A Survey of Big Data in Healthcare Industry

Indu Khatri and Virendra Kumar Shrivastava

Abstract "Big Data" are data that are big not only in terms of "Volume" but also in terms of "Value". The exploration of big data in healthcare is increasing at an unprecedented rate. The credit goes to the advanced technologies that help to collect medical data from various sources and to the initiatives that bring deeper insights from these data. This paper presents the exceptional work done by corporations, educational institutions and governments leveraging big data to solve the problems and challenges pertaining to healthcare industry. This paper addresses the ongoing researches; researches that are in initial stages or that are mentioned in the Press Releases to show the advancement of big data in healthcare industry. The paper also proposes a common platform for healthcare analytics, aimed to reduce the redundancy in the techniques that are required in any kind of medical research.

Keywords Big · Data · Cognitive computing · Image processing · Data mining · Brain imaging · Medical imaging · Drug discovery · Diseases detection · Machine learning and deep learning algorithm

1 Introduction

"More Data = More Power + More Benefits" is the theme of this era. The power to retrieve and study heterogeneous healthcare data helps healthcare providers to deliver right intervention to the right patient at the right time and right cost. The healthcare industry generates huge amount of data i.e. in petabyte from Electronic

I. Khatri (✉)
Department of IP Operations & Program Management,
Cerner Healthcare, Bangalore 560045, Karnataka, India
e-mail: indu.khatri@hotmail.com

V.K. Shrivastava
Department of Computer Science & Engineering, APIIT SD India,
Panipat 132103, Haryana, India
e-mail: virendra@apiit.edu.in

R.K. Choudhary et al. (eds.), *Advanced Computing and Communication Technologies*,
Advances in Intelligent Systems and Computing 452,
DOI 10.1007/978-981-10-1023-1_25

Health Records, Clinical Notes, Medical Images, Wearable sensors, Mobile Devices, Genomic Sequences and Social Media etc.

A study conducted in 2014 by EMC Digital Universe (with research and analysis by IDC) shows that healthcare data is increasing at the rate of 48 % per year, establishing Healthcare as one of the fastest growing segments of the market [1]. The data are estimated to increase up to 2,314 Exabytes (10^{18}) as per the study within next 5 years. The expanding volume of healthcare data unlocks novel opportunities to bring novel life changing insights while improving patient care [2].

The clinical data can be leveraged to fight against diseases by collecting small measurements and trying to find the diseases through machine centric scalable models than human-centric limited experience. Big data can be revolutionary for those who are struggling with deadly diseases. It takes a long time to detect the diseases. Using image processing and machine learning based classification models, diseases can be detected at very early stages. There are 6V's (Volume, Velocity, Variety, Veracity, Validity, and Volatility) of big data, evolving into value of data [3]. Many organizations have already started leveraging big data to solve their day-to-day problems. The other healthcare stakeholders can be encouraged with the means of this paper to accelerate the usage of big data for better insights in healthcare analytics. The quality of value can be increased while the costs can be reduced. The other set of industries such as Information and Technology, Electronics and Communication, and Analytics are bringing innovative technologies at a faster pace to help analyze these data.

The paper is organized as follows. Section 2 provides an overview of healthcare system. In Sect. 3, the paper discusses about big data initiatives in healthcare and recent innovations i.e. IBM Watson—an artificial intelligent computer system, Stanford and Google's collaboration on drug discovery, Calico, Frost and Sullivan's market research, Pittsburgh's Health Data Alliance and Human Brain Project. Section 4 talks about challenges in uplifting healthcare using big data. Section 5 explores future outlook: initiative—a common platform for healthcare big data and analytics. Last but not the least the study is concluded in Sect. 6.

2 Overview Healthcare System

Ferlie and Shortell [4, 5] have given healthcare system structure that is explained via four-level Model. According to this model, the healthcare system revolves around following four nested stages.

2.1 Patient

The first stage, patient is an element whose care defines the healthcare system. Lately, patients have started playing the role of active rather than passive consumers. An active patient is one who wants to be involved in the analysis, design,

implementation and maintenance (coordination) of his/her care. The latest technologies enable patients to share their physiological real-time data with physicians (even from remote location), accelerates the speed of diagnosis and treatment. This is the stage which is encircled by all other stages.

2.2 Care Providers

The next stage, patient is an element whose care defines the healthcare system categorizes everyone who provides care to the patient. Healthcare professionals such as physicians, doctors, nurses, pharmacies and even the patients' family members who contribute to the delivery of care are designated as "Care Team". The care providers analyze the patients' data to standardize care, stratify patients and synthesize best decisions/care for their health.

2.3 Organization

The third stage is the organization that offers the infrastructure and other necessary resources to physically perform the care related work. The hospitals, clinic, community hospitals, nursing homes etc. are part of this stage. According to Ferlie and Shortell, "organization is an acute level that sets the culture for change via decision-making systems, operating systems and human resources."

2.4 Environment

The final stage includes political and economic environment that sets the regulatory, market, and policy framework. Public and private regulators, insurers, healthcare purchasers, research funders are few elements who act as pillars of healthcare organizations and other stages (Figs. 1, 2 and 3).

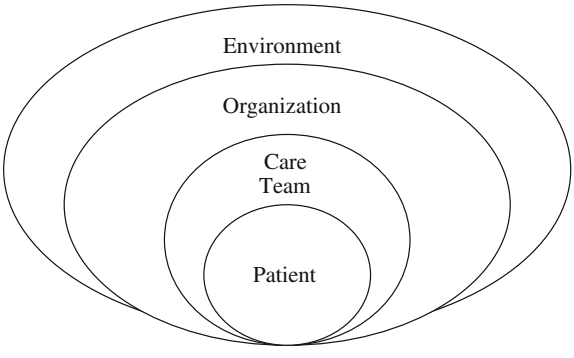

Fig. 1 Four-level healthcare model by Ferlie and Shortell shown in this sample caption [5]

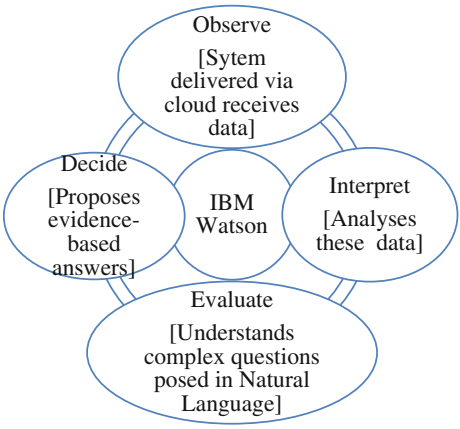

Fig. 2 Common cognitive framework used by humans and implemented for IBM Watson [8]

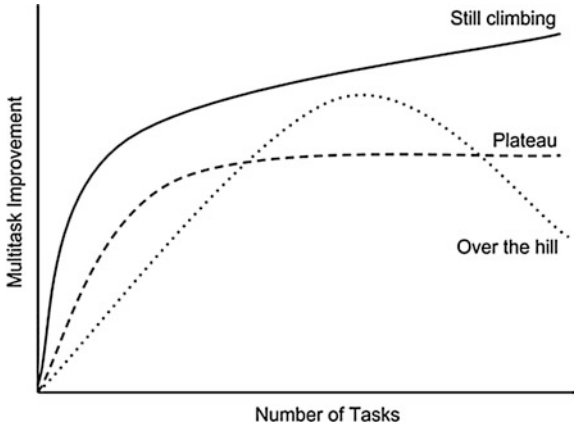

Fig. 3 Potential graph to show that prediction power improves as data; processes/tasks increase [12]

3 Big Data Initiatives in Healthcare

This paper provides a glimpse of few recent and innovative big data initiatives in healthcare. These initiatives are conducted either by corporations or by educational institutions. This section begins with mentioning what and how part of the challenges being addressed via each initiative.

3.1 IBM Watson—Replicating Human Cognition

Challenges being addressed

- Use either SI (MKS) or CGS as primary units. (SI units are Data collection from various sources and converts into structured form.

- Data analytics on top of structured data for medical insights.
- Applied advanced statistical, machine learning and natural language processing techniques for disease detection.

Description

The major challenge in medical field is to make smarter decisions at right time from very large volume of structured and unstructured data coming from heterogeneous sources. The doctors want to have an easy access to these data so that they can bring better insights for their patients (even for those who are in remote region). The clinicians want to shortlist eligible clinical trials for the patient [6]. The researchers want to study the data to develop effective drugs for the diseases [7].

IBM helps healthcare industry to overcome above cited challenges by providing centralized location to store heterogeneous data coming from different sources (electronic health records, mobile applications, clinical trials, social media etc.). These data can be de-identified, shared and combined with the frequently increasing observation of clinical, research and social health data (in a secure and private environment) via cloud computing. IBM Watson offers cognitive computing powers to be applied to the stored data. Watson's cognitive powers are similar to the powers that humans possess to inform their decisions: Observe; Interpret; Evaluate; and Decide [8].

IBM has recently joined hands with CVS Health to develop care management solutions for chronic diseases using predictive analytics and IBM Watson's cognitive computing. In the United States, chronic diseases such as heart disease, diabetes, hypertension and obesity are the leading cause of death and disability. The annual spending on these diseases represents 86 % of the United States' $2.9 trillion in total annual health spending [9]. With this collaboration, CVS Health's healthcare practitioners can use Watson's cognitive computing capabilities to achieve medical insights from blend of health information. Watson reads and understands information, interact in natural language and continuously learn by leveraging cognitive capabilities to bring patient centric primary care and meet health goals. The combined solution will focus on:

(a) Predict: Identifying patients at risk for declining health and conducting proactive and customized engagement programs for them.
(b) Adopt: Monitoring patients to adopt safe and healthy behaviors, adhering to the prescribed medicines and the healthy lifestyle.
(c) Suggest: Recommending proper use of cost-effective care.

3.2 *Google–Stanford: Machine Learning for Drug Discovery*

Challenges being addressed

- Create architecture for structured data.
- Obtain insights using Deep Learning, complex ML algorithms for drug discovery.

Description

Generally, the drug discovery takes about 12 years of research and costs almost over a billion dollars [10]. Even after such a time consuming and costly process, only a tiny fraction of drugs get approval for human use. Google Research in collaboration with Stanford's Pande Lab has conducted experiments by gathering a huge collection of publicly available data to achieve significant improvement over simple machine learning algorithms used for drug discovery and for decreasing costs involved in the process. The examination included how data from diverse sources can be explored to progress towards determining the effective chemical compounds that could be used as drug treatments for numerous diseases [11]. The research groups leveraged the experimental data from multiple diseases with multitask neural networks to improve the virtual drug screening effectiveness [12]. The teams gathered approximately 40 million measurements (from disparate sources) for over 200 distinct biological processes to achieve greater predictive power. Deep Learning algorithm provided a flexible approach for incorporating these data into predictive models. The groups trained multitasking neural architectures by providing a learning framework that gathers information from and allows information sharing across distinct biological sources for drug discovery. These techniques helped to discover the most effective drug treatments by leveraging more tasks and more data sets that can help in yielding better performance. With this whole process, the team concluded following major results of multitask framework:

- Better predictive accuracies than single-task methods.
- Predictive Power keeps improving as additional data and processes/tasks.
- Multitask improvement < −Total amount of data + Total number of tasks.
- Multi task networks permit restricted transferability to tasks that are not.

3.3 Calico—Control Your Age

Challenges being addressed

- Collects, visualizes and analyzes biological and genetic data.
- Gene structure analysis for delayed aging or increasing human life expectancy.

Description

Calico's mission is to harness advanced technologies to increase our understanding of the biology that controls lifespan [13]. With this stock pile of knowledge (big data), Calico unleashes ways with which people can live healthier and longer lives.

Calico has brought experts from different field of medicines and engineering together to perform researches by collecting, manipulating, analyzing and visualizing the biological big data. The recent research will be conducted via collaboration between AncestryDNA and Calico to understand the genetics of human lifespan. AncestryDNA plans to bring genetics data from its over one million genotyped customers. The team is planning to evaluate data with respect to public family trees. Just to elaborate, both the companies will together analyze and examine the genetics' role and influence in families enjoying unusual longevity. It is an endeavor to discover the genes that are serving few people to live longer. AncestryDNA will initially offer its huge volume of genomic data, tools and algorithms to undergo this analysis; Calico will later concentrate on formulating and marketing the potential therapeutics in case any emerge from the analysis [14].

Once genes for longevity are predicted, drug companies can develop genetically informed drugs. Big data is the magical phenomena that is helping and can help to an extent never imagined before. Imagine a social life where everybody is healthy and long living.

3.4 Frost and Sullivan—Big Data in Medical Imaging

Challenges being addressed

- Medical image data collection and processing for already existing images (CT Scans, X-ray reports, Magnetic Resonance Imaging).
- Real-time service oriented architecture development.

Description

Frost and Sullivan is especially interested in a special kind of biometric data, medical images for new studies [15]. Unlike big data in transaction data and human generated data, big data in medical imaging refers to the datasets (medical images and related datasets) that are in order of petabytes (10^{15}). US market for Big Data in medical imaging has been divided into Big Data Management and Big Data

Analytics. One of the major reasons for the increase in inflow of the medical imaging is practical implementation of multiple healthcare information technologies that have improved collection, inspection and even dissemination of medical images. Some of the imaging techniques that help in providing datasets are magnetic resonance imaging (MRI), X-ray, molecular imaging, ultrasound, computed tomography (CT) etc. To achieve plethora of information from medical images, a real-time approach via semantic technologies and service-oriented architecture model is being followed. Subsequently using advanced imaging techniques, this data will be debriefed and analyzed for clinical, operational and financial purposes. This increased data interoperability in a real-time environment will bring numerous medical advantages to healthcare. Through this combination of big data management and big data analytics solutions, medical images are expected to be interpreted with more precision. More correctness will lead to improved workflow efficiency, accurate diagnosis, appropriate treatment decisions and proper health management.

3.5 Pittsburgh Health Data Alliance—Carnegie Mellon University (CMU), The University of Pittsburgh and University of Pittsburgh Medical Center

Challenges being addressed

- Collection of raw clinical data, sensors and equipment's data and insurance records.
- Data analysis for optimization and higher efficiency in operations (service end).

Description

One of the major health data alliances of 2015 has been the Pittsburgh Health Data Alliance. The three Pittsburgh experts adore the impact of big data and are confident that together they can utilize the data to improve human health, revolutionizing the practice of medicine [16].

CMU contributes its distinctive capabilities with technology. University of Pittsburgh adds the health sciences proficiency and University of Pittsburgh Medical Center provides raw clinical data (electronic records, images, clinical notes, genomes, wearable sensors, and insurance records) to perform multiple researches. The move is towards fetching data, performing the analysis and implementing new healthcare technologies, products and services in real life scenarios immediately [17]. Center for Machine Learning and Health aims to solve healthcare related challenges using artificial intelligence and machine learning. Few examples include the techniques to identify the emergency room wait times of

patients; mobile applications to keep track of a person's calories intake, exercise and sleep timings to prevent the onset of any disease. Other example may include monitoring patients who are at higher readmission risks, saving costs etc. The other center-Center for Commercial Applications of Healthcare Data—focuses on inventing new technologies based on engineered big data solutions. These technologies are strategized to be used in commercial theranostics and imaging systems for both doctors and patients. Both the centers as part of Pittsburgh Health Data Alliance aim to bring the knowledge from the data to the patient-centric solutions.

3.6 Human Brain Project (HBP)

Challenges being addressed

- Imitating brain, its structure and functioning for building efficient systems.
- Detecting brain diseases using simulated brain models combined with data mining brain related images.

Description

One of the most fascinating researches of the 21st century is the Human Brain Project, HBP. This international project began its journey in 2013 with an intention to understand the human brain. The marvelous power of human brain compels medical researchers and scientists around the world to know more about this tiny body part that contains enormous data. These brain insights are promised to diagnose and treat lethal brain diseases using innovative computing technologies. The HBP deals with these three major science areas for research: neuroscience, neuromedicine, and computing [18]. Through neuroscience, the project wants to collaborate, understand and stimulate information about human brain. Through neuromedicine, project aims to collect information about brain diseases by aggregating medical records from multiple sources. Through computing, the project concentrates on developing brain-inspired computing systems such as high-performing hardware and software. Keeping these target goals, project execution starts with building following six information and communications technology platforms: (a) Neuroinformatics Platform [19], (b) Brain Simulation Platform [20], (c) High Performance Computing Platform [21], (d) Medical Informatics Platform [22], (e) Neuro-morphic Computing Platform [23], (f) Neuro-robotics Platform [24].

4 Healthcare Big Data Challenges

Although Big Data is leading to significant healthcare performance improvement however there are still several challenges which persist [25, 26]:

(1) Understanding unstructured clinical notes in the right context: Medical information is collected by various examiners, however in different context. Based on the symptoms the test varies and so are corresponding reports and each medical examiner examines according to their own way. Lack of standardized approach makes the data unstructured and contextually right while aggregating the information.
(2) Handling large volumes of medical imaging data efficiently and extracting potentially useful information and biomarkers: Too much information is also not good as it could lead to lot of noise. Furthermore, data stored in unstructured form even worsens the situation.
(3) Complexity of the data: Analyzing genomic data is a computationally intensive task and combining with standard clinical data adds additional layers of complexity.
(4) Capturing the patient's behavioral data through several sensors; their various social interactions and communications: Build user specific database at one location when the information is distributed at many sources.
(5) Existing Electronic Health Records are limited to data acquisition than analytics: EHRs have greatly simplified data acquisition, but don't have the ability to aggregate, transform, or create actionable analytics from it. Intelligence is limited to retrospective reporting, which is insufficient for forward-looking healthcare data analytics that hold the key to performance improvement.
(6) Adapting changes: Institutions are notoriously resistant to change. This is especially true as they grow larger and existing processes and procedures become "the way it's always been done". The shift to an analytics-based culture requires everyone in the organization to use a single source of truth to guide choices, stop making gut decisions, and avoid "data shopping" to find data that supports a conclusion that has already been made.

5 Future Outlook: Common Platform for Healthcare Analytics

Almost all the researches mentioned above go through a phase of data collection, data mining and data analysis. The areas of application as well as approaches could be different, but the initial steps are similar. Much of the effort and time goes in data collection, cleaning and data preparation for models [27]. Since, most of these parts are standard and used by all the analytics firms or organizations, a common platform to pursue these common tasks can save a lot of time and effort. Analysts can concentrate on data-driven performance improvements by intending to make the Data Analysis process iterative, progressing level by level and focusing on early

delivery. Maximum automation by leveraging Object Oriented Programming and robust design methodologies to handle any kind of data (textual, numeric etc.) can be utilized. Researchers can then spend more time on individual objective of building evidence for improved care delivery (based on brainstorming data from research, clinical care settings, and operational settings) [28]. Recently, a set of machine learning algorithms known as Deep Learning has solved this challenge.

Automatically extracting complex data representations (abstractions) is the basis for Deep Learning algorithms. The outcome of these algorithms is layered/hierarchical architecture of learning and representing higher-level data in terms of lower-level data. Hierarchical learning approach of human brain is the inspiration of these algorithms. Deep Learning provides the power to extract representations from unsupervised data to computers, making machines independent of human knowledge [29, 30].

These algorithms take the data in any format (text, sound, video, numerical data etc.) in parallel or individually and build models corresponding to the defined objective. In deep learning, humans or researchers do not perform feature extraction directly, but these models themselves extract relevant features and generate insights. These features act together to present an intelligent behavior [31].

This kind of approach can also be used in Healthcare analytics through which supervised and unsupervised models can be built for drug discovery, disease detection and more generic insights or better services and optimization at application end(irrespective of the data formats-clinical data and logs, images, sound etc.). Healthcare companies can extract crucial consumer/patient behavioral patterns (current lifestyle habits and the consequent correlated future health risks) based on the information gathered from social networking websites. Since, healthcare is one of the biggest sources of data in the realm of analytics without organized and structured data. Therefore a common platform approach will be highly beneficial for researchers and analysts.

6 Conclusion

The role of big data is beyond the description. Healthcare stakeholders have begun experiencing the immense power that data possess. The researches, inventions, innovations and discoveries in the field are incomplete without the medical practitioners realizing their necessities. The promising advantages of big data such as evidence-based diagnosis and drugs; personalized care and treatment; decreased costs; faster and effective decisions can bring value into the lives of not only patients but also caregivers. The healthcare's future is clearly in real-time intelligent decision making from the data. Finally, after looking at the challenges in processing the data by healthcare analysts and researchers, it is proposed that there is a need of

a common platform which can be leveraged by all the researchers to pursue common tasks of feature engineering and data preparation. This way more time will be spent on invention rather than on time-consuming task that can be automated.

References

1. EMC Digital Universe & IDC: The digital universe: driving data growth in healthcare, challenges and opportunities (2004)
2. Glaser, J.: Solving big problems with big data. In: Hospitals and Health Networks Daily, Dec (2014)
3. Khan, M.A., Uddin, M.F., Gupta, N.: Seven V's of big data. In: ASEE Zone 1, Proceedings of 2014 Zone 1 Conference of the American Society for Engineering Education (ASEE Zone 1) (2014)
4. Ferlie, E.B., Shortell, S.M.: Improving the quality of healthcare in the United Kingdom and the United States: a framework for change. **79**, 281–315 (2001)
5. National Academy of Sciences. http://www.ncbi.nlm.nih.gov/books/NBK22878/
6. Transforming clinical trial matching with cognitive computing. http://www.ibm.com/smarterplanet/us/en/ibmwatson/clinical-trial-matching.html. Accessed 12 Aug 2015
7. IBM Watson ushers in a new era of data-driven discoveries. https://www-03.ibm.com/press/us/en/pressrelease/44697.wss. Accessed 25 Aug 2015
8. What is watson. http://www.ibm.com/smarterplanet/us/en/ibmwatson/what-is-watson.html. Accessed 15 Aug 2015
9. CVS Health and IBM tap watson to develop care management solutions for chronic disease. http://www-03.ibm.com/press/us/en/pressrelease/47400.wss. Accessed 15 Aug 2015
10. Applications of physical simulation and Bayesian statistics/machine learning to biologically and biomedically important questions. http://pande.stanford.edu/projects/#ntoc1. Accessed 25 Aug 2015
11. Large-scale machine learning for drug discovery. http://googleresearch.blogspot.in/2015/03/large-scale-machine-learning-for-drug.html. Accessed March 2015
12. Massively multitask networks for drug discovery: a preliminary work on machine learning, unpublished. http://arxiv.org/pdf/1502.02072v1.pdf
13. Calico Labs: http://www.calicolabs.com/
14. AncestryDNA and calico to research the genetics of human lifespan. http://www.calicolabs.com/news/2015/07/21/. Accessed 25 Aug 2015
15. Frost, Sullivan: Big data opportunities in the US medical imaging market. Mounting data volumes in medical imaging driving the need for big data tools market engineering, April 2015
16. Pittsburgh Health Data Alliance: www.healthdataalliance.com
17. Pittsburgh Health Data Alliance: Pitt, CMU, UPMC form alliance to transform healthcare through big data, March 2015
18. Kasabov Nikola, K.: Springer handbook of bio-/neuro informatic, pp ix, ISBN:978-3-642-30574-0 (2014)
19. https://www.humanbrainproject.eu/neuroinformatics-platform1
20. https://www.humanbrainproject.eu/brain-simulation-platform1
21. https://www.humanbrainproject.eu/high-performance-computing-platform1
22. https://www.humanbrainproject.eu/medical-informatics-platform1
23. https://www.humanbrainproject.eu/neuromorphic-computing-platform1
24. https://www.humanbrainproject.eu/neurorobotics-platform1
25. Sun, J., Reddy, C.K.: Big data analytics for healthcare. In: SIAM International Conference on Data Mining (2013)

26. getting started with healthcare data analytics. http://www.mckesson.com/healthcare-analytics/healthcare-big-data-challenges
27. http://ventanaresearch.com/blog/commentblog.aspx?id=4716
28. Nambiar, R., Bhardwaj, R., Sethi, A., Varghese, R.: A look at challenges and opportunities of big data in healthcare. In: IEEE, International Conference on Big Data (2013)
29. http://www.journalofbigdata.com/content/2/1/1
30. http://www.iro.umontreal.ca/ ~ lisa/pointeurs/bengio+lecun_chapter2007.pdf
31. Bengio, Y., Goodfellow, I.J., Courville, A.: Deep learning. Book in Preparation for MIT Press, Chap. 12, Unpublished (2015)