

Health Data Integration with Secured Record Linkage and Trust-Level Security Based Authentication

Varsha Katiwal^{1*}, Nisha Balani², Priyanka Dudhe³

^{1,2,3}Dept. of Computer Science & Engineering, Jhulelal Institute of Technology, Nagpur, India

Corresponding Author: varshakatiwal3@gmail.com

Available online at: www.ijcseonline.org

Abstract— Discovering Knowledge from various health data domains requires the incorporation of healthcare data from diversified sources. Maintaining record linkage during the integration of medical data is an important research issue. Researchers have given different solutions to this problem that are applicable for developed countries where electronic health record of patients are maintained with identifiers like social security number (SSN), universal patient identifier (UPI), health insurance number, etc. These solutions cannot be used correctly for record linkage of health data of developing countries because of missing data, ambiguity in patient identification, and high amount of noise in patient information. We have proposed a privacy preserved secured record linkage architecture that can support constrained health data of developing countries such as Bangladesh. Our technique can unidentifiable private data of the patients while maintaining record linkage in integrated health repositories to facilitate knowledge discovery process. This concept motivates us to create a trust level security authentication. It means, this healthcare database will be fully secured using cryptography algorithm of encryption and decryption using AES algorithm and authentication will be controlled on “Trust Level Security”. It means that if any researcher or organization need to access this data, then he/she must have at least above average trust level. We score 1 as minimum trust level 10 as maximum trust level and 5 as an average trust level. Trust level will be calculated on the basis of how much other organizations and researchers trust on A researcher or organization.

Keywords— Data Security; Health Data Warehouse; Privacy Preserved Record Linkage; Data Mining;

I. INTRODUCTION

Healthcare data hubs are highly beneficial in many fields such as tracking health patterns, evidence-based medicine, personalized treatments, etc. Clinical diagnostic equipment creates a large amount of health records and related documents every day. These worthy healthcare data are reserved in different healthcare information systems such as Picture Archiving and Communications System, Hospital Information System, Radiology Information System, etc. in public hospitals, private clinics, and diagnostic centres. Data required making proper medical decisions are trapped within fragmented and heterogeneous health systems that are not properly integrated. So, the integration of these health records into a single warehouse is necessary. [1][2]. For maximum benefit from integrated health data repositories (IHDR), linkage of records is essential. Discovering effective knowledge (e.g., correlations among diseases) from medical dataset requires maintaining record linkage. Record linkage is the process of identifying record pairs from different information systems which belong to the same real-world entity. Given two repositories of records, the record-linkage process consists of determining all pairs that are similar to

each other. The similarity between two records is defined based on domain-specific similarities over individual attributes constituting the record. [3][4]. Health data containing protected health information (PHI) such as name, date of birth (DOB), and address can be made linkable easily with the help of PHI. But retaining PHI in healthcare data is very risky. These data are highly lucrative to hackers. Sell value of medical records containing PHI are 100 times more than credit card numbers and Social Security Numbers (SSN). [5] Protecting the privacy of patients while maintaining effective record linkage, that is Privacy Preserved Record Linkage (PPRL), is currently an important focus of the researchers [3][6][7]. This concept motivates us to create a trust level security authentication. It means, this healthcare database will be fully secured using cryptography algorithm of encryption and decryption using AES algorithm and authentication will be controlled on “Trust Level Security”. It means that if any researcher or organization needs to access this data, then he/she must have at least above average trust level. We score 1 as minimum trust level 10 as maximum trust level and 5 as an average trust level. Trust level will be calculated on the basis of how much other

organizations and researchers trust on A researcher or organization.

II. LITERATURE SURVEY

Here, we would like to present the previously existing systems and their methodologies that have tend to improve the efficiency and accuracy for recognizing methodologies. People do not have medical cards with unique health ID. Health care centres do not store National ID numbers or Social Security Numbers (SSN). Many people in real do not know their full name and unable to pronounce their name correctly even in the mother tongue. A person provides different versions of his name in health care centres. The problem of name ambiguity can be understood from Table I. Enormous people do not know their actual birth date because of lacking of birth registration. For several years, a person provides same age to hospitals and diagnostic centers. In all health centres, there are long queues of patients. Many necessary attribute values cannot be inputted for processing a high number of patients in limited time. Less qualified staffs are available and hence employed for inputting patients' data, which leads to unintentional wrong input data. So patients' health records available in Bangladesh and other developing countries contain more noisy data with more missing values. Moreover, unique patient identification numbers are unavailable.

The patient visit to different health service providers can be classified as follows.

a) Patient visits hospitals

There are two types of hospitals in Bangladesh, Government hospitals and private hospitals. According to Directorate General of Health Services (DGHS), a total number of government hospitals under DGHS is 592 [10],[11]. According to the list provided by Bangladesh Private Clinic and Diagnostic Owners Association (BPCDOA), the only Government approved association of private hospital owners, there are 2761 private hospitals in Bangladesh [12], [13].

Patients normally visit a hospital's outdoor or OPD unit, where the person in the reception notes down the basic information of the patient. Then the corresponding doctor checked the patient and wrote up the treatment notes. If necessary, the doctor gives some pathological tests that the patient performed in the diagnostic unit inside the hospital or any outside diagnostic centres. The test results are stored in the centres where a test is performed. In almost all hospitals, there is no patient tracking system with unique patient ID. The irony is that the number of times same patient visits same hospital for treatment or diagnosis, his or her records will be recorded each time as a different patient with different ID or serial number.

b) Patient visits Diagnostic Centres

According to Bangladesh Private Clinic and Diagnostic

Owners Association (BPCDOA, there are more than 8000 private diagnostic centres in Bangladesh registered by the Government. A patient may visit any diagnostic centre to perform some routine health checkups to know his health conditions. These tests include Blood Sugar, Cholesterol level test, etc. In almost all Diagnostic Centres (more than 99%), whenever same patient visits, he is treated as a new patient and his records are stored as a new entry with no relationship or linking with the previous records of the same patient.

c) Patient visits Personal Chamber of Doctors There are about 75700 Registered MBBS doctors and 6800 Dental doctors in Bangladesh [14]. Most of the doctors have private chambers where they consult patients after office hours. A patient can visit a doctor's chamber for treatment. The doctor may recommend some pathological tests. Here also the patients are not tracked with unique ID and no linkage is maintained among the test records of a single patient. In Bangladesh perspective, health records of a person are stored either in electronic form or hard copy format and thousands of different records of the same person are stored with thousand different identities.

In order to complete this research successfully, we have gone through following research papers to get ideas.

In this research we studied that, Data warehousing methodologies share a common set of tasks, including business requirements analysis, data design, architectural design, implementation and deployment. Clinical data warehouses are complex and time consuming to review a series of patient records however it is one of the efficient data repository existing to deliver quality patient care. The presented data warehouse architectures are practicable solutions to tackle data integration issues and could be adopted by small to large clinical data warehouse applications. [1]

In this research we studied that, Clinical data warehouses offer tremendous benefits as a foundation for data mining. By serving as a source for comprehensive clinical and demographic information on large patient populations, they streamline knowledge discovery efforts by providing standard and efficient mechanisms to replace time-consuming and expensive original data collection, organization, and processing and challenges related to warehouse development and implementation.[9]

Record linkage to integrate uncoordinated databases is critical in biomedical research using Big Data. Balancing privacy protection against the need for high quality record linkage requires a human-machine hybrid system to safely manage uncertainty in the ever changing streams of chaotic Big Data. [3]

Healthcare organizations a large amount of data in diverse health information systems. Potential and useful hidden knowledge can be discovered if integration of this huge medical data is performed in national level. The integration process requires linkage of patients' records among different heterogeneous sources. in this concept, we have provided a practical solution of privacy and security problems for developing national health data warehouse of Bangladesh. Our developed technique can unidentifiable private data of the patients while maintaining record linkage in national warehouse to facilitate knowledge discovery process. For this purpose, we have used encrypted mobile number, gender and name-value of patients to produce Patient Identification Key. Our system is being implemented to protect privacy of sensitive health data in health data warehouse. [8]

III. METHODOLOGY

After studying the literature we proposed a system which works on the following modules.

a. Designing

- I. In this module we will design the GUI for insert/update/delete of patient record, so as to develop Health Information System. This will be Frontend Web User-Friendly GUI. We will develop individual GUI for Admin, Users, Trust Level Authentication

b. Health Information System

- I. This system will contain data of patients belonging to different regions and religions of underdeveloped country like Bangladesh
- II. Data is differentiated according to attributes like name, mobile no, age, sex
- III. We will use MS SQL Server to maintain database of Health Information System

c. Record Linkage

Data required making proper medical decisions are trapped within fragmented and heterogeneous health systems that are not properly integrated. So, the integration of these health records into a single warehouse is necessary. For maximum benefit from integrated health data repositories (IHDR), linkage of records is essential.

Patient Identification Technique based on Secured Record Linkage (PITSRL)

- I. The process of NAMEVALUE generation
- II. Patient Identification Technique based on Secured Record Linkage (PITSRL)

d. Cryptography

- I. Data will be stored in encrypted format to improve data security. We will use AES Algorithm for encryption and decryption
- II. We will maintain key and share it according to trust level.

e. Trust – Level Authentication

- I. Higher the trust level higher the scope to get key.
- II. If A want to access data of patient belonging to B head, A will request key from B to decrypt and access data. B will go through trust score of A, if B satisfies then he will share his/her key to A.

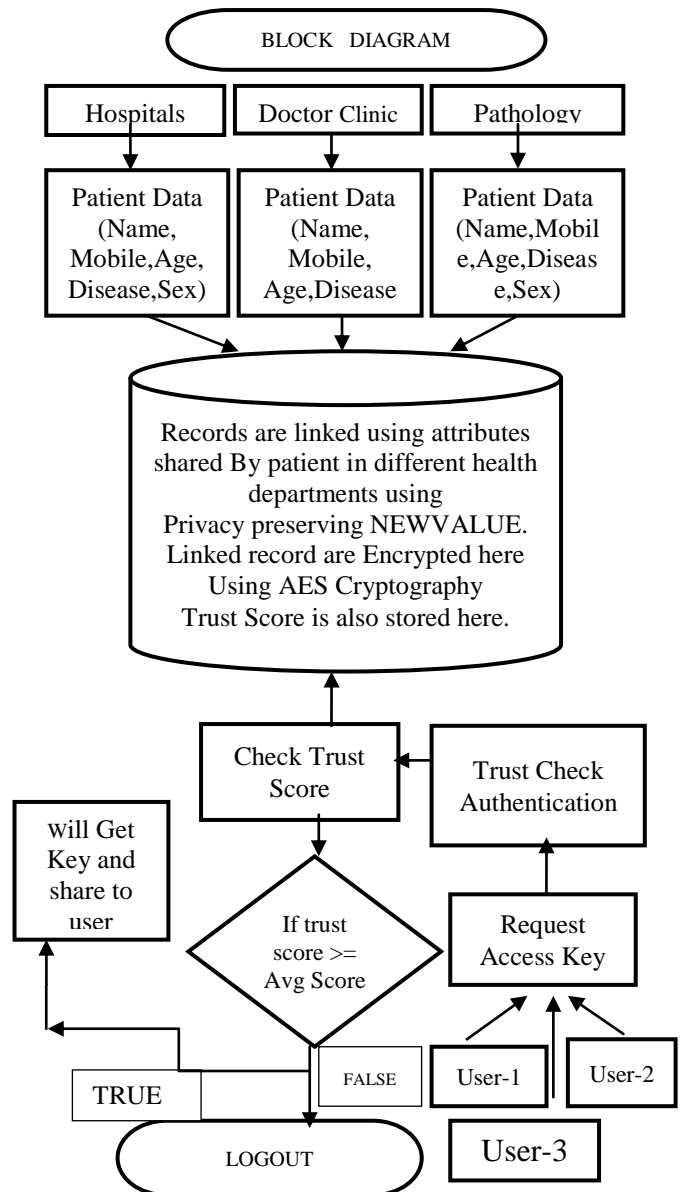


Figure 1. Architecture of project flow diagram

IV. EXPECTED OUTPUT

Using the above proposed model, **Health Information System**

This system will contain data of patients belonging to different regions and religions of underdeveloped country

like Bangladesh. Data is differentiated according to attributes like name, mobile no, age, ex.

In this System records will be linked using Record Linkage

Data required making proper medical decisions are trapped within fragmented and heterogeneous health systems that are not properly integrated. So, the integration of these health records into a single warehouse is necessary. For maximum benefit from integrated health data repositories (IHDR), linkage of records is essential.

To improve confidential information of patients we must use cryptography

Data is stored in encrypted format to improve data security. We will use AES Algorithm for encryption and decryption.

To Gain Access to confidential data we must use Trust – Level Authentication

It means that if any researcher or organization needs to access this data, then he/she must have at least above average trust level. We score 1 as minimum trust level 10 as maximum trust level and 5 as an average trust level. Trust level will be calculated on the basis of how much other organizations and researchers trust on A researcher or organization.

V. CONCLUSION AND FUTURE SCOPE

A problem with mobile number-based identification is that many people use multiple mobile numbers. A person with multiple mobile phones can provide one number in a health centre and another one in other health centre or the same centre in different time. Thus, the person's health data with two different mobile numbers will be treated as two different individual's data in the warehouse. It will impact on mining results. Another problem, though rarely, may occur due to change of mobile numbers by patients. For example, a child after getting an adult, own a mobile. He or she already has his/her records in the health warehouse with his guardians' mobile number. In future, we will go for solutions that are capable of addressing above mentioned issues. The fact that should be kept in mind that mobile/phone-based record linkage provides the opportunity of linking billions of currently available patient records of Bangladesh that cannot be properly linked otherwise.

This paper aims to provide a generalized platform to Preserving record linkage by retaining identifiable attributes in national health data warehouse plays a vital role for effective data extraction. Once Health data warehouse is developed with record linkage, we provide Trust Level Authentication process to access patient records. It means that if any researcher or organization needs to access this data, then he/she must have at least above average trust level.

Higher the trust level higher will be scope to access this data. Health data warehouse development is a complex and time-consuming process but is essential to deliver quality health services.

REFERENCES

- [1] T.R. Sahama, and P.R Croll, "A Data Warehouse Architecture for Clinical Data Warehousing," Australasian Workshop on Health Knowledge Management and Discovery, 2007.
- [2] J. H. Weber-Jahnke & C. Obyr "Protecting privacy during peer-to-peer exchange of medical documents", *Inf Syst Front* (2012), Springer, 14:87–104
- [3] H.C. Kum, A. Krishnamurthy, A. Machanavajjhala et. at., "Privacy preserving interactive record linkage (PIRL)," *J Am Med Inform Assoc* vol. 21, 2014, pp. 212–220.
- [4] J. Liang, L. Chen, and S. Mehrotra, "Efficient Record Linkage in Large Data Sets," In Proc. of the Eighth International Conference on Database Systems for Advanced Applications, 2003.
- [5] Your medical record is worth more to hackers than your credit card.
- [6] P. Christen, "Automatic record linkage using seeded nearest neighbour and support vector machine classification," In Proc. of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008.
- [7] S. I. Khan and A.S.M.L. Hoque, "Privacy and security problems of national health data warehouse: a convenient solution for developing countries," In Proc. of the International Conference on Networking Systems and Security (NSysS). IEEE, 2016.
- [8] C. F. Andrea, C. Danielle, T.M. Matthew et. al., "Development and evaluation of a de-identification procedure for a case register sourced from mental health electronic records," *BMC Medical Informatics and Decision Making*, Vol. 13, 2013, pp.13:71.
- [9] J.A. Lyman, K. Scully, and J.H. Harrison, "The development of health care data warehouses to support data mining," *Clin Lab Med*, 28,1 2008, pp. 55-71
- [10] P. Christen, "Automatic record linkage using seeded nearest neighbor and support vector machine classification," In Proc. of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008.
- [11] E.A. Sauleau, J. Paumier, and A. Buemi, "Medical record linkage in health information systems by approximate string matching and clustering," *BMC Med Inform Decision Making*, Vol. 5, 2005, pp.32– 44.
- [12] A.B. McCoy, A. Wright, Kahn M. et al., "Matching identifiers in electronic health records: implications for duplicate records and patient safety," *BMJ Qual Saf* Vol. 22, 2013, pp.219–24.
- [13] C. F. Andrea, C. Danielle, T.M. Matthew et. al., "Development and evaluation of a de-identification procedure for a case register sourced from mental health electronic records," *BMC Medical Informatics and Decision Making*, Vol. 13, 2013, pp.13:71.
- [14] N. K. Abel, P. C. John, L. J. Kathryn et al., "Design and implementation of a privacy preserving electronic health record linkage tool in Chicago," *Journal of the American Medical Informatics Association*, 2015, pp.1-9.

Authors Profile

Miss Varsha Katiwal pursued Bachelor of Engineering in Computer Engineering from Rashtrasant Tukadoji Maharaj Nagpur University, Nagpur in 2011 and is currently a Master of Technology scholar from Rashtrasant Tukadoji Maharaj Nagpur University.