Development of National Health Data Warehouse Bangladesh: Privacy Issues and a Practical Solution

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*—Healthcare organizations in Bangladesh own 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. To facilitate effective data mining, it is essential to preserve record linkage in health data warehouse by retaining identifiable attributes. On the other hand, identifiable health data have high risk to patient privacy and also increase the chance of attacks by cyber criminals. In this paper, we have provided a practical solution of privacy and security problems for developing national health data warehouse of Bangladesh. Our developed technique can anonymize identifiable 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.**

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

I. INTRODUCTION

Today we are living in a world where data are collected in ever-increasing amounts, behaviors of humans and machines are recorded and summarized. The field of data mining has made momentous and widespread advances over the past decades and has been successfully applied to diverse areas such as business, healthcare sectors, engineering, and social media [1], [2].

For efficient knowledge discovery from data, development of a data warehouse (DW) is required. A data warehouse is a subject-oriented, integrated, non-volatile, and time-variant collection of data to support management decisions. It joins the data spread throughout an organization into a central structure [3]. A health data warehouse is a data store used to analyze consolidated and historical health data [4], [5]. Development of a Health DW contains two key phases. Firstly, a conceptual view of the warehouse is specified according to the user requirements. Secondly, the associated data sources and the extraction, transform and load process are determined. During the warehouse operation, data must be refreshed on a regular basis such that data stored in the warehouse reflect the current operational state [3].

Acknowledgments: This research is performed in the Dept. of CSE, BUET and funded by the ICT Division, Ministry of Posts, Telecommunication and Information Technology, Government of Bangladesh.

Extensive amounts of knowledge and data stored in medical databases require the development of specialized tools for accessing the data, analysis, knowledge discovery, and effective use of stored knowledge and data. Modern hospitals are well equipped with monitoring and other data collection devices which provide relatively inexpensive means to collect and store the data in inter-hospital and intra-hospital information systems. Large collection of medical data is a valuable resource from which potentially new and useful knowledge can be discovered through mining. Patient records collected for diagnosis and prognosis typically encompass values of historical, clinical and laboratory parameters. Such datasets have characters such as missing values, incorrectness i.e. random noise in data and sparseness. [6].

In Healthcare sector, data mining can be used for solving descriptive and predictive Data Mining tasks. Descriptive Data Mining tasks are concerned with finding interesting patterns or clusters in data using association rule learning or clustering. Whereas predictive Data Mining starts from the entire data set and aims at inducing a predictive model that holds on the data and can be used for prediction or classification of yet unseen instances [7].

A huge amount of health records and related documents created by clinical diagnostic equipments are generated daily. These valuable data are stored in various medical information systems such as HIS (Hospital Information System), PACS (Picture Archiving and Communications System), RIS (Radiology Information System) in various hospitals, departments and diagnostic laboratories. Data required to make proper medical decisions are trapped within fragmented and heterogeneous health systems that are not properly integrated. So integration of these health records into a single warehouse is necessary.

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. Similarity between two records is defined based on domain-specific similarities over individual attributes constituting the record. Record linkage is essential when joining datasets based on entities that may or may not share a common identifier such as national id or social security number [8], [9]. For discovering effective knowledge such as correlations among diseases from medical dataset it is very essential to maintain record linkage. Health data have to

978-1-4673-9930-2/15/$31.00 ©2015 IEEE

be linkable in some way. But medical record linkage has inverse relation with patients' privacy. If a health dataset preserves record linkage means privacy of individuals are at risk. So, protecting privacy of patients, while maintaining effective record linkage, is an important research issue.

In this paper we have presented a brief overview of security and privacy risks of integrated healthcare information system raised worldwide. As the Government of Bangladesh is implementing the National Health Data Warehouse in the country, there rises a high risk of violation of privacy. Medical records of the Bangladeshi citizens will be a potential target of cyber criminals. We have provided a practical solution to protect the privacy of individuals. Our technique can anonymize the identifiable private data of the patients while maintaining record linkage for doctors and researchers. Our approach is suitable for Bangladesh and other developing countries where poverty and illiteracy rates are high.

The rest of this paper is organized as follows. In Section II we have briefly presented research works related to anonymization of medical data and record linkage. Section III describes importance of healthcare data security. In Section IV, some statistics of medical data breaches are presented. In Section V and VI, we have presented our technique: Patient De-Identification with Linkage Preservation (PDLP). Limitations of our research are presented in Section VII. Finally Section VIII concludes the paper.

II. Related Works

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]. The first step is to standardize and indexing elementary identity fields using blocking variables. The second is to match similar pair records and finally in the third step clusters of coherent related records are created, using graph drawing technique, agglomerative clustering methods and partitioning methods.

In [12] for five institutions de-identified record with an exact match of patient first and last names and dates of birth were retrieved. Numbers of patient records existing for the topmost 250 commonly occurring first and last name pairs were determined. The authors also identified methods for managing duplicate records.

The authors in [8] synthesize this literature to formalize a new framework for privacy preserving interactive record linkage (PPIRL) 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 system is designed to create dictionaries using patient identifiers (PIs) entered into dedicated source fields and then identify, match and mask them (with ZZZZZ) when they appear in medical texts.

The authors of [14] developed a software application that performs data cleaning, preprocessing, and hashing of patient identifiers to remove all protected health information. The application creates seeded hash code combinations of patient identifiers using an algorithm.

In summary, above cited works can not be effectively applicable in development of National Health Data Warehouse of Bangladesh. This is due to the unavailability of social security numbers or similar identification keys for the whole population. Another reason is the high illiteracy rates such that many people even do not aware of their date of births, full name etc.

III. Healthcare Data Security and Privacy Issues

As healthcare data are highly private, so protecting privacy of individual patients as well as security of the whole data repository is a major concern. Security of a Health Information System deals with protecting medical data from intruders, malwares, 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 healthcare sector.

A. Data security and privacy

Data Security refers to protective digital security measures that are applied to prevent unauthorized access to computers, databases and websites. It also protects data from corruption. Data security is also known as information security. Examples of data security technologies include software/hardware disk encryption, backups, data masking and data erasure [15].

Data or Information Privacy deals with the ability an organization or individual has to determine what data in a computer system can be shared with others. Information privacy is the privacy of personal information and usually relates to personal data stored on computer systems. It is considered an important aspect of information sharing. Wherever personally identifiable information is accumulated in any form.

A major challenge in data privacy is to share information in a way that personally identifiable data is protected. Information privacy may be applied in numerous ways, including encryption, authentication and data masking - each attempting to ensure that information is available only to authorized persons.

B. Medical data: highly private

Nobody likes his medical records to be revealed to others. We do not want others to know about our medical or psychological conditions or treatments. It may affect one's insurance coverage or employment. Security and authentication systems are often required for individuals that process and store medical records. Many countries developed standards for doctor-patient relationships that preserve confidentiality. These standards protect patients' dignity and ensure the patients' comfort to reveal accurate information to receive the correct treatment. Any data miner and healthcare researcher dealing with health data should respect the private

life of the persons concerned and must avoid using the data in a way which might cause undue offence.

C. Healthcare data: Lucretive to Cyber Criminals

There is a growing trend of hacking into medical records. Hackers' objective is to exploit personal information, which is a lucrative business to them. In U.S., a stolen Social Security Number might sell for 25 cents in the underground market, and a credit card number for $1. Whereas sell value of a comprehensive medical record may varies from 10$ to $1,000 in black market [16]. The data for sale includes names, birth dates, health policy numbers, diagnosis codes and billing information. Fraudsters use this data to create fake IDs to buy medical equipment or drugs. The hackers combine a patient number with a false provider number and file made-up claims with insurers, according to experts who have investigated cyber attacks on healthcare organizations.

Hospitals have low security, so it is relatively easy for the hackers to get large amount of medical data. Government sector like public health department systems that contain health-related data is also an increasing target of hackers for two main reasons. First, they are generally more vulnerable, as they are older systems running older, less secure software. Second, they are rich in data like personally identifiable information, healthcare, financial information [17].

Many healthcare organizations encrypt their transactional databases, but ignore encrypting database backups. That's why backups are often attackers' second target. A data-protection strategy must cover data everywhere it is stored, and at every stages, from creation and processing, to storage, backup and transmission.

D. National Health DW: Pros and Cons

There is no doubt that development of national health data warehouse is very much essential for every countries including Bangladesh, but it raise high risk to data security and privacy of citizens. After deployment of National Health DW, health service providers, and healthcare researchers can have access to private health data of millions of patients without bar.

Prior integration to health data warehouse, these sensitive and private data reside to a single organization such as hospitals or diagnostic centers. Only the particular organization, where a patient's health data is generated, is responsible to protect the data privately. If any leakage of a patient data or any other kind of privacy violation occurs, that particular organization is liable for this. The organization will be charged or case filed against it in the law enforcing agencies. So every health service provider is bound to protect the patient diagnostic test reports and other health data for its own interest.

Now the situation is far different in the case of National Health DW. Here, if a patient's sensitive and private data leaks, who will be blamed? Against whom or which organization to file Defamation-suit? So users of the data warehouse have to be careful enough to maintain security of National Health DW and protect the privacy of these highly sensitive data. It is not possible to guarantee that all the doctors, health researchers, health service providers will execute their responsibility and no one will violate in any way. So proper measure has to be taken so that individual patient cannot be identified from health database or warehouse and their privacy is safeguarded.

IV. Data Lickage of Health Information Systems

A data breach or leakage can be defined as any incident which involves loss or exposure of personal records digitally. Personal records means information about a person that cannot be obtained easily through other public means; and this information only known by an individual or by an organization under the terms of a confidentiality agreement. Cyber criminals recognize two critical facts about the healthcare industry:

- Healthcare organizations posses large and monetarily lucrative personal data
- Most of them do not have the resources and technologies to detect cyber attacks and effectively protect healthcare data.

According to 2015 Fifth Annual Benchmark Study on Privacy & Security of Healthcare Data which covered 90 healthcare organizations in U.S., More than 90% of healthcare service providers had a data breach, and 40% had more than five data breaches over the past two years [18]. According to the report, for the first time, criminal attacks are the *number one* cause of healthcare data breaches. Criminal attacks on healthcare organizations are 125% higher compared to five years ago. During interview, 45% of healthcare organizations say the root cause of the data breach was a criminal attack and 12% say it was due to a malicious insider. According to the Fifth Annual Study on 2014, medical identity theft nearly doubled in five years, from 1.4 million adult victims to over 2.3 million in 2014 [19].

In last ten years at least 18 health breach reported in Europe affected minimum 9,337,197 individual records [20]. The health records include details on the patients' conditions, names, home addresses and dates of birth. Not only Europe, healthcare networks and servers containing integrated health records are in high risk of cyber attacks all over the world.

Hacking is increasing in the Healthcare Servers at a shocking rate. From 21th October 2009 till data there are 1279 health data breach reported, among which only 190 is server attacks that is average 14.85%. But up to 1st August 2015, in last 12 months 58 of 255 are server attack, which is 22.74%. We have uncovered this insight by analyzing the data provided by U.S. Department of Health and Human Services [21]. Hackers are increasingly targeted to the health servers which are very alarming to national level health information system development.

No information system can be assumed to be completely protected from all kind of criminal and cyber attacks. Security can be more vulnerable in the case of large scale, national level health information systems where Internet communication has to be maintained for the sake of easy data

collection from far-most parts of the country. So integrated health information system should be designed in such a way that support following contradictory properties:

- There is enough data to maintain record linkage so that doctors, researchers can get useful insight from the system.
- If data breach occurs, individual patient's privacy will not be compromised.

We have provided a practical solution that is capable to provide above mentioned facilities. It used mobile phone number based identification. In this technique, we have used encrypted mobile phone numbers of the patients' to generate Patient Identification Key (PIK). Motivation for our technique is discussed in the next section.

V. Why Mobile Phone Based Identification Key?

We have used encrypted 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 [22] and there are 124705000 active mobile connections [23].
2. Approximately 78.43 active mobile connections per 100 peoples are available.
3. Almost every family irrespective of rich or poor, urban or rural posses at least one mobile phone.
4. Every person must go through some security verification to purchase mobile SIM. Recently biometric registration for SIM purchase is also inaugurated.
5. People already uses mobile for various identification and transaction purposes such as getting 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 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. Though encrypted mobile numbers can uniquely distinguish patient data, we have also stored name-value in the data warehouse because in some cases mobile number is insufficient for clustering patient data from health warehouse. For example, a father and his children may have same mobile number and GEO-CODE but different name-values. Details of our technique, with the explanation of name-value, are presented in the next section.

VI. Patient De-Identification with Linkage Preservation (PDLP) Technique

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. These privacy preserved linkable health records are stored into national health data warehouse as unified data format. To design the system we have followed a practical oriented approach suitable for Bangladesh. The rate of poverty and illiteracy among mass people is still high in Bangladesh. Developing national scale information system is a major challenge is these countries due to large population and fewer resources.

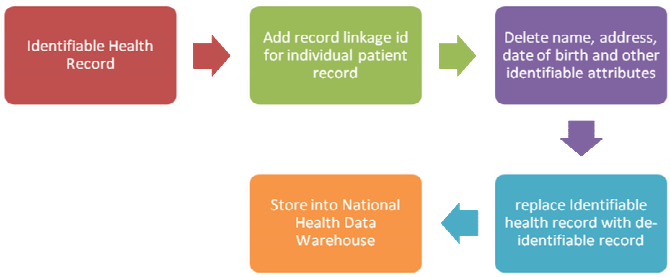

Fig. 1. Block diagram of Patient De-Identification with Linkage Preservation (PDLP) Technique.

The input of PDLP 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 is heterogeneous formats like Oracle, MS SQL or MySQL databases; CSV or MS Excel files etc. The detail architecture of the National Health Data Warehouse Bangladesh can be seen from [24], [25]. These raw health records contain attributes related to patient identification such as patient name, address, and mobile number. Our Patient Identification Key (PIK) algorithm works in the following two steps.

- In first step, a PIK is generated for each patient record using available patient identifiable data.
- In second step , all data those are capable of identifying individual patients are removed from the health record.

We have used three attributes to generate identification key; mobile number, name, and gender of a patient. Mobile number is stored in an encrypted format. Name is converted to name-value. Age is used to generate year of birth and age group. Location is used to generate GEO-CODE.

Name-value 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 pronounce to write, vowels are highly ambiguous and vowels can be written in many ways. This can be understood clearly from the following table. From Table 1 we can see that doctors or computer operator can write or entry following two patients in six or more ways. To remove ambiguity, vowels are discarded from significant portion of a name. Then the data is encrypted using simple encryption

technique so that real name cannot be understood by the health warehouse users.

TABLE I. UNAMBIGUOUS AND SIGNIFICANT NAME-VALUE SELECTION

Patient Name	Significant portion	Unambiguous significant portion	Encrypted name-value
Md. Kamal Uddin	Kamal Uddin	kml uddn	fhgsxcci
Mr. Kamal Uddin	Kamal Uddin	kml uddn	fhgsxcci
Muhammad Kamal Uddin	Kamal Uddin	kml uddn	fhgsxcci
Muradul Bashir	Muradul Bashir	mrdl bsr	hlcgsaml
Md. Muradul Bashir	Muradul Bashir	mrdl bsr	hlcgsaml
Moradul Bosir	Moradul Bosir	mrdl bsr	hlcgsaml

Mobile numbers of the patients are encrypted and concatenated before the name-value. Gender information is also concatenated to get the Patient identification key (PIK) that is shown in Fig.2.

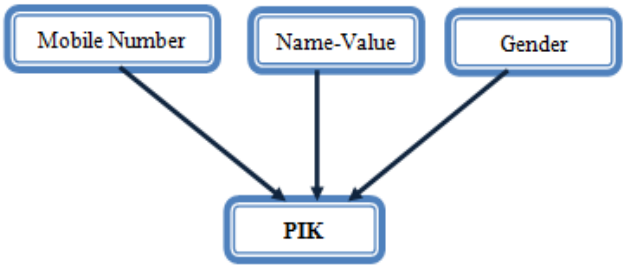

Fig. 2. Mobile number, Name-Value and Gender of patients contribute to generate Patient Identification Key (PIK)

GEO-CODE is the 8 character de-identified address of a patient generated from his health record. We have used the *concept hierarchy* technique to produce GEO-CODE from patients' address records [26]. This data is stored is the national health data warehouse to facilitate spatial data mining related to geographic location such as correlation of diseases and places. Generation of GEO-CODE for Bangladeshi citizens can be understood by Fig 3. This standard is used by Bangladesh government for passport and voter ID. It is also used by Directorate General of Health Services (DGHS), Ministry of Health & Family Welfare of Bangladesh Government [27], [28].

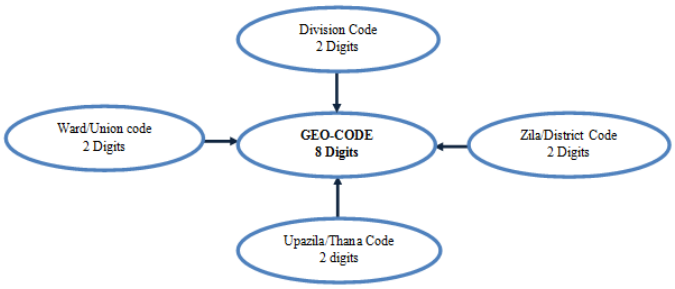

Fig. 3. GEO-CODE sequence of 8 digits used in Bangladesh contains Division Code, Zila Code, Thana Code and Union Code of 2 digits each

Our algorithm of Patient De-Identification with Linkage Preservation (PDLP) is presented below:

Input: Health record set including patient identifiable data
Output: De-identified Linkable patient record
Do

1. Encrypt mobile number
2. Convert
   a. Patient name to name-value
   b. Address to GEO-CODE
   c. Date of Birth / Age to birth-year
3. Generate Patient Identification Key (PIK) from encrypted mobile number, name-value and gender
4. Add PIK and GEO-CODE, birth-year to record set
5. Delete Patient name, Mobile number, Address, Date of Birth, Credit Card number data.

While: End of 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 own birth date is more or less same among poor and uneducated people around the world. In many cultures and jurisdictions, if a person's real birthday is not known (for example, if he or she is an orphan), then their birthday may be considered to be January 1. [29]. The main reason of this, it is very easy to remember. It is also easy to select in the online form's drop down menu. This is main reason, why we have not considered date of birth rather considered birth year in our patient de-identification with linkage preservation technique.

Another important reason is that in most hospitals and diagnostic centers, only patient's age is collected rather than 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.

When the National Health DW will be fully functional, every day more than 65 million records will be added in the fact tables of the warehouse and the DW will consume 12.50 GB memory/day [25]. For improved performance of such a big data repository, fragmentation can be done using vertical or horizontal technique such as [30], [31].

VII. LIMITATIONS AND FURTHER RESEARCH

A problem with mobile number based identification is many people use multiple mobile numbers. A parson 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 warehouse. It will impact on mining results. A simple solution is to develop social awareness by the Government so that citizens provide only one number among their available numbers when taking health and other citizen services.

Another problem, though rarely, may occur due to change of mobile numbers by patients. For example a child after getting adult, own a mobile. He or she has already records in the health warehouse with his guardians' mobile number. This kind of problem can be addressed by writing simple DML query that will replace old patient identification key with new one generated with changed mobile number.

One of the many future research directions can be to design efficient data mining algorithms that can cluster all records of individual patients properly from National Health Data Warehouse.

VIII. CONCLUSION

Development of health data warehouse in national level is essential to deliver quality health services and medical research. Preserving record linkage by retaining identifiable attributes in National Health Data Warehouse is required for effective data mining. But identifiable health data have high risk to patients' privacy. Integrated health systems are in top hit list of cyber criminals as medical data worth higher than credit card numbers in the underground market. In this paper, we have provided a practical solution: Patient De-Identification with Linkage Preservation (PDLP) that can anonymize the identifiable private data of existing billions of medical records while maintaining record linkage. We have used encrypted mobile number, gender and name-value of patients to produce anonymized and linkable Patient Identification Key. This system is being implemented in Bangladesh to develop National Health Data Warehouse. Using PDLP, patients' data can be shared and integrate among different government and private hospitals and diagnostic centers in Bangladesh.

REFERENCES

1. R. Khosla, and T. Dillon, "Knowledge discovery, data mining and hybrid systems," In Engineering Intelligent Hybrid Multi-Agent Systems, Kluwer Academic Publishers pp.143–177, 1997.
2. A. Azzalini, and B.Scarpa, Data Analysis and Data Mining An Introduction, Oxford University Press, 2012.
3. W. Inmon, Building the Data Warehouse, 4th edition, Wiley-New York 2005.
4. T.R. Sahama, and P.R Croll, "A data warehouse architecture for clinical data warehousing," In Proc. of the Australasian Workshop on Health Knowledge Management and Discovery, 2007.
5. 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
6. K. Cios, "Uniqueness of medical data mining," Artificial intelligence in medicine, Vol. 26, 2002, pp.1-24
7. O. Maimon, and L. Rokach, Data Mining and Knowledge Discovery Handbook, 2nd Edition, Springer, 2010.
8. H.C. Kum, A. Krishnamurthy, A. Machanavajjhala et. at., "Privacy preserving interactive record linkage (PPIRL)," J Am Med Inform Assoc vol. 21, 2014, pp. 212–220.
9. 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.
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.
15. F. T. Harold, K. Micki, Information Security Management Handbook, Volume 2, CRC Press
16. Why Hackers Are Targeting Health Data http://www.databreachtoday.asia/hackers-are-targeting-health-data-a-7024
17. Your medical record is worth more to hackers than your credit card, URL http://www.reuters.com/article/2014/09/24/us-cybersecurity-hospitals-idUSKCN0HJ21I20140924
18. Fifth Annual Benchmark Study on Privacy & Security of Healthcare Data, Ponemon Institute Research Report, May 2015
19. Fifth Annual Study on Medical Identity Theft 2014
20. Central European University, Reported Breaches of Compromised Personal Records in Europe http://cmds.ceu.edu/sites/cmcs.ceu.hu/files/attachment/article/663/databreachesineurope.pdf
21. U.S. Department of Health and Human Services https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf
22. Bangladesh Bureau of Statistics http://www.bbs.gov.bd/PageWebMenuContent.aspx?MenuKey=243
23. Bangladesh Telecommunication Regulatory Commission http://www.btrc.gov.bd/content/mobile-phone-subscribers-bangladesh-may-2015
24. S.I. Khan and A.S.M.L. Hoque, "Towards development of health data warehouse: bangladesh perspective", In Proc. of the 2nd International Conference on Electrical Engineering and Information Communication Technology (iCEEiCT), 2015.
25. S.I. Khan and A.S.M.L. Hoque, "Development of national health data warehouse for data mining," Database Systems Journal, Vol. VI, No. 1, 2015.
26. Y. Lu, Concept Hierarchy In Data Mining: Specification, Generation And Implementation, Doctoral dissertation, Simon Fraser University, 1997.
27. Directorate General Of Health Services http://app.dghs.gov.bd/bbscode/
28. http://www.bbs.gov.bd/PageWebMenuContent.aspx?MenuKey=150
29. http://www.syracuse.com/news/index.ssf/2011/01/on_new_years_day_wish_a_happy.html
30. S. Ceri, M. Negri, and G. Pelagatti, "Horizontal data partitioning in database design," In Proc. ACM SIGMOD, pp. 128–136, 1982.
31. S.I. Khan and A.S.M.L. Hoque "Scalability and performance analysis of crud matrix based fragmentation technique for distributed database, In Proc. of 15th International Conference on Computer and Information Technology (ICCIT), pp. 562- 567, 2012.