Asian J. Med. Biol. Res. 2019, 5 (4), 258-264; doi: 10.3329/ajmbr.v5i4.45262

Asian Journal of
Medical and Biological Research
ISSN 2411-4472 (Print) 2412-5571 (Online)
www.ebupress.com/journal/ajmbr

Review

An overview of data mining in medical informatics: Bangladesh perspective

Mst. Farzana Akter*, Tanjila Islam, Kaniz Fatema Trisha and Mohammad Ohid Ullah

Department of Statistics, Shahjalal University of Science and Technology, Sylhet-3114, Bangladesh

*Corresponding author: Mst. Farzana Akter, Department of Statistics, Shahjalal University of Science and Technology, Sylhet-3114, Bangladesh. Phone: +8801875750485; E-mail: farzanajui11@gmail.com

Received: 14 November 2019/Accepted: 26 December 2019/ Published: 31 December 2019

Abstract: Recently using information technology in the health care system is an important issue. Medical informatics is the combination of information science, computer science, and health care. As population is increasing rapidly, it is obvious to use medical informatics to save human lives and to treat people in efficient way. Therefore, we tried to explore an overview of the necessities and practical uses of data mining in administrative, clinical, research as well as educational aspects of medical informatics in Bangladesh. It is one the most populous countries in the world and the health care system including data mining in medical informatics is not so handy. Besides, for the effect of monsoon weather people of this country are affected by various diseases but poor investment and weak implementation make these diseases a burden. The study focuses on the needs of clinical data warehousing and the practice of examining these databases in order to improve various aspects of medical informatics in Bangladesh. The study suggests that government and private health care organizations need to take account to store their data and create a research wing in every hospital in Bangladesh as well as other developing countries in the world so that researchers and doctors may be able to find out the solution of their problems. For the greater benefits of the people, more research on medical informatics is essential and implementation of the research outputs must be done in medical treatment.

Keywords: data mining; medical informatics; Bangladesh

1. Introduction

Medical informatics is in its most sensational period because nowadays it is easy to handle big data by computer. Medical informatics is the study of information engineering and how to exercise it in health care field. This disciplines works jointly with information science, computer science, social science, behavioral science, management science, and others (Nadri *et al.*, 2017). As the population is increasing rapidly in the world mostly in developed countries and so the cost of health care, governments and other health organizations are mostly rely to save time, money and human lives on the use of health Informatics (Raghupathi, 2010). From an annual report of United States 44000 to 98000 patients died in hospitals because of medical error (Institute of Medicine (US) Committee on Quality of Health Care in America, 2000). A study shows that due to drug the adverse events and ranging from minor side effects to death of an individual costs $136 billion per year only in United States (Johnson and Bootman, 1995). Electronic patient records, technology based alerting, reminder, predictive system for administration of hospitals, and education and training for nurses and doctors can reduce medical error and financial costs of healthcare system (Raghupathi, 2010).

Data produced in healthcare system is an example of "Big data". The term "Big data" is designated newly but the process of analyzing large dataset is old enough (Everts, 2016). In hospitals, large amount of data generates from each patients such as datasets including MRI images or gene microarrays for each patient (Herland *et al*., 2014). As of 2011, health care organizations had generated over 150 Exabyte's of data (one Exabyte is 1000 petabytes). It is noted that as Health informatics generates large and growing amount of data, the healthcare industry can be protected by using data mining in Medical Informatics up to $450 billion each year just in the

United States (Herland *et al*., 2014; Yuan *et al*., 2013). Since 1971 the healthcare system of this country does not have any significant development though there is mass population. Government have taken many steps and set up medical forums to develop the healthcare system as effective and accountable. Though there is adequate number of qualified medical personnel still people is getting improper treatments, unsatisfactory care and losing confidence over this system. Lack of medical equipment, lack of adequate training of health personnel, lack of information about the patients and diseases make it more challenging (Al Mahdy, 2009). For the improvement of healthcare service it is necessary to establish clinical data warehousing system and the practice of examining these databases in various aspects of medical informatics. In this modern era developed countries like USA, Canada etc. can provide proper treatment to patients; retrieve patients' historical data, assist disease predictive and prevention system through this data mining process in health informatics (Alkhatib *et al*., 2015). In this study we discuss the present situation of applying data mining in various aspects of medical informatics in Bangladesh.

2. Applications of data mining on medical informatics in Bangladesh

Medical informatics contains mainly four subfields: Clinical care, Administration of health services, Medical research and Education & Training. Here we present the details of each subfield in Bangladesh perspective.

2.1. Clinical care

Normally doctors or nurses prescribe new medicines to the patients on the basis of the information of previous medical reports. But in Bangladesh most of the physicians prescribe medicines without checking previous history and family records as there is no centralized electronic patient record database system. Even most of the time they only prescribe medicines based on symptoms without rechecking those symptoms via tests. If health care system of Bangladesh uses centralized patients record database, by applying data mining in that database physicians can go beyond what is more appropriate to treat a patient. For similar case a new physician could refer more effective medicine to patient by query for the decision based on the centralized database (Raghupathi, 2010).

Clinical Decision Support System (CDSS) is a computer based software which is designed to help health care provider to make health care decisions. Search capabilities for medical queries, monitor inputs and check them for predetermined triggers, reminders for periodic tasks, suggestions based on medical knowledge and different prediction models like diagnosis and prognosis are the implementations of CDSS (Raghupathi, 2010). But CDSS is not available in Bangladesh yet. Knowledge based system and data mining based system are the main types of CDSS. A hybrid of these two systems attains high performance. This hybrid CDSS was proposed for rural Bangladesh but it has not applied yet (Iqbal, 2012). Figure 1 shows how CDSS works.

CDSS have some subsections like Health Evolution through Logic Processing (HELP) system, The Acute Physiology and Chronic Health Evolution (APACHE) series of models and the pneumonia Severity of Illness Index are currently using in clinics and hospitals of developed countries (Raghupathi, 2010). It is high time to use these technology in Bangladesh.

2.2. Administration of health services

In health care organizations administrators take large number of steps to make better the health status of the population. Those steps totally depends on the data related to the incidents for which decisions are going to be made. Administrators are going to take decisions about arranging more tools and equipment's for a specified time such as disease outbreak and decide whether or not any additional services requires in terms of cost and benefit. For taking these type of decisions in developed countries they use a computer based system to predict their needs accurately (Raghupathi, 2010). But in Bangladesh health care institutions does not have such kind of system and sometimes they cannot handle the pressure of patients and cannot treat them effectively because of scarcity of tools and beds at that period of time.

To detect disease outbreaks, Izadi and Buckeridge developed a system POMDPs (Partially Observable Markov Decision Processes) which can suggest a better solution to overcome the disease outbreaks in terms of cost and effects and predict about outbreaks though the amount of false detection of outbreaks by this system is not affordable for practical use (Raghupathi, 2010). In Bangladesh administrators need such kind of system to improve the health care facilities and to give better care to the patients but that kind of system need to have small fraction of error.

2.3. Medical research

Most basic areas covered by medical research include cellular and molecular biology, medical genetics, immunology, neuroscience, and psychology. Researchers aim to establish an understanding of the cellular, molecular and physiological mechanisms underpinning human health and disease. Data mining methods can help researchers to obtain penetrative patterns, cause and effect relationships, and prognostic scoring system from available data of medical patients. If clinical and administrative decision support systems are starts to apply in various clinics, hospitals, and research centers, this method could apply on those small or scatter data from clinics, hospitals and research centers (Raghupathi, 2010).

In Bangladesh, Bangladesh Medical Research Council (BMRC) a focal organization for health research established in 1972 as an autonomous body under the Ministry of Health and Family Welfare. Its target is to focus on the problems and issues related to medical and health sciences and determine priority areas in research on the basis of healthcare needs, fields of medicine, public health, reproductive health and nutrition and also make sure to arrange the application and utilization of the results of these researches. BMRC publishes a quarterly bulletin; a journal called Research Information and Communication on Health (RICH) twice a year; a newsletter twice a year; and a journal titled Current Awareness Service (CAS) once a year.

2.4. Education and training

Education and training fourth subfield of medical informatics is the application of Data Mining techniques to educational data. To provide knowledge about medical informatics to healthcare professionals like doctor, nurse, paramedic etc., to retrain them and to keep them up to date with modern technologies of medical science it is seen as the emerging interdisciplinary research area in the field of e-learning (Raghupathi, 2010; Romero and Ventura, 2010). Trainee, instructors and administrators can be benefitted from the data mining techniques as it monitor the learning paths, resources, materials of learning tasks, discover learning patterns, web based educational system, intelligent tutorial system and e-learning (Raghupathi, 2010; Sachin and Vijay, 2012). University of Alberta developed a centralized e-learning system and internet community named HOMER which provides medical students free online lifetime access with various learning materials of advanced medical research and knowledge (Raghupathi, 2010). A Naive Bayesian approach is applied by Leonard *et al*. to find cross-references between the symbol of genes and proteins and Medline articles which case study is considered as an overview of a relatively new data mining technique to find relevant reference articles for particular genes and proteins (Leonard *et al*., 2002). Flow chart of education and training system is shown in Figure 2.

But in Bangladesh medical students and trainee do not have any modern technological education advantages as if they can be benefitted with more successful learning experiences. As a result they always lag behind and administration deprived of the effectiveness of the modern educational program. So medical professionals treat with back dated medical treatment systems.

Bangladesh has achieved admirable progress in healthcare sector and socioeconomic development over the last few decades. As a developing country Bangladesh have quite insufficient budget resource allocation in healthcare sectors compared to other developing economies. A report shows that in 2014 less than 6% of the total expenditure was for health care sector. In 2017-18 fiscal years the budget allocation for health and family welfare sector was 5.2% which is much lower than the 15 percent budgetary allocation recommended by World Health Organization (WHO). For achieving the objectives of the Sustainable Development Goals (SDG), the expenditure for health is too inadequate. It also represents a poor amount of allocation for a country of over 160 million people. According to WHO, the ratio of doctors to nurses to technologists should be 1:3:5 but in Bangladesh the ratio stands at 1:0.4:0.24. This inappropriate ratio is the result of poor and insufficient allocation of budget in health care system.

3. Data warehouse

Data warehouse (DW) is a large, central electronic storage system of oriented, integrated, well-decorated and with all significant parts of data. Basically data warehousing was first used for business intelligence in context of sales, production, planning and in order to keep information of employees and was invented in 1988 by IBM researchers Barry Devlin and Paul Murphy. Typically in any organization data is kept for a short period of time but in the field of medical informatics in DW (Data Warehouse) data must be kept for the people's lifetime and till the cure of diseases and with time-variant (Pedersen and Jensen, 1998). It is helpful for getting information easily, in research purposes, and for making better decisions in healthcare sector. This system also can help to reduce the total cost on healthcare. Simultaneous access by multiple medical personnel and patients from multiple reliable sources, online processing for administrative and clinical initiative, the expenses for clinical proceedings, enhancing the quality of data and compatibility, security etc. are the advantageous features of data

warehousing (Raghupathi, 2010; Khan and Haque, 2015). There are some characteristics a proper DW should carry on: (1) This should be central, easily accessible, (2) Should have all significant parts of data or information, (3) Properly oriented by subject, (4) Perfectly integrated, (5) Well-decorated as any sector related with this data house could be benefited like researchers or others and (6) Confidential and security should maintain hardly. In Bangladesh healthcare systems are weak and not properly oriented like other developed country. As we are on a journey to build a digital Bangladesh so healthcare sector should also be developed and on the contrary central data warehouse with healthcare information is the first and foremost step on this. For having this the great rate of progress in the field of information technology, data mining, computing, information security and also with great approach and expansion of World Wide Web (WWW) are highly necessary. This can help to get information for all like junior or senior doctors, medical students, patients (limited as they required) and researchers also. To establish this whole storage this may could be a long and critical procedure and also may take time a little longer but this is highly needed to develop the healthcare system of this country.

We are suggesting a software based data warehouse system where information can be input directly into a software from heterogeneous sources as shown in the above flow chart of Figure 3. This software will be designed in such a way that the raw data will be converted into an oriented form and this organized information will be stored into different folders in a database. Peoples from different levels with respect to health care system e.g. patients, doctors, nurses, administrators, medical students, researchers etc. can query for necessary information from that database. This database system can reduce higher cost of implementing health data warehouse system.

4. Big data and data mining in health informatics

Data mining is a process of collecting useable information from large data set which is known as big data. In this process one or more software will be used to analysis these large raw dataset. In the middle of 1990's data mining was emerged as a new concept and new approach of analyzing data and knowledge discovery. In 1995 the first ACM conference on knowledge discovery and data mining was held in the USA. But in late 2009 the term "Data Mining" was first registered for the 2010 Medical Subject Headings (MeSH1) (Yoo *et al*., 2012). Big data has some characteristics which are: Volume, Velocity, Variety, Veracity and Value. Volume means size of gathered data; Velocity refers the new data generation speed; Variety pertains to the norm and nature of the data; Veracity measures the precision of the data and Value appraises the quality of the data in terms of the aimed result (Hilbert, 2016; 2015). Data for Health informatics research has some of the characteristics of big data. Large amount of data generates from each patients such as datasets including MRI images or gene microarrays for each patient. Big Velocity occurs when data generated at high speed which can be seen at time of observing real time events whether that can be observing a patient's existent condition through medical equipment or trying to track an epidemic through incoming web posts (such as from Facebook, Twitter etc.) Big Variety pertains to that datasets which includes large amount of different types of independent variable, datasets that collected from different sources or any dataset which is complex and thus need to be evaluate at many level of data throughout Health informatics. High Veracity of data arises in Health Informatics because of faulty clinical censors, gene microarrays or from individual patients data stored in database. High value of data can be gathered by traditional methods such as clinical settings (Herland *et al.*, 2014). The field of Health Informatics has various sub-fields including Bioinformatics, Image Informatics (e.g. Neuroinformatics), Clinical Informatics, Public Health Informatics, and also Translational Bioinformatics (TBI). Bioinformatics is not usually supposed as a part of Health Informatics but it becomes a vital source of health information. It is a field which uses various tools and develops method to describe biological data. It uses the knowledge of computer science, statistics, biology and engineering to analyze the biological data and interpret them to improve the health care condition (Lesk, 2011). Bioinformatics data such as DNA sequence of thousands of organisms is incessantly increasing which is the great example of Big Volume. McDonald created khemr which is a Bioinformatics suite of software which seeks to solve hardware computational problem and it helps to pre-process Big Volume genomic sequence by converting it into short fragmented sequence which can be stored in Bloom filter-based hash table to analyze the data effectively and efficiently (McDonald and Brown, 2013). Neuroinformatics is a subfield of Health Informatics which investigated only the brain image data to know how brain works, connection between various part of body and brain and find relation between brain image data and medical events information. Neuroinformatics works mainly with neuroscience and informatics research to improve and uses computer based tools in understanding the function and structure of brain. This subfield covers mainly – tools for analyzing, visualizing nerve system, theoretical, mathematical and simulation environments for describing the structure and function of brain. Clinical Informatics helps to predict that can

assist a physician to make a faster and more accurate decision about patients by analyzing their data. According to a research, the clinical research and the implication of that research in practice has about 152 years distance (Bennett and Doub, 2011). Nowadays decisions are mainly based on previous information or on what have been found by experts but if physicians embrace the findings of recent research which defines a new ways then the decisions would be more accurate and reliable about patients. Public health Informatics deals with population level data to achieve acuteness of medical state. Population level data collects from population via social media, poll or from hospitals, doctors, clinics and this type of data has Big Volume, Big Velocity and Big Variety. Translational Bioinformatics is a subfield of health informatics which tries to converge molecular bioinformatics, biostatistics, statistical genetics and clinical informatics. It mainly works with informatics methodology to rapidly increasing biomedical and genomic data to improve medical tools for efficiently using it in health care by doctors, experts etc. (https://www.amia.org/applications-informatics/translational-bioinformatics).

For using data mining to medical data firstly we have to know the algorithms or techniques of data mining. Usually there are two classifications of algorithms of data mining. Descriptive data mining for measuring the similarity between objects and identify the patterns in data so that a big data easily can understand. This includes clustering, association, summarization, and sequence discovery. Predictive data mining applies the methods to unpredicted data including classification, regression, time series analysis, and prediction (Yoo *et al*., 2012). To create an optimal result from a big data by data mining, there are 5 techniques those could apply, like; classify different data in different classes(classification analysis), using proper methods to identify relations between variables in database (association rule learning), detect or identify the observations that are failed to complete the expected requirements for having a place in the dataset (anomaly or outlier detection), clustering the observations by putting same characteristics observations in the same group or cluster and run analysis based on clustering which may help to find the degree of association between to objects and also helpful for making personal profile of patients (clustering analysis) and using of regression analysis to find out the dependency of variables and so on (regression analysis).

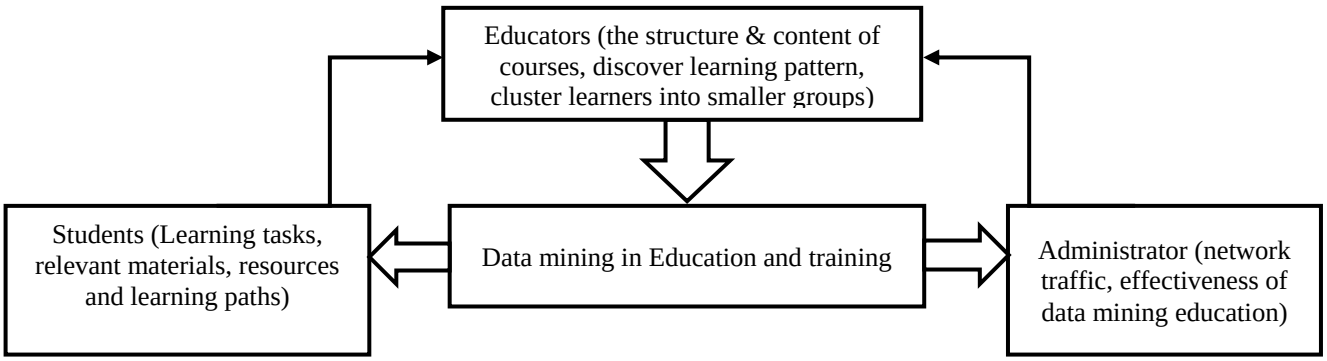

Figure 1. How clinical decision support system works.

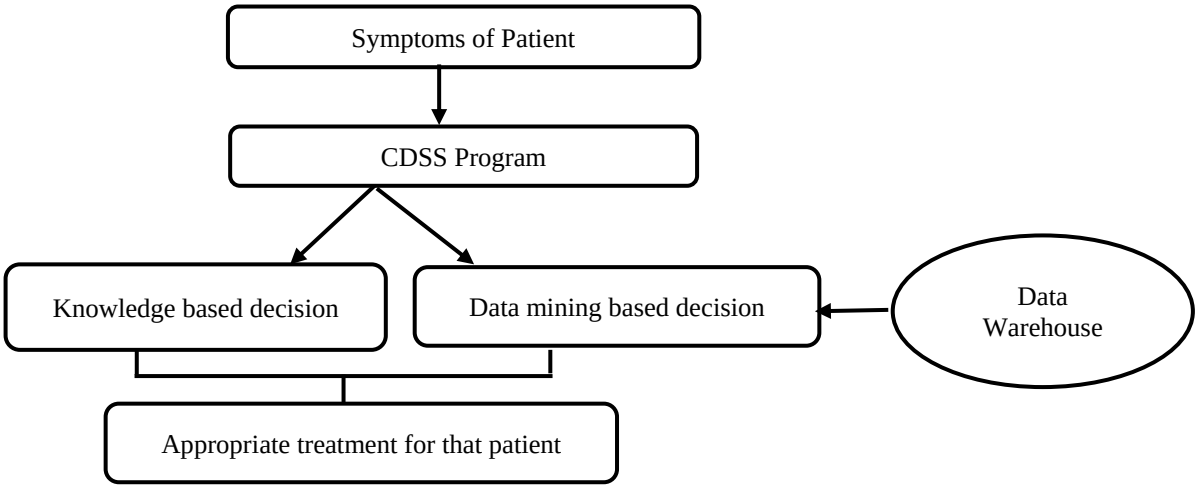

Figure 2. Flow chart of education and training system.

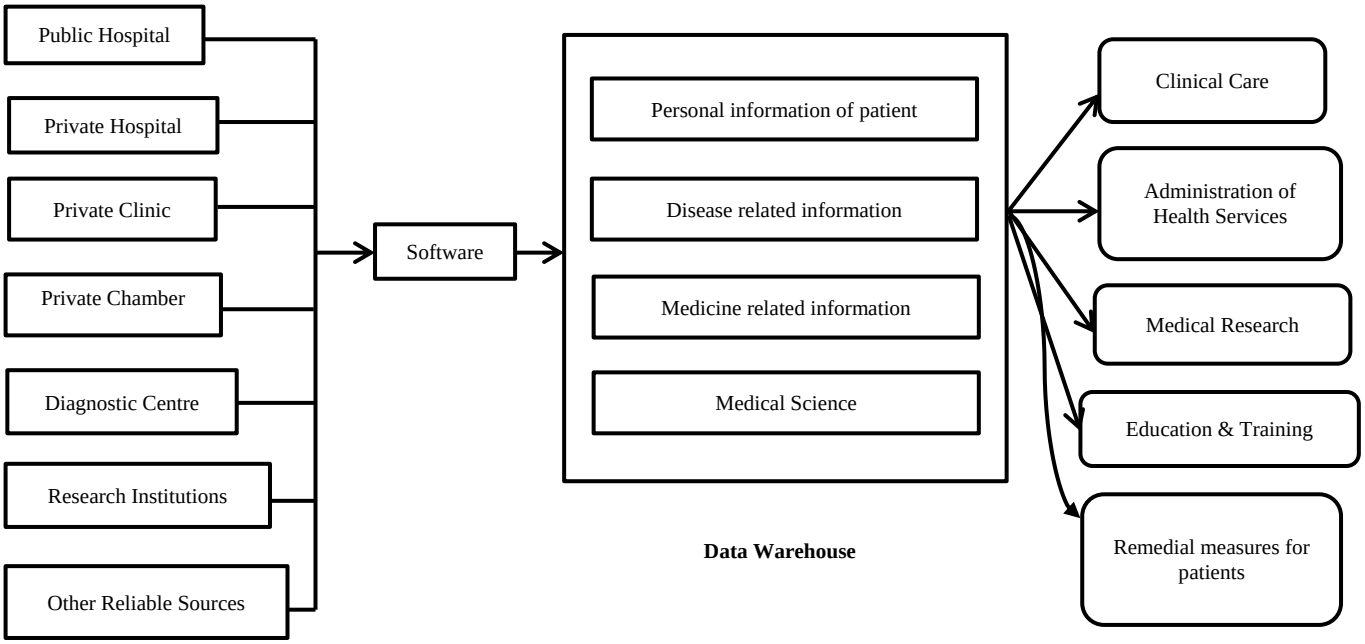

Figure 3. Data warehousing for health care system in Bangladesh.

5. Recommendations

We recommend that it needs to increase the allocation of budget for health and also improve rural health care services and resources, revise and develop health policies, develop existing technologies and include new necessary technologies as people's expectations to ensure relatively quick and proper health care service. In order to solve the problem arising with medical treatment and to improve the quality of health care system data mining has potential impact on medical informatics.

6. Conclusions

In this study we discoursed about medical informatics and its different levels, status of Data Warehousing, Data mining and its technique in Bangladesh. From this study we established that various research work has been done on different aspects of medical informatics but the finding of these research work doesn't use in real life applications. Data formed in health care sector should be easily accessible for students and researchers as they can improve the quality of findings. And also Government should consider about these findings to apply these in relative sectors.

Conflict of interest

None to declare.

References

Al Mahdy H, 2009. Reforming the Bangladesh healthcare system. Int. J. Health Care Qual. Assur., 22: 411-416.

Alkhatib MA, A Talaei-Khoei and A Hossein Ghapanchi, 2015. Analysis of Research in Healthcare Data Analytics. Proceedings of the Australasian Conference on Information Systems (ACIS), November 30-December 4 2015, Adelaide, Australia, pp. 1-26.

Bennett C and T Doub, 2011. Data mining and electronic health records: selecting optimal clinical treatments in practice. ArXiv, abs/1112: 1668.

Everts S, 2016. Information Overload. Distillations, 2: 26-33.

Herland M, TM Khoshgoftaar and R Wald, 2014. A review of data mining using big data in health informatics. J. Big Data, 1: 2.

Hilbert M, 2015. DT&SC 7-3: What is Big Data? Available: https://www.youtube.com/watch?v=XRVIh1h47sA

Hilbert M, 2016. Big data for development: a review of promises and challenges. Dev. Policy Rev., 34: 135-174.

Institute of Medicine (US) Committee on Quality of Health Care in America, 2000. To Err Is Human: Building a Safer Health System. Eds. Kohn LT, JM Corrigan, and MS Donaldson, National Academies Press (US), Washington (DC).

Iqbal RA, 2012. Hybrid clinical decision support system: An automated diagnostic system for rural Bangladesh. Proceedings of the International Conference on Informatics, Electronics & Vision (ICIEV), Dhaka, Bangladesh, pp. 76-81.

Johnson JA and JL Bootman, 1995. Drug-related morbidity and mortality: a cost-of-illness model. Arch Intern Med., 155:1949-1956.

Khan SI and ASL Haque, 2015. Development of national health data warehouse for data mining. Database Syst. J., 6: 3-13.

Leonard JE, JB Colombe and JL Levy, 2002. Finding relevant references to genes and proteins in Medline using a Bayesian approach. Bioinformatics, 18: 1515-1522.

Lesk AM, 2011. Bioinformatics | Science. Encyclopedia Britannica. Available: https://www.britannica.com/science/bioinformatics

McDonald E and CT Brown, 2013. khmer: Working with big data in Bioinformatics. ArXiv, abs/1303.2223.

Nadri H, B Rahimi, T Timpka and S Sedghi, 2017. The top 100 articles in the medical informatics: a bibliometric analysis. J. Med. Syst., 41: 150.

Pedersen TB and CS Jensen, 1998. Research issues in clinical data warehousing. Proceedings of the Tenth International Conference on Scientific and Statistical Database Management, July 01-July 03 1998, Washington (DC), United States of America, pp. 43-52.

Raghupathi W, 2010. Data mining in healthcare. In: Healthcare Informatics: Improving Efficiency and Productivity, Ed. Kudyba, S., pp. 211–223.

Romero C and S Ventura, 2010. Educational data mining: A review of the state of the art. IEEE Trans. Syst. Man. Cybern. Part C Appl. Rev., 40: 601-618.

Sachin RB and M Vijay, 2012. A Survey and Future Vision of Data Mining in Educational Field. Proceedings of the 2nd International Conference on Advanced Computing & Communication Technologies, January 07-January 08 2012, Washington (DC), United States of America, pp. 96-100.

Yoo I, P Alafaireet, M Marinov, K Pena-Hernandez, R Gopidi, JF Chang and L Hua, 2012. Data mining in healthcare and biomedicine: a survey of the literature. J. Med. Syst., 36: 2431-2448.

Yuan Q, EO Nsoesie, B Lv, G Peng, R Chunara and JS Brownstein, 2013. Monitoring influenza epidemics in China with search query from Baidu. PLoS ONE, 8: e64323.