International Journal of Computer Engineering in Research Trends
Multidisciplinary, Open Access, Peer-Reviewed and fully refereed
Review Paper Volume-6, Issue-05, May 2019 Regular E-ISSN: 2349-7084

Big data in healthcare: Challenges and approaches

P.Murthuja[1*]

[1*]*Associate professor, Prabhath Institute of Computer Science, Parnapalli village , Bandi atmakur, Kurnool District, Andhra Pradesh, India*
Email: pmurthuja.mca@hotmail.com .

Available online at: http://www.ijcert.org

Received: 16/05/2019, Revised: 23/05/2019, Accepted: 29/05/2019, Published: 06/06/2019

Abstract: - Now a day's huge volume of data is generated due to wide usage of social media, online shopping or transactions gives delivery to big data. Visual representation and analysis of this large volume of data is one of the major research topics today. Healthcare is one of the most promising areas for using big data for change. Big data healthcare has enormous potential to improve patient outcomes, obtain valuable information, prevent disease, reduce healthcare delivery costs and improve quality of life. In this paper i focus on challenges associated with healthcare big data and also explore the common approaches for analysing big data in health Care system.

Keywords: Healthcare, Big data analytics, Internet of things, personalized medicine.

1. INTRODUCTION

Big Data, the generic term for data sets of structured and unstructured data that are extremely large and complex so that the traditional software, algorithm, and data repositories are inadequate to collect, process, analyze, and store them has become an intensively studied area in recent years. With the development of the Internet, the mobile Internet, the Internet of things, social media, biology, finance, and digital medicine, the volume of data has increased dramatically. Big Data not only describes the large size of data as its name suggests but also implies rapid data processing ability and novel technology and approaches for handling the data. To imagine this size, we would have to assign about 45 zettabyte (ZB) of data to all individuals. This exemplifies the Phenomenal speed at which the digital universe is expanding.

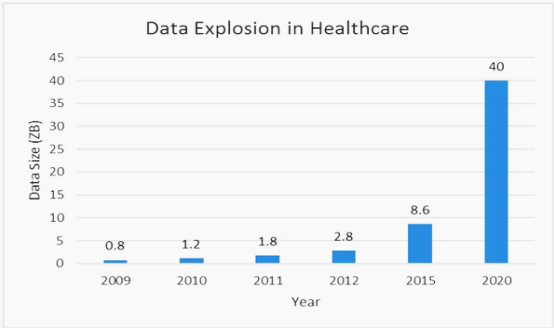

Figure 1. Data Exploration in Healthcare

The internet giants, like Google and Facebook, have been collecting and storing massive amounts of data. For instance, depending on our preferences, Google may store a variety of information including user location, advertisement preferences, list of

applications used, internet browsing history, contacts, bookmarks, emails, and other necessary information associated with the user. Similarly, Facebook stores and analyzes more than about 30 petabytes (PB) of user-generated data. Such large amounts of data constitute 'big data'. Over the past decade, big data has been successfully used by the IT industry to generate critical information that can generate significant revenue.

The term "big data" has become extremely popular across the globe in recent years. Almost every sector of research, whether it relates to industry or academics, is generating and analyzing big data for various purposes. The most challenging task regarding this huge heap of data which can be organized and unorganized, is its management. Given the fact that big data is unmanageable using the traditional software, we need technically advanced applications and software that can utilize fast and cost-efficient high-end computational power for such tasks. Implementation of artificial intelligence (AI) algorithms and novel fusion algorithms would be necessary to make sense from this large amount of data. Indeed, it would be a great feat to achieve automated decision-making by the implementation of machine learning (ML) methods like neural networks and other AI techniques. However, in absence of appropriate software and hardware support, big data can be quite hazy. We need to develop better techniques to handle this 'endless sea' of data and smart web applications for efficient analysis to gain workable insights. With proper storage and analytical tools in hand, the information and insights derived from big data can make the critical social infrastructure components and services (like health-care, safety or transportation) more aware, interactive and efficient [1]. In addition, visualization of big data in a user-friendly manner will be a critical factor for societal development. Figure 1 presents the main elements in big data lifecycle in healthcare.

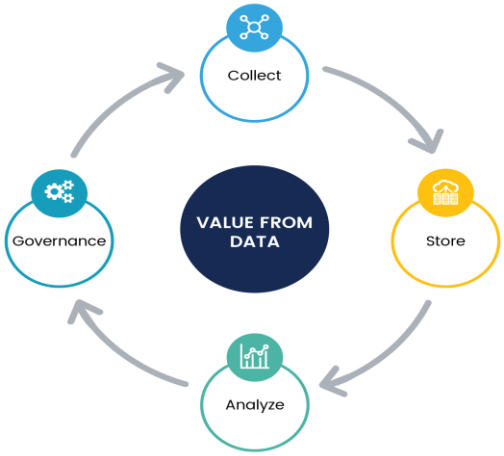

Fig 2 life cycle of big data in healthcare

Similar to EHR, an electronic medical record (EMR) stores the standard medical and clinical data gathered from the patients. EHRs, EMRs, personal health record (PHR), medical practice management software (MPM), and many other healthcare data components collectively have the potential to improve the quality, service efficiency, and costs of healthcare along with the reduction of medical errors. The big data in health-care includes the healthcare payer-provider data (such as EMRs, pharmacy prescription, and insurance records) along with the genomics-driven experiments (such as genotyping, gene expression data) and other data acquired from the smart web of internet of things (IoT). The adoption of EHRs was slow at the beginning of the 21st century however it has grown substantially after 2009 [2, 3]. The management and usage of such healthcare data has been increasingly dependent on information technology. The development and usage of wellness monitoring devices and related software that can generate alerts and share the health related data of a patient with the respective health care providers has gained momentum, especially in establishing a real-time biomedical and health monitoring system. These devices are generating a huge amount of data that can be analyzed to provide real-time clinical or medical care [4]. The use of big data from healthcare shows promise for improving health outcomes and controlling costs.

2. NATURE OF THE BIG DATA IN HEALTHCARE

EHRs can enable advanced analytics and help clinical decision-making by providing enormous

data. However, a large proportion of this data is currently unstructured in nature. An unstructured data is the information that does not adhere to a pre-defined model or organizational framework. The reason for this choice may simply be that we can record it in a myriad of formats. Another reason for opting unstructured for-mat is that often the structured input options (drop-down menus, radio buttons, and check boxes) can fall short for capturing data of complex nature. For example, we cannot record the non-standard data regarding a patient's clinical suspicions, socioeconomic data, patient preferences, key lifestyle factors, and other related information in any other way but an unstructured format. It is difficult to group such varied, yet critical, sources of information into an intuitive or unified data format for further analysis using algorithms to understand and leverage the patients care. Nonetheless, the healthcare industry is required to utilize the full potential of these rich streams of information to enhance the patient experience.

In the healthcare sector, it could materialize in terms of better management, care and low-cost treatments. We are miles away from realizing the benefits of big data in a meaningful way and harnessing the insights that come from it. In order to achieve these goals, we need to manage and analyze the big data in a systematic manner.

Major Date Types of Big Data in Health Care

Data type	Data name	Data description	Data acquisition
Big Data in medicine and clinics	Electronic health record (EHR)/ electronic medical record (EMR)	Standard data collection of medical and health information for patients and can be shared in different organizations Often comes from medical activities and public health data	Hospital information resource, surgery's work, activities of anesthesia, physical examination, radiography, magnetic resonance imaging (MRI), computer tomography (CT), information of patient, phamacy, treatment, medical imaging, imaging report, identification information of patient, clinical diagnosis, medicine scheme, notes from physician, sensor data patient demographics, clinic or inpatient notes, electronic reports
	Personal health record (PHR)	As its name suggests, it is the health-related data and information of patients and about people's lifelong health information. It is available for further use	Allergies and adverse drug reactions, chronic diseases, family history, illnesses and hospitalizations, imaging reports, laboratory test results, medications and dosing, prescription record, surgeries and other procedures, vaccinations and observations of daily living, and reported by patients .
	Medical images	Data that present visual information of interior human body	X-ray, CT, histology, positron- emission tomography (PET), radiography, MRI, nuclear medicine, elastography, tactile imaging, photo acoustic imaging, echocardiography ultrasonography, angiography

3. CHALLENGES ASSOCIATED WITH HEALTHCARE BIG DATA

Methods for big data management and analysis are being continuously developed especially for real-time data streaming, capture, aggregation, analytics (using ML and predictive), and visualization solutions that can help integrate a better utilization of EMRs with the healthcare. For example, the EHR adoption rate of federally tested and certified EHR programs in the healthcare sector in the U.S.A. is nearly complete [2]. However, the availability of hundreds of EHR products certified by the government, each with different clinical terminologies, technical specifications, and functional capabilities has led to difficulties in the interoperability and sharing of data. Nonetheless, we can safely say that the healthcare industry has entered into a 'post-EMR' deployment phase. Now, the main objective is to gain actionable insights from these vast amounts of data collected as EMRs. Here, we discuss some of these challenges in brief.

Storage: Storing large volume of data is one of the primary challenges, but many organizations are comfortable with data storage on their own premises. It has several advantages like control over security, access, and up-time. However, an on-site server network can be expensive to scale and difficult to maintain. It appears that with decreasing costs and increasing reliability, the cloud-based storage using IT infrastructure is a better option which most of the

healthcare organizations have opted for. Organizations must choose cloud-partners that understand the importance of healthcare-specific compliance and security issues. Additionally, cloud storage offers lower up-front costs, nimble disaster recovery, and easier expansion. Organizations can also have a hybrid approach to their data storage programs, which may be the most flexible and workable approach for providers with varying data access and storage needs.

Cleaning: The data needs to cleansed or scrubbed to ensure the accuracy, correctness, consistency, relevancy, and purity after acquisition. Tis cleaning process can be manual or automatized using logic rules to ensure high levels of accuracy and integrity. More sophisticated and precise tools use machine-learning techniques to reduce time and expenses and to stop foul data from derailing big data projects.

Unified format: Patients produce a huge volume of data that is not easy to capture with traditional EHR format, as it is knotty and not easily manageable. It is too difficult to handle big data especially when it comes without a perfect data organization to the healthcare providers. A need to codify all the clinically relevant information surfaced for the purpose of claims, billing purposes, and clinical analytics. Therefore, medical coding systems like Current Procedural Terminology (CPT) and International Classification of Diseases (ICD) code sets were developed to represent the core clinical concepts. However, these code sets have their own limitations.

Accuracy: Some studies have observed that the reporting of patient data into EMRs or EHRs is not entirely accurate yet [5-8], probably because of poor EHR utility, complex workflows, and a broken understanding of why big data is all-important to capture well. All these factors can contribute to the quality issues for big data all along its lifecycle. The EHRs intend to improve the quality and communication of data in clinical workflows though reports indicate discrepancies in these contexts. The documentation quality might improve by using self-report questionnaires from patients for their symptoms.

Image preprocessing: Studies have observed various physical factors that can lead to altered data quality and misinterpretations from existing medical records [9]. Medical images often suffer technical barriers that involve multiple types of noise and artifacts. Improper handling of medical images can also cause tampering of images for instance might lead to delineation of anatomical structures such as are some of the measures that can be implemented to benefit the purpose.

Data sharing: Patients may or may not receive their care at multiple locations. In the former case, sharing data with other healthcare organizations would be essential. During such sharing, if the data is not interoperable then data movement between disparate organizations could be severely curtailed. Tis could be due to technical and organizational barriers. Tis may leave clinicians without key information for making decisions regarding follow-ups and treatment strategies for patients. Solutions like Fast Healthcare Interoperability Resource (FHIR) and public APIs, Common Well (a not-for-profit trade association) and Care quality (a consensus-built, common interoperability framework) are making data interoperability and sharing easy and secure. The biggest roadblock for data sharing is the treatment of data as a commodity that can provide a competitive advantage. Therefore, sometimes both providers and vendors intentionally interfere with the flow of information to block the information flow between different EHR systems [10].

4. Common Approaches for Analyzing Big Data in Health Care

With the growing awareness of data as an asset, more and more data-mining approaches are adopted in order to gain insights from large volumes of data. In medicine and health care, a data-rich environment generates an enormous amount of data every day. Thus, we need to use data-mining approaches such as classification, clustering, regression analysis, and association rules to analyze big health care data.

Classification: Classification is the process of organizing data into categories for its most effective and efficient use. Classification is widely applied in mining health care data. There are some specific introductions in these areas.

Clustering: Clustering is the task of grouping a set of objects in such a way that objects in the same cluster are more similar to each other than those in other clusters. Clustering techniques are widely used for exploratory data analysis, with applications including patient segmentation, outlier health care data detection, disease prediction, and clustering of patients.

Regression analysis: Regression analysis is widely used in analyzing health care Big Data for estimating the relationships among variables or properties. The main research issues include trend features of data sequences, prediction of data sequences, and relationships between data.

Association rules: In a medical database, the most complete and detailed information is anamnesis data, which contain disease name, prescription, patient's detail information, etc. Through this method, it is possible to find the association rules between diseases

5. Conclusion and Future Scope

The potential of big data in healthcare is endless, e.g. There are many barriers to its true potential, including health research, knowledge discovery, clinical care and personal health management), technical issues, privacy and security issues, and skilled talent. Big data security and confidentiality are seen as a huge barrier for researchers in this field. In this paper, I explore different medical data types in big data, Challenges associated with healthcare big data and also presented common Approaches for Analyzing Big Data in Health system this context, our future direction will include the investigate and propose Data cleaning techniques over healthcare dataset for improving performance of chronic disease classification and predication.

References

[1] Hendrik Schuff, Jeremy Barnes, Julian Mohme, Sebastian Pado, and Roman Klinger," "Annotation, Modelling and Analysis of Fine-Grained Emotions on a Stance and Sentiment Detection Corpus " by" , Proceedings of the 8th Workshop on Computational Approaches to Subjectivity , Sentiment and Social media Analysis, pages 13-23, Copenhagen, Denmark, September 7-11, 2017.

[2] Sunidhi Sharma, D.K.Sharma, Supriti Sharma, "Text Analysis and Sentiment Analysis using Facebook in R Language: Case studies" , International Journal of Computer and Mathematical Sciences, ISSN 2347-8527, vol 6 Issue 12, December 2017.

[3] Edison Marrese-Taylor, Jorge A. Balazs, Yutaka Matsuo, "Mining fine-grained opinions on closed captions of YouTube videos with an attention-RNN" Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social media Analysis, pages 102-111, Copenhagen, Denmark, September 7-11, 2017.

[4] Kishaloy Halder, Lahari Poddar, Min-Yen Kan, "Modeling Temporal Progression of Emotional Status in Mental Health Forum : a Recurrent Neural Net Approach" Proceedings of Workshop on Computational Approaches to Subjectivity , Sentiment and Social media Analysis, pages 127-135, Copenhagen, Denmark, September 7-11, 2017.

[5] Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde, "Assessing State-of-the-Art Sentiment Models on State-of-the-Art Sentiment Datasets" Proceedings of the 8th Workshop on Computational Approaches to Subjectivity , Sentiment and Social media Analysis, pages 2-12, Copenhagen, Denmark, September 7-11, 2017.

[6] Keenen Cates, Pengcheng Xiao, *, Zeyhang, Calvin Dailey, "Can Emoticons Be Used To Predict Sentiment?" Journal of Data Science 355-376, April 04, 2018.

[7] Prabaharan Poornachandran,, "deepCybErNet at EmoInt-2017: Deep Emotion Intensities in Tweets" Proceedings of the 8th Workshop on Computational Approaches to Subjectivity , Sentiment and Social media Analysis, pages 102-111, Copenhagen, Denmark, September 7-11, 2017.

[8] Murphy G, Hanken MA, Waters K. Electronic health records: changing the vision. Philadelphia: Saunders W B Co;1999. p. 627.

[9] Shameer K, et al. Translational bioinformatics in the era of real-time biomedical, health care and wellness data streams. Brief Bioinform. 2017;18(1):105–24.

[10] Service, R.F. The race for the $1000 genome. Science. 2006;311(5767):1544–6.

