Health Data Integration with Secured Record Linkage A Practical Solution for Bangladesh and Other Developing Countries

Shahidul Islam Khan, Bangladesh University of Engineering and Technology
Health Data Integration with Secured Record Linkage
A Practical Solution for Bangladesh and Other Developing Countries

Shahidul Islam Khan; Abu Sayed Md. Latiful Hoque
Department of Computer Science and Engineering (CSE)
Bangladesh University of Engineering and Technology (BUET)
Dhaka, Bangladesh
nayeemkh@gmail.com; asmlatifulhoque@cse.buet.ac.bd

Abstract—Knowledge discovery from various health data repositories 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. Also, identifiable health data in electronic health repositories may produce a significant risk to patient privacy and also make the health information systems security vulnerable to hackers. In this paper, we have analyzed the practical problems of collecting and integrating healthcare data in Bangladesh for developing national health data warehouse. 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 anonymize identifiable private data of the patients while maintaining record linkage in integrated health repositories to facilitate knowledge discovery process. Experimental results show that our proposed method successfully linked records with acceptable accuracy for noisy data in the absence of any standard ID like SSN.

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

I. INTRODUCTION

Healthcare data warehouses are highly beneficial in many fields such as mining 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 centers. 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]-[6].

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 repository 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 [7], [8]. 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) [9]. 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 [7], [10]-[16].

A two-step approach to automatic record pair classification has been presented in [10]. In its first step, training examples of high quality are automatically selected from the compared record pairs, and used in the second step to train a support vector machine (SVM) classifier. A three step record linkage method is proposed in [11]. De-identified record with an exact match of the patient first and last names and date of birth were linked from five institutions [12]. Numbers of patient records existing for the topmost commonly occurring first and last name pairs were determined. The authors in [7] synthesize this literature to formalize a new framework for privacy preserving interactive record linkage with tractable privacy and utility properties and then analyze the literature using this framework. Development, implementation, and evaluation of a bespoke de-identification algorithm used to create the Mental Health register is discussed in [13]. The authors of [14] developed a software application that performs data cleaning, preprocessing and anonymization.

In summary, above works have their own merits and limitations, but they are not effectively applicable in development of National Health Data Warehouse of Bangladesh or other economically developing countries. This is primarily due to the unavailability of social security numbers or similar identification keys for the whole population. Though the Government of Bangladesh is providing National ID numbers for all its citizens, there is no practice of incorporation of this number with health records now and not shortly. Another reason is that healthcare data of

This research is performed in the Dept. of CSE, BUET and supported by the ICT Division, Ministry of Posts, Telecommunications and Information Technology, Government of the People's Republic of Bangladesh.

Author Version
Bangladesh contain higher error than that of developed countries. Many Bangladeshis people do not know their name and date of birth correctly.

In this paper, we analyzed different problems for healthcare data integration in Bangladesh to develop a national level health data warehouse for its citizens. We also proposed a PPRL architecture that can support constrained health data of developing countries such as Bangladesh. Experimental results show that our proposed technique can successfully link records in the absence of SSN or other National ID number and the presence of typical noisy data, e.g., misspelled patient names. For a noisy health dataset of 633609 patient records, we achieved 87% correct record linkage.

The rest of the paper is organized as follows. In Section II we have analyzed common socio-economic characteristics of developing countries and their impact on available digital health data. Global security and privacy issues related to digital health data are briefly presented in Section III. Section IV illustrates the scenario of health data generation and storage, and record linkage problem in Bangladesh perspective. In Section V and VI, we have presented our technique: Patient Identification Technique based on Secured Record Linkage (PITSRL) and analyzed its performance. The motivation behind our key concept and limitations of our research are presented in Section VII and VIII. Finally, Section IX concludes the paper.

II. COMMON CHARACTERISTICS OF DEVELOPING COUNTRIES THAT IMPACT HEALTH DATA

Developing countries are those with low, lower middle or upper middle incomes. There are some common socio-economic characteristics found in the developing countries of the world that have a similar impact on healthcare facilities and health data. These characteristics include Lower per-capita income, higher population growth rates, and low level of urbanization. This implies poor health, inadequate education, and more rural population [17], [18]. Above socio-economic conditions made an impact in the available health care data in the following ways.

- People do not have medical cards with unique health ID. Health care centers 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 centers. 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 centers, 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. These makes techniques of developed countries for record linkage inappropriate for developing countries and more specialized technique to address the situation.

III. Security and privacy issues related to health data

Security of a Health Information System deals with protecting medical data from intruders, malware, and frauds. It retains confidentiality and integrity of healthcare data. As medical systems are more interconnected and networked, security has become a huge challenge in the healthcare sector. Security weaknesses of a health information system can result in privacy losses as Protected Health Information (PHI) of patients exhibits a potential to be sold with great demand in the underground markets. It is an increasing target for hackers to break down the security of health systems and expose the private data of patients for money.

Nobody likes his medical records to be revealed to unauthorized persons. There is no doubt that development of national health data warehouse is very much essential for every country including economically developing and highly populated countries, but it raises the high risk to data security and privacy of citizens. After deployment of National Health Data Warehouse, health service providers, and healthcare researchers can have access to private health data of millions of patients. This situation leads to new security loopholes.

- As the integrated National health data repository is connected with its users by Internet and intranets, the integrated repository has more chance of being hacked.
- As more users will have access to the integrated system, there is more chance that some users may have the bad intention of selling the protected health information (PHI) of patients for money.

From the year of 2014, the extent of hacking over healthcare servers has increased by a significant margin. [19], [20]. The attackers’ motivation is to get huge PHI in a single successful hack. We have analyzed the data provided by U.S. Department of Health and Human Services and found that hackers are increasingly targeting healthcare servers which are very alarming to national level health information system development [21]. So developers of the data warehouse have to be careful enough to establish security of National Health DW and protect the privacy of these highly sensitive data.

<table>
<thead>
<tr>
<th>Actual Patient Name</th>
<th>Inputted Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sobuj Chowdhury</td>
<td>Mr. Sobuj Chowdhury</td>
</tr>
<tr>
<td></td>
<td>Sobuj Chowdhury</td>
</tr>
<tr>
<td></td>
<td>Mr. Sobuj</td>
</tr>
<tr>
<td></td>
<td>Sobuj Chy.</td>
</tr>
<tr>
<td></td>
<td>Sabuj Chowdhury</td>
</tr>
<tr>
<td></td>
<td>Sabuz Chaudhury</td>
</tr>
</tbody>
</table>
Proper measures have to be taken so that an intruder if somehow break the security of warehouse, cannot identify individual patients from health database or warehouse and privacy of the patients are safeguarded.

IV. HEALTH DATA GENERATION AND STORAGE IN BANGLADESH

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 [22], [23]. 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 [24], [25]. Concatenation

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 center. The test results are stored in the centers 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 or diagnostic unit, his or her records will be recorded each time as a different patient with different ID or serial number.

b) Patient visits Diagnostic Centers

According to Bangladesh Private Clinic and Diagnostic Owners Association (BPCDOA), there are more than 8000 private diagnostic centers in Bangladesh registered by the Government. A patient may visit any diagnostic center to perform some routine health checkups to know his health conditions. These tests include Blood Sugar, Cholesterol level test, etc. In almost all Diagnostic Centers (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 [26]. 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. These records are highly distributed in terms of time (e.g., doing pathological tests at different times), space (e.g., outdoor, indoor or lab), and locations (e.g., different hospitals). For extraction of fruitful knowledge from health data, it is the first requirement to accumulate health records from widely variable sources. While accumulation, these records cannot be mapped with the patients because of storing patient records every time with different identities. Record linkage the problem is to find an optimum reliable mapping of each patient to his health record throughout the lifespan.

V. A SOLUTION: PATIENT IDENTIFICATION TECHNIQUE BASED ON SECURED RECORD LINKAGE (PIKSRRL)

The block diagram of our system is shown in Fig. 1. Health Records with Patient identifiable attributes such as name, address, date of birth in heterogeneous format from various health service providers are inputted in the systems. These records are then de-identified preserving record linkage. Privacy preserved linkable health records are then stored in national health data warehouse as a unified format.

The input of PIKSRRL system is health records provided by different health care organizations such as government and private hospitals, diagnostic centers, research centers, health NGOs. These data are in heterogeneous formats like Oracle, MS SQL or MySQL databases; CSV or MS Excel files, etc. The detailed architecture of the National Health Data Warehouse Bangladesh can be seen from [27]. These raw health records contain attributes related to patient identification such as patient name, address, and mobile number. Fig. 2 shows the patient identification process in our system. Our PIKSRRL algorithm works in two steps.

- In step 1, a Patient Identification Key based on Secure Record Linkage (PIKSRRL) is generated for each patient record using available patient identifiable data.
- In step 2, all identifiable data, capable of identifying individual patients is removed from the health record.

We have used five attributes to generate identification key; mobile number, name, age, geocode, and gender. Mobile numbers are made secured through masking. Name is converted to NAMEVALUE. Age is used to generate the year of birth and age group.

NAMEVALUE is the encrypted text string generated by our developed Name-Value Algorithm using significant and
unambiguous characters contained in a patient’s name. We have treated salutations and titles as insignificant. In the practical situations at most health centers or doctors’ chamber, patients are asked, and they tell their information i.e. name, age verbally. From pronouncing to write, vowels are highly ambiguous, and vowels can be written in many ways. The situation can be understood clearly from Table I. We can see that doctors or computer operator can write or enter following two patients in six or more ways. To remove ambiguity, vowels are discarded from the significant portion of a name. Then the data is masked so that the health warehouse users can not understand the real name.

The process of NAMEVALUE generation is explained in Table II. Our algorithm of NAMEVALUE generation from patient name is presented next:

Algorithm 1: NAMEVALUE

Input: Patient name

Output: NAMEVALUE of the inputted name

Steps:
1. Delete Salutation
2. Delete a/A, e/E, i/I, o/O, u/U unless beginning of name or after white space
3. Map g/G/j/J/z/Z to same code
4. Map k/K/q/Q to same code
5. Map all other characters according to Code Table

TABLE II. SELECTION OF SIGNIFICANT, UNAMBIGUOUS NAMEVALUE

<table>
<thead>
<tr>
<th>Patient Name</th>
<th>Significant portion</th>
<th>Unambiguous significant portion</th>
<th>Masked NAMEVALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mr. Abul Hosain</td>
<td>Abul Hosain</td>
<td>abl hsn</td>
<td>tagsemi</td>
</tr>
<tr>
<td>Mr. Md. Abul Hosen</td>
<td>Abul Hosen</td>
<td>abl hsn</td>
<td>tagsemi</td>
</tr>
<tr>
<td>Mohammad Abul Hosain</td>
<td>Abul Hosain</td>
<td>abl hsn</td>
<td>tagsemi</td>
</tr>
</tbody>
</table>

The code table is the table that is used for data masking. Mobile numbers of the patients are masked and concatenated before the NAMEVALUE. Gender information is also concatenated to get the PIKSRL that is shown in Fig. 1.

Our algorithm of Patient Identification Technique based on Secured Record Linkage (PITSRL) is presented below:

Algorithm 2: PITSRL

Input: Health record set with patient identifiable data

Output: De-identified Linkable patient record

Steps:
Repeat
1. Encrypt mobile number
2. Convert patient name to NAMEVALUE
3. Convert Date of Birth or Age to BIRTHYEAR
4. Generate Patient Identification Key based on Secured Record Linkage (PIKSRL) from masked mobile number, NAMEVALUE, Birth year and Gender
5. Add PIKSRL Key to record set
6. Delete Patient name, Mobile number, Address, Date of Birth, Credit Card number data.

Until last record

In Bangladesh, many people do not know their date of birth. The case is most common to the aged rural people with less education. The major reasons are illiteracy, poverty, lack of social awareness, etc. The festival of celebrating one’s birthday is also quite new in Bangladeshi culture. Though birth registration certificate is the first official document for an individual, many people do not have it till last decade. The situation of not knowing self birth-date is more or less same among poor and uneducated people around the world. In various cultures and jurisdictions, if a person’s real birthday is unknown (e.g., in the case of an orphan), then his birthday is considered to be January 1 [28]. This is the main reason, why we have not considered the date of birth rather considered the birth year in our PITSRL technique.

Another important reason for considering birth year instead of date of birth is that in most hospitals and diagnostic centers, only patient’s age is collected rather than the date of birth. They have already millions of patients’ health data scattered among their enormous health system without patients’ date of birth but with their ages. From patients’ age data, their birth year can be easily calculated in the integrated warehouse data.

VI. RESULTS AND DISCUSSION

For the validation of our proposed technique, we have used real healthcare dataset consisting of 633609 patient records.
The time-span of the dataset is almost one year. The dataset provider collects data from twelve private hospitals and diagnostic centers on a regular interval. Among these records, 77021 records contain a null, 0, 00, in the patient phone/mobile attribute. 6173 records contain invalid strings in phone/mobile attribute. Dates of Birth or Age of 41251 records are missing.

We have implemented the PITSRL algorithm in Java and run in an HP brand computer of Graduate Lab of Dept. of CSE, BUET. The Computer has Intel Core i5 3.20 GHZ Processor and 4GB RAM and runs Windows 10 Pro. The program processed the dataset of 633609 patients in an average of 7 Seconds and 60 milliseconds (10 runs). Input file size was 24.1MB, and output file size was 14.4MB.

It was found that 87% of patient data contained a valid Mobile or Phone number thus PIKSRL key contained masked patient mobile/phone number. In combination with NAMEVALUE, birth year, and gender, the PIKSRL key for these cases can be treated as unique. In the case of 13% patients’ records were valid mobile/phone numbers of the patients were unavailable, PIKSRL key was generated from only NAMEVALUE, birth year, and gender; some records were also suffered from missing the date of birth information. So PIKSRL key for these 13% records cannot be used for proper record linkage.

To verify whether PITSRL can identify same patients from the integration of different healthcare data, we have used a training dataset of 100 patient records of 67 unique patients. The data repetition status (multiple records of same patients) is shown in Table III:

<table>
<thead>
<tr>
<th>Number of Repetition</th>
<th>No. of Patients</th>
<th>No. of Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>Four different records of same patients</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Three different records of same patients</td>
<td>9</td>
<td>27</td>
</tr>
<tr>
<td>Two different records of same patients</td>
<td>12</td>
<td>24</td>
</tr>
<tr>
<td>Patients with single health record</td>
<td>45</td>
<td>45</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>100</td>
</tr>
</tbody>
</table>

We have inputted the dataset to our system and found that the system generated 67 unique PIKSRL keys with 100% accuracy. The system also generated different PIKSRL keys for different patients with the same name. The system also generated the same key for same patients with slightly misspelled named and different registration number.

VII. Motivation of Using PITSRL in Developing Countries

We have used masked mobile phone numbers to distinguish individual patients because of the following reasons:
1. According to Bangladesh Bureau of Statistics total population of Bangladesh is 158,988,940 [29] and there are 124705000 active mobile connections [30].
2. Approximately 78.43 active mobile connections per 100 people.
3. Almost every family irrespective of rich or poor, urban or rural possess at least one mobile phone.
4. Every person must go through some security verification to purchase mobile SIM.
5. People already use mobile for various identification and transaction purposes such as getting a passport and national id card, performing financial transactions, etc.
6. Mobile numbers are easy to remember and tell within shortest possible time.
7. Every mobile number is unique.
8. Most importantly, almost all health care centers collect mobile numbers of patients for communication and billing purposes. So the mobile number is available with existing millions of health data.

No other identification number i.e. passport, national id, birth registration number has the above features. In all other SAARC countries, the situation is more or less same which can be seen from [31]-[34] and presented in Table IV. In January 2015, there were 3.65 billion unique mobile users in the world with 7.09 billion active mobile subscriptions among 7.258 billion people in the world according to [35].

<table>
<thead>
<tr>
<th>Country</th>
<th>Number of Mobile</th>
<th>Connection/100 citizens</th>
</tr>
</thead>
<tbody>
<tr>
<td>India</td>
<td>960,579,472</td>
<td>77.58</td>
</tr>
<tr>
<td>Pakistan</td>
<td>140,000,000</td>
<td>77</td>
</tr>
<tr>
<td>Bangladesh</td>
<td>124,705,000</td>
<td>78.43</td>
</tr>
<tr>
<td>Sri Lanka</td>
<td>22,123,000</td>
<td>107</td>
</tr>
<tr>
<td>Nepal</td>
<td>18,240,670</td>
<td>86.82</td>
</tr>
</tbody>
</table>

Though masked mobile numbers can uniquely distinguish patient data, we have also stored NAMEVALUE in the data warehouse because, in some cases, the mobile number is insufficient for clustering patient data from health warehouse. For example, father and his child have a same mobile number and same address but different NAMEVALUEs.

VIII. Limitations and Further Research

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 center and another one in other health center or the same center 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.
IX. CONCLUSIONS

Preserving record linkage by retaining identifiable attributes in national health data warehouse plays a vital role for effective data mining. But identifiable health data have a high risk to patients’ privacy and also make the health information systems vulnerable to hackers. Disclosure of medical information about a patient may be harmful to his personal life and also for his career. In this paper, we characterized the factors constraining healthcare data available in Bangladesh. We have discussed the problems of collecting and integrating healthcare data in Bangladesh to build up its National Health Data Warehouse. We have also provided a practical solution: Patient Identification Technique based on Secured Record Linkage (PITSRL) that can anonymize the identifiable private data of the patients while maintaining record linkage to facilitate knowledge discovery by healthcare providers and researchers. We have used the masked phone number, gender, year of birth and NAMEVALUE of patients to produce anonymizable and linkable Patient Identification Key. It was seen from experimental results that our developed technique can successfully link records in the presence of noisy data, e.g., misspelled patient names. For a noisy health dataset of 633609 patients, we achieved 87% correct record linkage key. For the complete dataset of 100 patient record, PITSRL achieved 100% accuracy. Using PITSRL, patients’ data can be shared and integrate among different government and private hospitals and diagnostic centers in Bangladesh. Our approach is also suitable for the developing countries where poverty and illiteracy rates are high among mass people.

REFERENCES

[9] Your medical record is worth more to hackers than your credit card, URL: http://www.reuters.com/article/2014/09/24/us-cybersecurity-hospitals-idUSKCN0URH2120140924
[18] Diana Kendall, Sociology in Our Times, cengage, 11th Ed, 2015