Big Data Knowledge System in Healthcare

Gunasekaran Manogaran, Chandu Thota, Daphne Lopez, V. Vijayakumar, Kaja M. Abbas and Revathi Sundarsekar

Abstract The health care systems are rapidly adopting large amounts of data, driven by record keeping, compliance and regulatory requirements, and patient care. The advances in healthcare system will rapidly enlarge the size of the health records that are accessible electronically. Concurrently, fast progress has been made in clinical analytics. For example, new techniques for analyzing large size of data and gleaning new business insights from that analysis is part of what is known as big data. Big data also hold the promise of supporting a wide range of medical and healthcare functions, including among others disease surveillance, clinical decision support and population health management. Hence, effective big data based knowledge management system is needed for monitoring of patients and identify the clinical decisions to the doctor. The chapter proposes a big data based knowledge management system to develop the clinical decisions. The proposed knowledge system is developed based on variety of databases such as Electronic Health Record (EHR), Medical Imaging Data, Unstructured Clinical Notes and Genetic Data. The proposed methodology asynchronously communicates with different data sources and produces many alternative decisions to the doctor.

G. Manogaran (✉) · D. Lopez
VIT University, School of Information Technology and Engineering,
Vellore, Tamil Nadu, India
e-mail: gunavit@gmail.com

C. Thota
Albert Einstein Lab, Infosys Ltd, Hyderabad, India

V. Vijayakumar
VIT University, School of Computing Science and Engineering, Chennai,
Tamil Nadu, India

K.M. Abbas
Department of Population Health Sciences, Virginia Tech, Blacksburg, USA

R. Sundarsekar
Priyadarshini Engineering College, Vellore, Tamil Nadu, India

C. Bhatt et al. (eds.), *Internet of Things and Big Data Technologies for Next Generation Healthcare*, Studies in Big Data 23,
DOI 10.1007/978-3-319-49736-5_7

Keywords Big data · Knowledge system · Health care system · Electronic Health Record (EHR) · Medical Imaging Data · Unstructured Clinical Notes · Genetic Data

1 Introduction

"Big Data" initially meant the volume, velocity and variety of data that becomes tricky to analyze by using conventional data processing platforms and techniques [1]. Nowadays, data production sources are improved rapidly, such as telescopes, sensor networks, high throughput instruments, streaming machines and these environments generate massive amount of data. Nowadays, big data has been playing a crucial role in a variety of environments such as healthcare, business organization, industry, scientific research, natural resource management, social networking and public administration. Big data can be categorized by 10V's as follows (Fig. 1). *Volume*: The big volume indeed represents Big Data. Recently, the data generation sources are augmented and it causes diversity of data such as text, video, audio and large size images. In order to process the enormous amount of data, our conventional data processing platforms and techniques has to be enhanced [2]. *Velocity*: The rate of the incoming data has increased dramatically this velocity indeed represents Big Data. The phrase velocity represents the data generation speed. The data explosion of the social media has changed and causes variety in data. Nowadays, people are not concerned in old post (a tweet, status updates etc.) and notice to most hot updates [2]. *Variety*: The variety of the data indeed represents Big Data. Nowadays, the collection of data types is also increased. For example, most organizations use the following type of data formats such as database, excel, CSV, which can be stored in a plain text file. Nevertheless, sometimes the data may not be in the anticipated format and it causes difficulties to process. In order to defeat this issue the organization has to be identified the data storage system which can analyze variety of data [2]. *Value*: The value of data indeed represents Big Data. Having continuous amounts of data is not helpful until it can be turned into value. It is essential to understand that does not always mean there is value in Big Data. The benefits and costs of analyzing and collecting the big data is more impartment thing when doing big data analytics. *Veracity*: This veracity of data indeed represents Big Data. Veracity represents the data understandability; it doesn't represent data quality. It is significant that the association should perform data processing to prevent 'dirty data' from accumulating in the systems. *Validity*: It is essential to ensure whether the data is precise and accurate for the future use. In order to take the right decisions in future the organizations should valid the data noticeably. *Variability*: Variablity refers to the data consistent and data value. *Viscosity*: Viscosity is an element of Velocity and it represents the latency or lag time in data transmit between the source and destination. *Virality*: Virality represents the speed of the data send and receives from various sources. *Visualization*: Visualization is used symbolize the Big Data in a complete view and determine the

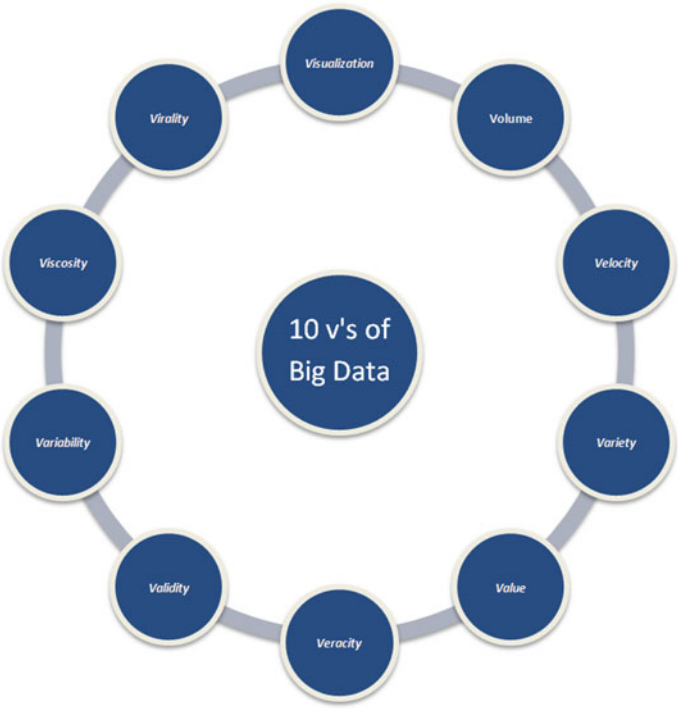

Fig. 1 10V's of Big Data

hidden values. Visualization is an essential key to making big data an integral part of decision making.

Big Data also impact more in healthcare. Nowadays, health care systems are rapidly adopting clinical data, which will rapidly enlarge the size of the health records that are accessible, electronically [3, 4]. A recent study expounds, six use cases of big data to decrease the cost of patients, triage and readmissions [5]. In yet another study, big data use cases in healthcare have been divided into number of categories such as clinical decision support, administration and delivery, user behavior, and maintain services. Jee et al. described that how to reform the healthcare system based on big data analytics to choose appropriate treatment path, improvement of healthcare systems, and so on [6]. The above use cases have utilized the following big data in health care implementation. (1) Patient-centered framework produced based on the big data framework to approximate the amount of healthcare (cost), patient impact (outcomes), and dropping re-admission rates [7]. (2) Virtual physiological human analysis combined with big data analytics to create robust and valuable solutions in silico medicine [8].

Table 1 Comparison of various databases for Big Data

Name	Spark SQL	HBase	Hive
Description	Spark SQL is a component on top of 'Spark Core' for structured data processing	Apache HBase is a scalable and distributed database is used to store the data on top of the HDFS	Apache Hive is a SQL interface and relational model for querying, analyzing and summarizing large size of datasets stored in HDFS
Database model	Relational DBMS	Wide column store	RDBMS
Technical documentation	spark.apache.org/-docs/-latest/-sql-programming-guide.html	hbase.apache.org	cwiki.apache.org/-confluence/-display/-Hive/Home
Developer	Apache software foundation	Apache software foundation	Apache software foundation
Initial release	2014	2008	2012
Current release	v2.0.0, July 2016	1.2.2, July 2016	2.0.0, February 2016
MapReduce	N/A	Yes	Yes
Database as a Service (DBaaS)	No	No	No
Implementation language	Scala, Java, Python, R	Java	Java
Supported programming languages	Java, Python, R, Scala	C, C#, C++, Groovy, Java, PHP, Python, Scala	C++, Java, PHP, Python
Data scheme	Yes	Schema-free	Yes
Typing	Yes	No	Yes
XML support	No	No	N/A
Secondary indexes	No	No	Yes
SQL	No	No	No
APIs and other access methods	JDBC, ODBC	Java API RESTful HTTP API Thrift	JDBC, ODBC, Thrift
Server-side scripts	No	Yes	Yes
Triggers	No	Yes	No
Transaction concepts, concurrency	No	No	No
Concurrency and durability	Yes	Yes	Yes
In-memory capabilities	No	No	N/A

Table 2 Comparison of various platforms for Big Data

Name	MapReduce	Strom	Spark streaming
Description	Hadoop MapReduce is a type of programming model used for processing huge size of data sets across a Hadoop cluster	Storm on YARN is powerful for scenarios requiring real-time analytics, machine learning and continuous monitoring of operations	Apache Spark follows in-memory database so that it can generate one hundred times quicker output for user queries on stream of data
Website	https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html	strom.apache.org	spark.apache.org/streaming/
Developer	Apache software foundation	Apache software foundation	Apache software foundation
Execution model	Batch	Streaming	Batch, streaming
Supported language	Java	Any language	Java, Python, R, Scala
Associated ML tools	Mahout	SAMOA	MLlib, Mahout, H2O
In-memory capabilities	No	Yes	Yes
Low latency	No	Yes	Yes
Fault tolerance	Yes	Yes	Yes
Enterprise support	No	No	Yes

2 Overview of Big Data Tools and Technologies

This section describes various tools and technologies for big data. The comparison of various databases and various platforms for big data are depicted in Tables 1 and 2 respectively.

2.1 Hadoop Architecture

Apache Hadoop consists of master/slave architecture that uses namenode and datanode to process the huge data. The namenode performs as the master and datanodes act as slaves. The namenode manages the access of all datanodes. The main accountability of datanodes is to administrate and store the huge data across multiple nodes. User configures the number of block replication in Hadoop architecture (Fig. 2).

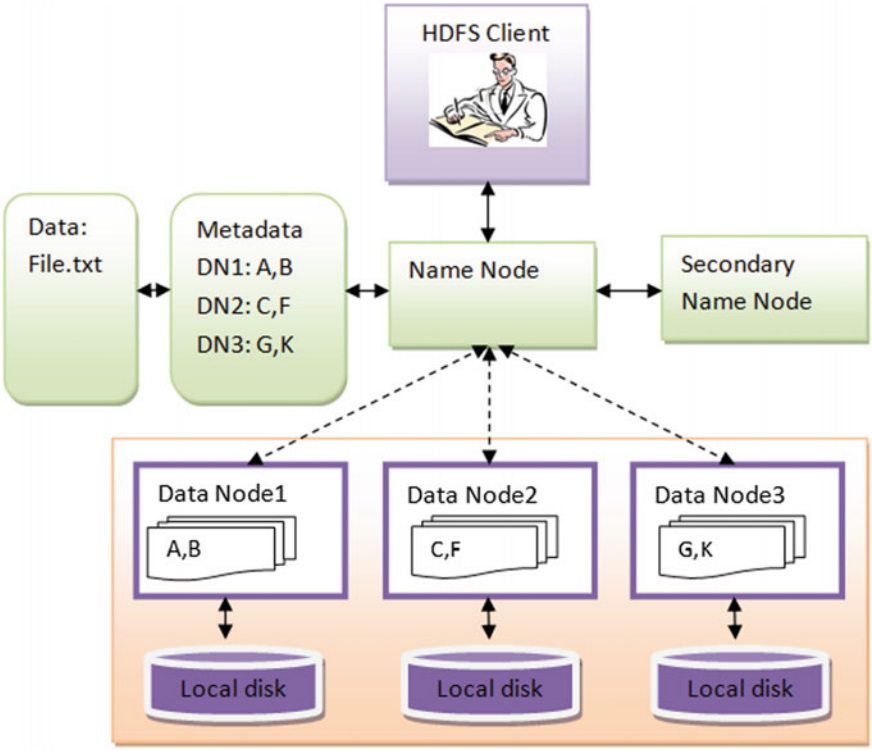

Fig. 2 HDFS architecture

2.2 *Hadoop Components*

Hadoop Distributed File System (HDFS): The HDFS is initially intended to process on cluster of nodes. HDFS stores data in distributed manner and used for many applications those have large data. The typical HDFS architecture is depicted in Fig. 1.

HDFS architecture does the following tasks:

- Distributed File System (DFS) always makes getting new data as simple as adding a new file to the folder, which contains the master dataset
- Distributed file system distributes the huge size of data across a cluster of commodity hardware. As more number of computers is added, the storage space and I/O throughput increase
- Distributed file system uses MapReduce framework to process the huge data in parallel manner
- Distributed file system restricts the users to delete or modify files in the master dataset folder. This feature protects the master data against human mistakes or bugs.

Namenodes: The namenode is always serves as the master server and it does the following tasks:

- Controls the file system namespace
- Periodically stores information about the metadata of the data blocks
- Data blocks' location is stored on the data node
- Name also perform following functions such as opening files and directories, renaming and closing, Once the namenode of the system crashes, then the entire Hadoop system goes down.

Datanode: In general, every node in the Hadoop cluster maintains at least one datanode. These nodes focus on managing the data storage of their system and are accountable of the following tasks:

- Performs write/read functions on the Hadoop file systems based on the client's request
- According to the instructions of the namenode it perform operations such as block creation, deletion, and replication
- Periodically send the blocks information to the namenode.

Secondary Namenode: This node is used to make a copy of name node. In other words, this makes a secondary copy of namenode.

JobTracker: This node used to track all the data nodes. It includes scheduling, monitoring of all task.

TaskTracker: TaskTracker is always runs on the datanodes of the hadoop cluster to run map task and reduce task. This node does the following tasks:

- Performs write/read functions on the Hadoop file systems based on the client's request
- According to the instructions of the namenode it perform operations such as block creation, deletion, and replication
- Periodically send the blocks information to the namenode.

2.3 Hadoop MapReduce

Hadoop MapReduce is a type of programming model used for processing huge size of data sets across a Hadoop cluster. Hadoop framework also provides the scheduling, distribution, and parallelization services to process the big data. Hadoop MapReduce programming consists of following features:

- MapReduce programming languages C++, Java or Python can be chosen by programmers developers
- MapReduce programming model is an ability to process petabytes of data, stored in Hadoop cluster

- MapReduce Parallel processing is an ability to process the huge size of data in minutes
- MapReduce manages node failure on its own, hence, if any one machine fails, an additional machine is take care of the node failure
- MapReduce model is also used to increase the processing speed and reduce the network I/O patterns.

2.4 Apache Sqoop

Apache Sqoop can extract data from Hadoop Distributed File System (HDFS) and export it into external structured data stores (relational databases). Apache Sqoop consists of the following functions to incorporate bulk data transfer between Hadoop and relational databases:

- Performs transformation of huge data between Hadoop Distributed File System and relational databases
- It consists of improved compression and light-weight indexing technique for efficient query performance
- Used to transfer data from EDWs and external storage into Hadoop file system for cost-effectiveness of combined data processing and storage
- Faster performance and better resource utilization
- Transferring huge data from external storage into Hadoop system
- Schema-on-read data lake is used in the Sqoop to merge structured data with unstructured data, so that effectiveness of the data analysis is enhanced
- Provide excessive storage to other systems and load processing.

2.5 Apache Flume

Apache Flume is used for transferring batch files, log files and high-volume streaming data into HDFS for storage. Flume consists of the following functions:

- Enables stream data from numerous sources into Hadoop system for analysis and storage
- It follows channel-based transactions to assure reliable data delivery. For example, when a message is transferred from one machine to another, two transactions are happening concurrently, one is represented on the destination side and the other one is on the source side
- It follows horizontal scaling to consume most recent data streams and additional storage.

2.6 Apache Pig

Apache Pig maintains the generation of batch views. This query approach consists of numerous functions together in a single pipeline; so it decreases the number of data scanning. Apache Pig also supports the traditional data functions like joins, filters, ordering, etc. and nested data types like tuples, maps, and bags on structured, semi-structured, or unstructured data. In general, Apache Pig often used while joining new incremental data with the previous data results.

2.7 Apache Hive

Apache Hive is a SQL interface and relational model for querying, analyzing and summarizing large size of datasets stored in HDFS. HiveQL is a type of query language for hive, which converts normal SQL-like queries into MapReduce jobs executed on Hadoop Distributed File System (HDFS).

2.8 Cloudera Impala

Cloudera Impala is used to provide fast response to the user queries, instead of long batch jobs previously related to SQL-on-Hadoop methodologies. Impala has incorporation with Apache Hive metastore database so that user can distribute the databases and tables between both components.

2.9 Apache Mahout

Apache Mahout is used to provide more accurate result for the user queries. In general, machine learning is an artificial intelligence that allows computers to learn based on data alone; it provides better performance as more data is analyzed. It provides several scalable data mining techniques such as clustering, classification, filtering, dimensionality reduction, pattern mining and so on.

2.10 Apache Hadoop Yarn

Apache Hadoop Yarn is used to distribute the big data analytics jobs by Map Reduce and HDFS. YARN consists of following features for Hadoop framework

such as security, resource management and data governance tools. As its architectural center, YARN improves Hadoop compute cluster in the following ways:

- It provides open-source or proprietary tools to use Hadoop system for real time and batch processing
- Apache YARN follows dynamic allocation of system resources that increases resource utilization compared to static MapReduce model rules used in previous Hadoop versions
- Apache YARN is capable handling petabytes of data across hundreds of nodes in the Hadoop cluster
- Apache YARN also process existing MapReduce applications without any disruption.

2.11 Apache Parquest

Apache Parquest supports master data management when user needs columnar storage. In addition, Apache Parquest doesn't store the complete data into the memory; as an alternative it stores those data which are actually required, thus dropping the required space in the memory as well as raising the speed.

2.12 Apache Spark Streaming

Apache Spark is fundamentally works based on cluster computing framework. Unlike Hadoop's MapReduce paradigm, Apache Spark follows in-memory database so that it can generate one hundred times quicker output for user queries on stream of data. Spark streaming has been provided as a part of Spark, which finds its application in real-time for example to monitor, control the access of end users on a website and fraud detection in real time.

2.13 Apache HBase

Apache HBase is a scalable and distributed database is used to store the data on top of the HDFS. It has been identified after the Google's BigTable was developed and can store millions of rows and columns. In view of the fact that HBase maintains Master/Slave architecture, it is extremely accessible to all nodes in the cluster.

3 Proposed Big Data Knowledge System in Healthcare

3.1 Role of Knowledge System in Healthcare

Knowledge is the combination of information, data, and experience. It is developed based on the trainings, analysis and various work experience. This knowledge is used to develop decisions at emergency situations and complex problems. Nowadays, knowledge developed from various experience are often used in critical healthcare problems and disease diagnosis. In addition, clinical management, surgical environment and drug recommendation are also used knowledge system to get desired output. In addition, more knowledge is developed from past issues and mistakes. In general, knowledge is classified into two type's namely tacit and explicit knowledge based on the generation sources. Explicit knowledge is easy to collect, format, and distribute with various persons. Hospital and medical procedures and disease diagnosis are considered as some of the example for explicit knowledge [9]. Alternatively, tacit knowledge is developed based on the individuals' experience [10]. Because of the difficulty, subjectivity, and objectivity, tacit knowledge is very complex to collect, format, and distribute to other individuals.

3.2 Types of Knowledge in Healthcare

Knowledge can be further classified in the three types [11].

Provider Knowledge: Medical experts have both tacit and explicit knowledge. In general, every doctor is required to identify typical medical diagnosis or details from various available sources. Years of experience in medical diagnosis is used to take better decisions.

Patient Knowledge: Tacit knowledge is developed from the patients and it is considered "health status". Generally speaking, practitioners and doctors may not know about the current and past medical conditions of the patients.

Organizational Knowledge: Organizational Knowledge is also a vital role in patient treatments and diagnoses for preventative maintenance and illnesses. Most medical organizations have other familiar resources that are available for doctors and patients to contact. Organizational Knowledge is developed from text-based materials, medical diagnostic systems, and other sources.

3.3 Sources of Big Data in Healthcare

The most familiar big data sources in medical environments include Electronic Health Record (EHR), Medical Imaging Data, Unstructured Clinical Notes and Genetic Data.

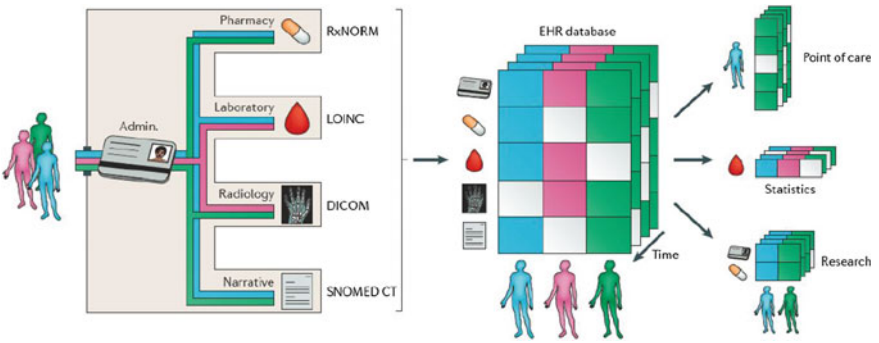

Fig. 3 Various types of Electronic Health Records (EHRs)

Electronic Health Records (EHRs): The following information is generally available in all EHRs are: laboratory results, billing data, medication records, and test details [12]. In most of the cases laboratory results and billing data are in the structured "name-value pair" data. Recently, more number of researchers is trying to develop big data based electronic phenotype algorithms to identify diseases from the EHR. Figure 3 represents the various types of Electronic Health Records (EHRs) [13].

(a) *Billing data*: Billing data are uses various codes to document the patients' symptoms, clinical records and lab results. International Classification of Diseases (ICD) and Current Procedural Terminology (CPT) are often used to document the billing data. This codes and data derived from ICD is most often used for research purposes. Summary of the ICD and CPT are depicted in Tables 3 and 4 respectively.
(b) *Laboratory data*: Laboratory data and vital signs are mostly in the structured format. It follows coding scheme to store the huge amount of lab related data.

Table 3 Summary of ICD

	ICD
Ease of use	• High
Format	• Structured
Advantages	• Simple to work and good prediction
Disadvantages	• Disease codes frequently used for all cases • Less accuracy

Table 4 Summary of CPT

	CPT
Ease of use	• High
Format	• Structured
Advantages	• Easy to work and high precision
Disadvantages	• Data is not accurate

Table 5 Summary of laboratory data

	Laboratory data
Ease of use	• High
Format	• Mostly structured
Advantages	• Data validity is high
Disadvantages	• May require to calculate cumulative dissimilar variations

Nowadays, many dictionaries and various algorithms are developed to reduce the complexity if laboratory data [9–11]. Summary of laboratory data are depicted in Table 5.

(c) *Medication records*: Medication records are used to identify accurate phenotype characterization. In addition, medication records also used to improve the disease diagnosis and drug recommendation in healthcare industry. This record is also used to avoid unwanted lab test and clinical care for individuals who are not actually affected by the disease. In addition, medical records are also used to identify the significance and importance of various drugs for number of diseases. Nowadays, format and variety of the medication records are increasing noticeably, it would helps to identify the number of hospital stays and reduce fault diagnosis rate [14]. Summary of medication records are depicted in Table 6.

Table 6 Summary of medication records

	Medication records
Ease of use	• Medium
Format	• Structured and unstructured
Advantages	• High data validity
Disadvantages	• Need to develop communication platform between inpatient and outpatient data
Summary	• Useful for disease diagnosis and clinical care

Table 7 Summary of clinical notes

	Clinical notes
Ease of use	• Medium
Format	• Unstructured
Advantages	• More details about doctors' judgment
Disadvantages	• Difficult to process without human intervention • Precision is fully depends on processing method • Cut, copy and paste are often affect the quality of the data
Summary	• Clinical documents are often used to identify common and well known diseases

Unstructured Clinical Note: Clinical documentation is often in the form of unstructured and it is widely used to improve the disease diagnosis [15]. Clinical notes are also considered as big data and scalable algorithms are used to process such huge size of data. For example, natural language processing and text search algorithms are widely used to process such huge size of clinical notes. Normally, clinical notes are created with the help of dictated and transcribed or computer-based documentation (CBD) systems. Summary of the clinical notes are depicted in Table 7.

Medical Imaging Data: Medical images are most often used for diagnosis, planning and therapy assessment [16, 17]. Recently, imaging techniques are increased such as Computed Tomography (CT), X-ray, molecular imaging, Magnetic Resonance Imaging (MRI), ultrasound, photoacoustic imaging, fluoroscopy and mammography. Nowadays, size of the medical videos and CT scans are also increased rapidly [18, 19, 20]. Such data requires huge storage space and fast algorithms to process and disease diagnosis [21, 22]. Medical imaging consists of different image acquisition methodologies generally utilized for various clinical applications. For example, visualizing blood vessel structure can be done using CT, MRI, photoacoustic imaging, and ultrasound. The main challenge with the image data is that it is not only large size, but also complex and multi dimensional [15].

Documentation from Reports and Tests: Nowadays, the cost to sequence the human genome (encompassing 30,000–35,000 genes) is quickly decreasing with the improvement of high-throughput sequencing tools and methods. Nowadays, it is difficult to process huge size of genome data and compute results. This would require advance scalable algorithms to process such huge size of clinical records. In order to overcome this issue, researchers are developed P4 medicine paradigm (i.e. predictive, preventive, participatory, and personalized health) with omics outline [15].

(d) *Wearable Sensor Devices*: Nowadays, more wearable medical devices are developed for patients' continuous health monitoring [23, 24]. These devices generate huge amount of health data continually. The typical communication among wearable things is depicted in Fig. 4. More familiar used wearable sensors and it functionalities is depicted in the Table 8. Security requirements and solutions in wearable medical devices are depicted in Table 9.

4 Big Data Genomics and Its Requirements

4.1 Acquisition for Big Genomics Data

The huge growth in genomic data is creates an opportunities decrease the overall prices, increase throughput and accuracy. The advancement in big data analytics is used in genomes sequencing to progress the accuracy and decrease the time taken

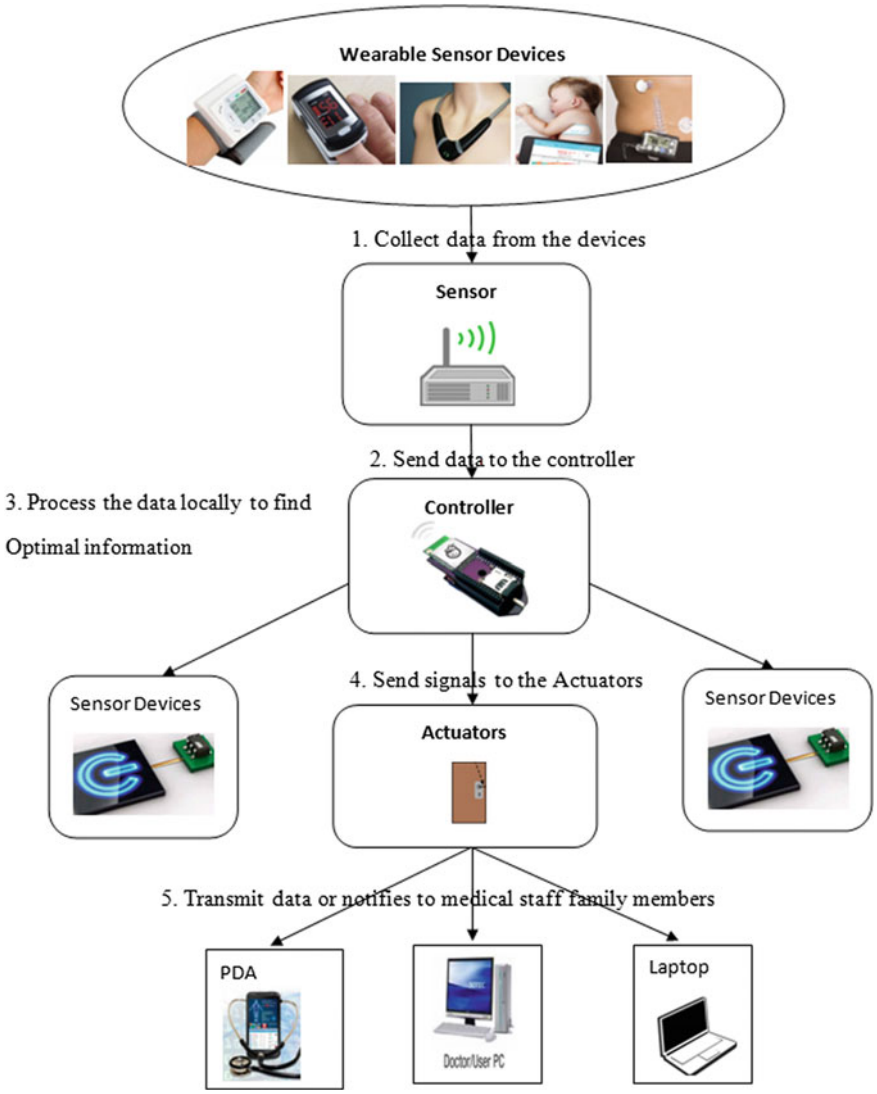

Fig. 4 Communication among wearable things in healthcare

for rapid diagnosis. Huge big genomics data provides high value insights and meaningful results for better prediction in healthcare.

Table 8 Sensors in human body

SNO	Name of the sensor	Sensor use	Sensor placement
1	Accelerometer	Measuring the human energy expenditure	Wearable
2	CO_2	Measuring the carbon dioxide level from mixed gas	Wearable
3	Electrocardiogram sensor	Measuring the electrocardiograph signal	Wearable
4	Accelerometer	Measuring the angular velocity with respect to the body axis	Wearable
5	Moisture sensor	Measuring the sweating rate	Wearable
6	Blood monitoring sensor	Measuring the percentage of oxygen saturation in blood	Wearable
7	Stress sensor	Measuring the pressure changes of the underside of foot	Wearable/ surrounding
8	Breath monitoring sensor	Measuring the rate of breathing	Wearable
9	Heat sensor	Measuring the rate of body temperature	Wearable
10	Image sensor	Capturing the motion, length, location, and area	Wearable/ surrounding
11	Blood pressure monitoring sensor	Measuring the systolic and diastolic pressure	Wearable
12	Heart rate monitoring sensor	Measuring the heart rate	Wearable
13	Blood sugar monitoring sensor	Sensors record glucose levels continuously around the clock	Wearable

4.2 Storage for Big Genomics Data

In healthcare industry there is a need to develop the storage system for large size of genomics data. Recently, 3-D memory and scalable methodologies are invented to increase the scalability and computing features of genomes and omics data. The above mention technologies are five time faster than the traditional optical switching technologies. Nowadays, compression and indexing systems are rapidly increased to store big genomes and omics data. For example, scalable MapReduce based algorithmic technologies are used to compare one genome to many others in an efficient way. In addition, many of the researchers are developing streaming methods to make on-the-fly comparisons for genome sequencing applications.

Table 9 Various security requirements and solutions in components of wearable healthcare system

Components in the wearable healthcare system	Vulnerabilities	Types of threats and attacks	Available security requirements and solutions
Physical objects	• Physical layer devices have limited communication, calculation and storage resources • Physical objects are distributed in various regions. Hence, unauthorized user can accesses the devices and performs damages and illegal actions such as reprogram the device, extract security keys and information.	• DoS/DDoS attacks • Physical attacks • Integrating WSNs • Integrating RFID • Unauthorized access control and data access	• Encryption/Cryptographic techniques • Continuously evaluates the suspicious nodes' behaviour can reduce the influence of malicious user access • Authentication • Authorization • Access control • Identification
Communication technologies	• IoT is a dynamic network infrastructure • Power issues • Network issues • Selection of security technique and its challenges	• Wireless WAN communications • Wireless LAN/PAN communications • Secure IoT communication protocols in constrained resources environment • Secure transmitted data	• Encryption/decryption is used to provide confidentiality service • Strong authentication also used to provide security solutions • Backup solution is used when network fails • Authorized access and availability

(continued)

Table 9 (continued)

Components in the wearable healthcare system	Vulnerabilities	Types of threats and attacks	Available security requirements and solutions
Applications	• Data coverage • Cloud computing • Security issues in web application • Secure communication	• DoS • XSS attack • CSRF attack • SQL injection • Data protection • Data access • PHRs attacks • Malicious user attacks • Sharing data in different environments • Real-time information processing • Sharing the same sensed data by several applications	• Encryption/decryption mechanisms • Secure data access • Scheduling techniques • Assuring identification • Assuring authentication • Firewall and antivirus • Intrusion detection

4.3 Distribution for Big Genomics Data

Cloud computing technologies are most often used for distributing genome sequences at a population scale. These technologies reduce the data movement and increases code federation. For example, Google, Amazon, and Facebook uses distributed computing framework to store large amount of data in distributed manner. Nowadays, cloud computing technologies are used for large scale genomic data querying and sequencing. For example, TCGA and BGI-cloud are uses cloud computing based platforms to store huge genomes and omics data in distributed manner. Though, efficient storage and processing methodologies are available to process such huge amount of genome data. There is a need to provide authentication, encryption, and other security frameworks to make certain that genomic data remain private.

4.4 Analysis for Big Genomics Data

The vital role of genome sequencing is to measure and observe the changes of DNA mutations and find the other molecular measurements relate to various diseases. In order to achieve the above task, there is a need to develop the scalable computing methodologies for processing such huge amount of genomics data. R, Mahout, and Hadoop machine learning algorithms are most often used to process such huge size of genome data.

5 Big Data in Descriptive Epidemiology

The Global Burden of Disease, Injuries and Risk Factors (GBD) study exemplifies an ongoing project on big data in descriptive epidemiology. The GBD studies generate estimates and trends of epidemiological metrics—morbidity, mortality and risk factor rates by age, sex, cause, year and geography [25]. A suite of analytical and statistical methods are used in the estimation process, and Disability Adjusted-Life Year (DALY) is used as an objective comparative metric for 310 diseases and injuries, and 79 behavioral and risk factors for 188 countries from 1990 to 2015. DALYs combine the morbidity metric of Years of Life lost due to Disability (YLD) and mortality metric of Years of Life Lost due to premature mortality (YLL); that is, DALY = YLD + YLL.

The GBD studies are led by the Institute of Health Metrics and Evaluation [26], in collaboration with more than 1700 collaborators from 125 countries. Dynamic data visualizations enable user-specific analysis and derive new insights [27], and data is publicly accessible through the Global Health Data Exchange [28]. Figure 5

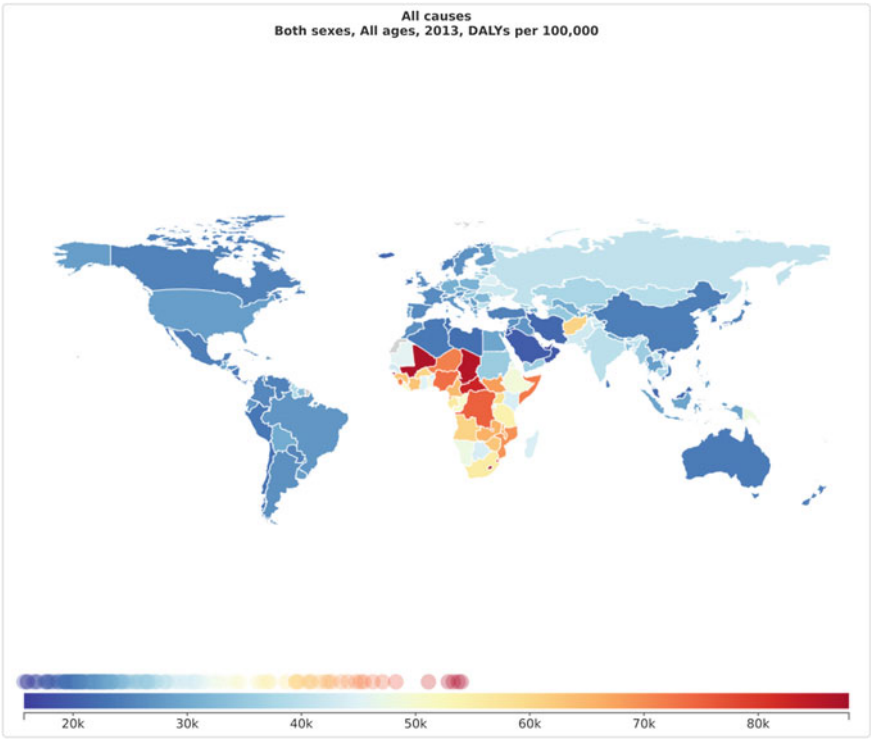

Fig. 5 Global burden of disease in 2013. Global burden of disease (DALYs per 100,000 people) due to all causes for both sexes and all ages in 2013

illustrates the global burden of disease due to all causes for both sexes and all ages in 2013, expressed through the metric of DALYs per 100,000 people.

The GBD studies follow the Guidelines for Accurate and Transparent Health Estimates Reporting (GATHER), which is a checklist and standard of 18 best practices for health estimates to improve transparency, accuracy and reliability [29, 30]. The GBD studies compile data from multiple health databases and epidemiological studies across different countries, and analyze over a billion data points. The GBD analytic method enables regular updates with data from new epidemiological studies. Policymakers of different countries, including China, India, Mexico, United Kingdom, and other countries are adopting the GBD approach to measure and analyze the population health of their respective countries.

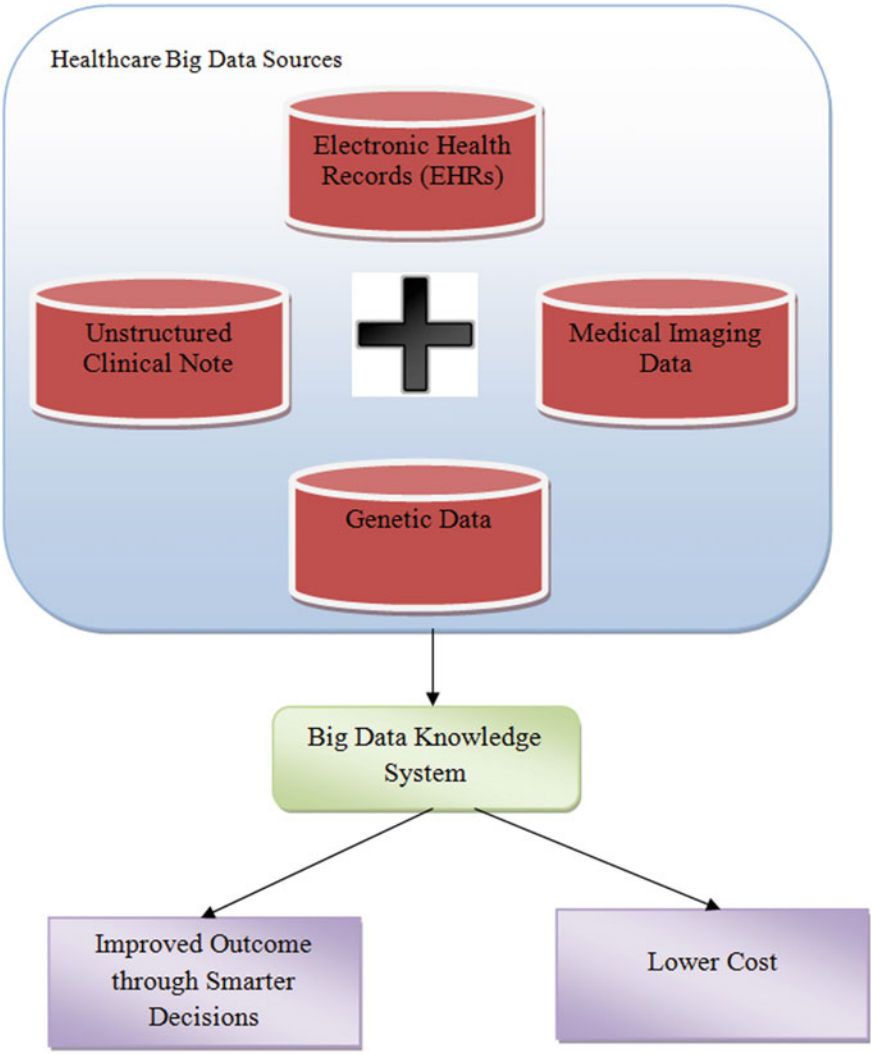

Fig. 6 Proposed Big Data knowledge system

6 Functionalities of Proposed Knowledge System

Big Data base knowledge system is shown in the Fig. 6. It consists of variety of databases such as EHR, Medical Imaging Data, Unstructured Clinical Notes and Genetic Data.

6.1 Identifying Different Decisions Using Levels of Healthcare System

The Healthcare Leadership Alliance (HLA) organized the healthcare system into four general levels or themes of analysis: consumers, employees, organizations, and environment [31].

Consumers: In general, many organizers in healthcare and medical industry have at least some knowledge in the patient position. This type of knowledge is used to enable the manager with a helpful structure of indication but also generate sightless. In general, users frequently tend to over generalize one patient's experience to those of others and in the procedure miss individual differences. In order to develop an efficient knowledge system the organization should understand the patient's need.

Employees: The employees of the healthcare organization are considered as the well knowledgeable individuals in the medical industry; nevertheless, the environment of proficient instruction creates for composite tapestry of interrelationships. In general, the goals, income and organizational power vary based on the healthcare occupation and spot.

Organization: In general, healthcare services directly connected with the end users. Hence, it is completely different from other organizations and industries. This creates "receiving it right the first time" for various business organizations; due to this reason the healthcare organization is one of the most regulated of all the organizations in the globe. In general, it is very important to understand the systems context for professionals joining into healthcare organization from other trade. For example, individuals from the financial departments are normally adapted to working under significant authoritarian oversight; though, those individuals feel difficult to face with various work forces. In addition, Individuals from production environments are familiar with skill set for increasing the operational efficiency, but they may not expertise in developing a healthy product.

Environment: In general, healthcare organizations are classified by various levels and trades. The changes in the economy and awareness toward fitness of the country would affect the healthcare delivery. More interaction among these forces is helpful to develop better knowledge in healthcare.

6.2 Developing Knowledge in the Healthcare System

Developing knowledge in the healthcare system consists of following ways [31]:

Knowledge Development in Consumers:

- Conduct interview with patient families
- Contribute in meetings and/or society outreach procedures and programs.

Knowledge Development in Employees:

- Monitoring the healthcare providers activities
- Participate staff meetings from other professions/departments
- Study applicable professional information and publications sources.

Knowledge Development in Organization:

- Increase mentoring relationships with head of the departments (e.g., finance/billing, legal affairs, community affairs, public health department)
- Contribute in community relations events and groups.

Knowledge Development in Environment:

- Study the healthcare division of business and common interest periodicals
- Examine relevant scholarly and employment journals (e.g., Health Affairs, Healthcare Executive)
- Participate seminars and workshops provided by special interest groups
- Volunteer in community activities.

7 Limitations and Future Work

This chapter discusses only about Electronic Health Record (EHR), Medical Imaging Data, Unstructured Clinical Notes and Genetic Data. The other sources of big data in healthcare are not discussed. The future work of this chapter is to combine the various other sources of big data in healthcare such as social media, web searches and mobile devices and developing the knowledge system.

8 Conclusion

Nowadays, health care systems are rapidly adopting clinical data, which will rapidly enlarge the size of the health records that are accessible, electronically. This chapter studies the characteristics and challenges for big data in healthcare, and proposes a big data based knowledge system. The proposed knowledge system is developed based on variety of databases such as EHR, Medical Imaging Data, Unstructured Clinical Notes and Genetic Data.

References

1. Manogaran, G., Thota, C., Kumar, M.: MetaCloudDataStorage architecture for Big Data security in cloud computing. Procedia Comput. Sci. **87**, 128–133 (2016)
2. Victor, N., Lopez, D., Abawajy, J.: Privacy models for big data: a survey. Int. J. Big Data Intell. **3**, 61 (2016)

3. Lopez, D., Sekaran, G.: Climate change and disease dynamics—a Big Data perspective. Int. J. Infect. Dis. **45**, 23–24 (2016)
4. Lopez, D., Gunasekaran, M., Murugan, B.S., Kaur, H., Abbas, K.M.: Spatial Big Data analytics of influenza epidemic in Vellore, India. In: IEEE International Conference on Big Data, pp. 19–24, Oct 2014
5. Bates, D., Saria, S., Ohno-Machado, L., Shah, A., Escobar, G.: Big Data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Aff. **33**, 1123–1131 (2014)
6. Jee Kim, G.: Potentiality of Big Data in the medical sector: focus on how to reshape the healthcare system. Healthc. Inform. Res. **19**, 79 (2013)
7. Chawla, N., Davis, D.: Bringing Big Data to personalized healthcare: a patient-centered framework. J Gen. Intern. Med. **28**, 660–665 (2013)
8. Viceconti, M., Hunter, P., Hose, R.: Big Data, big knowledge: Big Data for personalized healthcare. IEEE J. Biomed. Health Inform. **19**, 1209–1215 (2015)
9. Lopez, D., Gunasekaran, M.: Assessment of vaccination strategies using fuzzy multicriteria decision making. In: Proceedings of the Fifth International Conference on Fuzzy and Neuro Computing (FANCCO-2015), pp. 195–208. Springer International (2015)
10. Kothari, A., Hovanec, N., Hastie, R., Sibbald, S.: Lessons from the business sector for successful knowledge management in health care: a systematic review. BMC Health Serv. Res. **11**, 173 (2011)
11. Chen, E.T.: An observation of healthcare knowledge management. Commun. IIMA **13**, 95–106 (2013)
12. Denny, J.: Chapter 13: mining electronic health records in the genomics era. PLoS Comput. Biol. **8**, e1002823 (2012)
13. Jensen, P., Jensen, L., Brunak, S.: Mining electronic health records: towards better research applications and clinical care. Nat. Rev. Genet. **13**, 395–405 (2012)
14. Poon, E., Keohane, C., Yoon, C., et al.: Effect of bar-code technology on the safety of medication administration. Obstet. Gynecol. Surv. **65**, 629–630 (2010)
15. Belle, A., Thiagarajan, R., Soroushmehr, S.M., Navidi, F., Beard, D.A., Najarian, K.: Big data analytics in healthcare. BioMed. Res. Int. **10**, 1–16 (2015)
16. Virmani, J., Dey, N., Kumar, V.: PCA-PNN and PCA-SVM based CAD systems for breast density classification. In: Applications of Intelligent Optimization in Biology and Medicine, pp. 159–180. Springer International Publishing (2016)
17. Bhattacherjee, A., Roy, S., Paul, S., Roy, P., Kausar, N., Dey, N.: Classification approach for breast cancer detection using back propagation neural network: a study. In: Biomedical Image Analysis and Mining Techniques for Improved Health Outcomes, p. 210 (2015)
18. Suri, J., Dey, N., Bose, S., Das, A., Chaudhuri, S.S., Saba, L., Shafique, S., Nicolaides, A.: 2084743 Diagnostic preservation of atherosclerotic ultrasound video for stroke telemedicine in watermarking framework. Ultrasound Med. Biol. **41**(4), S133 (2015)
19. Dey, N., Mukhopadhyay, S., Das, A., Chaudhuri, S.S.: Analysis of P-QRS-T components modified by blind watermarking technique within the electrocardiogram signal for authentication in wireless telecardiology using DWT. Int. J. Image, Graph. Sign. Process. **4**(7), 33 (2012)
20. Acharjee, S., Ray, R., Chakraborty, S., Nath, S., Dey, N.: Watermarking in motion vector for security enhancement of medical videos. In: IEEE International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), pp. 532–537, July 2014
21. Bose, S., Acharjee, S., Chowdhury, S.R., Chakraborty, S., Dey, N.: Effect of watermarking in vector quantization based image compression. In: IEEE International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), pp. 503–508, July 2014
22. Pal, A.K., Dey, N., Samanta, S., Das, A., Chaudhuri, S.S.: A hybrid reversible watermarking technique for color biomedical images. In: IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pp. 1–6, Dec 2013

23. Nandi, S., Roy, S., Dansana, J., Karaa, W.B.A., Ray, R., Chowdhury, S.R., Chakraborty, S., Dey, N.: Cellular automata based encrypted ECG-hash Code generation: an application in inter human biometric authentication system. Int. J. Comput. Netw. Inf. Secur. **6**(11), 1 (2014)
24. Biswas, S., Roy, A.B., Ghosh, K. and Dey, N.: A biometric authentication based secured ATM Banking System. Int. J. Adv. Res. Comput. Sci. Softw. Eng. ISSN, 2277
25. Murray, C.J., Barber, R.M., Foreman, K.J., Ozgoren, A.A., Abd-Allah, F., Abera, S.F., Aboyans, V., Abraham, J.P., Abubakar, I., Abu-Raddad, L.J., Abu-Rmeileh, N.M.: Global, regional, and national disability-adjusted life years (DALYs) for 306 diseases and injuries and healthy life expectancy (HALE) for 188 countries, 1990–2013: quantifying the epidemiological transition. The Lancet **386**(10009), 2145–2191 (2015)
26. Healthdata.org: Global Burden of Disease (GBD)| Institute for Health Metrics and Evaluation. http://www.healthdata.org/gbd (2016). Accessed 8 Aug 2016
27. Healthdata.org: GBD Data Visualizations| Institute for Health Metrics and Evaluation. http://www.healthdata.org/gbd/data-visualizations (2016) Accessed 8 Aug 2016
28. Healthdata.org.: Global Health Data Exchange (GHDx)| Institute for Health Metrics and Evaluation. http://www.healthdata.org/about/ghdx (2016) Accessed 8 Aug 2016
29. Stevens, G., Alkema, L., Black, R., Boerma, J., Collins, G., Ezzati, M., Grove, J., Hogan, D., Hogan, M., Horton, R., Lawn, J., Marušić, A., Mathers, C., Murray, C., Rudan, I., Salomon, J., Simpson, P., Vos, T., Welch, V.: Guidelines for accurate and transparent health estimates reporting: the GATHER statement. PLoS Med. **13**(6), e1002056 (2016)
30. Stevens, G., Alkema, L., Black, R., Boerma, J., Collins, G., Ezzati, M., Grove, J., Hogan, D., Hogan, M., Horton, R., Lawn, J., Marušić, A., Mathers, C., Murray, C., Rudan, I., Salomon, J., Simpson, P., Vos, T., Welch, V.: Guidelines for accurate and transparent health estimates reporting: the GATHER statement. The Lancet (2016)
31. Garman, A.N., Tran, L.: Knowledge of the healthcare environment. J. Healthc. Manag. **51**, 152–155 (2006)