

Big Data in Medical Research and Application

Teng Yilin¹, Cao Gaofang¹, and Wang Rui^{1,*}

¹ Binzhou Medical University Binhai Road 346#, Laishan District, Yantai City, Shandong Province

Abstract. With the advanced instruments and information technology integrated in biomedical science more and more extensively, the advent of the big data era has had a significant impact on biomedical research, making human awareness of themselves and diseases more profound. The future medicine tends to combine data and medicine, to master gene database and medical human disease data, then to apply data statistics and analysis and application in healthcare. New techniques of big data in medicine are bound to make medical research and application more predictable. This paper introduces the main sources and characteristics of medical big data, points out the necessity of big data research in medical field, summarizes the current research in medical big data and its application in disease prediction, clinical assistance and pharmaceutical research and development. In other aspects, it analyses the problems of medical big data in applied research.

1 Introduction

With the development of information technology, the era of big data came since the rapid growth of data volume. At the beginning, big data refers to a set of data that are too large and complex for traditional data processing methods to deal with [1]. Later, big data tends to be the use of predictive analytics, user behaviour analytics, or certain other advanced data analytics methods that extract value from data, while seldom to a particular size of data set [2]. Nowadays, data resources become the same important strategic resources as natural resources, and it became the source of new inventions and new technologies as well. Although the data is priceless, but its value is like to extract the essential from a large mass of material. It means data fusion will produce far greater value than value of the single data [3].

The significance of big data is to provide "big insights": collect information from different sources, and then analyse the information to reveal trends that cannot be found by traditional methods. Among all the industries which discover value from big data, medicine industry is likely to achieve the greatest return so that it is an important field about big data application. With big data, healthcare providers can not only know how to improve profitability and operational efficiency, but also find a way to promote human health level directly. How to develop and utilize the medical big data has become the focus of attention.

2 Medical Big Data

According to the International Data Companies (IDC), the big data market in China will increase five times in 2012 - 2016, where the largest portion of it will be concentrated in the four major industries of government, banking, medicine and telecommunications [4]. Groves P, Kayyali B et al. [5] put forward that big data will have an important influence on the American medical industry, and many potential values are gradually revealed. Eric [6] pointed out that big data would lead "creative destruction" to the entire medical system, including management and diagnostic. This kind of destruction may take place in medical clinical diagnosis data, such as medical images, records, examination, diagnosis and treatment expenses, etc.; it may also take place in drug research and development, epidemiological investigation, pathogenesis, physiological and pathological changes, etc. Thus big data in medicine tends to be direction of future medical development.

2.1 The concept of medical big data

In a 2001 research report and related lectures, META group [7] used "3Vs" model for describing big data, i.e. increasing volume (amount of data), velocity (speed of data in and out), and variety (range of data types and sources). Then De Mauro [8] added that "Big Data represents the Information assets characterized by

* Corresponding author: oucwangrui@163.com

such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value". IBM cooperation expanded the 3Vs of big to more other complementary characteristics of:

Volume: big data doesn't sample; it just observes and tracks what happens

Velocity: big data is often available in real-time

Variety: big data draws from text, images, audio, video; plus it completes missing pieces through data fusion

Value: the value of the data is low relative. It means huge amounts of information, but low density value.

Veracity: data needs to be accuracy and trustable, that is, the quality of data.

On the premise of satisfying these basic characteristics of big data, data related to the medical and health service process and results can be called the medical big data.

2.2 The source of medical big data

The source of medical big data can be divided into medical data and health data. Medical data includes hospital clinical data, research data of institutions, etc.; health data includes public health monitoring data, personal health behaviour data, etc. **Table 1** shows the specific types of data sources

Table 1. source of medical big data.

	source	types
Medical data	Clinical medicine	Electronic medical records, medical images, inspection and detection reports, medical equipment testing, etc
	Pharmaceutical r&d	Clinical trials, drug research and development, medical equipment r & d
	Biomedical Science	Genomics, metabolomics, proteomics, etc
Health data	Public health	Population basic information, life death record, disease detection, nutrition health survey, health risk factors investigation, etc
	Personal health	Wearable equipment, household testing instruments, fitness sports records, drug purchase records and so on
	Health network	Health website or app behavior record, search engine, wechat micro blog, instant messaging, etc

3 Application of medical big data

3.1 Disease prediction

The trend of big data development is to overcome the shortcomings of traditional sampling survey which is

unable to fully and accurately predict the results. Therefore, medical big data can be well applied to real-time biological monitoring and public health monitoring, public health monitoring includes infectious disease surveillance, chronic non-communicable diseases and related risk factors monitoring. In addition, the epidemic situation monitoring can be carried out through the large data of electronic medical records, and the epidemic of infectious diseases can be predicted through the network information.

In 2009, Google used to predict the spread of winter flu, and a wave of big data changes in public health. Recently, Google announced that Google X hatched a medical health project called "baseline", using big data to prevent cancer. The project is thought to be Google's most ambitious and most difficult project ever. "Baseline" uses the massive storage and computing power to bombard a very large amount of the gene and molecular information in the human body. According to the data statistics, the human body structure data of healthy people are obtained, and then the abnormality of the molecular level of the individual is found.

It can be noticed that multiple data fusion will lead to a variety of diseases prediction. In molecular biology, "Omics" mainly includes Genomics, Proteomics, Metabolomics, transcriptomics, lipidomics, Immunomics, glycomics, RNAomics and so on. The root "-ome" translation is a collection of individual systems, such as genomics, which is a combination of all the genes in an organism. the subject is to study the genes and the relationships between these genes.

From public health and personal health data to the level of individual cell biology "omics".The use of big data mining, data fusion and data analysis methods can provide a comprehensive, new understanding of the occurrence and prevention of disease, also conduct to the prevention of individual diseases in addition.

The 2011 e. coli O104 incident in Germany caused worldwide concern. In the field of bacterial infectious diseases, new and emergent bacterial infections are particularly arrested. Genomics provides a new powerful technology in Screening and identification of infectious diseases, tracing of the source of infectious diseases, and variation of pathogenic factors. Data analysis for infectious disease shows that big data played a good performance.

Another case is Human Connectome project, which is a comprehensive map of neural connections in the brain. More broadly, a connectome would include the mapping of all neural connections within an organism's nervous system.

With the help of artificial intelligence, neuroscientists using machine learning system abroad, draw a map of the outermost equivalent unprecedented human brain, the subdivision of the cerebral cortex, identified the lines on each hemisphere of the brain, and split it into 180 cortex region, of which 97 area has not been described before, but now clarify the architecture, function, and the different between them.

The new map of the human brain by Glasser, M. F. et al. [9] using magnetic resonance imaging (MRI) scan of 210 healthy adults, using a variety of image data, combined with machine learning system. Figure 1 shows the human brain divided into discrete areas based on structure and function.

Research of “Human Connectome” could investigate the association between brain network and cognitive change caused by dementia, schizophrenia or related genetic characteristics and clinical symptoms. It enhances the understanding of the pathogenesis of neuropsychiatric diseases, the characteristics of brain damage and the mechanism of brain reorganization, and provides a scientific basis for the prevention, diagnosis and early treatment of these diseases by utilizing the big data of MRI and other materials.

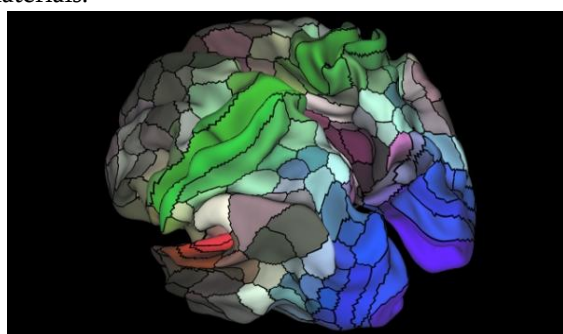


Figure 1: Map of human brain

3.2 Clinical assistance

Medical big data can assist clinical decision, it can improve diagnosis efficiency and treatment quality. The most advanced application of medical big data assistance is “precision medicine”[10]. Precise medical treatment, also called personalized medical treatment, refers to a customized medical model, which is based on personal genomic information, combining with relevant environmental information such as proteomics and metabonomics, and designing the most proper treatment scheme for patients, so as to achieve the maximum therapeutic effect and minimize side effects. Figure 2 shows that medical big data of Genomics can predict the treatment effect. It can be found that when there are two “G”s in the gene, the

treatment would be effective, and only one “G” would lead to the non-effectiveness. What’s more, if there is none “G”, the patients would suffer severe side effect.

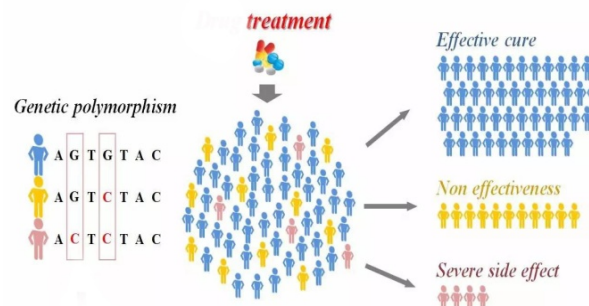


Figure 2. Treatment effect prediction

The evidence-based method of traditional medicine in our country is watch ,hear, ask and touch, even though the birth of medical image provides a deeper basis for doctors' diagnosis, it is still a sampling method. Doctors can get information only from the surface of individual patients, thus the experience is likely to be more important on the accuracy of disease judgment. Furthermore, the existing technology accurate medical image resolution to 3mm, but it is hard for human eyes to extract more widely distributed information. In other hand, the big data techniques can realize the same homogenous disease affect data reconstruction, such as low resolution reconstruction, 3D reconstruction, feature (texture, histogram, image entropy) analysis, then obtain more accurate information of the focus.

We take cancer treatment as an example; doctors could only see the institutional information from CT films, while this information reflected a little. And the accuracy rate is about 30%. But if we use big data according to the clinical pathology and image data of more than 500 cases, extraction characteristics, modelling analysis, the accuracy of front-end data prediction can reach 70 %. In other words, the doctor can reduce the past mistakes from 70 % to 30 %. At the same time, big data can aid doctor’s select which is the most effective method for cancer treatment.

3.3 Pharmaceutical R & D

Using data from a group of people with a particular disease can quickly identify biomarkers associated with the occurrence, prognosis, or therapeutic effects of the disease. Big data makes the pathogenesis of ethology and disease more in-depth understanding, full use and is able to accelerate drug screening process.

The prediction model of biological processes and drugs becomes more complex and widely used. By

using molecular and clinical data, predictive modelling can help identify potential new molecules that are highly likely to be developed successfully as drugs. Meanwhile there are more channels to recruit accuracy patients into clinical trials, such as social media. Screening criteria for clinical trials will also take into account more factors such as genetic information to target specific populations, so that clinical trials can be smaller, shorter, and less costly and more effective.

4 The problem of medical big data

The research and application of medical big data in medical treatment is very necessary, while at the same time, there are some problems in the application, which can be summarized as follows:

4.1 Data sharing

Data sharing is the basic of the application of medical big data, it should be paid attention that how to break the data island and avoid data use for only a few people or organization. It is necessary to realize the standardization of medical data, which is the premise of data sharing. Data sharing can effectively carry on data integration and data fusion, and obtain more valuable information.

4.2 Data security

How to store and manage data effectively and safely. The volume of medical data is large, the update speed is fast, and because of individual differences, the data is not repeatable, and involves the patient privacy, data security and proper management is strongly demanded.

4.3 Data analysis

How to effectively analyse, integrate and mine data, especially the semi-structured and unstructured image data are the difficulties and challenges of big data analysis. Big data applications require team collaboration, besides technology, data sources. It needs the inter-disciplinary talent who can not only master data mining, parallel computing algorithms, but also understand of biomedical knowledge. Thus, to promote the development of the interdisciplinary of computer science and biomedicine is conducive to the future development of medical big data application that

5. Conclusion

The arrival of big data era has penetrated into varies industries of life. There are a huge amount of data in medical field, its significance not only lies in the amount of data, but also in the deeper value obtained through the processing of these data. The development of medical treatment is also constantly shifting to data-driven, to obtain the optimal decision and program. In general, the development of medical big data is still in the early stage, how to share, standardize, manage and develop these valuable data is the key. Medical big data will change the medical practice mode, improve the quality of medical and health service, and ultimately benefit the medical purpose of individualized treatment and group prevention.

References

1. Snijders, C.; Matzat, U., Reips, U.-D... 'Big Data': Big gaps of knowledge in the field of Internet. International Journal of Internet Science(2012)
2. Viktor Mayer-Schonberger, Kenneth Cukier. *Big Data: A Revolution That Will Transform How We Live, Work and Think*. John Murray Publishers, London (2012)
3. Zhao Guodong, Yi Huanhuan, Mi Wanjun. *Big Data Revolution: How Data Science Changes the World* Tsinghua university press , Beijing.(2013)
4. Groves P, Kayyali B, Knott D, et al.. *The big data revolution in healthcare*. Accelerating value and innovation. Mckinsey & Company.(2013)
5. Inmon W..*Building the data warehouse*. John Wiley & Sons Inc, New York (2002)
6. Eric J. Topol *The Creative Destruction of Medicine Basic Books*,New York (2012)
7. Laney, Douglas. *3D Data Management: Controlling Data Volume, Velocity and Variety* . Gartner.
8. De Mauro, Andrea; Greco, Marco; Grimaldi, Michele. "A Formal definition of Big Data based on its essential Features". Library Review.(2016)
9. Glasser, M. F. et al. Nature (2016)
<http://dx.doi.org/10.1038/nature18933> .
10. National Research Council, Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease (NRC,2011)