An Analysis of the Problems for Health Data Integration in Bangladesh

Shahidul Islam Khan, Bangladesh University of Engineering and Technology
Abu Sayed Md. Latiful Hoque, Bangladesh University of Engineering and Technology

Available at: https://works.bepress.com/shahidulislam-khan/8/
An Analysis of the Problems for Health Data Integration in Bangladesh

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; asmlatifhoque@cse.buet.ac.bd

Abstract— Discovering the hidden knowledge from different health data repositories requires the integration of health data from widely diversified sources. While integration, maintaining record linkage is a critical research issue for developing countries such as Bangladesh. 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), etc. These algorithms cannot be used for integrating health data in Bangladesh because of many missing data, different ID used for the same patient in same or different systems and high amount of noise in other patient information. In this paper, we have defined the constraining factors of healthcare data available in Bangladesh and other developing countries. We have analyzed the practical problems of collecting and integrating healthcare data in Bangladesh to build up its National Health Data Warehouse (NHDW). We have also provided some important recommendations to boost the integration process and also to support record linkage.

Keywords—Data Integration; Health Data Warehouse; Record Linkage; Bangladesh

I. INTRODUCTION

Integrated health data repositories or health data warehouse are proven to be beneficial in many fields such as mining health patterns, evidence-based medicine, personalized treatments etc. [1]-[5]. Significantly large volume of medical records and associated documents are generated by health investigative equipment 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 requires for making proper medical decisions are trapped within the fragmented and heterogeneous healthcare delivery systems that are not integrated appropriately. So it is very much essential to integrate these health records into one place—a single data warehouse.

To get most out of Integrated Health Data Repositories (IHDR), linkage of records is essential. Uncovering fruitful knowledge (e.g., finding correlation among diseases) from health dataset requires maintaining record linkage. Record linkage is a mechanism to find out record pairs from multiple information systems which refer to the same object in reality. Given two databases of records, the record-linkage procedure determines all pairs that are similar to each other. The likeness of two records is defined using domain-specific matching over individual attributes. Record linkage is indispensable to join datasets based on entities that may or may not share common identifiers [6], [7].

Numerous techniques have been proposed by the researchers for record linkage in the medical domain. These techniques are shown to be efficient for IHDR of the modern world where health records typically contain PID, correctly inputted name, date of birth, ZIP codes of the patients. But in the cases of developing countries like Bangladesh current linkage techniques cannot be implemented properly. Some common problems of developing countries are illiteracy and dense population. Thus available health data suffers from lacking patient identity in any form (e.g., medical card no., SSN, NID) as well as correct names and DOB of the patients.

Bangladesh government took the initiative to develop National Health Data Warehouse (NHDW) in 2009 with the help of German Donor GIZ. The objective of the warehouse is to build a digital data storage that will eliminate the disparity between the available medical recordsets and will also support interoperability. Healthcare data from different organizations under Directorate General of Health Services (DGHS) of Bangladesh Government are gathered using DHIS2 and OpenMRS; both are open source software. The warehouse provides a gangplank for reporting and data mining functions that could be used by the decision makers at different levels to monitor health improvements at all levels down to Upazila health complexes [8], [9].

In this paper, we have analyzed different problems for healthcare data integration in Bangladesh to develop a national level health data warehouse for the citizen. We have also provided some guidelines to boost up the integration process.

II. CONSTRAINTS OF HEALTH DATA IN BANGLADESH AND OTHER DEVELOPING COUNTRIES

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 [10], [11]. This implies poor health, inadequate education, and the majority of the people lives in...
the rural area. Above socio-economic conditions made an impact in the available health care data of Bangladesh and other developing countries in the following ways.

a) Health records without unique Patient ID: people do not have medical cards with unique health ID. Health care centers do not have provision to store National ID numbers or Social Security Numbers (SSN).

b) Misspelled names: Many people in real do not know their full name and unable to pronounce their name correctly even in the mother tongue. The same person provides a different version of their name in the health care facilities. The problem can be understood from the table below.

<table>
<thead>
<tr>
<th>Actual Patient Name</th>
<th>Inputted Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abul Hossain</td>
<td>Mr. Abul Hossain</td>
</tr>
<tr>
<td></td>
<td>Mr. Md. Abul Hosen</td>
</tr>
<tr>
<td></td>
<td>Mohammad Abul Hossain</td>
</tr>
<tr>
<td></td>
<td>Abul Hosen</td>
</tr>
<tr>
<td></td>
<td>Abul Hossen</td>
</tr>
<tr>
<td></td>
<td>Mr. Abul</td>
</tr>
<tr>
<td></td>
<td>Md. Hosen</td>
</tr>
<tr>
<td></td>
<td>Md. Hossain</td>
</tr>
</tbody>
</table>

- No actual date of birth: Enormous people do not know their actual birth date because of lacking of birth registration. For several years, they provide same age (e.g., 43 years) to hospitals and diagnostic centers. A lot of people do not know their actual date of birth in Bangladesh. This is a very common scenario for aged rural people with less education.

d) Missing attribute values: As Bangladesh and other developing countries have the dense population and inadequate facilities, in all health centers there are long queues of patients. So many necessary attribute values cannot be inputted for processing a high number of patients in limited time.

e) Error in data: less qualified staffs for inputting patients’ data. This leads to unintentional wrong input data.

So patients’ health records in Bangladesh contain more noisy data with more missing values and without unique patient identification numbers. The situation is depicted in Fig. 1. These makes techniques of developed countries for record linkage inappropriate for Bangladesh and more specialized technique to address the situation.

III. HEALTH DATA GENERATION AND STORAGE IN BANGLADESH

The patient visit cycle to different hospitals, diagnostic centers, and private practitioners’ chamber is illustrated in Fig.2 to Fig. 5. The patient visit to different health service providers can be grouped as follows.

1) Patient treatment cycle in Bangladesh
   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), the total number of government hospitals under DGHS is 592 [12], [13]. 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 [14], [15].
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 write 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 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 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%), every time when 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 (Fig. 4).

2) **Record linkage problem**

Based on the patient cycles as described above, different cases arise.

Case 1: Patient $P_1$ visits hospital $H_1$ at timestamp $T_1$ for event $e_1$ with ID $\text{concat}(P_1, H_1, T_1, e_1)$

Case 2: Patient $P_1$ visits hospital $H_1$ at timestamp $T_2$ for event $e_1$ with ID $\text{concat}(P_1, H_1, T_2, e_1)$

................................................................................

Case i: Patient $P_1$ visits hospital $H_k$ at timestamp $T_i$ for event $e_1$ with ID $\text{concat}(P_1, H_k, T_i, e_1)$

................................................................................

Case n: Patient $P_1$ visits hospital $H_k$ at time stamp $T_n$ for event $e_1$ with ID $\text{concat}(P_1, H_k, T_n, e_1)$

Now the question is how many possible records that can be evolved in the lifetime of a patient?

Let $RL$ is the total health records for the lifetime of a single patient.

So $RL \subseteq H \times T \times e$

We can estimate an upper limit of $RL$ as follows:

Let life span of a person = $y$ years

Average visit to any health care facility per month = $v$

Visit per year = $12v$

Total health care visits in life span of a person $T= 12vy$

Average Life expectancy in Bangladesh: Male-70years and Female-72years [17].

If we consider $y=71$ and $v=3$/month

So $T= 12 \times 3 \times 71 = 2556$ times

If one visit creates one record, $RL = 2556$
In Bangladesh perspective, health records of a person are stored either in electronic form or hard copy format and 2556 different records of the same person are stored with 2556 different identities. These records are highly distributed in terms of time (e.g., doing pathological tests in different times), space (e.g., outdoor, indoor or lab), data type (e.g., MS Excel or Oracle), and locations (e.g., different hospitals).

For exaction of fruitful knowledge from health data, it is the first requirement to accumulate health records from widely variable sources. Privacy of patients needs to be preserved as health data have high sell values [18]. While accumulation, these records cannot be mapped with the patients because of storing patient records every time with different identities. Record linkage problem is to find an optimum reliable mapping of each patient to his health record throughout the lifespan.

IV. RECOMMENDATIONS

Followings are our recommendations to improve the usability, performance, and security of the National Health Data Warehouse (NHDW) of Bangladesh.

1. It is very important to include all healthcare providers, regardless public or private, in the NHDW soon. If the coverage of the national data repository reaches 100%, then the Govt. will be able to take more fruitful administrative and management decisions.
2. Data quality is very important because utilization of the data will improve the service delivery. Policy makers can be benefited by using quality data for taking better decisions. There is scope to improve the quality of the available data in the warehouse. For future data collection, data entry personnel at Upazila level should be given more training.
3. Patient information collection forms in all hospitals, clinics, diagnostic centers, research centers etc. need to be standardized and necessary fields should be included.
4. As health records are stored in different medical centers in different formats (e.g., MS Excel, Oracle, MySQL, etc.), this heterogeneity needs to be addressed properly by adopting efficient Extract, Transform, and Load (ETL) mechanism.
5. For privacy reason, no medical record can be stored at any level, from diagnostic centers to NHDW, with personally identifiable attributes of the patients.

V. CONCLUSION

Health data warehouse development is a complex and time-consuming process but plays a vital role in delivering quality health services. To extract knowledge from different health data sources requires the integration of healthcare data from various diversified sources like hospitals, clinics, diagnostic centers, research centers etc. Preservation of record linkage using identifiable attributes of patients in national health data warehouse is very important for knowledge discovery. Researchers have given different solutions of health data integration and record linkage that are applicable for developed countries where electronic health record of patients are maintained with different identifiers. Their solutions are inappropriate for integrating health data in Bangladesh because of many missing data, different ID used for the same patient in same or different systems and high amount of noise in other patient information. Bangladesh government is developing Health Data Warehouse at the national scale. In this paper, we have characterized the factors constraining health care date 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 guidelines to boost the integration process and also to support record linkage.

REFERENCES