHRs) for patients and health information exchanges (HIEs) to assist doctors share patient data. Data theft is a possibility when a huge amount of data is shared over a network.
- **User Error in Technology Adoption** – Once the user receives the information from healthcare providers, during sending over the internet and network forms, they have to take care of their records and personal information.
- **Rise of Hackers** - By breaking down the protected system, hackers also steal the health care data and access the personal data with a social security number and other details.
- **Healthcare's utilization of digital and mobile technology** – The healthcare industry generates an abundant of data which requires more space for storage. As a result, healthcare providers are opting the cloud technology to store the data. This causes the vulnerability and misuse of data.
- **Outdated Technology usage in hospitals –** Running hospitals is an expensive things with adopting new technologies up to date. Thus, hospitals slow down the process of the security updates time to time.

4. HEALTHCARE DATA PRIVACY AND SECURITY IN INDIA:

Several countries are applying different data privacy and security standards in healthcare. In India also many privacy and security methods are implemented to protect the healthcare data. The privacy and security standards are related each other. Any health record system includes protections to ensure that the information is accessible when necessary and that the information is not improperly used, disclosed, retrieved, changed, or removed during storage or retrieval or transmission. To safeguard the data, the security and privacy standards work together. Healthcare organizations need to consider many issues before acceptance of security measures. While adopting security measures, an organization must consider various factors includes size, capabilities, hardware and software requirements, the cost, and infrastructure also. The Ministry of Health and Family Welfare initiated the project called Digital Information Security in Healthcare Act (DISHA) with motive to secure the healthcare data in India, giving people complete ownership of their health data. DISHA offers three main objectives:

- Setting up a national and state level digital health authority.
- Enforcing electronic health data protection and security controls.
- Regulating storage and exchange of electronic health records.

➢ **Purpose of the Security Standards**

Healthcare providers are required by safety regulations to incorporate flexible and adequate administrative, physical and technological measures to:

- Ensure the privacy, security, and availability of all electronic protected health information (e-PHI) they create, send, receive, or store.
- Protection of e-PHI from threats and ensures security and integrity.
- Protect against e-PHI activities or disclosures that are not allowed or permitted under the Privacy Standards.
- Ensure that their employees follow their security policies and procedures.

➢ **Security Technical Standards**

In order to secure the e-PHI treated by a healthcare provider, as part of its safety plan, the provider must incorporate technological protections. Technical protections refer to the use of e-PHI security technology by regulation of access to it. Therefore, the following requirements must be discussed, with an emphasis on their functionality. It should be noted that they would need to use an EHR/EMR solution.

Sl. No	Guidelines for	Standards	Methods
1	Security and Privacy Requirements	ISO/TS 14441	a) Authentication b) Automatic log-off
2	Information security management	ISO 27799	a) Access Control b) Access privileges
3	Privilege Management	ISO 22600	a) Audit log
4	Audit trails	ISO 27789	a) Encryption b) Integrity
5	Digital Certificates	ISO 17090	Public key Infrastructure

Table.1.Technical standards for security.

HIPAA – Health Insurance Portability and Accountability Act.

The Health Insurance Portability and Accountability Act of 1996 (HIPAA) is a federal law that mandates the creation of national rules to protect the release of sensitive patient health information without the patient's consent or knowledge. The US Department of Health and Human Services (HHS) issued the HIPAA Privacy Regulation in order to implement HIPAA's requirements. The HIPAA Privacy Rule protects a portion of the information covered under the Privacy Rule. HIPAA has improved the way health care, insurance, life sciences, and other industries see and address health, safety, and confidentiality issues by establishing precise patient data security and confidentiality criteria. From a technological standpoint, HIPAA covers a wide range of topics, including websites, medical equipment, electronic health records, and medical imaging. The emergence of virtualization, cloud computing, smartphones, and mobile apps, among other recent technological innovations, has created additional barriers for payers and health care providers to comply with HIPAA. Companies will be able to meet HIPAA regulations while also improving their overall security and risk management.

HIPAA Privacy Rules

The Privacy Rule's criteria apply to the use and disclosure of "protected health information" from individuals by organisations that are subject to the Privacy Rule. "Data controllers" refers to these individuals and organisations. Individuals' rights to recognise and monitor how their health information is used are similarly protected by the Privacy Rule. The Privacy Rule's major goal is to secure individuals' health information while allowing for the flow of health information needed to provide and facilitate high-quality health care and protect the public's health and well-being. The Privacy Rule finds a balance between requiring necessary data uses while protecting the privacy of people seeking medication or recovery.

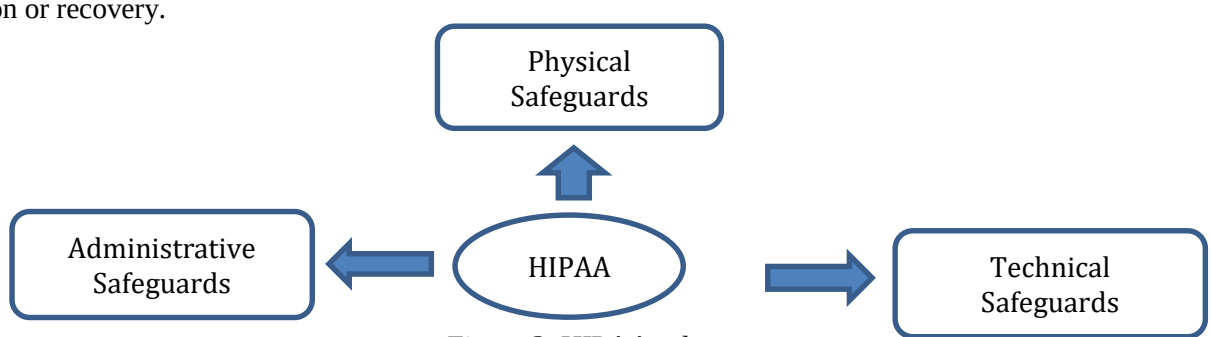


Figure.3. HIPAA rules.

Healthcare Information security and Privacy: Data Mining

Huge amounts of data have been generated as information technology has become more commonly used to assist essential healthcare activities. Different organisations require access to this information in order to use knowledge discovery strategies. Data mining presents unique challenges, particularly when data is gathered from different sources. For a variety of reasons, such as census, public health, and finance, hospitals and doctors are often required to divulge such information. Patient identification, ZIP code, race, date of birth, gender, service date, diagnosis codes (ICD9), treatment codes, physician identification number, physician Postcode, and total fees are also included. Compilations of this data have been made available to industry and scholars. Because such compilations do not include the patient's name, address, phone number, or social security number, they are considered de-identified and pose little risk to the patient's privacy. Processes like data mining can associate specific diagnoses with a person by cross-linking this data with other publically accessible databases [6]. To decrease the risk of harm to patients, their organisations, or themselves, data miners should have a basic understanding of the privacy and protection of health information.

The challenge of auditing data access and consumption, as well as evaluating intrusion detection and access control, has been solved using data mining approaches. Commercial technologies are available that automatically correlate and compare suspicious data obtained from numerous locations in computer systems, draw conclusions, and respond to potential assaults and security breaches. The major purpose of data mining privacy security is to create algorithms that modify the data such that personal data and private information stay private even after the mining process.

Detecting Fraud and Abuse – This entails creating regular patterns, then finding unexpected patterns of medical claims submitted by clinics, physicians, laboratories, and others. This tool can also be used to spot erroneous referrals or prescriptions, as well as insurance fraud and false medical claims. Although data mining is beneficial to healthcare, it has also raised some privacy concerns. Patients are concerned that large volumes of patient data will be transferred during the data mining process, and that their medical information will fall into the wrong hands. Experts, on the other hand, believe that this is a risk worth taking.

Data Breaches in Healthcare sector

Breach occurrences are widespread in the healthcare industry, and they can be caused by a variety of factors, including credential-stealing malware, an insider who purposely or mistakenly releases patient data, or missing laptops or other equipment. Personal health information (PHI) is more valuable on the black market than credit card credentials or conventional personally identifiable information. PHI is advantageous since it can be exploited by thieves to scare victims with fraud and scams based on the victim's medical problems or settlements. It can be used to make bogus insurance claims, allowing for the purchase and resale of medical equipment. Other criminals exploit PHI to gain unauthorized access to medications for their personal use or sales.

6. PROPOSED METHODOLOGY:

Every healthcare organizations need to store and maintain the data in the form of a Health Information System (HIS) electronically. The HIS documents are warehoused in the kind of Electronic Medical Record (EMR) or Electronic Health Record (EHR). Healthcare data in EHR can be categorized into three types: unstructured, semi-structured, and structured. Typically, structured data stored in standard way of accessing information easily and efficiently. It includes patient details, disease details, and treatment particulars and so on. Semi-structured data generally has the flow chart format. Unstructured data requires more processing and analysis methods for processing. This unstructured data contains clinical notes, discharge summary record, and laboratory reports. Figure.4. Depicts the EHR data processing approach.

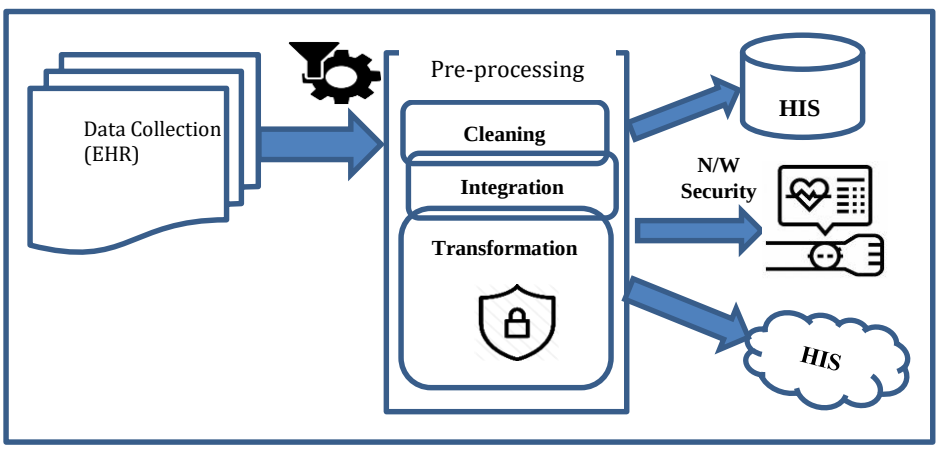


Figure.4. EHR processing.

Data Collection-Electronic Health Records are included the patient medical history, lab reports, clinical notes, scanning images, medicine prescriptions, insurance details, billing details, patient's personal information. EHRs are collected from various sources with different format containing noisy, outlier and inconsistent data [7]. The gathered data has to be cleansed using pre-processing techniques such as data cleaning, integration and data transformation.

Data Cleaning-The collected EHR data, which contains noisy, inconsistent, incomplete data, should be correcting the data inconsistency, smoothing noise, and filling the defaults. When collecting EHR data, particular data features may be lost due to human errors and system failure. For filling default value, there are numerous methods are available. If missing value records are less in numbers, just ignore them otherwise fill default value by column averages, possible value.

Data Integration- EHRs are gathered from heterogeneous sources or from different datasets. Integration deals with the assimilation of various EHR datasets into a single dataset that contains all the details essential for the analysis. This process improves the data mining speed and accuracy.

Data Transformation- The transformation includes the conversion of a variety of datasets into a unified format suitable for data mining. Data transformation process includes the data aggregation and data normalization.

Privacy protection- Compared to paper based medical records, the introduction of Electronic Health Record has significantly encouraged the development of health care, whereas it has also brought a lot of security problems. EHR consists personal information about patient's privacy, thus healthcare organization must ensures the protection for sensitive information related to patients. There are many ways to protect EHR privacy includes HIPAA privacy rules, privacy protocols, encryption, anonymization, and access control methods. EHR data to be store in several ways includes database, cloud, wearable devices etc.

6.1 Data Privacy in Transformation Phase

Data transformation ensures the unified form of data came from various resources. Preserving privacy during the EHR transformation needed many methods to be involved [8]. Transforming a variety of data types into a single standard form includes many techniques.

K-Anonymity is a resource that is managed by anonymized data. The release of data with scientific promises, given personal-specific field-structured data, means that it is impossible to re-identify the individuals who are the subjects of the information while the information remains technically relevant. The release of data is said to have a k-anonymity property if the information for each person contained in the release cannot be identified from at least k - 1 individual whose information also appears in the release.

Identifying variables: These include data that can relate to respondent recognition and can be further classified as:

Direct identifiers disclose the respondent's identity clearly and unambiguously. Names, passport numbers, social identification numbers and addresses are examples. Prior to publication, direct identifiers should be deleted from the dataset. Direct identifier removal is a straightforward method and is often the first step in generating a secure access set of data. However, elimination of direct identifiers is often not appropriate. Instead of removing direct identifiers masking technique can be used to hide the details.

Quasi-identifiers (or key variables) provide information that can lead to re-identification of the respondents when linked with other quasi-identifiers in the dataset. This is particularly the case when other external information or data may be used to match the information. Examples of quasi-identifiers are, birth date, sex, race, and marital status, which can be easily combined or connected to external information accessible to the public and make identification possible. The combinations of many quasi-identifiers are referred as key variables. Quasi-identifier values alone often do not relate to identity (e.g. male/female). However, a combination of multiple quasi-identifier values may make a record distinctive and thus identifiable (e.g. male, 14 years, married). To solve the problem, it is not generally advisable to simply delete quasi-identifiers from the data. For any sensible analysis, they will be critical parameters in several instances. In practice, as a quasi-identifier, any parameter in the dataset may be used. The statistical disclosure control solves this by defining and anonymizing variables as quasi-identifiers while also keeping the information for release in the dataset.

Non-identifying variables are variables that are unable to be used for respondent re-identification. This could be because no other data files or other external sources hold these variables which are not observable to a hacker. In the process, non-identifying variables are nevertheless essential, as they can contain confidential/sensitive information that may potentially be harmful if disclosure occurs as a result of identity disclosure based on variables being identified.

6.2 EXPERIMENTATION:

The data set of 10000 records were collected from online source emrbots.org website to conduct an experiment. The collected data contains the missing value, inconsistency, and noisy. The data pre-processed using data cleansing methods such as handling missing value and noisy value. After pre-processing 10000 records are reduced and get finally

8177 records with 8 attributes. Data integrated to united format, then during data transformation preserving privacy is very essential.

PID	PName	Gender	DateOfBirth	Marital_Status	PatientRace	Patientpopulatationperbelowpoverty	Diseases (ICD-10)
P100121	John	Male	20-01-1956	Married	African	15.16	B9719
P100127	Amith	Female	10-08-1989	Married	White	12.11	B9739
P200101	Dilux	Female	30-12-1999	Single	White	13.36	A001
P200110	Deol	Male	27-05-1975	Married	Asian	10.98	B570
P200129	Smith	Male	16-04-2000	Single	African	9.34	A1811

Table.1. Sample Dataset.

R programming is used to handle the different variables and process anonymization.

Direct Identifiers Quasi Identifiers Sensitive

PID	P.Name	Gender	Date Of Birth	Marital Status	Patient Race	Patient population perc. below poverty	Diseases (ICD-10)
P100121	John	Male	20-01-1956	Married	African	15.16	B9719
P100127	Amith	Female	10-08-1989	Divorced	White	12.11	B9739
P200101	Dilux	Female	30-12-1999	Single	White	13.36	A001
P200110	Deol	Male	27-05-1975	Married	Asian	10.98	B570
P200129	Smith	Male	16-04-2000	Single	African	9.34	A1811

Table.2. Classification of attributes based on the information

Handling Direct Identifiers – Either remove or masking the attribute values. In Table.2, patient identification number and patient names are considered as a direct identifier. The patient name is not necessary to identify the patients, thus we can remove the name column from the table. Nevertheless, patient_id is the key to identify the patients, which can be deal with by masking it.

PID
P10****
P20****
P21****

Quasi Identifiers suppressed using anonymization methods include k-anonymity and l-diversity by setting the threshold value for each column. After the suppression of the dataset, the dataset view changes. The suppression technique will help to preserve the data from intruders.

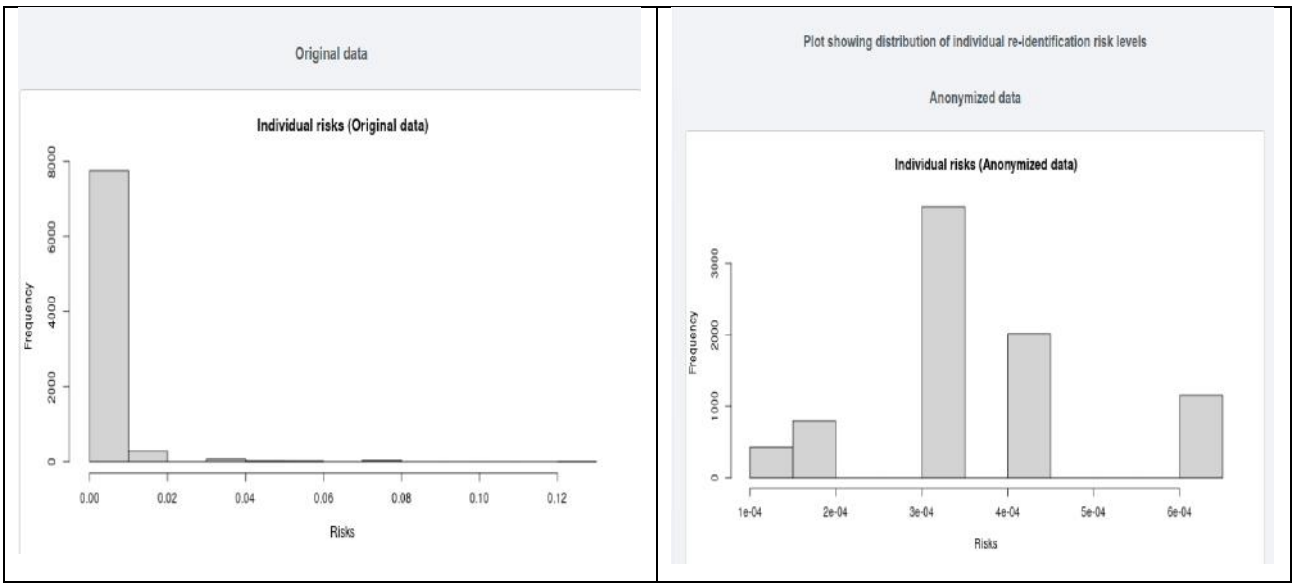
7. Result and Discussion :

Dataset with 8177 records suppressed with minimum threshold value for different attributes show modified records. The risk factors measures the risk of each attribute in the dataset. We can suppress the attribute having high risk by threshold value.

Fig.5. Example dataset suppression

In the experiment, patient gender, race, and marital status are considered as s key variable and applied the local suppression method to avoid disclosure of high-risk attributes. Thresholds set to obtain the modified data from original data for each attribute 0.002, 0.003, and 0 respectively. We considered the patient marital status to be the high-risk measure. Therefore threshold value will be 0. Overall expectation of the re-identification is 11.33(0.14%) in the population compared to 30 (0.37%) re-identification in the original data.

Risk measures-based on the individual re-identification risk value is 2.98 re-identification (0.04%) in the anatomized dataset. In the original data set expected risk is 30 (0.37%) re-identifications.



(a) Original data risk (b) Anonymised data risk

8. CONCLUSION:

Data mining approach promises the data security and privacy in healthcare industry. In the context of healthcare data privacy and security, also discussed privacy and security challenges at each stage of the data lifecycle, as well as the benefits and drawbacks of existing solutions. This paper mostly looked at recent privacy preservation strategies in healthcare and explored how anonymization methods have been applied. This work looked into the security and privacy issues in data mining by presenting various existing approaches and techniques for establishing security and privacy, which are expected to be very useful to healthcare companies. Local suppression will aid in the preservation of data while reducing the possibility of re-identification.

REFERENCES:

1. Srinivas & Arpita Biswas (2012). Protecting Patient Information In India: Data Privacy Law And Its Challenges. *A Journal of the NUJS Law Review*, Volume 5 Issue 3.
2. Hsinchun Chen (2005). Medical Informatics-Knowledge Management and Data Mining in Biomedicine. In Ted Cooper & Jeff Collman, *Managing Information Security and Privacy in Healthcare Data Mining* (pp. 95-137). New York, NY: Springer Science+Business Media, Inc.
3. Viswanathan Ganapathy and Daniel Logan (2016). *The Heart Of Healthcare Data Security-De-Risking Test And Production Environments*. Healthcare IT Solution Of TCS White Paper.
4. Anastasiia Pika and Et Al (2020). Privacy-Preserving Process Mining In Healthcare. International Journal of Environmental Research and Public Health. 17(5):1612.
5. Karim Abouelmehdi and Et Al (2018). Big Healthcare Data: Preserving Security and Privacy. *A Journal of Big Data*. Springer Publication.
6. Goodwin, L.K. And Prather, J.C. (2002). Protecting Patient Privacy In Clinical Data Mining. *A Journal of Healthcare Information Management.* 16(4):62-7.
7. Uma K and M Hanumanthappa (2017). Data Collection Methods and Data Preprocessing Techniques for Healthcare Data Using Data Mining. *International Journal of Scientific & Engineering Research*. Volume 8, Issue 6.
8. Meany, M.E. (2001). "Data Mining, Data surveillance, And Medical Information Privacy," In *Privacy In Health Care*. J, Humber, Ed., Humana Press, Pp. 145-164.

Web References:

- https://healthitsecurity.com/news/5-healthcare-data-security-challenges-and-solutions
- https://www.nhp.gov.in/data-privacy-and-security_mtl
- https://searchhealthit.techtarget.com/definition/HIPAA
- https://www.usfhealthonline.com/resources/healthcare-analytics/data-mining-in-healthcare