Research based on big data analysis of medical industry

lijun chen\textsuperscript{1} Jiaying lin\textsuperscript{2} Zhang yi\textsuperscript{3}

South China Institute of Software Engineering of Guang Zhou Universitity, Guangzhou, China

372158286@qq.com, ljq@sise.com.cn, zy@sise.com.cn

Abstract: In recent decades, institutions in every country and every place have produced a large amount of structured, unstructured and semi-structured data with different (heterogeneous) structures. The healthcare sector faces the need to manage big data from a variety of sources known for generating large amounts of heterogeneous data. In the medical and health industry, various big data analysis tools and technologies have been developed to deal with these massive data. In this article, we will discuss the impact of big data on healthcare and health, as well as the various big data processing tools available in the Hadoop system. We also explored the conceptual architecture of medical big data analysis, including data collection histories from different branches, genomic databases, electronic health records, text/image, and clinical decision support systems.

1. INTRODUCTION
With the wide application of health care informatization, massive data sets are generated in the process of medical service, health care and health management to form health care big data [1]. Health and medical big data can usually be divided into the following aspects: health and medical service data based on electronic health records, electronic medical records, medical imaging, examination and examination, etc.; Biomedical data such as gene sequence and proteome; Data of new rural cooperative medical care, basic medical insurance for urban workers, and basic medical insurance for urban residents; Drug clinical trials, drug screening, centralized procurement of essential drugs, electronic supervision of drugs and vaccines in medical institutions and other pharmaceutical r&d and management data; Public health data such as disease surveillance, public health emergency surveillance and infectious disease reporting; Behavioral and emotional data of patients, such as behavioral performance, health product purchase records, fitness information, etc.; Health resources and medical services survey, family planning statistics and other statistical data; Resident marriage, family, family planning registration and other population management data; Air pollutants, climate conditions and other environmental data closely related to human health [2].

Big data of health care will bring creative changes to clinical diagnosis and treatment, drug research and development, health monitoring, public health, policy formulation and implementation, comprehensively improve the governance ability and level of health care, and create great value [3]. According to McKinsey's prediction, the effective use of big data in health care can bring more than $30 billion worth of value to the United States every year [4].

2. BIG DATA ANALYTICS IN HEALTH INFORMATICS
The main difference between traditional health analysis and big-data health analytics is the execution of computer programming. In the traditional system, the healthcare industry depended on other
industries for big data analysis. Many healthcare shareholders trust information technology because of its meaningful outcomes—their operating systems are functional and they can process the data into standardized forms. Today, the healthcare industry is faced with the challenge of handling rapidly developing big healthcare data. The field of big data analytics is growing and has the potential to provide useful insights for the healthcare system. As noted above, most of the massive amounts of data generated by this system is saved in hard copies, which must then be digitized [5]. Big data can improve healthcare delivery and reduce its cost, while supporting advanced patient care, improving patient outcomes, and avoiding unnecessary costs [6]. Big data analytics is currently used to predict the outcomes of decisions made by physicians, the outcome of a heart operation for a condition based on patient’s age, current condition, and health status. Essentially, we can say that the role of big data in the health sector is to manage data sets related to healthcare, which are complex and difficult to manage using current hardware, software, and management tools. In addition to the burgeoning volume of healthcare data, reimbursement methods are also changing [7]. Therefore, purposeful use and pay based on performance have emerged as important factors in the healthcare sector. In 2011, organizations working in the field of healthcare had produced more than 150 exabytes of data [8], all of which must be efficiently analyzed to be at all useful to the healthcare system. The storage of healthcare related data in EHRs occurs in a variety of forms. A sudden increase in data related to healthcare informatics has also been observed in the field of bioinformatics, where many terabytes of data are generated by genomic sequencing. There are a variety of analytical techniques available for interpreting medical, which can then be used for patient care. The diverse origins and forms of big data are challenging the healthcare informatics community to develop methods for data processing. There is a great need for technologies that combine different data sources. Many conceptual approaches can be used to identify irregularities in large amounts of data from different data sets. The structure of medical data analysis is as follows:

2.1 Predictive Analytics of Healthcare
In the past two years, predictive analytics has been considered one of the main business intelligence methods, but its application in the real world goes far beyond the business context. Big data analysis includes a variety of methods, including text analysis and multimedia analysis. However, one of the most critical categories is predictive analytics, which includes statistical methods such as data mining and machine learning that examine current and historical facts to predict the future. Predictive methods currently used in hospitals to determine if a patient is at risk of re-admission. These data can help doctors make important patient care decisions. Predictive analysis requires understanding and use of machine learning, and machine learning is widely used in this method.

2.2 Machine Learning in Healthcare
The concept of machine learning is very similar to data mining [9], which scans data to identify patterns, rather than extracting data based on human understanding, in data mining applications; machine learning uses these data to improve program understanding; machine learning recognition Corresponding data mode, then changed the program function.

2.3 Electronic Health Records
EHR represents the broadest range of health applications for big data in the medical field. Each patient has his or her own medical records, including medical history, allergy diagnoses, symptoms, and laboratory test results. Patient records are shared with health care providers in the public and private sectors through a secure information system. These files can be modified because doctors can modify and add new medical test results over time without the need to do paperwork or copy data [10].

3. Four Vs of Big Data in Healthcare
Four primary attributes (shown in Fig. 1) that are associated with big data: volume, velocity, variety, and veracity.
3.1 Volume
Big data first is that the amount of data must be large.

3.2 Velocity
The data growth speed is fast, the processing speed is also fast, the timeliness request is high.

3.3 Variety
Diversity of data types and sources.

3.4 Veracity
The accuracy and reliability of the data, that is, the quality of the data. Data are a reflection of objective reality, and they can be deceptive, which can be amplified when data volumes are small.
4 Impact of Big Data on the Healthcare System

The potential of big data is that it could revolutionize outcomes regarding the most suitable or accurate patient diagnosis and the accuracy information used in the health informatics system. As such, the investigation of huge amounts of information will have a powerful effect on medicinal services framework in five respects, or “pathways” (shown in Fig. 2). Improving outcomes for patients with respect to these pathways, as described below, will be the focus of the healthcare system and will directly impact the patient.

4.1 Right Living
A right life means that the patient has a better and healthier life. Through the right life, patients can make the best decisions, in this paper, we have provided for themselves by using information mining to better choose and enhance their well-being. By choosing the right daily health path (about their diet, preventive health care, exercise, and other activities of daily living), patients can reach a healthy life.

4.2 Right Care
This pathway ensures that patients receive the most appropriate treatment available and that all providers obtain the same data and has the same objectives to avoid redundancy of planning and effort. This aspect has become more viable in the era of big data.

4.3 Right Provider
Healthcare providers in this pathway can obtain an overall view of their patients by combining data from various sources such as medical equipment, public health statistics, and socioeconomic data. The accessibility of this information enables human service providers to conduct targeted investigations and develop the skills and abilities to identify and provide better treatment options to patients.

4.4 Right Innovation
This pathway recognizes that new disease conditions, new treatments, and new medical will continue to evolve. Likewise, advancements in the provision of patient services, for example, upgrading medications and the efficiency of research and development efforts, will enable new ways to promote wellbeing and patient health via national social insurance system. The availability of early trial data is important for stakeholders. This data can be used to explore high-potential targets and identify techniques for improving traditional clinical treatment methods.

4.5 Right Value
To improve the quality and value of health-related services, providers must pay careful and ongoing attention to their patients. Patients must obtain the most beneficial results identified by their social insurance system. Measures that could be taken to ensure the intelligent use of data includes, for example, identifying and destroying data misrepresentation, manipulations, and waste, and improving resources.

5 Hadoop-Based Applications for Health Industry
In light of the fact that healthcare data exists primarily in printed form, there is a need for the active digitization of print form data. The majority of this data is also unstructured, so it is a major challenge for this industry to extract meaningful information regarding patient care, clinical operations, and research. The collection of software utilities known as the Hadoop ecosystem can help the healthcare sector to manage this vast amount of data. The various applications of the Hadoop ecosystem in the healthcare sector are as follows:
5.1 Treatment of Cancer and Genomics
We all know that human DNA contains 3 billion base pairs. In order to fight cancer, it is very important to effectively organize big data. The pattern of cancer mutations varies from individual genetics, which explains the incurability of certain cancers. Oncologists have determined that it is important to provide a specific treatment for a particular cancer based on the patient's genetic makeup when determining the pattern of cancer. Hadoop MapReduce helps map 3 billion DNA base pairs to determine the appropriate cancer treatment for each particular patient. Arizona State University is working on a project to develop a medical model that uses individual genomic data and selects treatment options based on patient cancer gene recognition. This model provides a basis for big data analysis and treatment, and increases the chances of saving patients' lives.

5.2 Monitoring of patient vital
Now, hospital workers around the world are using big data technology to link their work together. They are the basic use of Hadoop components, such as: Hadoop distributed file system (HDFS), common public components, HBase, zookeeper, spark, and the sensor produced by the large amounts of unstructured data, namely the patient's vital signs, every minute heartbeat, blood pressure, blood sugar levels and respiratory rate, all together. Without Hadoop, these healthcare workers would not be able to analyze the unstructured data generated by patient healthcare systems.

5.3 Hospital Network
In the United States, several hospitals use the Hadoop ecosystem's NoSQL database to collect and manage large amounts of their real-time data from a variety of sources related to patient treatment, care, and costs, which helps them identify high-risk patients while reducing overhead.

5.4 Healthcare Intelligence
Hadoop technology also supports the healthcare intelligence applications used by hospitals and insurance companies. Hadoop ecosystem's Pig, Hive, and MapReduce technologies process large datasets related to medicines, diseases, symptoms, opinions, geographic regions, and other factors to extract meaningful information (e.g., desired age) for insurance companies.

5.5 Prevention and Detection of Frauds
In the early days of big data application, health-based insurance companies used multiple ways to identify fraudulent behaviors and establish methods to prevent medical fraud. Hadoop big data can also prevent fraud related to medical claims at an early stage by using real-time health applications based on Hadoop, real claims billing, weather forecasts, voice data, and other data sources.

6 Big data analysis framework of health information
At present, the main content of big data analysis is to conduct in-depth insight and understanding of big data, rather than to collect big data. Data analysis involves developing and applying algorithms to analyze complex data to extract meaningful information. In recent years, researchers have begun to consider the appropriate architecture for big data analytics healthcare systems. One architecture adopts a four-tier architecture consisting of a transformation layer, a data source layer, a big data platform layer and an analysis layer. In this hierarchical system, data comes from different data sources and has different formats and storage systems. Each layer has specific data processing capabilities and can perform specific tasks on HDFS using the MapReduce processing model. Other layers perform other tasks, namely report generation, query delivery, data mining processing, and online analysis processing. The main requirement of big data analysis and processing is to package data at a high speed to minimize the packaging time.

Big data technology, broadly speaking, refers to scientific innovations that mimic those used in large data sets. The first component deals with requirements for large data sources; In the second component cluster, the cluster of components with a centralized big data processing infrastructure is at
the pinnacle of high performance. It has been observed that tools mainly used for big data analysis and processing provide data security, scalability and manageability with the help of MapReduce paradigm. In the third component, the big data analysis application has a storage domain for integrating accessed databases using different applications. In the fourth component are the most popular big data analysis applications in healthcare systems, including reporting, online analytics processing (OLAP), query, and data mining.

As shown in figure 3, medical data comes from a wide range of sources, including EHRs, genomic databases, genomic data files, texts and images (unstructured data sources), clinical decision support systems, government-related resources, medical testing laboratories and pharmacies, and health insurance companies. This data is typically available in different scheme tables and stored in different locations in ASCII/text format. In the next section, we describe the various processing tools based on bigdata hadoop that support the development of health-based applications for the health industry.

7 Big data tools and technologies
In order to manage unstructured big data, we cannot use traditional relational databases. IT departments have developed a variety of ways to document, organize, and analyze this type of data using the Hadoop platform. More efficient tools are needed to extract meaningful output from big data. Most tools are implemented in Apache Hadoop architecture, including MapReduce, HDFS, Pig, Hbase, Oozie, Mahout, Hive, Zookeeper, Yarn, Sqoop, etc.

8 Conclusion
In this paper, we have carried out an in-depth description and a brief overview of the application of big data and medical system. Big data plays an important role in medical informatization and has a huge impact on the four v's of medical system and medical big data. We also propose a conceptual architecture for solving medical problems, using big data and using the term Hadoop based, including using big data, analyzing the development of different levels of medical data, and obtaining answers to medical problems. The combination of big data and medical analysis could lead to effective treatments for specific patients by prescribing the right drugs for everyone, rather than the most. As we all know, big data analysis is still in the early stage of development, and the existing tools and methods cannot solve the problems related to big data. Big data can be seen as a big system, facing huge challenges. Therefore, solving the problems faced by the health care system requires a lot of research in this field.

Acknowledgments
The project was supported by the Guangdong Province Promotional Service Trade Innovation and Development Project (Fund Code: 2160699-87) and the C++ Programming Practice Teaching Reform (Project No. SYJG201803).

References
[1] A. Gandomi and Beyond the hype: Big data concepts, methods and analytics, International Journal of Information Management, vol. 35, no. 2, pp. 136–146, 2015.
[2] A. O’Driscoll and R. D. Sleator, “Big Data”, Hadoop and cloud computing in genomics, Journal of Biomedical Informatics, vol. 46, no. 5, pp. 773–782, 2013.
[3] C. L. P. Chen, Data-intensive applications, challenges, techniques and technologies: A survey on big data, Information Sciences, vol. 275, pp. 313–348, 2014.
[4] M. Herland, and R. Wald, a review of data mining using big data in health informatics, Journal of Big Data, vol. 1, no. 1, p. 1-3, 2014.
[5] D. H. Shin, Ecological views of big data: Perspective and issues, Telematics and Informatics, vol. 32, no. 2, pp. 310–321, 2015.
[6] B. Saraladevi and P. Dhavachelvan, Big data and Hadoop-A study in security perspective, Procedia Computer Science, vol. 50, pp. 595–602, 2015.
[7] X. Wu and W. Ding, Data mining with big data, IEEE transactions on Knowledge and Data Engineering, vol. 26, no. 1, pp. 96–108, 2014.
[8] S. Sharmas and V. Mangats, Technology and trends to handle big data: Survey, in Proc. 5th International Conference on Advanced Computing & Communication Technologies, 2015, pp. 265–272.

[9] R. Mehmoodl and G. Grahamn, Big data logistics: A healthcare transport capacity sharing model, Procedia Computer Science, vol. 64, pp. 1106–1115, 2015.

[10] D. P. Augustines, Leveraging big data analytics and Hadoop in developing India healthcare services, International Journal of Computer Applications, vol. 89, no. 16, pp. 43–51, 2014.