Big Data in Healthcare: Prospects, Challenges and Resolutions

Iroju Olaronke
Department of Computer Science, Adeyemi College of Education
Ondo, Nigeria
irojuolaronke@gmail.com

Ojerinde Oluwaseun
Department of Computer Science, Federal University of Technology
Minna, Nigeria
o.ojerinde@futminna.edu.ng

***Abstract*—The healthcare system consists of large volumes of data which are usually generated from diverse sources such as physicians' case notes, hospital admission notes, discharge summaries, pharmacies, insurance companies, medical imaging, laboratories, sensor based devices, genomics, social media as well as articles in medical journals. Healthcare data are however very complex and difficult to manage. This is as a result of the astronomical growth of healthcare data, the high speed at which these data are generated as well as the diversity of data types in healthcare. The capturing, storage, analysis and retrieval of health related data are rapidly shifting from paper based system towards digitization. However, the vast volume as well as the complexity of these data makes it difficult for the data to be processed and analyzed by traditional approaches and techniques. Consequently, technologies such as cloud computing and virtualization are now gradually used for processing massive data effectively and securely in healthcare. Hence, the healthcare system is swiftly becoming a big data industry. Thus, this paper examines the concept of big data in healthcare, its benefits and attendant challenges. This paper revealed that the fragmentation of healthcare data, ethical issues, usability issues as well as security and privacy issues are some of the factors impeding the successful implementation of big data in healthcare. This paper therefore suggests that ensuring security and privacy of healthcare data, the adoption of a standardized healthcare terminology, education strategy and the design of usable systems for processing large volumes of data are some of the ways of successfully implementing big data in healthcare.**

Keywords—big data; information and communication technology; healthcare

I. INTRODUCTION

The healthcare system is collaborative in nature. This is because it consists of a large number of stakeholders such as physicians of diverse specialties, nurses, radiologists, laboratory technologists and pathologists that work together to achieve the common goals of reducing medical cost and errors as well as providing quality and improved healthcare services. Each of these stakeholders generate data from heterogeneous sources such as physical examination, clinical notes, patients interviews and observations, laboratory tests, imaging reports, treatments, therapies, surveys, bills and insurance. However, in recent times, the advancement in Information and Communication Technologies (ICT) has resulted in the development of new sources of data in healthcare such as sensor based devices, social media posts, micro blogging sites, genomic sequences, smart meters, log files, videos, Radio Frequency Identification devices (RFID) and Global Positioning System (GPS). The rapid increase in the rate at which data is generated from heterogeneous sources increases the volume of healthcare data [1]. Consequently, it becomes difficult to store, process and analyze health related data with traditional data processing applications [2]. However, innovative tools and techniques as well as powerful computing technologies are now used to store, process, analyze and extract values from voluminous and heterogeneous healthcare data in a real time manner [3]. Hence, the healthcare system is fast becoming a big data industry.

Big data provides a wide range of opportunities for the healthcare industry. For instance the Harvard Business Review [4] revealed that the adoption of big data in healthcare has led to the simplification of Information Technology (IT), evidenced based and value conscious medicine, better preventive care and personalized treatment. Furthermore, big data reveals patterns and trends in data which helps in the process of diagnosing and treating patients. Thus, the deployment of big data in healthcare has led to the improvement of patients care at a lower cost and increased patient satisfaction. However, one of the major challenges of big data in healthcare is security and privacy problems. This is because electronic health records are highly susceptible to inappropriate access, data breeches and fraud, medical identity theft, compromised data integrity and widespread unauthorized distribution [5]. In addition, the complex and fragmented nature of healthcare data, the diverse schemas and standards underlying the data as well as the rapid growth of incompatible vocabularies and ontologies in healthcare are some of the major factors hindering the effective implementation of big data in healthcare. Thus, the healthcare system is characterized by high error rates and high cost which result in a high rate of mortality [6]. Consequently, this paper examines the concept of big data in the context of healthcare, the tools and techniques used for the implementation of big data in healthcare as well as the benefits and challenges of big data in healthcare. This paper also suggests the possible ways of facilitating the successful implementation of big data in healthcare.

II. OVERVIEW OF BIG DATA IN HEALTHCARE

In a broad sense, big data can be defined as a collection of large and complex data sets which are difficult to manage using common database management tools or traditional data

978-1-5090-4171-8/16/$31.00 ©2016 IEEE

processing applications [7]. Big data is also defined as a large volume of high velocity, complex and variable data which require advanced techniques and technologies for capturing, storing, distributing, managing and analyzing information [8]. In addition, Burghard [9] viewed big data as a new generation of technologies and architectures that are designed to economically extract value from very large volumes of a wide variety of data by enabling high velocity capture, discovery and analysis. Similarly, in the context of healthcare, big data refers to a collection of large and complex electronic health data which are difficult to process, distribute and analyze with traditional approaches and techniques. Big data in the context of healthcare can also be defined as a collection of tools, technologies, methods and procedures which are used to create, store, process, analyze and retrieve large sets of electronic health data in an efficient manner. However, the concept of big data is not only used for large sets of data but also the ability to produce useful and valuable information from a large collection of data set using innovative tools and technologies. The sources of big data in healthcare include the following:

A. Machine Generated Data

These are data that are obtained from machines that are used in the healthcare system. Examples include data obtained from remote sensors, wearable devices, smart meters and other devices that measure vital signs.

B. Biometric Data

These are data that are obtained from individuals' physical characteristics such as finger prints, genetics, signature, retinal scans, heart rate, blood pressure, pulse and pulse-oximetry readings as well as x-ray and other medical images.

C. Human Generated Data

These include data generated by human beings in the healthcare system. These type of data include unstructured and semi structured clinical data such as case notes, laboratory results, hospital admission records, discharge summaries and electronic mails. Human generated data also include structured Electronic Health Record (EHR) data.

D. Transactional Data

These include data from healthcare claims and billing records.

E. Behavioural Data

These include data generated from social interactions and communication tools such as websites and social media sites such as Twitter and Facebook.

F. Epidemological Data

These data include vital statistical data, health surveys and disease registries.

G. Publication Data

These include data from clinical researches and medical reference materials.

III. Characteristics of Big Data in Healthcare

Big data in healthcare has been characterized by diverse authors as 3V's [7], [10], [11]. These characteristics typically include volume, variety and velocity. However, as a result of the change in the nature of healthcare data, this paper characterizes big data in healthcare into 5V's. These include volume, velocity, variability, variety and veracity.

A. Volume

Volume refers to the amount of data generated by the stakeholders in the healthcare industries. Healthcare data are usually complex, noisy, heterogeneous, longitudinal and voluminous [7]. For instance in 2011, the data in the United States of American healthcare system was about 150 Exabyte [8]. Furthermore, in 2012, digital healthcare data worldwide was estimated to be equivalent to 500 Petabytes and it is expected to reach 25,000 Petabytes by 2020 [7].

B. Velocity

This refers to the speed or rate at which data are created, stored, analyzed, visualized and exchanged amongst healthcare providers. The healthcare system requires the seamless, secured and meaningful exchange of health information for effective and efficient patient care. This is because patients usually receive care from diverse healthcare providers at different geographical locations. This information is required to be accessed by healthcare practitioners in a real time manner for decisions making [12]. An error or omission in this process can lead to the untimely death of a patient. Hence, velocity is required in healthcare because it allows healthcare providers to exchange and use data in a timely manner.

C. Variability

Variability refers to changes in data rate, data format, data structure and data semantics. Data in the healthcare system come in different formats such as flat files, relational tables, images and texts. In addition, a concept in the healthcare system can have multiple meanings. For instance the term mass denotes breast mass which is a form of breast cancer while mass in a radiological report of the chest denotes mass in lung [13]. In addition, the healthcare system is composed of the same abbreviations that denote different concepts. For example, the acronym APC might mean Activated Protein C, Advanced Pancreatic Cancer, AlloPhyCocyanin and Antibody Producing Cells [14].

D. Variety

The healthcare system consists of diverse sources of data which could be structured, semi-structured and unstructured. Diverse sources of data in healthcare include multimedia, social media, blogs, web server logs as well as financial transactions**.** Furthermore, healthcare providers may generate data from different geographical locations and data can be stored in numerous legacy and application systems such as transaction processing system and diverse databases.

978-1-5090-4171-8/16/$31.00 ©2016 IEEE

E. Veracity

Veracity refers to the quality of data produced. The quality of data produced in the healthcare system has been a major cause of concern [15]. For instance, unstructured data are composed of diverse grammatical structures, varied expressions expressed in diverse natural languages as well as the use of single concept to denote multiple terms. Hence, the healthcare system is characterized by ambiguity which results in high cost and high error rates. However, the quality of data produced within the healthcare system is very vital because life and death decisions depend on accurate information and high quality of healthcare data which assist healthcare providers to make critical decisions at the right time. Hence, healthcare data must be relevant, reliable and error-free.

Table 1 summarizes the characteristics of big data in healthcare while the concept of big data is depicted in Fig. 1.

TABLE I. CHARACTERISTICS OF BIG DATA: THE 5 V'S

Characteristics	Meaning
volume	This refers to a large amount of data generated by various stakeholders in the healthcare system
velocity	This refers to the speed at which health data are created, stored, analyzed visualized and shared amongst healthcare providers
variability	This refers to changes in data rate, data format, data structure and data semantics
variety	This refers to the diverse sources of health data
veracity	This refers to the quality of health data produced

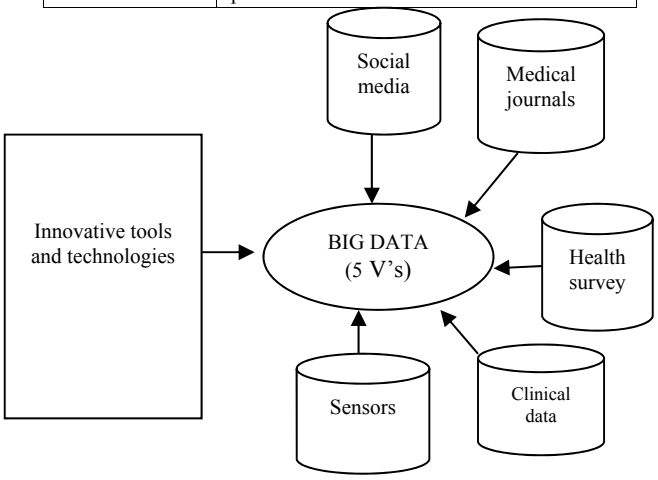

Fig. 1. The concept of big data

IV. TOOLS AND TECHNOLOGIES FOR ANALYZING BIG DATA IN THE HEALTHCARE SYSTEM

With the advancement of Information and Communication Technology, most especially the Internet of Things (IoT), healthcare data has soared from the Exabyte to the Petabyte and it is gradually approaching the Zettabyte and Yottabyte [8]. With this astronomical rate of data growth, the high speed at which the data is generated and the complexity of healthcare data, it becomes difficult for traditional data architectures to efficiently handle data sets in healthcare. Hence, this section appraises some of the sophisticated tools, technologies, applications, and platforms for analyzing big data in healthcare. These tools are generally referred to as big data analytics. Typical examples of big data analytics include the following:

A. Google Big Query

Google Big Query uses Google's cloud infrastructure to store and query massive datasets in few seconds. Data are protected with multiple layers of security in Google Big Query.

B. Map Reduce

Map Reduce is a software tool that is made up of two basic functions which include Map and Reduce. Map provides an interface for the distribution of sub-tasks while Reduce collects the work and resolves the results into a single value. Map Reduce according to Youssef [2] allows a vast amount of data to be processed in large clusters.

C. Jaql

Jaql is a functional and declarative query language that is designed to process large volumes of data sets. Jaql uses Map Reduce tasks to convert high-level queries into low-level queries in order to facilitate parallel processing [2].

D. Hadoop

Hadoop is an open source software framework that processes large amounts of data across massively parallel clusters of servers. It is a non-relational Database Management System. The key component of Hadoop is the Hadoop Distributed File System (HDFS) which manages data across different servers. HDFS stores data of diverse types and structures.

E. Non Relational Databases

In relational databases, data are stored in rows and columns and can be accessed through Structured Query Language (SQL). However, non-relational databases do not rely on SQL to retrieve data. Rather, data is stored and retrieved through key-value pairs which provide links to where files are stored on disks. Non-relational databases do not have strong mechanisms that ensure adequate security [2]. Hence, they cannot be widely adopted for healthcare applications.

F. Cloud Based Services

The cloud provides virtualization for computing resources over the Internet. The cloud also provides adequate access to computing resources. Cloud-based services come in diverse forms which include software-as-a-service (SaaS), platform-as-a-service (PaaS) and infrastructure-as-a-service (IaaS). SaaS allows a vendor to provide the hardware, application software, operating system, and storage. PaaS provides a basic platform such as Oracle Cloud Computing, Microsoft Windows Azure, and Google App Engine. The IaaS allows the vendor to provide the raw computing power and storage.

Table 2 clearly presents a detailed comparison of the tools and technologies used for analyzing big data in the healthcare system.

978-1-5090-4171-8/16/$31.00 ©2016 IEEE

TABLE II. A COMPARATIVE ANALYSIS OF TOOLS FOR ANALYSING BIG DATA

Tools	Platforms	Type of Databases	Advantages	Limitations
Google Big Query	It is an open source and cloud based platform	It is a columnar database	It allows data to be replicated across diverse data centers.	*Google Big Query does not support indexes*
Map Reduce	It is an open source and cloud based platform	It is a non-relational database	It works well with unstructured and semi structured data such as audio and visual data	It lacks indexing capabilities of modern database systems
Hadoop	It is an open source and cloud based platform	It is a non-relational database	It stores data of any structure such as Web logs	It lacks technical support and security
Jaql	It is a proprietary query language	It is a query language for JavaScript object notation	It supports both structured and semi structured data	There are no user defined types, this implies that schema information is only a constraint on possible values of a domain
Microsoft Windows Azure	It is a public cloud based platform	It is a relational database	It allows users to make relational queries against structured, semi-structured and unstructured documents	The size of the database is limited, hence huge databases are not possible on Microsoft Windows Azure

V. BENEFITS OF BIG DATA IN HEALTHCARE

In recent times, the healthcare system has witnessed a rapid increase in the use of electronic healthcare systems to improve the quality of patients' care. However, the astronomical increase in the volume of healthcare data has made it difficult for health related data to be effectively processed by traditional data processing applications. In order to effectively extract values from healthcare data, it is expedient to understand the importance of big data in healthcare. Consequently, this section examines the need for analyzing big data in healthcare with sophisticated tools and technologies.

A. Evidenced Based Care

The standard medical practice is shifting from relatively ad-hoc and subjective decision making towards evidence-based healthcare [16]. Evidence medicine is a system in which the treatment of patients depends largely on available scientific evidences. Big data provides evidenced based care by aggregating data sets from diverse sources. The trends and pattern in the data will provide enough evidence for diagnosing and treating patients. Smaller data sets might not provide sufficient evidence to determine if statistical differences are present in the data sets [16].

B. Reduced Cost of Healthcare

One of the major challenges that the healthcare industry is facing is increasing costs. For example, in the United States of America, the cost of healthcare was about 16.9 % of the Gross Domestic Product in 2012 [17]. However, a recent survey carried out by a Health Research Institute revealed that more readily available information could cut the cost of healthcare [4]. The survey also revealed that patients would likely choose non-traditional forms of healthcare, such as at-home urinalysis tests using a device attached to a smart phone if they cost less. Hence, Priyanka and Kulennavar [3] estimated that analyzing data with innovative tools and techniques would result in the savings of $300 billion per year in the United States of America. This reduction in healthcare cost is about 8% of the national healthcare expenditures [3].

C. Increases Patients' Participation in the Care Process

Big data ensures that patients have access to accurate and up-to-date information. This will enable the patients to understand their choices, make decisions concerning their care as well as improve their lifestyle to avoid chronic diseases.

D. Improves Public Health Surveillance

Analyzing healthcare data with innovative tools helps in analyzing disease patterns, tracking disease outbreaks and transmission. This increases public health surveillance, education and speed response.

E. Reduces Mortality Rate

Big data ensures the early detection and identification/diagnosis of diseases. This guarantees that right decisions on the treatment of a particular disease are taken in an effective and timely manner. This reduces patients' morbidity and mortality.

F. Increases Communication between Healthcare Providers and Patients

Big data enhances effective communication amongst healthcare providers and patients. For instance, patients with similar health challenges as well as healthcare providers with similar specialties across the globe can exchange ideas on the prevention and cure of a particular disease on social media. This process facilitates interoperability across healthcare institutions.

G. Early Detection of Security Threats and Fraud in Healthcare

Big data can be used to easily identify patterns and irregularities that indicate the presence of security threats and fraud in healthcare.

H. Improves the Quality of Care

Big data improves the quality of care delivered to patients by ensuring that decisions are based on large volumes of relevant and up-to-date data.

978-1-5090-4171-8/16/$31.00 ©2016 IEEE

VI. Limitations of Big Data in Healthcare

There is no doubt that big data has positive impacts on the healthcare system. However, the major challenge confronting big data in healthcare is not the lack of data but the lack of information to support decision making, planning and strategy [18]. Hence, health related data needs to be validated, processed and integrated in order to extract value from it. This section therefore outlines the challenges of big data in healthcare.

A. Resistance to Change

The healthcare system tends to be laggard in adopting technology unlike other sectors like the banking sector and the oil and gas industry. This is because of the inadequate administrative support for Information Technology (IT) and related practice changes, lack of trust, legal issues as well as the lack of necessary and appropriate skills for the operation of the ICT tools [12]. Hence, the paper based system is still deployed in the healthcare system. The paper based system however does not support the integration of data from diverse sources in the healthcare system.

B. Fragmentation of Healthcare Data

The challenges of big data in healthcare are compounded by the fragmentation and dispersion of data which are stored in proprietary heterogeneous systems across healthcare organizations. In addition, healthcare data are also stored in legacy systems (usually electronic medical record systems) which have limited interoperability capabilities. It is however difficult to integrate health data to form big data because there are different schemas, formats, metadata and standards underlying the data.

C. Ethical Challenges

Ethical challenges such as data privacy, confidentiality, control of access to patients' information, the commercialization of de-identified patients' information, ownership and governance of patients' information are factors that hinder the effective exchange of information amongst patients and healthcare providers. Consequently, the integration of healthcare data from diverse sources becomes a challenge. Hence, the access to a detailed and complete picture of a patient during care in a timely manner becomes a problem.

D. Proliferation of Healthcare Standards

Standards are agreed-upon specifications that allow disparate systems, tools, technologies and platform to work together. However, healthcare institutes do not conform to a single standard. For instance, the titles and codes of case reports, drugs, diseases, and examination vary in different hospitals [12]. Hence, the lack of a common standard in the healthcare system hinders interoperability amongst heterogeneous systems.

E. Security and Privacy Issues

One of the major challenges facing the integration of diverse sources of healthcare information into big data is security and privacy issues. Healthcare information is susceptible to security threats such as the improper disclosure of patients' information, unauthorized use of patients' information and unauthorized destruction of patients' data. Hence, healthcare providers become discouraged to share health information using electronic healthcare systems.

VII. Resolutions

In order for the healthcare system to benefit from the numerous advantages of big data in healthcare, the following resolutions are hereby made:

A. Security and Privacy Policies

Healthcare providers and developers of big data analytics should ensure that healthcare data are adequately protected and secured. Tools that will ensure the confidentiality, integrity, and availability of protected health information should be used for analyzing big data in healthcare. In addition, healthcare data should be adequately protected with physical security, data encryption, user authentication, and application security. The use of audit trail systems should also be encouraged.

B. Usability of Big Data Analytics

Big data analytics should possess well designed user interfaces. Big data analytics should also be designed in a way that it would be easy to learn and use.

C. Establishment and Adoption of a Unified Standard

The establishment and adoption of a standard terminology/vocabulary, that is, a common language for describing medical terminology is an essential task that is required for the successful implementation of big data in healthcare. This is to ensure the consistency, reusability and the sharability of healthcare information.

D. Education Strategy

The healthcare system is laggard in the adoption of ICT unlike other systems such as the banking system. Hence, healthcare providers should be educated on the importance of adopting innovative ICT tools for analyzing and processing big data in healthcare. In addition, healthcare providers should be encouraged to acquire knowledge and technical skills on how health related data should be collected, analyzed and interpreted.

VIII. Conclusion

Big data are referred to as large volumes of high velocity, complex, and variable data which require advanced techniques and technologies for capturing, storing, distributing, managing and analyzing data. The basic goal of big data is to economically extract values from very large volumes of a wide variety of data. In healthcare system, big data is required to facilitate seamless communication amongst healthcare providers and patients, it increases patients' participation in the care process, it results in evidenced based care and it also facilitates the early detection of security threats and fraud. However, in spite of the numerous benefits of big data in healthcare, factors such as resistance to change from traditional mode of care to the use of ICT based care, the longitudinal nature of health information, lack of healthcare standard as well as security challenges hinder the effective adoption of big data in healthcare. Consequently, this paper suggests that the use of a standard vocabulary, the design of

978-1-5090-4171-8/16/$31.00 ©2016 IEEE

effective tools that have well designed interfaces and the implementation of security based policies will enhance the use of big data in healthcare.

REFERENCES

[1] W. Raghupathi, and V. Raghupathi, "Big data analytics in healthcare: promise and potential," Health Information Science and Systems, vol. 2, pp. 1-10, 2014.

[2] A.E. Youssef,"A framework for secure healthcare systems based on big data analytics in mobile cloud computing environments," International Journal of Ambient Systems and Applications , vol.2, pp. 1-11, 2014.

[3] K. Priyanka, and N. Kulennavar, "a survey on big data analytics in health care," International Journal of Computer Science and Information Technologies,"vol. 5 , pp. 5865-5868, 2014.

[4] Harvard Business Review,"How Big Data Impacts Healthcare," Harvard Business Review, 2014.

[5] O. Iroju, and R Ikono, " A security based framework for interoperability of healthcare systems," International Journal of Applied Information Systems, vol. 1, pp. 23-31, 2013.

[6] C. Bock, L. Carnahan, S. Fenves, M. Gruninger, V. Kashyap, B. Lide, J. Nell, R. Rama, and R. Sriram, "Healthcare strategic focus area: clinical informatics," National Institute of Standards and Technology, Technology Administration, Department of Commerce, United States of America, pp.1-33, 2005.

[7] J. Sun, and C.K. Reddy, "Big data analytics for healthcare, " Tutorial presentation at the SIAM International Conference on Data Mining, Austin, Texas, 2013.

[8] IHTT, "Transforming health care through big data strategies for leveraging big data in the health care industry,"' http://ihealthtran.com/ wordpress/2013/03/iht%C2%B2-releases-big-data-research reportdownload-today/, 2013.

[9] C. Burghard, "Big data and analytics key to accountable care success," IDC Health Insights, 2012.

[10] H. J. Watson, "Tutorial: Big data analytics: concepts, technologies, and applications," Communications of the Association for Information Systems, vol. 34, pp. 1-24, 2014.

[11] C.J. McCall, "Big data in healthcare: Hype or hope," A White Paper by Carol J McCall, 2015.

[12] O. Iroju, A. Soriyan, I. Gambo, and J. Olaleke, "Interoperability in healthcare: benefits, challenges and resolutions," International Journal of Innovation and Applied Studies, vol. 3,pp. 262-270, 2013.

[13] C. Friedman, and S.B. Johnson, Natural language and text processing in biomedicine, United States of America: Springer:, 2006, pp.312-343.

[14] O.G. Iroju, and J.O. Olaleke, "A systematic review of natural language processing in healthcare," International Journal of Information Technology and Computer Science, vol.8,pp. 44-48, 2015

[15] B, Feldman, E.M. Martin, and T. Skotnes,"Big data in healthcare hype and hope," Bonnie, October 2012.

[16] S. Piai, and M. Claps, "Bigger data for better healthcare," IDC Health Insights,2013.

[17] OECD Health Statistics, "How does the united states compare, http://www.oecd.org/health/healthdata, 2014.

[18] Cognizant, "Big data is the future of healthcare," Cognizant 20-20 Insights, 2012.

978-1-5090-4171-8/16/$31.00 ©2016 IEEE