Awareness Survey of Anonymisation of Protected Health Information in Pakistan

Muhammad Usman Shahid[1]
Faculty of Computer Science
IBA
Karachi, Pakistan

Waqas Mahmood[3]
Faculty of Computer Science
IBA
Karachi, Pakistan

Saman Hina[2]
Department of Computer Science & Software Engineering
NED University of Engineering & Technology
Karachi, Pakistan

Hamda Usman[4]
Saulat Institute of Pharmeceutical Sciences & Drug Research,
Quaid-e-Azam University,
Islamabad, Pakistan

***Abstract*—With the growing advancement of science and technology, research has become the vital step in every educational field. This research survey sheds light on the methods of de-identification and anonymisation for protecting the privacy of the patients, practitioners and nurses. Researchers require huge amounts of patient data for carrying out different analyses. Patient information must therefore be preserved while ensuring that the applied privacy policies do not render the data less valuable. De-identification and anonymisation techniques masks the patient identity through various methods such as suppression, randomisation, shuffling, creating pseudonyms, generalisation, adding noise, scrambling, masking, encoding and encryption, etc. The dataset having critical information is called protected health information (PHI) through which an individual can be identified. Thus, PHI must be preserved through an appropriate means to make data valuable and at the same time, protect the data from hackers. This paper presents the importance of securing PHIs in Pakistan by analysing the results of an awareness survey.**

***Keywords*—Anonymisation; De-Identification; Protected health information; Patient data**

I. Introduction

As in the case of all other fields of study, research is increasingly being done in the field of healthcare also. Researchers now work on evidence-based studies for which they require real world data sets. In the field of heath care, such data sets contain original patient medical cases. To maximise the utility of contained information, the data needs to be in a readily useful form. In addition, protection of the information is also critically important. This is because the information contained in such data sets could be leaked as data breaches and then such data could be used for false purposes. Data must thus be presented in de-identified/anonymised form in order to protect privacy of the patient without disclosing any identifiable information of the patient and other related personnel. When it comes to the healthcare domain, sharing of data without any ethical review can cause serious consequences. This is because, generally speaking, every individual is concerned about his/her personal health information and do not want to share it with anyone other than his/her healthcare professional. In parallel to this fact, it has also been observed that patient's data is a very useful source to develop decision making systems by applying artificial intelligence techniques. This data can provide unseen facts and can be of great help for researchers in predicting useful information in healthcare domain.

De-identification is the technique to remove identifiers form patient records, thus minimising the risk of unintended disclosure of personal information. On the other hand, Anonymisation is the method of de-identification through which data cannot be reverted to its original form [5]. In order to secure patient's health information researchers have developed automatic systems to secure data for research purposes (Neamatullah, 2008).

Protected health information (PHI) has been described as the evidence with the help of which an individual can be recognised [1]. HIPAA (Health Information Portability and Accountability Act) Privacy Rule from the US Department of Health and Human Services is the principal recommendation for de-identifying personal health information (PHI) (TransCelerate 2013). HIPAA provides eighteen standard PHI categories for the de-identification of clinical data which are used in place of original data [5].

This paper discusses various methods of de-identification and anonymisation providing privacy to patient's personal information and minimising the risk of data being leaked while ensuring that the data is available to the public in its most utilisable form.

The format of this paper is made such as this section is followed by the literature review that contains summary of research articles of the related domain followed by a section pertaining to an awareness survey conducted for the given case study. Finally, case study is summarised and concluded in the closing sections of this paper.

II. Background to Design Survey

HIPAA is the main act given by the US government for preserving privacy of patient clinical data. The main work discussed in the following text summaries how real patient data is converted into encoded form through the use of eighteen categories described by the HIPAA privacy act. These categories mainly include the names, age, zip codes and IDs, etc. These categories are either removed or set as blank through different de-identification techniques to make data unrecognisable to a researcher using the data. Thus, in this way privacy is preserved and data is made indistinguishable as well. In this section, we have discussed the summary of each article that was selected for research survey to cover the section of individual articles summary review. This research survey would circulate around the idea presented in these articles. Table 1 contains the lists of these articles.

TABLE. I. Articles Summarised

Sr No.	Articles considered to design survey
1.	*Practical Implications of Sharing Data: A Primer on Data Privacy, Anonymisation, and De-Identification.*
2.	*HIDE: An Integrated System for Health Information DE-identification.*
3.	*Data De-identification and Anonymisation of Individual Patient Data in Clinical Studies- A Model Approach.*
4.	*An Intelligent Framework for Protecting Privacy of Individuals* *Empirical Evaluations on Data Mining Classification.*
5.	*An Innovative Approach for the Protection of Healthcare Information Through the End-to-End Pseudo Anonymisation of End-Users.*

In the first paper used for designing the survey, the importance of protecting healthcare data has been a major concern. Healthcare data is vulnerable to data breach because it is easier to target. Healthcare data can be used to file false tax returns, open lines of credit, or claim medical benefits or to acquire prescription [3].

Protection of data must be done on the basis of sensitivity of data. Protected data within firewall is no longer considered as protected and secure to use.

In the struggle to make healthcare data protected, HIPPA privacy rule has been practiced in the US. HIPPA protects healthcare related information by either outlining appropriate uses and disclosures of PHI or as authorised by the individual subject of information.

HIPPA uses two mechanisms: HIPPA safe harbour method and statistical or expert determination methods.

HIPPA safe harbour requires removal of eighteen data elements from data to be de-identified without destroying the key of hidden identifiers. This method is not suitable for protecting data against advanced methods of re-identification. On the other hand, expert determination or statistical method requires finding professionals experienced with the rules governing identifiable information.

To make our data more secure, we must use appropriate techniques of de-identification. Identifiable of data is measured in order to make our data accessible yet securing the individual identities. Data identifiability model have 5-levels such as:

Level-1: Readily identifiable data.

Level-2: Masked data contains modified ‘identifying’ variables through randomisation and creating reversible or irreversible pseudonyms.

Level-3: Exposed data contains masked identifying variables as well as quasi-identifiers.

Level-4: Managed data contains least personal information.

Level-5: Aggregate data that cannot physically identify individuals.

De-identification and anonymisation are the methods used to protect personally identifiable data. De-identification focuses on removing identifiers from data set to minimise the risk of exposure of personal identity and information, while anonymisation is a process where fields that relate to individuals are removed from data set so that it cannot be linked back to original data set (TransCelerate 2013).

De-identification methods involve the demarcation of direct and quasi-identifiers in order to apply appropriate technique.

Continuous control of privacy is to be done through pro-activeness. In practice, researchers tried to cover typical process to overcome the gap like training of HIPPA, Access control of data, DBA training and such type of learning to enhance security.

There can be many breaches like using GPS system, position can be determined if one used GOOGLE API, unencrypted data recovery, online communication used for data transfer, etc. Such activities happen on daily basis so keeping track and identifying such data and measures against it should be done to make the data maximally secured. These are the main methods of protection like physical protection, encryption or cryptography, password management, protected data.

De-identification of both structured and unstructured data is reported by Sweeny. The major hurdle in data anonymisation is preservation of identifiable information while giving sufficient/optimal information to researches. This work shows that removing identifiers was not useful as it was linked to attacks [7]. Privacy protection was provided by using techniques such as generalisation, suppression (removal), permutation and swapping of certain data values, all following k-anonymity dominantly [1].

Other efforts of data de-identification include de-identification of medical text document that focuses on subset of HIPPA identifiers (e.g. name only). Some efforts focus on differentiating protected health information from non-protected health information.

HIDE is a prototype system for de-identification of structured and unstructured data. It is a two-step system. It involves data linking, in which structured person centric identifiers view is generated in which identifying attributes are linked to each individual. Identification and sensitive information extraction is the next component which used named entity extraction technique specifically conditional random fields (CRF) that extracted identifying and sensitive

information form unstructured data efficiently [4]. Anonymisation involved suppression and generalisation of identifiers view through different option of full, partial or statistical de-identification based on k-anonymisation [7].

Protected health information (PHI) is defined by HIPPA as individually identifiable health information [1]. Identifiable information means data through which an individual's identity could be traced. Personal identifiers include both direct identifiers and indirect identifiers.

Privacy models of de-identification have three forms:

- Full de-identification is done if all the identifiers are removed. As a result, it becomes nearly impossible to identify individuals in the data.
- Partial de-identification: According to HIPPA, suppression of direct identifiers is done and indirect identifiers are left unchanged.
- Statistical de-identification: In this privacy model as much privacy is protected as possible in such a way that it is sufficient to use for research purposes as it provides most of the useful data while optimising security to patient information.

The framework presented in this paper has number of components that were de-identified from heterogeneous data space using advanced anonymisation. Firstly, data is processed through data linking and identifying sensitive information simultaneously in cyclic form. This is followed by anonymisation to get the output. All of the HIPPA attributes are used for de-identification.

The technique used in this paper for extraction of attributes, is built on training data set produced by tagging done through a tagging software. In the second step, classification of terms was done. In the third step, data was processed for extraction. Unique function of this work is iterative process using one hundred pathology reports for experiment. The reports were tagged manually with identifiers like name, medical records, date of birth and age. After checking the accuracy, data was retagged as and when required. Lastly, the data was linked with de-identification through k-anonymisation [1].

In this paper individual patient data (IPD) is protected through the use of techniques such as Safe harbour method in addition to expert determination methods. The approach outlined here in this paper is primarily based on the enhanced safe harbour method [8].

Both these method follows a general principle of recognising the direct and quasi-identifiers as the first step and then applying the appropriate de-identification technique.

De-identification begins with the process of de identifying identifiers; individual privacy is maintained by generating/creating a new random code. The investigation is also given a new random code and participants from one investigator are assigned the same code to maintain relationship between them. All contact numbers and names are removed. In case of extension of main study, both the main study and its extension must utilise the new random code generated.

Dates present in any dataset are de-identified using two methods namely "offset date" and "relative study date". In offset date method, all the dates such as visit date, date of birth and date of adverse events are replaced with a new date for each participants. Complete study could be given a single new date but in order to achieve better privacy it is recommended to assign a new date to each individual.

In the relative study date method, the date of birth and age must follow the HIPPA privacy rules using the safe harbour method. Any age less than 89 years must be displayed through variable and anonymised age is greater than 89 years. Categories can also be made using a five year class gap such as <25 years, 25-29 years, 30-39 years, 85-89 years, >89 years, etc. Medical dictionaries are used by data providers to code diseases and medications. A medical dictionary such as MedDRA is used for adverse events and diseases. On the other hand, WHO drug dictionary is used for medication widely.

MedDRA allows all five levels of coding including system organ class, high-level group term, preferred term and lowest term. WHO Drug provides trade names and ingredients encoding medication.

Data providers must mention name and version number of each dictionary that is used so that a researcher can use suitable dictionary to code data set. Extra attention must be given to lowest level terms and product names of low frequency as they need more appropriate/proper aggregation to maintain privacy. In order to secure the privacy of free-text verbatim fields, de-identification is done in such a way that the original data set containing personal information is anonymised and written in the form that reflects original context of the document. This can be done by replacing personal information with data which do not reflects identity of any individual.

Data that contains rare diseases, rare vents, genetic information, extreme values (height, weight, BMI) or sensitive information must be mentioned as "redacted" or alternative techniques such as "adding noise" (offset method for dates) or aggregating data (defining age bands) is recommended to preserve patient privacy with maximum data utility for researchers [8].

Quality control is the main game changer of the whole de-identification technique. Data provider must confirm the de-identification method before the key identifier is removed because it cannot be reverted once lost.

Enhanced save harbour approach works to remove all of the eighteen HIPPA identifiers as well as additional information. Thus providers must not rely on automated system of de-identification and manual reviews must be done.

The paper highlights the need for patient privacy through utilisation of advanced technologies such as data mining specifically "Privacy preserving data mining (PPDM)." Privacy can be labelled as "distributed" and "centralised" according to privacy preserving data mining technique [6].

In case of distributed privacy, data is not published and only the required final output is achieved as end result. Privacy is preserved through the use of cryptogenic techniques. In case of centralised privacy, data is circulated to public after it has

been handled through various techniques including, anonymisation, perturbation, condensation, randomisation and fuzzy-based method. Although the data is not encrypted, protection of patient data before data is being published to the public is a prime concern. Generalisation technique shows its effects on every data field causing data accuracy issue. On the other hand suppression technique alters few tuples of the table thus rendering data incomplete. K-anonymity is the most authentic technique among all other techniques to preserve patient privacy [7]. It is based on generalisation and suppression which can overcome the problems of linking attack. The major drawback of prevailing algorithms is that these can cause information leakage due to accuracy and completeness of data. Secondly, the background knowledge attack cannot be handled in this case [6].

Privacy preserving data mining thus implemented adaptive utility-based anonymisation that has the ability to fight disclosure risk. The table created is called micro-data table. It has four attributes as follows.

- Explicit identifiers: These can instantly identify the individual such as name, ID, etc. They are usually hidden or their values are hidden.
- Quasi Identifiers: These when attached with the other information can identify an individual e.g., Date of Birth, Gender, etc.
- Sensitive attributes: These are person specific sensitive information; For instance, disease, income, etc. Protection of this attribute is the major focus of privacy preserving data mining.
- Non-sensitive attributes: When leaked, this attribute presents no problem, thus are least useful for attackers. Several attacks can be done on data. Few are discussed in the following text;

Linking attacks occurs when attackers recognises individual sharing information in many public data bases.

Homogeneity attack: occurs when there is lack of diversity in sensitive attribute.

Background knowledge attack: occurs when attackers already have some background knowledge about an individual.

The adaptive utility- based anonymisation (AUA) model works to overcome these attacks. It works on 2 step namely filtering based on association mining and anonymisation based on the utility of data.

Filtering involves dividing QI data set into frequent QI set and non-frequent set. Non-frequent attributes set are more prone to disclosure risk. Anonymisation based on utility of data generates different groups of anonymisation models following suppression mostly rather than generalisation [6].

Experimental setup involved generation of anonymised version of data with user preference having four different attributes. These attributes were checked through classifiers naive bayes, Zero R and random forest. Among which Zero R gave best results. The study proves that adaptive utility based anonymisation (AUA) method is effective for privacy presentation providing minimum disclosure risk of individuals.

Data protection and maintenance of anonymity is of paramount importance in healthcare system. It is a major challenge that is faced on daily basis and needs to be continually addressed. Authors presented an idea based on the conceptual architecture and approach of SHIELD. SHIELD was deployed within the framework of FI-STAR (Future Internet Social Technological Alignment in Healthcare) project. It was also included in the FI-STAR project consortium [2].

As presented by Gouvas, SHIELD targets the protection of healthcare data through the pseudo-anonymisation of the end-users. The paper has repeatedly highlighted that SHIELD is a novel network as well as software architecture that provides high quality pseudonymised context-aware services. The paper mainly highlights that SHIELD will give a holistic framework that will guaranty anonymity of end-users as well as protection of personal data. Furthermore, the paper presented that if SHIELD is used it will provide value added services that will not only hide the identity of the end-user but will also implement security of logging and will keep a check on all access to healthcare services and applications. Moreover, it will give authority to the parties to resolve the association between real identity and pseudonym. Finally, the paper concludes by mentioning that within the FI-STAR and FI-WARE platforms, SHIELD software and architecture can be used to provide advanced pseudonymised services that will support the protection of data in healthcare.

III. Survey Preparation

The methodology opted to adopt after critically analysing the literature review presented above was to identify the implementation of de-identification and anonymisation techniques on the health care system of Pakistan so that the researches going on in Pakistan or being done on the data obtained from Pakistan, used by foreign researchers could be as effective as possible maintaining maximum data protection.

To check the possibility of implementing the data de-identification and anonymisation techniques, authors first generated the idea to determine whether the general, professionals or personals had known about this technique or not. To build up this initiative, a survey questionnaire was generated containing various questions which helped me to generate my consensus about how much percentage of people knew about this specific technique or how much of them had at least an idea about it.

The design survey questionnaire contained 20 questions which had been asked some responsible professional personnel such as doctors, IT professionals and paramedical staff etc about their knowledge regarding general concepts and ideas about HIPAA and PHI. Eighty survey questionnaires were collected and analysed to conclude the awareness of de-identification techniques before sharing personal health information with the research community in Pakistan. These questions mainly focused on the awareness of data protection and privacy of any individual. Researchers and professionals that are working with any human-related information were

investigated about how they keep data about any individual and what are the consequences of sharing personal information with/without anonymisation.

IV. Results and Discussion

The results obtained after inspecting the survey questionnaire being filled by different professionals, we came across the decision that among the respondents most of them were not able to understand what HIPAA and PHI were exactly. Very few of the respondents knew about HIPAA as shown in Fig. 1. In their opinion data de-identification and anonymisation was a great way of making health care data valuable as well as sustaining the data from any security threat.

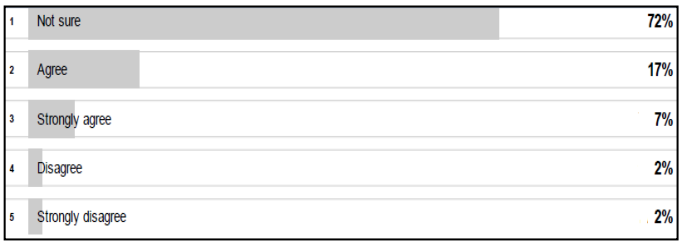

Fig. 1. Awareness percentage score of conducted survey

HIPAA is basically an act pursued in the USA to safe guard the privacy of patient health care data used in the research for the purpose of scientific development. HIPAA is an international standard that is helpful for researchers in the USA. Through this system a researcher could access patient data and can utilise it without being fearful about the leakage of any data [5].

Since it is an effective and successful method of data de-identification, it is explicit for this system to be deployed in Pakistan so that it could also be useful in the local setup.

Very few renowned institutions contain Ethical Review process before sharing data with any external organisation but mostly limited to paper work. For the implementation of this system in developing country such as Pakistan, first step is to educate the professionals and researchers about this act, importance of data protection; how this act works, what its key benefits and how it be beneficial for them as well as how our data could be utilisable for research in other countries. Sharing of data with foreign researchers would be a great step towards the success and achievement of any educational as well as developmental program in our country in collaboration with the foreign world.

References

[1] Gardner, James, and Li Xiong 2008. "HIDE: an integrated system for health information DE-identification." In Computer-Based Medical Systems, 2008. CBMS'08. 21st IEEE International Symposium on, pp. 254-259.

[2] Gouvas, Panagiotis, AnastasiosZafeiropoulos, KonstantinosPerakis, and ThanasisBouras 2015. "An Innovative Approach for the Protection of Healthcare Information Through the End-to-End Pseudo-Anonymization of End-Users." In Internet of Things. User-Centric IoT, pp. 210-216.

[3] LaVigne, Nancy, and Julie Wartell 2015. "Robbery of Pharmacies". Problem-Oriented Guides for Police, Problem-Specific Guide No. 73. Washington, DC: Office of CommunityOriented Policing Services.

[4] Nadeau, David, and Satoshi Sekine 2007. "A survey of named entity recognition and classification." Lingvisticae Investigationes 30, no. 1: 3-26.

[5] Nelson, Gregory S. 2015. "Practical Implications of Sharing Data: A Primer on Data Privacy, Anonymization, and De-Identification."

[6] Panackal, Jisha Jose, and Anitha S. Pillai 2014. "An intelligent framework for protecting privacy of individuals empirical evaluations on data mining classification." In Hybrid Intelligent Systems (HIS), 14th IEEE International Conference on, pp. 67-72.

[7] Sweeney, Latanya 2002. "k-anonymity: A model for protecting privacy." International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10, no. 05: 557-570.

[8] TransCelerate BioPharma Inc 2013. ”Data De-identification and Anonymization of Individual Patient Datain Clinical Studies– A Model Approach.” http://www.transceleratebiopharmainc.com/wp-content/uploads/2015/04/CDT-Data-Anonymization-Paper-final.pdf

[9] Neamatullah, I. (2008). Automated de-identification of free-text medical records. BMC Med Inform Decis Mak, 8, 32.