ISSN- 2394-5125 VOL 07, ISSUE 01, 2020

# HEALTHCARE DATA MANAGEMENT, ANALYSIS AND FUTURE PROSPECTS: BIG DATA

**Dr Versha Prasad**
Assistant Professor
University Institute of Health Sciences C.S.J.M. University Kanpur

**Abstract**
Various public and private sector industries generate, store, and analyse big data with an aim to improve the services they provide. In the healthcare industry, various sources for big data include hospital records, medical records of patients, results of medical examinations, and devices that are a part of internet of things. Biomedical research also generates a significant portion of big data relevant to public healthcare. This data requires proper management and analysis in order to derive meaningful information. Otherwise, seeking solution by analysing big data quickly becomes comparable to finding a needle in the haystack. There are various challenges associated with each step of handling big data which can only be surpassed by using high-end computing solutions for big data analysis. That is why, to provide relevant solutions for improving public health, healthcare providers are required to be fully equipped with appropriate infrastructure to systematically generate and analyse big data. An efficient management, analysis, and interpretation of big data can change the game by opening new avenues for modern healthcare. That is exactly why various industries, including the healthcare industry, are taking vigorous steps to convert this potential into better services and financial advantages. With a strong integration of biomedical and healthcare data, modern healthcare organizations can possibly revolutionize the medical therapies and personalized medicine.

Key words: Health care data, Health Records, Clinical Data, Medical Records.

## INTRODUCTION

Information has been the key to a better organization and new developments. The more information we have, the more optimally we can organize ourselves to deliver the best outcomes. That is why data collection is an important part for every organization. We can also use this data for the prediction of current trends of certain parameters and future events. As we are becoming more and more aware of this, we have started producing and collecting more data about almost everything by introducing technological developments in this direction. Today, we are facing a situation wherein we are flooded with tons of data from every aspect of our life such as social activities, science, work, health, etc. In a way, we can compare the present situation to a data deluge. The technological advances have helped us in generating more and more data, even to a level where it has become unmanageable with currently available technologies. This has led to the creation of the term 'big data' to describe data that is large and unmanageable. In order to meet our present and future social needs, we need to develop new strategies to organize this data and derive meaningful information. One such special social need is healthcare. Like every other industry, healthcare organizations are producing data at a tremendous rate that presents many advantages and challenges at the same time. In this review, we discuss about the basics of big data including its management, analysis and future prospects especially in healthcare sector.

## DATA OVERLOAD

Every day, people working with various organizations around the world are generating a massive amount of data. The term "digital universe" quantitatively defines such massive amounts of data created, replicated, and consumed in a single year. International Data Corporation (IDC) estimated the approximate size of the digital universe in 2005 to be 130 exabytes (EB). The digital universe in 2017 expanded to about 16,000 EB or 16 zettabytes (ZB). IDC predicted that the digital universe would expand to 40,000 EB by the year 2020. To imagine this size, we would have to assign about 5200 gigabytes (GB) of data to all individuals. This exemplifies the phenomenal speed at which the digital universe is expanding. 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.

These observations have become so conspicuous that has eventually led to the birth of a new field of science termed '*Data Science*'. Data science deals with various aspects including data management and analysis, to extract deeper insights for improving the functionality or services of a system (for example, healthcare and transport system). Additionally, with the availability of some of the most creative and meaningful ways to visualize big data post-analysis, it has become easier to understand the functioning of any complex system. As a large section of society is becoming aware of, and involved in generating big data, it has become necessary to

define what big data is. Therefore, in this review, we attempt to provide details on the impact of big data in the transformation of global healthcare sector and its impact on our daily lives.

**DEFINING BIG DATA**

As the name suggests, 'big data' represents large amounts of data that is unmanageable using traditional software or internet-based platforms. It surpasses the traditionally used amount of storage, processing and analytical power. Even though a number of definitions for big data exist, the most popular and well-accepted definition was given by Douglas Laney. Laney observed that (big) data was growing in three different dimensions namely, volume, velocity and variety (known as the 3 Vs).[1] The 'big' part of big data is indicative of its large volume. In addition to volume, the big data description also includes velocity and variety. Velocity indicates the speed or rate of data collection and making it accessible for further analysis; while, variety remarks on the different types of organized and unorganized data that any firm or system can collect, such as transaction-level data, video, audio, text or log files. These three Vs have become the standard definition of big data.

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 that 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 healthcare, safety or transportation) more aware, interactive and efficient.[2] In addition, visualization of big data in a user-friendly manner will be a critical factor for societal development.

**HEALTH DATA - MULTIDIMENSIONAL DATA**

Healthcare is a multi-dimensional system established with the sole aim for the prevention, diagnosis, and treatment of health-related issues or impairments in human beings. The major components of a healthcare system are the health professionals (physicians or nurses), health facilities (clinics, hospitals for delivering medicines and other diagnosis or treatment technologies), and a financing institution supporting the former two. The health professionals belong to various health sectors like dentistry, medicine, midwifery, nursing, psychology, physiotherapy, and many others. Healthcare is required at several levels depending on the urgency of situation. Professionals serve it as the first point of consultation (for primary care), acute care requiring skilled professionals (secondary care), advanced medical investigation and treatment (tertiary care) and highly uncommon diagnostic or surgical procedures (quaternary care). At all these levels, the health professionals are responsible for different kinds of information such as patient's medical history (diagnosis and prescriptions related data), medical and clinical data (like data from imaging and laboratory examinations), and other private or personal medical data. Previously, the common practice to store such medical records for a patient was in the form of either handwritten notes or typed reports.[3]

With the advent of computer systems and its potential, the digitization of all clinical exams and medical records in the healthcare systems has become a standard and widely adopted practice nowadays. In 2003, a division of the National Academies of Sciences, Engineering, and Medicine known as Institute of Medicine chose the term "*electronic health records*" to represent records maintained for improving the health care sector towards the benefit of patients and clinicians. Electronic health records (EHR) as defined by Murphy, Hanken and Waters are computerized medical records for patients any information relating to the past, present or future physical/mental health or condition of an individual which resides in electronic system(s) used to capture, transmit, receive, store, retrieve, link and manipulate multimedia data for the primary purpose of providing healthcare and health-related services.[4]

**ELECTRONIC HEALTH RECORDS**

EHRs have introduced many advantages for handling modern healthcare related data. Below, we describe some of the characteristic advantages of using EHRs. The first advantage of EHRs is that healthcare professionals have an improved access to the entire medical history of a patient. The information includes medical diagnoses, prescriptions, data related to known allergies, demographics, clinical narratives, and the results obtained from various laboratory tests. The recognition and treatment of medical conditions thus is time efficient due to a reduction in the lag time of previous test results. With time we have observed a significant decrease in the redundant and additional examinations, lost orders and ambiguities caused by illegible handwriting, and an improved care coordination between multiple healthcare providers. Overcoming such logistical errors has led to

reduction in the number of drug allergies by reducing errors in medication dose and frequency. Healthcare professionals have also found access over web based and electronic platforms to improve their medical practices significantly using automatic reminders and prompts regarding vaccinations, abnormal laboratory results, cancer screening, and other periodic checkups. There would be a greater continuity of care and timely interventions by facilitating communication among multiple healthcare providers and patients. They can be associated to electronic authorization and immediate insurance approvals due to less paperwork. EHRs enable faster data retrieval and facilitate reporting of key healthcare quality indicators to the organizations, and also improve public health surveillance by immediate reporting of disease outbreaks. EHRs also provide relevant data regarding the quality of care for the beneficiaries of employee health insurance programs and can help control the increasing costs of health insurance benefits. Finally, EHRs can reduce or absolutely eliminate delays and confusion in the billing and claims management area. The EHRs and internet together help provide access to millions of health-related medical information critical for patient life.

**DIGITALISATION OF HEALTH CARE DATA**

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 healthcare 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. [5]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.[6] The use of big data from healthcare shows promise for improving health outcomes and controlling costs.

**DATA IN BIOMEDICAL RESEARCH**

A biological system, such as a human cell, exhibits molecular and physical events of complex interplay. In order to understand interdependencies of various components and events of such a complex system, a biomedical or biological experiment usually gathers data on a smaller and/or simpler component. Consequently, it requires multiple simplified experiments to generate a wide map of a given biological phenomenon of interest. This indicates that more the data we have, the better we understand the biological processes. With this idea, modern techniques have evolved at a great pace. For instance, one can imagine the amount of data generated since the integration of efficient technologies like next-generation sequencing (NGS) and Genome wide association studies (GWAS) to decode human genetics. NGS-based data provides information at depths that were previously inaccessible and takes the experimental scenario to a completely new dimension. It has increased the resolution at which we observe or record biological events associated with specific diseases in a real time manner. The idea that large amounts of data can provide us a good amount of information that often remains unidentified or hidden in smaller experimental methods has ushered-in the '*-omics*' era. The '*omics*' discipline has witnessed significant progress as instead of studying a single '*gene*' scientists can now study the whole '*genome*' of an organism in '*genomics*' studies within a given amount of time. Similarly, instead of studying the expression or '*transcription*' of single gene, we can now study the expression of all the genes or the entire '*transcriptome*' of an organism under '*transcriptomics*' studies. Each of these individual experiments generate a large amount of data with more depth of information than ever before. Yet, this depth and resolution might be insufficient to provide all the details required to explain a particular mechanism or event. Therefore, one usually finds oneself analyzing a large amount of data obtained from multiple experiments to gain novel insights. This fact is supported by a continuous rise in the number of publications regarding big data in healthcare. Analysis of such big data from medical and healthcare systems can be of immense help in providing novel strategies for healthcare. The latest technological developments in data generation, collection and analysis, have raised expectations towards a

**INTERNET**

Healthcare industry has not been quick enough to adapt to the big data movement compared to other industries. Therefore, big data usage in the healthcare sector is still in its infancy. For example, healthcare and biomedical big data have not yet converged to enhance healthcare data with molecular pathology. Such convergence can help unravel various mechanisms of action or other aspects of predictive biology. Therefore, to assess an individual's health status, biomolecular and clinical datasets need to be married. One such source of clinical data in healthcare is 'internet of things' (IoT).

ISSN- 2394-5125 VOL 07, ISSUE 01, 2020

In fact, IoT is another big player implemented in a number of other industries including healthcare. Until recently, the objects of common use such as cars, watches, refrigerators and health-monitoring devices, did not usually produce or handle data and lacked internet connectivity. However, furnishing such objects with computer chips and sensors that enable data collection and transmission over internet has opened new avenues. The device technologies such as Radio Frequency IDentification (RFID) tags and readers, and Near Field Communication (NFC) devices, that can not only gather information but interact physically, are being increasingly used as the information and communication systems. This enables objects with RFID or NFC to communicate and function as a web of smart things. The analysis of data collected from these chips or sensors may reveal critical information that might be beneficial in improving lifestyle, establishing measures for energy conservation, improving transportation, and healthcare. In fact, IoT has become a rising movement in the field of healthcare. IoT devices create a continuous stream of data while monitoring the health of people (or patients) which makes these devices a major contributor to big data in healthcare. Such resources can interconnect various devices to provide a reliable, effective and smart healthcare service to the elderly and patients with a chronic illness.[7]

**ADVANTAGES OF INTERNET IN HEALTH CARE**

Using the web of IoT devices, a doctor can measure and monitor various parameters from his/her clients in their respective locations for example, home or office. Therefore, through early intervention and treatment, a patient might not need hospitalization or even visit the doctor resulting in significant cost reduction in healthcare expenses. Some examples of IoT devices used in healthcare include fitness or health-tracking wearable devices, biosensors, clinical devices for monitoring vital signs, and others types of devices or clinical instruments. Such IoT devices generate a large amount of health related data. If we can integrate this data with other existing healthcare data like EMRs or PHRs, we can predict a patients' health status and its progression from subclinical to pathological state. In fact, big data generated from IoT has been quiet advantageous in several areas in offering better investigation and predictions. On a larger scale, the data from such devices can help in personnel health monitoring, modelling the spread of a disease and finding ways to contain a particular disease outbreak.

The analysis of data from IoT would require an updated operating software because of its specific nature along with advanced hardware and software applications. We would need to manage data inflow from IoT instruments in real-time and analyze it by the minute. Associates in the healthcare system are trying to trim down the cost and ameliorate the quality of care by applying advanced analytics to both internally and externally generated data.

**MOBILE HEALTH (mHealth)**

In today's digital world, every individual seems to be obsessed to track their fitness and health statistics using the in-built pedometer of their portable and wearable devices such as, smartphones, smartwatches, fitness dashboards or tablets. With an increasingly mobile society in almost all aspects of life, the healthcare infrastructure needs remodeling to accommodate mobile devices.[8] The practice of medicine and public health using mobile devices, known as mHealth or mobile health, pervades different degrees of health care especially for chronic diseases, such as diabetes and cancer.[9] Healthcare organizations are increasingly using mobile health and wellness services for implementing novel and innovative ways to provide care and coordinate health as well as wellness. Mobile platforms can improve healthcare by accelerating interactive communication between patients and healthcare providers. In fact, Apple and Google have developed devoted platforms like Apple's ResearchKit and Google Fit for developing research applications for fitness and health statistics.[10] These applications support seamless interaction with various consumer devices and embedded sensors for data integration. These apps help the doctors to have direct access to your overall health data. Both the user and their doctors get to know the real-time status of your body. These apps and smart devices also help by improving our wellness planning and encouraging healthy lifestyles. The users or patients can become advocates for their own health.

**NATURE OF DATA IN HEALTH CARE**

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 format 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.

**MANAGEMENT AND ANLYSIS OF BIG DATA**

Big data is the huge amounts of a variety of data generated at a rapid rate. The data gathered from various sources is mostly required for optimizing consumer services rather than consumer consumption. This is also true for big data from the biomedical research and healthcare. The major challenge with big data is how to handle this large volume of information. To make it available for scientific community, the data is required to be stored in a file format that is easily accessible and readable for an efficient analysis. In the context of healthcare data, another major challenge is the implementation of high-end computing tools, protocols and high-end hardware in the clinical setting. Experts from diverse backgrounds including biology, information technology, statistics, and mathematics are required to work together to achieve this goal. The data collected using the sensors can be made available on a storage cloud with pre-installed software tools developed by analytic tool developers. These tools would have data mining and ML functions developed by AI experts to convert the information stored as data into knowledge. Upon implementation, it would enhance the efficiency of acquiring, storing, analyzing, and visualization of big data from healthcare. The main task is to annotate, integrate, and present this complex data in an appropriate manner for a better understanding. In absence of such relevant information, the (healthcare) data remains quite cloudy and may not lead the biomedical researchers any further. Finally, visualization tools developed by computer graphics designers can efficiently display this newly gained knowledge.

Heterogeneity of data is another challenge in big data analysis. The huge size and highly heterogeneous nature of big data in healthcare renders it relatively less informative using the conventional technologies. The most common platforms for operating the software framework that assists big data analysis are high power computing clusters accessed via grid computing infrastructures. Cloud computing is such a system that has virtualized storage technologies and provides reliable services. It offers high reliability, scalability and autonomy along with ubiquitous access, dynamic resource discovery and composability. Such platforms can act as a receiver of data from the ubiquitous sensors, as a computer to analyze and interpret the data, as well as providing the user with easy to understand web-based visualization. In IoT, the big data processing and analytics can be performed closer to data source using the services of mobile edge computing cloudlets and fog computing.

**Hadoop**

When working with hundreds or thousands of nodes, one has to handle issues like how to parallelize the computation, distribute the data, and handle failures. One of most popular open-source distributed application for this purpose is Hadoop. [11]Hadoop implements MapReduce algorithm for processing and generating large datasets. MapReduce uses *map* and *reduce* primitives to *map* each logical record' in the input into a set of intermediate key/value pairs, and *reduce* operation combines all the values that shared the same key.[12] It efficiently parallelizes the computation, handles failures, and schedules inter-machine communication across large-scale clusters of machines. Hadoop Distributed File System (HDFS) is the file system component that provides a scalable, efficient, and replica based storage of data at various nodes that form a part of a cluster.[13] Hadoop has other tools that enhance the storage and processing components therefore many large companies like Yahoo, Facebook, and others have rapidly adopted it. Hadoop has enabled researchers to use data sets otherwise impossible to handle. Many large projects, like the determination of a correlation between the air quality data and asthma admissions, drug development using genomic and proteomic data, and other such aspects of healthcare are implementing Hadoop. Therefore, with the implementation of Hadoop system, the healthcare analytics will not be held back.

**Apache Spark**

Apache Spark is another open source alternative to Hadoop. It is a unified engine for distributed data processing that includes higher-level libraries for supporting SQL queries (*Spark SQL*), streaming data (*Spark Streaming*), machine learning (*MLlib*) and graph processing (*GraphX*).[13] These libraries help in increasing developer productivity because the programming interface requires lesser coding efforts and can be seamlessly combined to create more types of complex computations. By implementing Resilient distributed Datasets (RDDs), in-memory processing of data is supported that can make Spark about 100× faster than Hadoop in multi-pass analytics (on smaller datasets).[14,15]This is more true when the data size is smaller than the available memory.[16] This indicates that processing of really big data with Apache Spark would require a large amount of memory. Since, the cost of memory is higher than the hard drive, MapReduce is expected to be more cost effective for large datasets compared to Apache Spark. Similarly, Apache Storm was developed to provide a real-time framework for data stream processing. This platform supports most of the programming languages. Additionally, it offers good horizontal scalability and built-in-fault-tolerance capability for big data analysis.

**Machine learning for information extraction, data analysis and predictions**

In healthcare, patient data contains recorded signals for instance, electrocardiogram (ECG), images, and videos. Healthcare providers have barely managed to convert such healthcare data into EHRs. Efforts are underway to digitize patient-histories from pre-EHR era notes and supplement the standardization process by turning static images into machine-readable text. For example, optical character recognition (OCR) software is one such approach that can recognize handwriting as well as computer fonts and push digitization. Such unstructured and structured healthcare datasets have untapped wealth of information that can be harnessed using advanced AI programs to draw critical actionable insights in the context of patient care. In fact, AI has emerged as the method of choice for big data applications in medicine. This smart system has quickly found its niche in decision making process for the diagnosis of diseases. Healthcare professionals analyze such data for targeted abnormalities using appropriate ML approaches. ML can filter out structured information from such raw data.

**Extracting information from EHR datasets**

Emerging ML or AI based strategies are helping to refine healthcare industry's information processing capabilities. For example, natural language processing (NLP) is a rapidly developing area of machine learning that can identify key syntactic structures in free text, help in speech recognition and extract the meaning behind a narrative. NLP tools can help generate new documents, like a clinical visit summary, or to dictate clinical notes. The unique content and complexity of clinical documentation can be challenging for many NLP developers. Nonetheless, we should be able to extract relevant information from healthcare data using such approaches as NLP.

AI has also been used to provide predictive capabilities to healthcare big data. For example, ML algorithms can convert the diagnostic system of medical images into automated decision-making. Though it is apparent that healthcare professionals may not be replaced by machines in the near future, yet AI can definitely assist physicians to make better clinical decisions or even replace human judgment in certain functional areas of healthcare.

**Image analytics**

Some of the most widely used imaging techniques in healthcare include computed tomography (CT), magnetic resonance imaging (MRI), X-ray, molecular imaging, ultrasound, photo-acoustic imaging, functional MRI (fMRI), positron emission tomography (PET), electroencephalography (EEG), and mammograms. These techniques capture high definition medical images (patient data) of large sizes. Healthcare professionals like radiologists, doctors and others do an excellent job in analyzing medical data in the form of these files for targeted abnormalities. However, it is also important to acknowledge the lack of specialized professionals for many diseases. In order to compensate for this dearth of professionals, efficient systems like Picture Archiving and Communication System (PACS) have been developed for storing and convenient access to medical image and reports data .[17] PACSs are popular for delivering images to local workstations, accomplished by protocols such as digital image communication in medicine (DICOM). However, data exchange with a PACS relies on using structured data to retrieve medical images. This by nature misses out on the unstructured information contained in some of the biomedical images. Moreover, it is possible to miss an additional information about a patient's health status that is present in these images or similar data. A professional focused on diagnosing an unrelated condition might not observe it, especially when the condition is still emerging. To help in such situations, image analytics is making an impact on healthcare by actively extracting disease biomarkers from biomedical images. This approach uses ML and pattern recognition techniques to draw insights from massive volumes of clinical image data to transform the diagnosis, treatment and monitoring of patients. It focuses on enhancing the diagnostic capability of medical imaging for clinical decision-making.

A number of software tools have been developed based on functionalities such as generic, registration, segmentation, visualization, reconstruction, simulation and diffusion to perform medical image analysis in order to dig out the hidden information. For example, Visualization Toolkit is a freely available software which allows powerful processing and analysis of 3D images from medical tests, while SPM can process and analyse 5 different types of brain images (e.g. MRI, fMRI, PET, CT-Scan and EEG) .[18] Such bioinformatics-based big data analysis may extract greater insights and value from imaging data to boost and support precision medicine projects, clinical decision support tools, and other modes of healthcare. For example, we can also use it to monitor new targeted-treatments for cancer.

**Big data from omics**

The big data from "omics" studies is a new kind of challenge for the bioinformaticians. Robust algorithms are required to analyse such complex data from biological systems. The ultimate goal is to convert this huge data into an informative knowledge base. The application of bioinformatics approaches to transform the biomedical and genomics data into predictive and preventive health is known as translational bioinformatics. It is at the forefront of data-driven healthcare. Various kinds of quantitative data in healthcare, for example from laboratory measurements, medication data and genomic profiles, can be combined and used to identify new meta-data that can help precision therapies.[19] This is why emerging new technologies are required to help in analysing this

digital wealth. In fact, highly ambitious multimillion-dollar projects like "*Big Data Research and Development Initiative*" have been launched that aim to enhance the quality of big data tools and techniques for a better organization, efficient access and smart analysis of big data. There are many advantages anticipated from the processing of '*omics'* data from large-

**Commercial platforms for healthcare data analytics**

In order to tackle big data challenges and perform smoother analytics, various companies have implemented AI to analyze published results, textual data, and image data to obtain meaningful outcomes. IBM Corporation is one of the biggest and experienced players in this sector to provide healthcare analytics services commercially. IBM's Watson Health is an AI platform to share and analyze health data among hospitals, providers and researchers. Similarly, Flatiron Health provides technology-oriented services in healthcare analytics specially focused in cancer research. Other big companies such as Oracle Corporation and Google Inc. are also focusing to develop cloud-based storage and distributed computing power platforms. Interestingly, in the recent few years, several companies and start-ups have also emerged to provide health care-based analytics and solutions.

**AYASDI**

Ayasdi is one such big vendor which focuses on ML based methodologies to primarily provide machine intelligence platform along with an application framework with tried & tested enterprise scalability. It provides various applications for healthcare analytics, for example, to understand and manage clinical variation, and to transform clinical care costs. It is also capable of analyzing and managing how hospitals are organized, conversation between doctors, risk-oriented decisions by doctors for treatment, and the care they deliver to patients. It also provides an application for the assessment and management of population health, a proactive strategy that goes beyond traditional risk analysis methodologies. It uses ML intelligence for predicting future risk trajectories, identifying risk drivers, and providing solutions for best outcomes.

**Linguamatics**

It is an NLP based algorithm that relies on an interactive text mining algorithm (I2E). I2E can extract and analyse a wide array of information. Results obtained using this technique are tenfold faster than other tools and does not require expert knowledge for data interpretation. This approach can provide information on genetic relationships and facts from unstructured data. Classical, ML requires well-curated data as input to generate clean and filtered results

**IBM Watson**

This is one of the unique ideas of the tech-giant IBM that targets big data analytics in almost every professional sector. This platform utilizes ML and AI based algorithms extensively to extract the maximum information from minimal input. IBM Watson enforces the regimen of integrating a wide array of healthcare domains to provide meaningful and structured data. In an attempt to uncover novel drug targets specifically in cancer disease model, IBM Watson and Pfizer have formed a productive collaboration to accelerate the discovery of novel immune-oncology combinations. Combining Watson's deep learning modules integrated with AI technologies allows the researchers to interpret complex genomic data sets. IBM Watson has been used to predict specific types of cancer based on the gene expression profiles obtained from various large data sets providing signs of multiple druggable targets. IBM Watson is also used in drug discovery programs by integrating curated literature and forming network maps to provide a detailed overview of the molecular landscape in a specific disease model.

In order to analyse the diversified medical data, healthcare domain, describes analytics in four categories: descriptive, diagnostic, predictive, and prescriptive analytics. Descriptive analytics refers for describing the current medical situations and commenting on that whereas diagnostic analysis explains reasons and factors behind occurrence of certain events, for example, choosing treatment option for a patient based on clustering and decision trees. Predictive analytics focuses on predictive ability of the future outcomes by determining trends and probabilities. These methods are mainly built up of machine leaning techniques and are helpful in the context of understanding complications that a patient can develop. Prescriptive analytics is to perform analysis to propose an action towards optimal decision making. For example, decision of avoiding a given treatment to the patient based on observed side effects and predicted complications. In order to improve performance of the current medical systems integration of big data into healthcare analytics can be a major factor; however, sophisticated strategies need to be developed. An architecture of best practices of different analytics in healthcare domain is required for integrating big data technologies to improve the outcomes. However, there are many challenges associated with the implementation of such strategies.

**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. 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. This 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, 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.[20,21,22,23] The documentation quality might improve by using self-report questionnaires from patients for their symptoms.

**Image pre-processing**

Studies have observed various physical factors that can lead to altered data quality and misinterpretations from existing medical records.[24]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 veins which is non-correlative with real case scenario. Reduction of noise, clearing artifacts, adjusting contrast of acquired images and image quality adjustment post mishandling are some of the measures that can be implemented to benefit the purpose.

**Security**

There have been many security breaches, hackings, phishing attacks, and ransomware episodes that data security is a priority for healthcare organizations. After noticing an array of vulnerabilities, a list of technical safeguards was developed for the protected health information (PHI). These rules, termed as HIPAA Security Rules, help guide organizations with storing, transmission, authentication protocols, and controls over access, integrity, and auditing. Common security measures like using up-to-date anti-virus software, firewalls, encrypting sensitive data, and multi-factor authentication can save a lot of trouble.

**Meta-data**

To have a successful data governance plan, it would be mandatory to have complete, accurate, and up-to-date metadata regarding all the stored data. The metadata would be composed of information like time of creation, purpose and person responsible for the data, previous usage (by who, why, how, and when) for researchers and data analysts. This would allow analysts to replicate previous queries and help later scientific studies and accurate benchmarking. This increases the usefulness of data and prevents creation of "data dumpsters" of low or no use.

**Querying**

Metadata would make it easier for organizations to query their data and get some answers. However, in absence of proper interoperability between datasets the query tools may not access an entire repository of data. Also,

different components of a dataset should be well interconnected or linked and easily accessible otherwise a complete portrait of an individual patient's health may not be generated. Medical coding systems like ICD-10, SNOMED-CT, or LOINC must be implemented to reduce free-form concepts into a shared ontology. If the accuracy, completeness, and standardization of the data are not in question, then Structured Query Language (SQL) can be used to query large datasets and relational databases.

**Visualization**

A clean and engaging visualization of data with charts, heat maps, and histograms to illustrate contrasting figures and correct labeling of information to reduce potential confusion, can make it much easier for us to absorb information and use it appropriately. Other examples include bar charts, pie charts, and scatterplots with their own specific ways to convey the data.

**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. This could be due to technical and organizational barriers. This 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, CommonWell (a not-for-profit trade association) and Carequality (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.[25]

The healthcare providers will need to overcome every challenge on this list and more to develop a big data exchange ecosystem that provides trustworthy, timely, and meaningful information by connecting all members of the care continuum. Time, commitment, funding, and communication would be required before these challenges are overcome.

**Big data analytics for cutting costs**

To develop a healthcare system based on big data that can exchange big data and provides us with trustworthy, timely, and meaningful information, we need to overcome every challenge mentioned above. Overcoming these challenges would require investment in terms of time, funding, and commitment. However, like other technological advances, the success of these ambitious steps would apparently ease the present burdens on healthcare especially in terms of costs. It is believed that the implementation of big data analytics by healthcare organizations might lead to a saving of over 25% in annual costs in the coming years. Better diagnosis and disease predictions by big data analytics can enable cost reduction by decreasing the hospital readmission rate. The healthcare firms do not understand the variables responsible for readmissions well enough. It would be easier for healthcare organizations to improve their protocols for dealing with patients and prevent readmission by determining these relationships well. Big data analytics can also help in optimizing staffing, forecasting operating room demands, streamlining patient care, and improving the pharmaceutical supply chain. All of these factors will lead to an ultimate reduction in the healthcare costs by the organizations.

**Quantum mechanics and big data analysis**

Big data sets can be staggering in size. Therefore, its analysis remains daunting even with the most powerful modern computers. For most of the analysis, the bottleneck lies in the computer's ability to access its memory and not in the processor.[26,27] The capacity, bandwidth or latency requirements of memory hierarchy outweigh the computational requirements so much that supercomputers are increasingly used for big data analysis .[28,29] An additional solution is the application of quantum approach for big data analysis.

**Quantum computing and its advantages**

The common digital computing uses binary digits to code for the data whereas quantum computation uses quantum bits or *qubits*.[30] A *qubit* is a quantum version of the classical binary bits that can represent a zero, a one, or any linear combination of states (called *superpositions*) of those two qubit states.[31] Therefore, qubits allow computer bits to operate in three states compared to two states in the classical computation. This allows quantum computers to work thousands of times faster than regular computers. For example, a conventional analysis of a dataset with $n$ points would require $2^n$ processing units whereas it would require just $n$ quantum bits using a quantum computer. Quantum computers use quantum mechanical phenomena like superposition and quantum entanglement to perform computations.[32,33]

Quantum algorithms can speed-up the big data analysis exponentially.[34] Some complex problems, believed to be unsolvable using conventional computing, can be solved by quantum approaches. For example, the current encryption techniques such as RSA, public-key (PK) and Data Encryption Standard (DES) which are thought to be impassable now would be irrelevant in future because quantum computers will quickly get through them.[35] Quantum approaches can dramatically reduce the information required for big data analysis. For example,

quantum theory can maximize the distinguishability between a multilayer network using a minimum number of layers.[36] In addition, quantum approaches require a relatively small dataset to obtain a maximally sensitive data analysis compared to the conventional (machine-learning) techniques. Therefore, quantum approaches can drastically reduce the amount of computational power required to analyse big data. Even though, quantum computing is still in its infancy and presents many open challenges, it is being implemented for healthcare data.

**Applications in big data analysis**

Quantum computing is picking up and seems to be a potential solution for big data analysis. For example, identification of rare events, such as the production of Higgs bosons at the Large Hadron Collider (LHC) can now be performed using quantum approaches.[37] At LHC, huge amounts of collision data (1PB/s) is generated that needs to be filtered and analysed. One such approach, the quantum annealing for ML (QAML) that implements a combination of ML and quantum computing with a programmable quantum annealer, helps reduce human intervention and increase the accuracy of assessing particle-collision data. In another example, the quantum support vector machine was implemented for both training and classification stages to classify new data.[38] Such quantum approaches could find applications in many areas of science. Indeed, recurrent quantum neural network (RQNN) was implemented to increase signal separability in electroencephalogram (EEG) signals.[39] Similarly, quantum annealing was applied to intensity modulated radiotherapy (IMRT) beamlet intensity optimization.[40] Similarly, there exist more applications of quantum approaches regarding healthcare e.g. quantum sensors and quantum microscopes.[41]

## CONCLUSION AND FUTURE PROSPECTS

Nowadays, various biomedical and healthcare tools such as genomics, mobile biometric sensors, and smartphone apps generate a big amount of data. Therefore, it is mandatory for us to know about and assess that can be achieved using this data. For example, the analysis of such data can provide further insights in terms of procedural, technical, medical and other types of improvements in healthcare. After a review of these healthcare procedures, it appears that the full potential of patient-specific medical specialty or personalized medicine is under way. The collective big data analysis of EHRs, EMRs and other medical data is continuously helping build a better prognostic framework. The companies providing service for healthcare analytics and clinical transformation are indeed contributing towards better and effective outcome. Common goals of these companies include reducing cost of analytics, developing effective Clinical Decision Support (CDS) systems, providing platforms for better treatment strategies, and identifying and preventing fraud associated with big data. Though, almost all of them face challenges on federal issues like how private data is handled, shared and kept safe. The combined pool of data from healthcare organizations and biomedical researchers have resulted in a better outlook, determination, and treatment of various diseases. This has also helped in building a better and healthier personalized healthcare framework. Modern healthcare fraternity has realized the potential of big data and therefore, have implemented big data analytics in healthcare and clinical practices. Supercomputers to quantum computers are helping in extracting meaningful information from big data in dramatically reduced time periods. With high hopes of extracting new and actionable knowledge that can improve the present status of healthcare services, researchers are plunging into biomedical big data despite the infrastructure challenges. Clinical trials, analysis of pharmacy and insurance claims together, discovery of biomarkers is a part of a novel and creative way to analyse healthcare big data.

Big data analytics leverage the gap within structured and unstructured data sources. The shift to an integrated data environment is a well-known hurdle to overcome. Interesting enough, the principle of big data heavily relies on the idea of the more the information, the more insights one can gain from this information and can make predictions for future events. It is rightfully projected by various reliable consulting firms and health care companies that the big data healthcare market is poised to grow at an exponential rate. However, in a short span we have witnessed a spectrum of analytics currently in use that have shown significant impacts on the decision making and performance of healthcare industry. The exponential growth of medical data from various domains has forced computational experts to design innovative strategies to analyse and interpret such enormous amount of data within a given timeframe. The integration of computational systems for signal processing from both research and practicing medical professionals has witnessed growth. Thus, developing a detailed model of a human body by combining physiological data and "-omics" techniques can be the next big target. This unique idea can enhance our knowledge of disease conditions and possibly help in the development of novel diagnostic tools. The continuous rise in available genomic data including inherent hidden errors from experiment and analytical practices need further attention. However, there are opportunities in each step of this extensive process to introduce systemic improvements within the healthcare research.

One can clearly see the transitions of health care market from a wider volume base to personalized or individual specific domain. Therefore, it is essential for technologists and professionals to understand this evolving situation. In the coming year it can be projected that big data analytics will march towards a predictive system. This would mean prediction of futuristic outcomes in an individual's health state based on current or existing

data (such as EHR-based and Omics-based). Similarly, it can also be presumed that structured information obtained from a certain geography might lead to generation of population health information. Taken together, big data will facilitate healthcare by introducing prediction of epidemics (in relation to population health), providing early warnings of disease conditions, and helping in the discovery of novel biomarkers and intelligent therapeutic intervention strategies for an improved quality of life.

**References**

1.Laney D. 3D data management: controlling data volume, velocity, and variety, Application delivery strategies. Stamford: META Group Inc; 2001
2. Gubbi J, et al. Internet of Things (IoT): a vision, architectural elements, and future directions. Future Gener Comput Syst. 2013;29(7):1645–60.
3.Doyle-Lindrud S. The evolution of the electronic health record. Clin J Oncol Nurs. 2015;19(2):153–4.
4. Reisman M. EHRs: the challenge of making electronic data usable and interoperable. Pharm Ther. 2017;42(9):572–5.
5. Murphy G, Hanken MA, Waters K. Electronic health records: changing the vision. Philadelphia: Saunders W B Co; 1999. p. 627.
6. Reiser SJ. The clinical record in medicine part 1: learning from cases*. Ann Intern Med. 1991;114(10):902–7.
7. Yin Y, et al. The internet of things in healthcare: an overview. J Ind Inf Integr. 2016; 1:3–1
8.Moore SK. Unhooking medicine [wireless networking]. IEEE Spectr 2001; 38(1)
9.Nasi G, Cucciniello M, Guerrazzi C. The role of mobile technologies in health care processes: the case of cancer supportive care. J Med Internet Res. 2015;17(2): e26.
10.Apple, ResearchKit/ResearchKit: ResearchKit 1.5.3. 2017.
11. Shvachko K, et al. The hadoop distributed file system. In: Proceedings of the 2010 IEEE 26th symposium on mass storage systems and technologies (MSST). New York: IEEE Computer Society; 2010. p. 1–10.
12.Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. Commun ACM. 2008;51(1):107–13.
13. Zaharia M, et al. Apache Spark: a unified engine for big data processing. Commun ACM. 2016;59(11):56–65.
14. Gopalani S, Arora R. Comparing Apache Spark and Map Reduce with performance analysis using K-means; 2015.
15. Ahmed H, et al. Performance comparison of spark clusters configured conventionally and a cloud servicE. Procedia Comput Sci. 2016; 82:99–106.
16. Saouabi M, Ezzati A. A comparative between hadoop mapreduce and apache Spark on HDFS. In: Proceedings of the 1st international conference on internet of things and machine learning. Liverpool: ACM; 2017. p. 1–4.
17. Strickland NH. PACS (picture archiving and communication systems): filmless radiology. Arch Dis Child. 2000;83(1):82–6.
18. Schroeder W, Martin K, Lorensen B. The visualization toolkit. 4th ed. Clifton Park: Kitware; 2006.
19. Li L, et al. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. Sci Transl Med. 2015;7(311):311ra174.
20. Valikodath NG, et al. Agreement of ocular symptom reporting between patient-reported outcomes and medical records. JAMA Ophthalmol. 2017;135(3):225–31.
21. Fromme EK, et al. How accurate is clinician reporting of chemotherapy adverse effects? A comparison with patient-reported symptoms from the Quality-of-Life Questionnaire C30. J Clin Oncol. 2004;22(17):3485–90.
22. Beckles GL, et al. Agreement between self-reports and medical records was only fair in a cross-sectional study of performance of annual eye examinations among adults with diabetes in managed care. Med Care. 2007;45(9):876–83.
23. Echaiz JF, et al. Low correlation between self-report and medical record documentation of urinary tract infection symptoms. Am J Infect Control. 2015;43(9):983–6.
24. Belle A, et al. Big data analytics in healthcare. Biomed Res Int. 2015; 2015:370194.
25. Adler-Milstein J, Pfeifer E. Information blocking: is it occurring and what policy strategies can address it? Milbank Q. 2017;95(1):117–35.
26. Or-Bach, Z. A 1,000x improvement in computer systems by bridging the processor-memory gap. In: 2017 IEEE SOI-3D-subthreshold microelectronics technology unified conference (S3S). 2017.
27. Mahapatra NR, Venkatrao B. The processor-memory bottleneck: problems and solutions. XRDS. 1999;5(3es):2.
28. Mahapatra NR, Venkatrao B. The processor-memory bottleneck: problems and solutions. XRDS. 1999;5(3es):2.

29. Dollas, A. Big data processing with FPGA supercomputers: opportunities and challenges. In: 2014 IEEE computer society annual symposium on VLSI; 2014.
30. Saffman M. Quantum computing with atomic qubits and Rydberg interactions: progress and challenges. J Phys B: At Mol Opt Phys. 2016;49(20):202001.
31. Nielsen MA, Chuang IL. Quantum computation and quantum information. 10th anniversary ed. Cambridge: Cambridge University Press; 2011. p. 708.
32. Raychev N. Quantum computing models for algebraic applications. Int J Scientific Eng Res. 2015;6(8):1281–8.
33 Harrow A. Why now is the right time to study quantum computing. XRDS. 2012;18(3):32–7.
34. Lloyd S, Garnerone S, Zanardi P. Quantum algorithms for topological and geometric analysis of data. Nat Commun. 2016; 7:10138.
35. Buchanan W, Woodward A. Will quantum computers be the end of public key encryption? J Cyber Secur Technol. 2017;1(1):1–22.
36. De Domenico M, et al. Structural reducibility of multilayer networks. Nat Commun. 2015; 6:6864.
37. Mott A, et al. Solving a Higgs optimization problem with quantum annealing for machine learning. Nature. 2017; 550:375.
38. Rebentrost P, Mohseni M, Lloyd S. Quantum support vector machine for big data classification. Phys Rev Lett. 2014;113(13):130503.
39. Gandhi V, et al. Quantum neural network-based EEG filtering for a brain-computer interface. IEEE Trans Neural Netw Learn Syst. 2014;25(2):278–88.
40. Nazareth DP, Spaans JD. First application of quantum annealing to IMRT beamlet intensity optimization. Phys Med Biol. 2015;60(10):4137–48.
41. Reardon S. Quantum microscope offers MRI for molecules. Nature. 2017;543(7644):162.