Potential of Big Data Analytics in Bio-Medical and Health Care Arena: An Exploratory Study

By Mandava Geetha Bhargava, Modugula T S Srinivasa Reddy, Shaik Shahbaz, P. Venkateswara Rao & V.Sucharita

KL University

Abstract- With the leveraging emerging big Data in every industry, Big Data can amalgamate all data related to patient to get a complete view of patient to analyze and predict the outcomes. Using big data analytics as tools. It can enhance development in new drugs, health care financing process and clinical approaches and extends a lots of benefits such as better health care quality and efficiency, fraud detection and early disease detection by means of analytics of big data. This paper provides a general survey of current progress and advances in research arena of big data, bio-medical and health care and some major challenges of big data concept and characteristics, this concerns includes big data from bio-medical and health care arena, benefits of big data, its applications and opportunities, Methods and technology progress about big data in bio-medical and health care and challenges of big data in both bio-medical and healthcare are also discussed.

Keywords: big data; health care; genomics; big data analytics.

GJCST-C Classification: J.3

Strictly as per the compliance and regulations of:

Potential of Big Data Analytics in Bio-Medical and Health Care Arena: An Exploratory Study

Mandava Geetha Bhargava [α], Modugula T S Srinivasa Reddy [σ], Shaik Shahbaz [ρ] , P. Venkateswara rao [ω] & V.Sucharita[¥]

Abstract- With the leveraging emerging big Data in every industry, Big Data can amalgamate all data related to patient to get a complete view of patient to analyze and predict the outcomes. Using big data analytics as tools. It can enhance development in new drugs, health care financing process and clinical approaches and extends a lots of benefits such as better health care quality and efficiency, fraud detection and early disease detection by means of analytics of big data. This paper provides a general survey of current progress and advances in research arena of big data, bio-medical and health care and some major challenges of big data concept and characteristics, this concerns includes big data from bio-medical and health care arena, benefits of big data, its applications and opportunities, Methods and technology progress about big data in bio-medical and health care and challenges of big data in both bio-medical and healthcare are also discussed.

Keywords: big data; health care; genomics; big data analytics.

I. Introduction

Big Data can be termed as massive data or complex that exceeds the processing capacity of traditional data processing applications and challenges are Acquire, Process, Manage, Generate, Capture, storage, sharing and visualize. Now a days Big Data is processed for analytics of various parameters in each and every field of work like Research, Education, I.T, Banking, Bio-Medical, Health Care, Construction, and Manufacturing etc. With help of some big data technologies it is been processed and characteristics of big data are 6V's i.e. Volume, Variety, Velocity, Value, Veracity, Variability as shown in fig.1.

Volume: This represents the amount of data size where it is shown through Zettabytes (Approximately 10^{15} Megabytes), Petabytes (Approximately 10^9 Megabytes), and Terabytes (Approximately 10^6 Megabytes), etc.

Velocity: This represents the generation of data with respect to time, in other term data generated with speed*.*

Variety: This represents the different types of data such as unstructured data from email message, net clicks, social media streamed videos and audios and structured data from relational tables and semi-structured data from key-value web clicks, etc.

Value: Value can be defined as the collected data can bring added value. It refers to the data having value to create knowledge i.e. there is some valuable information within the collected data.

Variability: this represents the changes of data during processing and throughout lifecycle. Increase in variety and variability characterized data also increase attractiveness of data.

Veracity: this represents the two aspects of data as follows: data consistency and data trustworthiness. Sometimes the data can be in doubt due to incompleteness, ambiguities, uncertainty and deception, etc.

Big data in bio-medical and health care arena mostly denoted as electronic health data where it is large size and complex in nature to manage and process with traditional processing applications. Some data are distinguished by a means of rightness and timeliness for example: the data can generated from

Author α, σ: M.Tech Scholars of Computer Science and Engineering Department, KL University, Vaddeswaram, Guntur Dst 522502, Andhra Pradesh, India. e-mail: m.gbhargava@gmail.com

Author ρ: B.Tech Scholars of Computer Science and Engineering Department, KL University, Vaddeswaram, Guntur Dst-522502, Andhra Pradesh, India.

Author ω: Professor, Dept of CSE NEC, Gudur.
e-mail: jesuchi78@yahoo.com

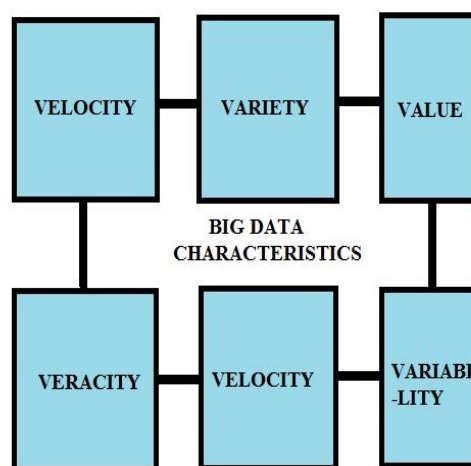*Author ¥: Professor, Dept of CSE, NEC, Gudur.*
e-maii: jesuchi78@yahoo.com

wearable or implantable sensors like Bio-Metric, Blood Pressure and ECG etc. is gathered and analyzed in real time. The data in bio-medical and health care arena can be differentiate as follows: genomic data where it consists of DNA expression, genotyping and gene expression etc., clinical data where it consists of structured and unstructured data such as X-ray images, test reports, patient discharge reports and laboratory data etc., behavior and patient sentiment data consists of web data, social media data, streamed data such as telehealth and endoscopy etc., clinical reference and health publication data consists of practice guidelines, journals , medical reference material and health products etc., business, administrative and external data consists of financial data ,scheduling, billing and biometric data and other important data such as patient feedback and device data etc. as shown in figure.2.
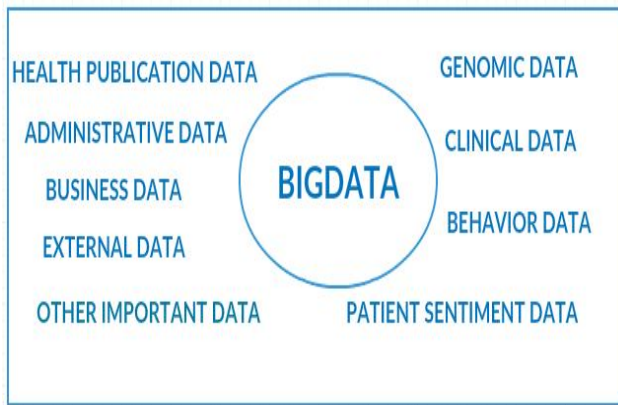

Fig.2: Different data in Bio-Medical and Health care Arena

II. Opportunities and Applications of Big Data in Biomedical and Health Care

Big data can avail support over all aspects of biomedical and health care. Big data analytics has gained traction in analytics of fraud detection and prevention, clinical outcome, genomics, epidemic disease prediction, pharmaceutical development and personalized patient care, etc. there are potential applications in biomedical and health care as they are discussed below:

a) Fraud detection and prevention analytics
Detecting, predicting and reducing fraud can be executed by using advanced analytics systems for fraud detection and checking the consistency and accuracy of claims. Big data predictive modelling can be used by health care users and payers for fraud prevention.

b) Clinical outcome Analytics
Clinical analysis can be implemented through merging financial, operational and clinical data for efficient clinical assessments. Clinical data can be utilize to reduce the manage and predict the health risks and improve clinical outcomes with cost of care.

c) Genomics analytics
The data about genes and DNAs can be analyzed for predicting and reducing the rate of disease of patients and it is becoming critical to the complete patient record to merge the both genomics data and clinical data helps to cure perilous diseases such as cancer, etc.

d) Epidemic disease prediction
In public and population health, continuously analyzing and aggregating public health data helps identifying and managing potential disease out breaks by means of analytics of social media and web-based data the disease outbreak can be known based on social content ,query activity and consumer search.

e) Design and manufacturing of medical devices
Tools of big data allows a broader set of device materials, tissue interactions, delivery methods, and anatomical configurations to be analyzed. Big data and computational methods can play an important role in design and manufacturing of medical devices.

f) Pharmaceuticals development
By analytics of pharmaceuticals data, the pharmaceutical companies can increase their ability to continue bringing new life enhancing medicines to patients in a timely manner, on basis of management of big data which was generated during all phases of pharmaceutical development, the cost of pharmaceutical product will be cost effective.

III. Research Background

Kiyana zolfaghar et al [1] done research on solutions for predicting risk of readmission for congestive heart failure patients by means integrating data of national impatient dataset and patient dataset and developed a datamining predictive model by means of integrated data and concluded that effectiveness of quality, scalability, efficiency by means of big data infrastructure on the predictive model.

Sean D.Young et al [2] done research on approaches of utilising real time social media technologies for identification and remote monitoring of HIV outcomes through negative binomial regression of tweets and concluded that the feasibility of using real time social media data to detect HIV risk-related communications, geographically map the location of those conversations and link them to national HIV outcome data for additional analysis

Priya Nambisan et al [3] done research on ruminating behavior of depression through social media, big data and public health informatics through tweets from micro blogging sites by means of screening the

vocabulary of tweets and shows sleep, pain and suicidal thoughts as they do offline and concluded that the characteristic can be used to detect and diagnose depression using the tweets in a much more effective and efficient way.

Zhendong Ji [4] done study on analysis of big data application in the medical industry and potential of its commercial value for the health care industry and concluded that by big data analysis in the medical industry provides future and promoting continuous development of medical industry through meta-analysis of gathered data.

Quan Zou et al [5] done a study on map reduce frame operation in bioinformatics through different applications and mechanisms of MapReduce and concluded that Hadoop framework has capable of handling bioinformatics data and traditional bioinformatics resources will be redesigned to support Hadoop MapReduce for high performance computing.

Liang y et al [6] studied on big data science and its applications in health care and medical research and concluded that big data offers new opportunities and promising with challenge in every field .the collaborative network, nurturing environments, team science approach with highly trained with computational skills, domain/disease expert and interdisciplinary are crucial.

Lidong Wang et al [7] done a study on big data in medical applications and health care by means of big data concept and characteristics, health care data and major issues of big data and concluded that big data can improve the research and development, translation of new therapies and has great potential to improve medicine, guide clinicians in delivering value based care.

IV. Methods and Technology Progress in Big Data

In health care /Bio-medical arena, massive amount of data about patient's medical histories, diagnosis and responses, symptomatology to treatments and therapies are gathered. Data mining techniques can implemented to derive knowledge from the gathered data in order to either examine reporting practices or to identify new patterns in infection control data. Moreover, predictive models can be utilized as detection tools can be utilize as electronic patient record gathered for every individual person of the area.

For big data healthcare/biomedical systems, the combination of Hadoop-MapReduce framework and R language is uniquely capable of storing and analyzing wide range of healthcare data types including genomic data, financial data, electronic medical records and claims data etc. the combination of two frameworks offers availability, high scalability, statistical analysis and reliability than traditional data processing systems. In

addition, intelligent functional components can be built such as surveillance, detection, notification, diagnosis and recognition etc. figure 3 shows a general framework of big data analytics.

Visual analytics presents a new area of research with big data by conceptualizing the output of complex processes. The appropriate visualization solutions to the big data examples such as real time interactive visualization and metrics dash boarding [9]. Unstructured data can be converted into form of tables is to put attributes of exchangeable image file (EXIF) tags or place analyzed data where it leads to easier at visualization process. Big data can be processed through cloud technologies where it provide us operationally, insights-clinically and in research [10]. The concept of STAAS (storage as a service), is a one of the services provided by cloud computing, which provides health care center with a massive amount of storage for processing on basis of demand at low cost. [11]. Beside general infrastructure of cloud (compute, storage, virtual machine management), the following services are required to handle big data [12]. Hadoop related frameworks and tools, specialist data analytics tools, Cluster services, massively parallel processing databases, databases /servers SQL, NoSQL and security infrastructure. Organizations used various methods to de-identification of the distance data from personal identities and preserve individual's privacy. De-identification has been seen as an important security measure to be taken under the data security and accountability principle [13].

V. Proposed Approach and System

From studies of literature, the mechanisms and methodologies are basis upon Hadoop-MapReduce Framework in above literatures where it does only analysis and processing, having disadvantage of statistics analysis where it can't does and can't stores the data. Authors proposing a system which consists of RHadoop platform, it contains both R Language and Hadoop-MapReduce framework where it can process the different types of data with statistical data and streams the data after and before analysis through separate statistical package. It can visualize the processed data as output and figure 4 shows the architecture of Hadoop and data analysis tool.

RHadoop is an open source project developed by revolution analytics, provides client side integration of Hadoop and R. It allows running a MapReduce jobs within R. Needs some packages for integration of R and
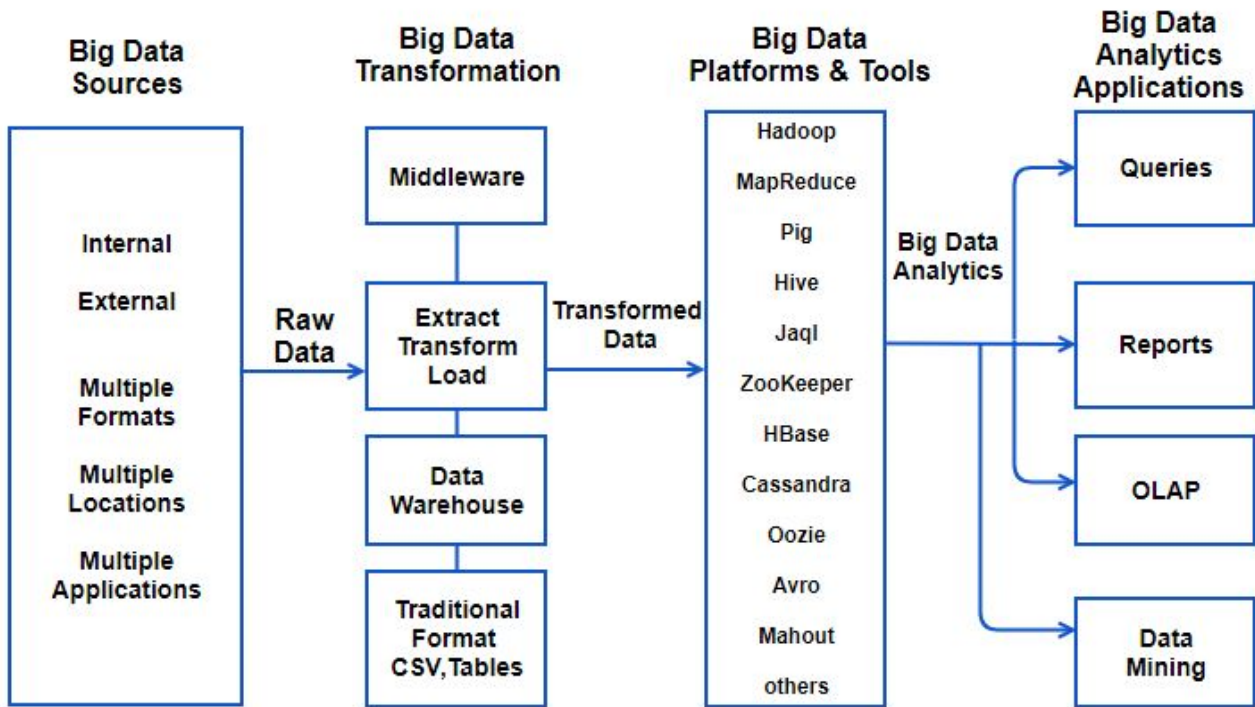
27

28

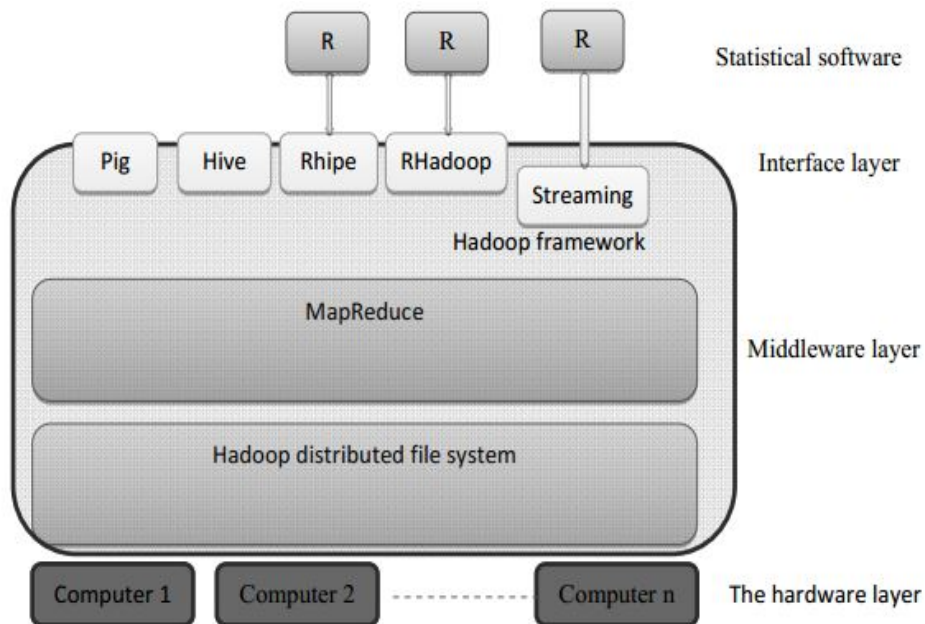*Fig.3:* General Framework of Big Data Analytics [8].

Fig.4: Architecture of Hadoop and Data Analysis

Hadoop such as plyrmr, rmr, rdfs, rhbase, etc. while setting up the RHadoop, it won't be complicated task because of dependencies on other R package.it has a wrapper R script called from streaming that invokes user defined mapper and reducer functions. Script that uses RHadoop as shown below.

1. library(rmr)
2. map<-function (k, v) {…}
3. reduce<-function (k,vv) {….}
4. mapreduce (input ="ffg.txt", output="output", textinputformat=rawtextinputformat,
   map =map, reduce=reduce)

firstly, the library called rmr is loaded into the memory as shown in line 1 and then it follows to definition of the map function which receives a key and value pair as input as shown in line 2. The reduce function in line 3 is called with a key and a list of values as arguments for each unique map key. Finally the script setup and runs the map reduce job as shown in line 4 in above script.

There are several layers in architecture of Hadoop and data analysis as shown in figure.4 as follows:

a) Hardware Layer
b) Middleware Layer
c) Interface Layer
d) Statistical software

a) Hardware Layer

In this layer, the components which are in this layer acts as multiple nodes on the network for analysing, processing, storing and related operations on the data by means of Hadoop and map reduce framework with Hadoop Distributed File System (HDFS) figure.5 depicts about the overview of Hadoop
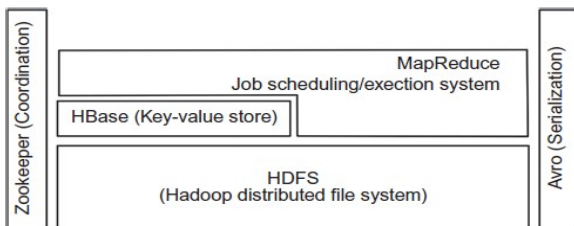

Fig.5: Overview of Hadoop

b) Middleware Layer

This layer consists of bunch of Hadoop and map reduce framework with Hadoop Distributed File System for sharing data from one node to different nodes, for processing and statistically analysis of data. Firstly the HDFS client interacts with namenode A for metadata of data stored in some datanode which are allocated to namenode A then the data can be operated by means of read or write as shown in figure.5
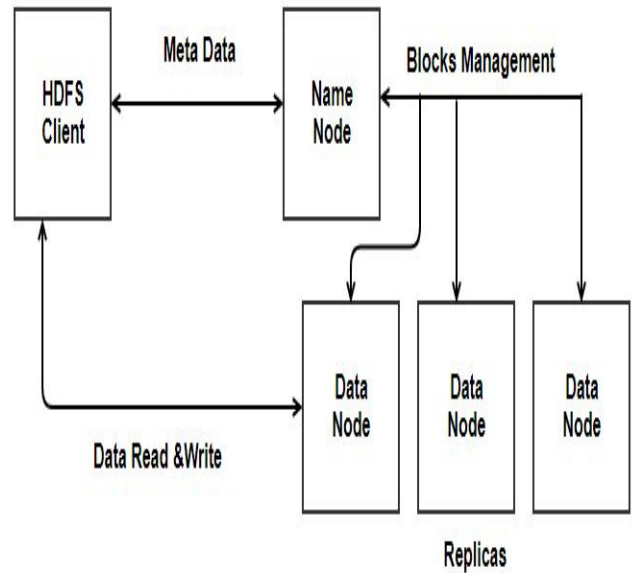

Fig.6: Architecture of Hadoop Distribution File System

c) Interface Layer

In this layer, the components consists of connectors and interfaces between Statistical analysis software and Hadoop-MapReduce framework. In this phase the data can be streamed to statistical software for analyzing and visualizing by means of direct and indirect streaming options where the direct streaming option leads to stream the data through connectors to statistical software and indirect streaming option leads to stream the data through Hadoop frameworks such as Hive, Hbase, Pig and etc. for storing and upgrading of data purpose

d) Statistical Software

The statistical software which handles data to analyze statistically through by means of visualization of graph. The software's such as SAS, R and etc.

VI. Challenges of Big Data in Biomedical and Health Care

Variety, volume and velocity characteristics of big data have brought challenges in retrieval, data storage ,visualization and search .veracity and variability of big data indicate data uncertainty and instability, which often makes big data analytics difficult and major challenges of big data in bio-medical and health care are as follows:

1) It is difficult to analyse and aggregate unstructured data such as test results, scanned documents, visual data and progress notes in patient electronic health record, etc.
2) The data in many health care providers are often segmented. Clinical data such as patient electronic

health record consists of test results, images and progress notes. Quality and outcomes data such as patient's falls, surgical site infections, etc. are in risk or quality management department where it needs standards for validating, consolidating and processing data are needed.

3) Privacy issues in the patients data such as health records , insurance details,etc. even if the privacy of the patient is protected , many health care providers are unwilling to share data due to market competition

4) Collected data can be damaged or leaked through hackers.

VII. Discussion and Conclusion

Big data is based on data generated from whole process of diagnosis and response of each case. It can lead to develop predictive models to determine which patients and health care users are mostly benefit from care management plan. By means of data analytics, it offers disease prevention, reduce in medical errors and better outcomes. it can improves and develops new therapies and research and development and has great potential to improve guide clinicians and medicines in delivering value based care. Big data has challenges in arena of bio-medical and health care to overcome such as information security, lack of infrastructure, data privacy and leakage, etc.

Utmost of all challenges can be scope for future research topics, the following topics may have a chance of future research: medical data confidentiality and interoperability, indexing and processing of continuous data, analyzing and aggregating of unstructured biomedical and health care data, security of health care data, etc. this paper focused on study on potential of big data in biomedical and health care arena and their application, challenges and opportunities.

References Références Referencias

1. Kiyana Zolfagha, Naren Meadem, Ankur Teredesai, Senjuti Basu Roy, Brian Muckian(2013), Big Data Solutions for Predicting Risk of Readmission for Congestive Heart Failure Patients, IEEE International Conference of big data,64-71

2. Sean D Young, Caitlin Rivers, Bryan Lewis (2014), methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes, preventive medicine volume 63, 112-115.

3. Priya Nambisan, Zhihui Luo, Akshat Kapoor, Timothy B Patrick, Ron A Cisler, (2015) Social media, big data and public health informatics: Ruminating behaviour of depression revealed through twitter, 48th Hawaii International Conference on system sciences, 2906-2913.

4. Zhendong ji, (2015) Applications analysis of big data analysis in the medical industry, international journal of database theory and application, vol.8, No.4, 107-116.

5. Quan Zou, Xu-Bin Li, Wen-Rui Jiang, Zi-Yu Lin, Gui-Lin Li and Ke Chen (2013), Survey of Mapreduce frame operation in bioinformatics, briefings in bioinformatics, Vol.15, No.4, 637-647.

6. Liang Y and Kelemen A (2016), Big Data Science and its Applications in Healthcare and Medical Research: Challenges and Opportunities .Austin Biom and Biostat, 3 (1).

7. Lidong Wang and Cheryl Ann Alexander (2015). Big Data in Medical Applications and Health Care, American Medical Journal, 6 (1).

8. Raghupathi, W. and V. Raghupathi, (2014). Big data analytics in healthcare: Promise and potential. Health Inform. Science Syst., 2: 1-10. DOI: 10.1186/2047-2501-2-3.

9. Schultz, T., (2013). Turning healthcare challenges into big data opportunities: A use-case review across the pharmaceutical development lifecycle. Bull. Association Inform. Sci. Technol., 39: 34-40. DOI:10.1002/bult.2013.1720390508.

10. Shrestha, R.B., (2014). Big data and cloud computing. Applied Radiology.

11. Hsieh, J.C., A.H. Li and C.C. Yang, (2013). Mobile, cloud and big data computing: Contributions, challenges and new directions in tele cardiology. Int. J. Environ.Res. Public Health, 10: 6131-6153.DOI: 10.3390/I jerph10116131.

12. Turk, M., (2012). A chart of the big data ecosystem, take 2.

13. Tene, O. and J. Polonetsky, (2013). Big data for all: Privacy and user control in the age of analytics. Northwestern J. Intellectual Property, 11: 239-272.