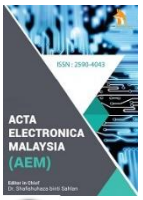
ZIBELINE INTERNATIONAL™ PUBLISHING

ISSN: 2590-4043
CODEN: AEMCDV

Acta Electronica Malaysia (AEM)

DOI: http://doi.org/10.26480/aem.01.2023.04.07



CrossMark

RESEARCH ARTICLE

DATA ANONYMIZATION IN HEALTH CARE INDUSTRY SURVEY PAPER

Monika Singh

Faculty of information technology, Gopal Narayan Singh University.
**Corresponding Author Email: singhmoni@gmail.com*

ARTICLE DETAILS

Article History:

Received 07 December 2022
Revised 11 January 2023
Accepted 14 February 2023
Available online 23 February 2023

ABSTRACT

As generations go by, our world is moving towards a technical dimension where information is becoming the most valuable resource. As a result, security and privacy must be taken more seriously as we develop. Therefore, when considering the most important human facilities needed for us to lead happy and comfortable lives, the health sector is a necessity. The data within hospitals needs proper protection and supervision by the government because research and medical studies depend on it. The numerous sorts of approaches that are accessible, as well as their benefits and drawbacks, were covered in this review paper.

KEYWORDS

Anonymization, E-Healthcare, Data Privacy, Metadata, Electronic Health Records, Doctors.

1. INTRODUCTION

Electronic health records (EHR) contain information that can be used by malevolent parties to quickly identify people and expose sensitive information about them. To examine these records and provide valuable findings and conclusions from the provided records, analysis must be carried out. Given the sensitivity of this information, protecting the privacy of medical data is crucial. If it gets into the wrong hands, it might have grave repercussions. Maintaining the usefulness of medical records is also important in order to use this data for surveys, analyses, and other purposes that will help to raise the standard of healthcare delivered. Our approach makes an effort to carefully balance these two goals so that neither data privacy nor utility are jeopardised. Techniques for anonymization are applied to accomplish this. These methods aid in safeguarding the data's privacy while maintaining the datasets' usefulness. A method that can be used to conceal data and restrict access to it is pseudo anonymization.

2. LITERATURE SURVEY

The authors of this research have put out a brand-new model for authorization and privacy in electronic health records. Three sections make up this model: The proposed approach, "Pseudonymization and anonymization with extendable Access Control Markup Language," is applied to a network architecture. A central database, an attribute server, and a data server comprise the network model's three server structures. The attributes are segregated from the data, and the data server stores attribute-less data. On the attribute server, the attributes are kept. For requests to this network, the central server serves as the gateway. The central server receives a request from the client. The central server sends an authorization request to the attribute server if the request is legitimate. The user is then given permission by the attribute server using a policy decision point model. The data server receives a request from the authentication server if the user may be authorized, and if the various values it needs are provided and meet the requirements, the data server responds with a pseudonym and the necessary data. The user is then given access to the dataset by the attribute server, which then replies to the client server. The usage of two policy decision points has improved authorization control.

In order to address the issues with privacy in medical records, random pseudonyms were used in this paper. The accessible medical records don't have any information about the patients' names or other explicit identities. They suggested a 4-dataset paradigm to boost privacy. Patients' traits are contained in one dataset, pseudonyms in another, user policies in the third, and users' data in the fourth. A user role and a randomly generated internal user number are both included in the pseudonym of a new user when they are introduced to the system. These are used for verification purposes only; they are not transmitted. These pseudonyms can be used to examine user data without exceeding privileges because they are associated with specific users. In this paradigm, there are seven main categories of authorization methods. There are four protocols, with three being for indirect users (Alz et al., 2019).

There are three steps in this technique for anonymizing client-provided data. Initialization, client- and server-side anonymization are the three. After startup, the grouping technique that creates anonymous attributes that meet the a, k requirements is employed. By adopting a technique based on the UPGMA cluster combination method to compress the data acquired from client-side anonymization, the communication cost is reduced on the server side. Here, anonymization enters its second step. On the client side, the original attributes are divided into groups using a bottom-up grouping mechanism. The objective is to create a pairwise dissimilarity over the dataset using the quasi-identifier attributes to assess the dissimilarity. The closest differences are then selected to form a group, and they are changed to satisfy the client's requirement for anonymity. The subsequent steps guarantee that each group's sensitive features adhere to the a,k anonymity criterion. The centroids are set up so that there is minimal information loss as new data is added. All of the generated groups' quasi-identifier properties are determined by centroids.

Data received from the client is compressed as described above on the server side. The closest group pairings are then created by calculating the Euler distance, and these pairs are subsequently joined. The values of the distances between these recently established groupings of characteristics are updated to create the distance matrix. This procedure is repeated until all values have reached level 2 anonymity. On both the client-side and the server-side, anonymity is protected in this way. When there are big datasets, this approach gets more anonymity. Additionally, information

Quick Response Code	Access this article online	
	Website: www.actaelectronicamalaysia.com	**DOI:** 10.26480/mecj.01.2023.04.07

Cite The Article: Monika Singh (2023). Data Anonymization in Health Care Industry Survey Paper. *Acta Electronica Malaysia, 7(1): 04-07.*

loss continues to decline as dataset size grows (Li et al., 2018). A differential privacy strategy for electronic medical records is proposed in this research. The level of access is often either complete access or none at all. Users in this article have control over the access levels. Four levels can be used to categories the access level.

3. BACKGROUND

There is no access to the information:

Data description: The requester is provided with a description of the data in the dataset.

Sample: Data access is granted with limited operation access for a predetermined period of time. Although access to unmodified data is occasionally provided, most data is altered to provide protection.

Full access: Unprotected access to raw data is provided.

In addition to this, obfuscation and K-anonymity are used to increase anonymity.

K-anonymity: To safeguard the original data, similar patient datasets are combined.

Data is obfuscated by adding noise, which obscures the data. Depending on the desired final utility of the data, the degree of noise can be altered.

Protected data is displayed as a range of values rather than as a set of specific numbers. Exponentially weighted moving average: The data is subjected to a straightforward EMA calculation, and a standard weight is selected for this reason.

Probability distribution: Instead of providing the data requester with the raw data, the median and standard deviation are provided.

Average trends: Using pairs of all nearby values, the average of two points in this time series data is calculated and presented on the graph.

Data that has been altered: A minor amount of modification is made to the data in order to preserve the original values, and this small amount of change has no appreciable impact on the data's utility.

The data has been obfuscated. This will alter the results of the aforementioned calculations, giving the user more privacy. The data is further cleansed of any identifying information using k-anonymity (Sahi et al., 2017).

In this research, a six-stage paradigm for anonymizing E-healthcare systems is provided.

User registration: To use the E-health system, a user must first register by entering their name, username, password, and email address.

Login: Using the credentials they have already provided, the user logs into the E-healthcare system.

By filling in both private and shareable attributes, the user can now construct their own medical dataset by uploading a file. The server conceals the private fields and makes the remaining fields anonymous after receiving the medical records upload. The Data Encryption Standard (DES) algorithm is then used to encrypt all of this data before it is saved on the cloud server.

Users have the opportunity to download their datasets as files by using the file upload feature. The server sends the user the requested files when the user requests a file for download. By providing the password for the DES algorithm that was used in the File upload process, the user must decrypt the file themself.

Verification by a Third-Party Auditor: The auditor looks for any potential dangers in the security of the cloud server as well as the data that is leaving it. The final user, who in this situation may be either the uploader themselves or their doctor with the users' permission, receives the data and can download it.

When seen separately, pseudonymization and anonymization each have advantages and disadvantages. So, a framework has been suggested in this paper.

There are 6 steps in the framework:

- Distinguish between non-critical and critical information in these documents, then store them without any security.
- Give the crucial components pseudonyms and use them as identifiers; you can also use them as access authorization tokens. Every new user that has access to a pseudonym generates a new one.
- Export the metadata for the document as an XML file. This metadata includes the document's structure, description, and key phrases.
- By encrypting the metadata with a key that the document owner supplies, it is ensured that only someone with the password or prior knowledge of the document will be able to identify the connections between the document's fragmented sections.
- The metadata of the fragmented document is supplied to the person who requested access to it so they can review the structure described there and realign the document to its original condition. The resulting XML can be used to search the documents, and the metadata is then uploaded to this person's database so it can be used again.

To connect two distinct parts of a text, pseudonyms are created at random and aren't based on any characteristic or thing. By using pseudonyms, you may prevent the actual identity of the document owner and the entirety of the document from being compromised in the event of a document leak.

A layer-based security model is used in this model. Each layer has a phase in the data access process assigned to it. The inner key that was used to encrypt the inner symmetric key is decrypted using a user-supplied key (private key). The metadata has been encrypted using this inner symmetric key.

Each user has an own metadata database. Every time a new file is added, its metadata is encrypted using this user's keys and then saved in this user's storage.

The authors suggest using a smartcard or pre-authorized token system with a user-known pin for improved security. This offers a solution for two-factor authentication. The encryption of any values in this framework uses either the RSA algorithm or the AES technique.

If the entire framework is installed locally, the smartcard hardware is where authentication and authorization take place. The smart card only keeps the first outer key that was used to encrypt the inner key in a server deployment scenario (Heurix et al., 2012).

Numerous other models that have been applied in various fields are based on this concept. With the help of this document, users are given the option to control who has access to their data and how much of it they can view or edit.

Each user who has registered with the system has been given a role. Patient, relative, doctor, and operator are examples of roles. The patient may grant access to a relative if they so choose. Relatives may have access that is either very limited or nearly identical to that of the record owner. The doctor has access to all of the patient's data. With the patients' consent, the operator has access to edit that data.

The Hull architecture is the foundation of this framework. Any of the inner hulls cannot be accessed without having access to the outer hull since each hull needs a different key to be decrypted. Patients' access levels can be pictured as being on the innermost hull, followed by the operator, the doctor, and then any place between the innermost and the outermost hull by families.

The patient's private key is backed up to avoid losing their key, which would render the data unusable and cut off fresh access to practically everyone. The key is cut into pieces and distributed to the people with access. The original key can be built if enough pieces are collected.

The user or patient reports the loss of a smartcard to the operator, who then tells the other operators. The user's private key is then retrieved by assembling the pieces from the other operators. This occurs when the system key is used to encrypt the patient identify before sending it to other operators to notify them of the loss. Any encrypted data can then be decrypted using the keys that were created from the gathered fragments after a new key pair was created and the old ones were replaced. The new private key will be programmed or stored on the users' new smart card, and this data will be encrypted using the new keys.

The users' private information will be first isolated and then encrypted

with the doctors' internal public key before the entire set of data is anonymized. then the patient's public key was used for the opposite. The creation of two new random numbers serves as a pseudonym between the patient and the doctor. The doctor receives this information and uses their private key to decode it along with the alias. When patients receive information from the doctor, they perform the reverse of this operation. The link between separated data and private information is created via the interchange of pseudonyms. The patient's number of pseudonyms determines how much data will be de-anonymized on the client-side storage, therefore whenever data needs to be stored, the amount of anonymization that needs to be applied to the data is dependent on the patient. Once they are added to the database, all datasets may be easily pseudonymized (Riedl et al., 2008).

The purpose of this study is to discuss how an Internet of Things (IOT)-based system can protect the privacy and security of health data. It uses a variety of methods and techniques for cryptography, anonymization, and pseudo-anonymization. Various of these techniques are used with all types of data, including data that is in use, at rest, and in motion. The OCARIoT programme is a dynamic tool that connects with a variety of organisations, including teachers, parents, healthcare providers, technology companies, the government, etc. It gathers data about schoolchildren and their parents. It is a versatile instrument used to record data about kids from all facets of their lives. Given the sensitive kid and health information, protecting all of this information is absolutely essential. Data is appropriately anonymised and pseudo-anonymized so that only the relevant authorities may understand it. For instance, teachers shouldn't be able to interpret a child's data because doing so would make it simple for them to identify the youngster. Given the interconnected nature of an IOT-based system, this process is incredibly arduous. The study emphasises how crucial it is to carefully examine each component of this IOT-based system to make sure that privacy and security norms are upheld. Given the interwoven nature of this ecosystem, any vulnerability might jeopardise the entire application. For OCARIoT to be a success, a comprehensive security plan with an understanding of all individual components and how they interact is required (Ribeiro and Nakamura, 2019).

Each year, a large number of biomedical journals and papers are published, each one featuring cutting-edge developments in medical practise and study. The most esteemed and well-known journals are The Lancelot and The New England Journal of Medicine. They release datasets containing private information about people while simultaneously sharing material that is helpful to humanity. In order to prevent reidentification of people, this article suggests anonymizing the data in order to protect individual privacy. Nevertheless, it's important to preserve a lot of data so that it can be utilised for analysis and predictive modelling. To make sure of this, classification metrics and metrics for information loss are both employed. ARX, an open source anonymization tool for de-identified datasets, has been expanded to allow for the development of strong statistical classifiers as well as in-depth performance analysis. The major goal is to allow medical professionals and specialists to openly distribute these statistical classifiers, which are privacy-protected, for their use. Using k-anonymity, the data sets are first made anonymous. Then, the information loss is calculated using different metrics, such as the Information Loss Metric and Iyengars Loss Metric, to exceptionally correctly anticipate the results of health. For instance, this research was able to precisely estimate the cost and length of a patient's stay at a hospital using the patient's vitals, such as blood pressure, heart rate, etc. (Prasser et al., 2017).

Electronic health records (EHR) contain information that can be used by malevolent parties to quickly identify people and expose sensitive information about them. To examine these records and provide valuable findings and conclusions from the provided records, analysis must be carried out. Given the sensitivity of this information, protecting the privacy of medical data is crucial. If it gets into the wrong hands, it might have grave repercussions. The usefulness of medical records must also be preserved in order to use them for surveys, analyses, and other purposes that will enhance the standard of healthcare delivered. Our approach makes an effort to carefully balance these two goals so that neither data privacy nor utility are jeopardised.

Techniques for anonymization are applied to accomplish this. These methods aid in achieving utility while protecting data privacy. The most popular strategy for accomplishing this in the medical sector is generalisation. In this study, we put into practise a utility-preserving strategy for disseminating data while protecting privacy (PPDP). Three main steps or categories make up the approach. The utility-preserving model is the subject of the first. Then we put the fake records in. The bogus records are then categorised. Full domain generalisation is used here. The disadvantage of earlier techniques like suppression and relocation is that they cannot handle huge datasets. Our suggested method, when measured across all criteria, exhibits a smaller information loss than the currently used methods while maintaining utility (ILM). Knowledge discovery and predictive modelling are only possible until the information loss is kept to a reasonable level.

4. Proposed Methods

This analysis combines logical explanatory, curio-building pseudonymization, and research approaches for evaluating historical rarity. The article starts out by exploring the security insurance tools that are currently available, such as encryption, anonymization, and pseudonymization, by looking at and analysing relevant work. In light of these findings and the acknowledged shortcomings, the pseudonymization strategy is described and evaluated using techniques for a danger analysis.

Since the records have already been stored in a pseudonymized form, factual requests that protect security may proceed without further anonymization. The following results were obtained from testing that involved searching for and recovering pseudonymized records: As knowledge The real calculation, which was rather simple in terms of CPU requirements, and cryptography both added a tiny burden to the overall recovery as a result of the estimation exhibitions. The implementation of the smart card as a cryptographic environment becomes a limiting factor on the consumer side. Even though the validation activity is only performed once for each meeting, meeting key encryption had a significant impact on the overall recovery activity, especially when scrambling reports. In this sense, workstations with based encryption are more appropriate than smart cards. The effect on security is therefore minimal because the fundamental internal symmetric key is still kept secret inside the HSM (Neubauer and Heurix, 2010).

Electronic health records (EHR) contain information that can be used by malevolent parties to quickly identify people and expose sensitive information about them. To examine these records and provide valuable findings and conclusions from the provided records, analysis must be carried out. Given the sensitivity of this information, protecting the privacy of medical data is crucial. If it gets into the wrong hands, it might have grave repercussions. The usefulness of medical records must also be preserved in order to use them for surveys, analyses, and other purposes that will enhance the standard of healthcare delivered. Our approach makes an effort to carefully balance these two goals so that neither data privacy nor utility are jeopardised.

Techniques for anonymization are applied to accomplish this. These methods aid in achieving utility while protecting data privacy. The most popular strategy for accomplishing this in the medical sector is generalisation. In this study, we put into practise a utility-preserving strategy for disseminating data while protecting privacy (PPDP). Three main steps or categories make up the approach. The utility-preserving model is the subject of the first. Then we put the fake records in. The bogus records are then categorised. Full domain generalisation is used here. Suppression and relocation techniques from the past are scalable to big datasets. Our suggested method, when measured across all criteria, exhibits a smaller information loss than the currently used methods while maintaining utility.

4.1 Proposed Approach

There are many different encryption algorithms available, including AES-256, which is currently the strongest of all. The author also suggested using k-anonymization, one of the most effective and widely used data identification techniques, but there are some minor flaws that could make it vulnerable. These two parameters are anonymized data and a strong patient anonymity level (Shah et al., 2018).

5. Conclusion

In order to allow security assistant use of the prosperity records inclinical assessments without requiring further anonymization steps, this paper presents a new method for the pseudonymization of clinical data that stores data that is unconnected from the looking at comprehension recognising information. Instead of clinical evaluations, which are not necessary to the individual people, protection organisations, and managers of the economic status of persons, for example, likely insurance or occupation applicants. If hospitals successfully apply this notion, their earnings will significantly increase since they will be able to use data to accurately anticipate the cost and length of a patient's stay without jeopardising the patient's privacy. In terms of security conservation and information security for e-medical services, the current protection

conservation solutions have a few flaws. The primary goal of this work was to identify those flaws because, according to the suggested thinking about the requirements, none of these systems had the three obligatory bounds.

The issue of EHR privacy in the cloud needs special consideration from the scientific community. They have implemented a number of algorithms to safeguard the privacy of the user. We have discussed a method for data anonymization in this paper. For data privacy, we propose a fixed interval technique. Other similar systems can employ this strategy. We compared the utility of the fixed interval technique versus generalisation. The approach's primary tenet is that the quasi-identifiers found in EHRs should be correctly categorised in accordance with predetermined intervals, and that the original values should then be replaced with the mean of those original values. This approach is prolificandefficient.

REFERENCES

Alz, M., Zhang, Z., and Zhang, J., 2019. PAX: Using Pseudonymization and AnonymizationtoProtectPatients'IdentitiesandDatayntheHealthcareSystem.InternationalJournalofEnvironmentalResearchandPublicHealth.16.1490.10.3390/ijerph16091490.

Heurix, J., Karlinger, M., and Neubauer, T., 2012. Pseudonymization with Metadata Encryption for Privacy-Preserving Searchable Documents. 2012 45th Hawaii International Conference on System Sciences, Maui, HI, Pp. 3011-3020, doi:10.1109/HICSS.2012.491.

Li, H., Guo, F., Zhang, W., Wang, J., and Xing, J., 2018. (a,k)-AnonymousScheme for Privacy-Preserving Data Collection in IoT-based Healthcare Services Systems. Journal ofMedicalSystems, 42 (10).1007/s10916-018-0896-7.

Neubauer, T., and Heurix, J., 2010. A methodology for the pseudonymization of medical data. International journal of medical informatics. 80.190-204.10.1016/j.ijmedinf.2010.10.016.

Prasser, F., Eicher, J., Bild, R., Spengler, H., and Kuhn, K.A., 2017. A Tool for Optimizing De-identified Health Datafor Use in Statistical Classification. 2017 IEEE 30th International Symposiumon Computer-Based Medical Systems (CBMS), Thessaloniki, Pp. 169-174. doi: 10.1109/CBMS.2017.105.

Ribeiro, S.L., and Nakamura, E.T., 2019. Privacy Protection with Pseudonymization and Anonymization In Health IoT System: Results from OCAR IoT," 2019 IEEE 19th International Conference on Bioinformatics and Bio engineering (BIBE), Athens, Greece, pp. 904-908. doi:10.1109/BIBE.2019.00169.

Riedl, B., Grascher, V., Fenz, T., and Neubauer, T., 2008. Pseudonymization for improving the Privacy in E Health Applications, "Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS2008), Waikoloa, HI, 2008, Pp. 255-255. doi:10.1109/HICSS.2008.366.

Sahi, M., Abbas, H., Saleem, K., Yang, X., Derhab, A., Orgun, M., Iqbal, W., Rashid, I., and Yaseen, A., 2017. Privacy Preservation in e-Healthcare Environments: A Review. IEEE Access, 6, Pp. 464-478. 10.1109/ACCESS.2017.2767561.

Shah, A., Abbas, H., Iqbal, W., and Latif, R., 2018. Enhancing E-Healthcare Privacy Preservation Framework through L Diversity, 2018 14th International Wireless Communications & Mobile Computing Conference (IWCMC), Limassol, Pp. 394-399. doi: 10.1109/IWCMC.2018.8450306.