Medical Data Mining: Health Care Knowledge Discovery Framework Based On Clinical Big Data Analysis

Dalia AbdulHadi AbdulAmeer*

* University Of Information Technology and Communications - Iraq

Abstract- "Big Data" is a new paradigm that is introduced in the field of computer science to abstract the size of data. It refers to huge quantities of data which need special processes according to business requirements. To get exact information and gain knowledge, professionals must apply intelligent tools, technologies and methodologies in the field's major areas. Yearly a lot of people lost their life because of medical mistakes. Diagnosing and treating patient's case is crucial matter to the medical team. Providers should deliver more accurate and personalized clinical data to improve the quality and efficiency of care. This paper discusses big data term in medical sector and how to analyze this clinical big datasets to discover knowledge to use it in clinical prediction. We discuss the deployment and the new trends with big data and explore two paradigms on building real data infrastructure for future studies and review some of on-line health care applications. Afterward, this paper propose framework which figure-out how to analyze big data and how to discover knowledge from the extracted information. Knowledge discovery and data mining are correlated terms; data mining is a step in knowledge acquisition processing journey, it is about applying data analysis and discovery algorithms, which should convey the widespread growth of data collection. As conclusion, big data analysis is expected to reveal the knowledge structure which guided the decisions making.

Index Terms- Big Data, Clinical data analysis, Data Mining, Data Warehouse, Healthcare knowledge.

I. INTRODUCTION

IT professionals in need of organizational data to assist stakeholders in doing informed decisions. These data comes through organizations' operational systems records and data warehouse [1]. Large amounts of data that come from business information systems known as big data. Big data is a common term in clinical area publications but in computer science means big datasets which distributed on many servers and require specific methodologies to process this data [2].

Health care industry has a big challenge on what to do about the collected and saved clinical data and how to analyze it according to specific perspective and different factors depends on health care foundations[3]. Medical or healthcare big data comes from electronic health records (EHRs) so health care organizations will help in mining information and reports from the EHRs using the proper tools, techniques and infrastructure. The collected data diverse between structured and unstructured data, it's important to mention that doctors' notes are a kind of unstructured data. In fact, big data term called on the application of specialized methodologies, technique and processes that mange a very large groups of datasets[3, 4].

In the past five years big data term is numerously mentioned on most clinical publications and reports. It refers to data that stored from clinical software packages. Big data is raw data, it's really big and will getting bigger day after day, and becomes complex more and more[5]. Therefore the processing of this big data is urgent and critical to get meaningful subjects[6]. Healthcare professionals are the people who put standards and identify the industries' needs based on million cases, which help on doing accurate decisions and managing the populations' health at management level.
Big data analytics main goal is to help stakeholders in concerned organizations to make informed and right decisions. As well as, big data applications looking for solving the problems of heterogeneity, time wasting, delay, privacy manner and high costs to maintain the processed data[3].

Medical studies need to process and analyze a large volume of data to handle it. So, the big challenge is how to turn the big data into knowledge or to help pharmaceutical companies to identify side effects of drugs or any process in direct contact with medical institutes to add more to the body of knowledge[7].
Knowledge in general is all about information, getting useful ones by understanding the patterns. In this paper when mention knowledge means how to implement and guide the electronic health records to improve health organizations outcomes according to clinical trends; using artificial methods to convert clinical data into statistical or logical information. These intelligent outcomes varies among diseases detection on early stages, monitoring patients who has a chronic disease such as diabetes or congestive heart failure, autism discovery and follow-up or may be it helps on addressing some parameters to improve people with autism disorder, registering

the behavior of group of candidates after taking a certain medication or undergoing treatment for a particular style or observe a side effects of specific medication. Also it helps people to figure out the best health plan for them [3, 6, 8].

II. DEPLOYMENT AND NEW TRENDS

A. BIG DATA

Big Data terminology referred to a large collection of datasets that is processed using traditional computer techniques and algorithms[2]. So big data related to data's size, speed and volume which need advanced computer science techniques and technologies to manage and analyze this data[9]. Big data can be categorized into six categories, as shown in Fig.1, according to the sources that it comes from such as web logs, call and medical records, surveillance of military purposes, video and photography archives[3].

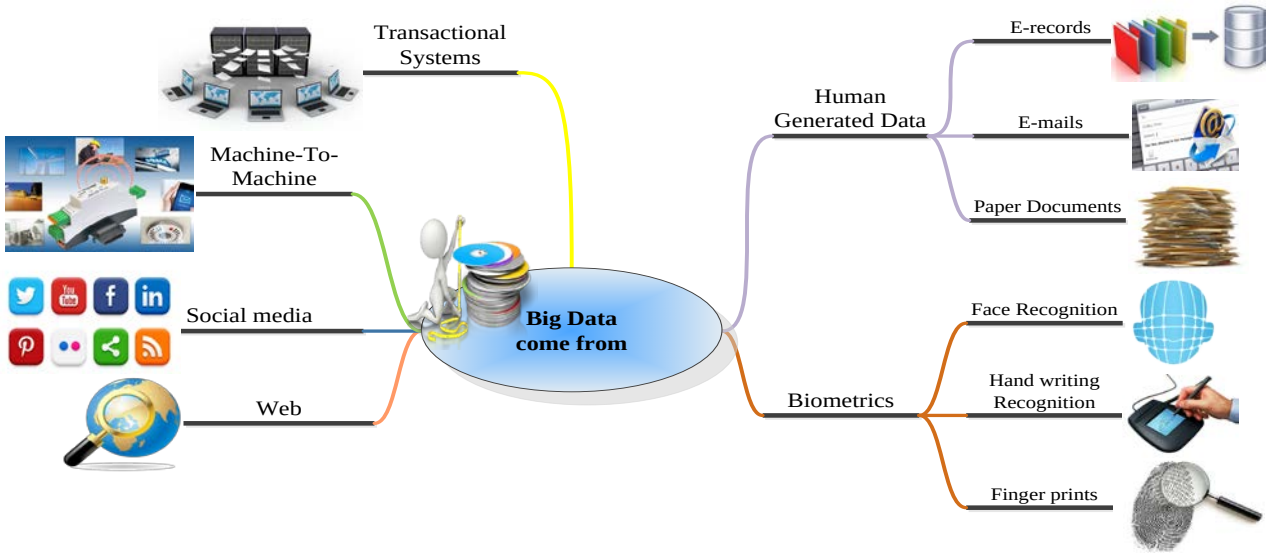

Figure 1: Big Data Resources

In other words, big data is everywhere and it growing exponentially at every moment. Big data in health care branch depends on real world sight to form an actual view to the digitized patients' records so big data in health care are adopted from real time electronic health records[10]. Experts claim that they drive their knowledge to combine, organize and analyze clinical big data to improve clinical care activities and outcomes in order to assist health care organizations, but they aware that this data is not organized in the right way and forms[4, 7]. Big data benefits in health care sector are abstracted in Fig. 2.

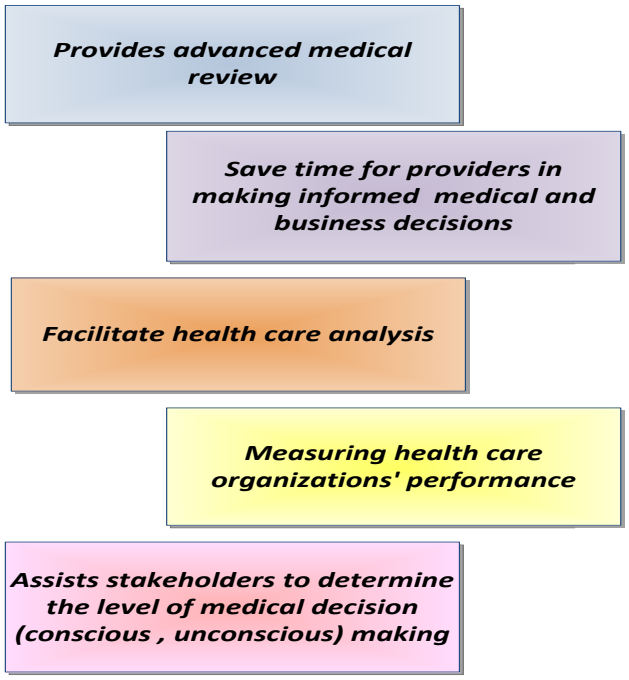

Figure 2: Big Data uses in healthcare

B. REAL WORLD HEALTH CARE STORIES ON BUILDING BIG DATA

In real world when says big data we speak about data that its size on terabyte or on petabytes (one petabyte= 1024 terabytes). It is worth mentioning that Facebook alone store about 100 petabyte of images, videos and other materials. Although there are three common patients online applications submitted by the biggest information technology companies in the world, but the world still needs more. Those three web sites are dossia.com which constructed by Intel, AT&T, Wal-Mart in December 2006. Second online health website is announced by Microsoft in October 2007 this website is HealthVault.com and the last one is Google Health which established in 2008[11].

According to International Data Corporation (IDC) Canada, prediction based on IT research associations big data is ranked as one of the top matters that the world's deal with in 2013. Since that the world get more interest and start focuses more on big data and apply it on different fields. Again big data comes from historical data, billing systems records, web blogs, medical records, image and video archives, military monitoring, voice materials and etc.

The first trend of Big data is by establishing an organized big data, for example what United Kingdom's National Health Service (NHS) done, is to build a paradigm of national research database that support researchers, doctors and technicians to get more understanding to the genetic causes to cancer. The construction of this national database planned to take around three to five years and the volunteers' information will be protected; this known as population health care management. The population health care management needs a modern IT tools, algorithms and techniques besides the availability of effective analytic methods to formalize the available EHRs.

The second new trend in big data categorization at real world applications is Medical history of a family is one of the important records that visual data analysis and health informatics starts dealing with; which is known as medical family tree[3, 12]. This history includes illnesses, symptoms, and allergies causes that family members share them based on their genetics and lifestyle. The mentioned factors are important keys to understand the medical conditions and disorders among relatives[6]. Professionals in health care sector through data mining techniques can notice patterns of their medical history to determine whether individuals or even future generation may be at risk of a particular indication or can detect the percentage value of a chronic disease may occur. Here the basic idea of big data analysis is to present the data in different forms which allows humans to use his mental inceptions and prove it on big datasets though visual data mining techniques, big data can be used to identify patients and populations at risk for various conditions. In this case, visual data clarification is very useful because we know little about the data and the clarification goal is vague.

C. HEALTH CARE BIG DATA APPLICATIONS

Recently, most of patients like to share their stories, experiences and knowledge through the networks, this obvious by observing social media, web blog or even in daily life. For future those materials should be utilized and directed on the right way[13]. This paper listed some of clinical platforms that used big data technology in one way or another, as shown below:

NextBio.com

NextBio.com is a big data technology consists of pharma and medical parts. NextBio has many packages of applications based on private and public experimental datasets. NextBio.com uses patients' genomic data to do medical decisions. These applications are developed to assist clinical and researches platforms. NextBio clinical platform enables experts to apply recent analysis techniques and methods to discover and translate clinical trials. On the other hand NextBio research platform provides tools to enhance pharma and medical parts by accessing and mining more than 10,000 clinical studies that help in analyze disease profiles, drug mechanisms discovery, optimize typical medical experiments and validate drug results.

Treato.com

Treato.com is an online health care blog that is generated by users. Treato.com portal consist of about 2-3 million of patients, advisers and caregiver around the world who share their knowledge, experiences and real stories about symptoms, medications and other conditions to provide real platform to patients and stakeholders that support decision making. Besides this blog has another section known as Treato IQ it is an intelligent platform from the users generated content; this platform helps stakeholders to get more understanding about the solutions and decisions about business targets. It's worth mentioning that treato.com manipulated by Hadoop.

Healthx.com

Healthx.com is a cloud-based health care IT services and health plan solutions which is looking for customer satisfaction and improving the business impacts, this is evident through its mission statement which is "Transforming healthcare through innovative digital engagement".

Short idea on Healthx: is an active member of Health Care Administrators Association (HCAA) and Self-Insurance Institute of America (SIIA). Healthx cover more than 200 million of American people diverse between health plans in private and general sectors to careers. Also Healthx provides a library of resources such as whitepapers, articles, case studies, infographics and practices.

Vitals.com

From its mission statement can understood its goal which is "find the best medical care with millions of doctor reviews". In other words, Vitals.com searches over millions of doctors around the world according to the branch of medicine they are belong to and access the reviews and find the best care according to different solutions. Those solutions stating consumers, healthcare plans, insurance plans, providers and advisements.

III. KNOWLEDGE DISCOVERY FRAMEWORK BASED ON BIG DATA ANALYSIS

Recently the usage of information technology in health care sector plays an important role to improve experts' decisions [5]. Delivering information through raw big medical data is a considerable challenge and a complicated process. Health care organizations with different denominations are looking for constructing big data platform to improve their performance and to enhance patients' safety, but it still a wish. Building big data analysis infrastructure will eliminate and reduce the manual processes and this will impact on costs, time and accuracy [1, 3, 7].

Big data analysis process to deploy knowledge will go through five stages to reach the goal. This paper presents a framework represented in Fig. 3, which generally explains how to lead raw data to discover knowledge. First step is data collection from EHRs on its private and public forms the next process is to create Clinical Datasets (CDSs) from the collected data. Clinical datasets is a warehouse that builds for the purpose of analysis, mining and reporting. Clean CDSs and extract the important data is the next step, Cleansing the raw big data its important stage by using data cleaning algorithms; in other words detect and remove errors or delete duplication or whatever process to improve the quality of data and to avoid processing time consuming. Clinical data warehouse (CDW) is a repository that is collected from multiple clinical software packages and it used to link, analyze, report and search within stored medical information. CDW aimed to provide health investigator by obtaining data from enormous clinical systems. Those Clinical Systems used in health and health related organizations.

Then this data have to be categorized according to stakeholder's perspectives, this process called information retrieval or data mining level. To get information cleansed big data can be categorized into statistical information or analytical information or tactical information or any different category that aid senior managers and decision maker to do exact decisions. The analysis procedures needs a health plans with a comprehensive view from patients, doctors, physicians, pharmacists, technicians and other intended people.

To implement big data analysis and knowledge deployment there are many techniques to produce predictive outcomes. The predictive model needs to be evaluated and validated, and then monitor the results after that the model can be deployed. From the IT expert perspective the data mining stage is more in data technologies and algorithms so it should to answer many questions and address many important factors. On the other hand from stakeholders' perspective they have to identify the business problems and differentiate data and business understandings. Knowledge discovery from the categorized information help experts on prediction, managing and guiding the extracted information.

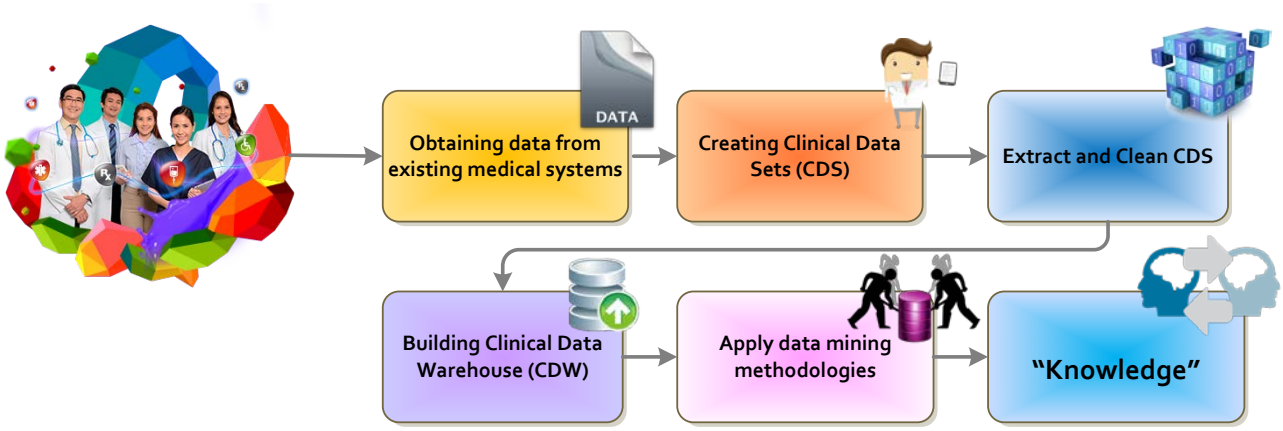

Figure 3: knowledge discovery through Clinical Big data analysis proposed framework

IV. CONCLUSION AND FUTURE WORKS

Lack of patients' records is the most important challenge that experts and researchers suffer from, which almost motivates healthcare organizations to build patients portals to help stakeholders in their business. Those portals are web-based health care information systems. Healthcare online systems could help to improve the communication between doctors and patients on the other hand to improve the quality of care which may lead to reduce medical errors and costs. In order to get better reviews patients in interest with share, save, manage, and retrieve their medical data, such as their medical history, medications, allergies, x-rays and test results. Accordingly building these online big repositories give them an opportunity to interact with doctors, physicians and pharmacists, but IT experts should take in mind patents' privacy and polices risk.

To be concluded that big data is about real time data. Healthcare management follow three main and important stages the first stage is data collection: this data may come from EHRs, hospitals or clinical billing systems, lab and imaging systems or any data attached with patients. The second stage is about cleaning and extracting useful data this more in data warehouse technology and the third is management and knowledge stage is data analysis using data mining techniques. From data scientists perspective in big data analysis target is to retain and to analyze more and more data with increasing the analysis speed, besides publishing accurate outcomes with low cost. Last to say that "Big Data" analysis in healthcare will provide more and more to the body of knowledge.

REFERENCES

[1] A. McAfee and E. Brynjolfsson, "Big data: the management revolution," *Harvard business review*, pp. 60-6, 68, 128, 2012.
[2] S. Sagiroglu and D. Sinanc, "Big data: A review," in *Collaboration Technologies and Systems (CTS), 2013 International Conference on*, 2013, pp. 42-47.
[3] M. Cottle, S. Kanwal, M. Kohn, T. Strome, and N. Treister, "Transforming health care through big data. Strategies for leveraging big data in the health care industry. New York: Institute for Health Technology Transformation," 2013.
[4] D. W. Bates and E. Zimlichman, "Finding patients before they crash: the next major opportunity to improve patient safety," *BMJ quality & safety*, pp. bmjqs-2014-003499, 2014.
[5] S. Mohanty, M. Jagadeesh, and H. Srivatsa, *Big Data Imperatives: Enterprise 'Big Data' Warehouse,'BI' Implementations and Analytics*: Apress, 2013.
[6] A. Rahimi, S.-T. Liaw, P. Ray, J. Taggart, and H. Yu, "Ontological specification of quality of chronic disease data in EHRs to support decision analytics: a realist review," *Decision Analytics*, vol. 1, pp. 1-31, 2014.
[7] M. De Mul, P. Alons, P. Van der Velde, I. Konings, J. Bakker, and J. Hazelzet, "Development of a clinical data warehouse from an intensive care clinical information system," *Computer methods and programs in biomedicine*, vol. 105, pp. 22-30, 2012.
[8] A. A. Aljumah, M. G. Ahamad, and M. K. Siddiqui, "Application of data mining: Diabetes health care in young and old patients," *Journal of King Saud University-Computer and Information Sciences*, vol. 25, pp. 127-136, 2013.
[9] P. Groves, B. Kayyali, D. Knott, and S. Van Kuiken, "The 'big data' revolution in healthcare," *McKinsey Quarterly*, 2013.
[10] J.-s. Li, H.-y. Yu, and X.-g. Zhang, *Data Mining in Hospital Information System*: INTECH Open Access Publisher, 2011.
[11] R. Steinbrook, "Personally controlled online health data-the next big thing in medical care?," *New England Journal of Medicine*, vol. 358, p. 1653, 2008.
[12] B. Kayyali, D. Knott, and S. Van Kuiken, "The big-data revolution in US health care: Accelerating value and innovation," *Mc Kinsey & Company*, 2013.
[13] J. Friedlin, M. Mahoui, J. Jones, and P. Jamieson, "Knowledge Discovery and Data Mining of Free Text Radiology Reports," in *Healthcare Informatics, Imaging and Systems Biology (HISB), 2011 First IEEE International Conference on*, 2011, pp. 89-96.

AUTHORS

First Author – Dalia AbdulHdi AbdulAmeer, MSc. in Information Technology, IT Dept., University Utara Malaysia, DaliaAl_ubaidi@uoitc.edu.iq

Correspondence Author – Dalia AbdulHdi AbdulAmeer, DaliaAl_ubaidi@uoitc.edu.iq, +9647901875292.