www.rspsciencehub.com Volume 03 Issue 07S July 2021

INTERNATIONAL RESEARCH JOURNAL ON ADVANCED SCIENCE HUB
e-ISSN : 2582 - 4376
Open Access

RSP SCIENCE HUB
(The Hub of Research Ideas)
Available online at www.rspsciencehub.com

Special Issue of Second International Conference on Advances in Science Hub (ICASH 2021)

Big Data Analysis and Management in Healthcare

Madhamsetty Charitha[1], *Nagaraj G Cholli*[2]

[1]*Department of Information Science and Engineering, RV College of Engineering, Karnataka, India.*

[2]*Assistant Professor, Dept. of Information Science and Engineering, RV College of Engineering, Karnataka, India.*

madhamsettyc.is18@rvce.edu.in[1], *nagaraj.cholli@rvce.edu.in*[2]

Abstract

Basically, Big Data means large volumes of data that can be used to solve problems. It has piqued people's attention over the past two decades because of the enormous potential it holds. Big data is generated, stored, and analyzed by a variety of public and private sector industries in order to enhance the services they provide. Hospital reports, patient medical records, medical test outcomes, and internet of things applications are all examples of big data outlets in the healthcare industry. Biomedical research often produces a large amount of big data that is pertinent to public health. To extract useful information from this data, it must be properly managed and analyzed. Otherwise, finding solutions by analyzing big data quickly becomes impossible. The ability to identify trends and transform large amounts of data into actionable information for precision , medicine and decision makers is at the heart of Big Data's potential in healthcare. In a variety of areas, the use of Big Data in healthcare is now offering solutions for optimizing patient care and creating value in healthcare organizations. In this paper, some big data solutions are provided for healthcare. Big Data Analytics strategies to mitigate covid-19 health disparities are provided. Finally we analyse some of the challenges with big data in healthcare.

Keywords: Big Data, Electronic Health Records(EHRs), Healthcare, Internet of things(IOT), Machine Learning, Hadoop, Apache Spark, Medical Imaging, COVID-19

1. Introduction

1.1 What exactly is Big Data?

The word "big data" refers to the data which is so massive, quick, or complex that processing it with conventional methods is difficult or impossible. The practice of accessing and storing vast volumes of data for analytics has a long history but the "big data" concept gained momentum in 2000's. Big data can be described in five V's i.e there are five characteristics of big data. Volume – Volume refers to size of data. The term "BigData" itself refers to enormous size. So, to consider some data as big data, it's volume must be large. The data collected by an organization is generally huge. They collect data from various sources like business transactions, videos, social media, industrial equipment etc. Example: The global mobile traffic, estimated in the year 2016, was 6.2 billion GB of data each month. By the year 2020 it is 40000 billion GB of data. Velocity – The term "Velocity" refers to the rapid accumu- lation of data. Data comes in at a high rate from machines, networks, social media, cell phones, and other outlets in Big Data velocity. A vast and rapid flow of data exists. This influences the data's potential, or how quickly data is produced and analyzed in order to fulfil demands. Example: On google, almost 3.5 billion inquiries are made in the world each day. Likewise, clients of FaceBook are expanding by 23% every year.[1-4].

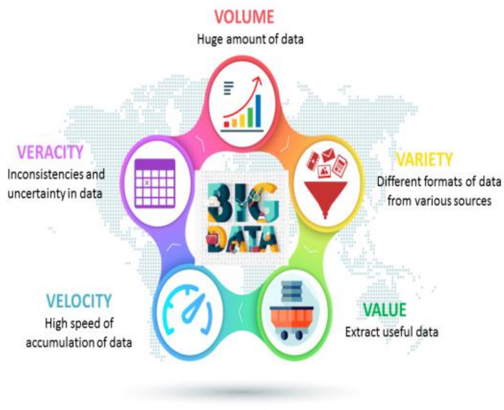

Fig.1. 5 V's of Big Data

Variety – The term "Variety" refers to various forms of data. It may also apply to a variety of sources. Data arrival can be from a variety of sources both within and outside a firm. Data can be structured, semi-structured, and unstructured.

Structured data is simply data that has been structured. It usually refers to data that has been specified in terms of length and format. Semi-structured data is a type of data that is semi-organized. It's a kind of data that doesn't follow the traditional data structure. This form of data is represented by log files. Unstructured data is simply data that has not been organized. It usually refers to data that is not feasible to be stored in a relational database's conventional row and column structure. Unstructured data includes things like text, images, and videos. Veracity - It refers to data discrepancies and ambiguity, i.e., available data can become messy at times, and consistency and accuracy are difficult to monitor. Because of the numerous dimensions of data arising from various diverse data forms and sources, big data is often variable. For instance, a large amount of data can cause confusion, while a smaller data can convey half or Inco. Value – The final V to be considered is Value! The majority of data with no value is useless to the organization unless it is converted into some useful data. Data is of little value or significance in and of itself; it must be turned into something useful in order to obtain information. As a result, it is claimed that Value! is the most significant of the five Vs.[5-8].

1.2 Big Data in Healthcare System

Healthcare System is a framework created for the main purpose of preventing, diagnosing, and treating human health conditions or impairments. Health practitioners(doctors, caretakers), health services (clinics, hospitals for providing medications and other diagnostic or treatment technologies), and a financial structure, assisting health practitioners and services, are the main components of a healthcare system. Dental, medical, midwifery, nursing, counseling, physiotherapy, and other health professions are represented among the health pro- fessionals. Various levels of healthcare are required depending on the severity of the case.

1. Primary care - It is used by physicians as an initial point of contact.
2. Secondary care - Critical care involving qualified professionals.
3. Tertiary care - Advanced medical investigation and
4. Quaternary care - Specialised surgical treatments.

Fig.2. Data Explosion in healthcare

Health practitioners and organizations are in charge of various types of information generated at different stages of healthcare systems, including administrative data such as billing and management data, medical data of patients such as diagnosis details, tests and lab reports and medication data. Earlier, patient's medical records were handwritten or typed. Also medical test reports were held in a paper filing system. In fact, the earliest case reports can be found in an Egyptian papyrus text dating back to 1600 BC. All clinical evaluations and medical records are digitised. This has become a common and largely accepted procedure these days, thanks to the advent of electronic systems and their potential. Figure 2 shows the growth of data generated by healthcare activities from 2012 to 2020.

1.3 Electronic Health Records (EHRs)

In healthcare, EHRs are the widely used form of big data. EHR is a patient's digital record that contains information such as his/her medical history such as diagnosis, health practitioners, hospital details etc, radiology reports, socio-economic information, lab test reports, among other things. The Institute of Medicine, a branch of the National Academies of Sciences, Engineering, and Medicine, coined the word "Electronic Health Records" in 2003 to describe medical records kept for the purpose of enhancing the health-care sector for the benefit of patients and health practitioners. An electronic health record is a computerised replica of a patient's handwritten records or charts). EHRs are patients' real-time records which can be accessed by only authorised users. Along with medical information of a patient, it is designed to include several other details. Figure 3 shows all the components of EHR. EHRs are exchanged through information systems which are secure and are accessible to both private and public sector providers. Every record is made up of a single editable file, allowing doctors to make alterations over time without any paperwork and no risk of data duplication. As a result of the shorter lag time between previous test results, recognizing and treating medical conditions is more effective. There has been substantial reduction in unnecessary and extra tests, as well as in uncertainties caused by unreadable handwriting. There has been better care coordination among different healthcare providers, over time. Getting beyond those logistical blunders reducing medication dosage and frequency errors, the number of drug allergies has decreased. Healthcare professionals have also discovered that they can significantly improve their medical practises by making use of automatic alerts and prompts for periodic checkups, vaccinations etc via web-based and electronic platforms. By encouraging contact among health service providers and patients, there will be more quality of treatment and timely interventions. They're linked to digitized authorization, so there will be faster approval of insurance policies because there's no paperwork involved. Through EHRS, data analyses and generating reports on crucial healthcare quality metrics is quicker. EHRs can also reduce or completely eliminate the discrepancies and delays in the billing system. EHRs,along with the internet can provide health-related information at any time and anywhere which is crucial for a patient's well being.As shown in figure 3, an Electronic Medical Record(EMR) is a component of EHR which contains medical and clinical data of patients exclusively. Another component of EHR is Personal Health Record which contains information about the vital signs such as body temperature, pulse rate, blood pressure etc. EHRs, EMRs, PHRs and other big data components in healthcare have the ability to increase healthcare quality, costs and performance while reducing medical errors.

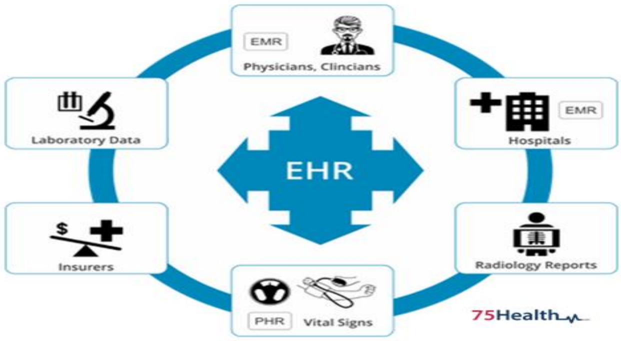

Fig.3. Components of EHR

1.4 Role of Big Data in Biomedical Studies

A biological system shows a complex interplay of molecular and physical events. Usually, data is gathered, experimenting on a smaller component of a biological system to identify and understand various interdependencies among several components of the system during biomedical experiments. As a result, generating a broad map of a biological phenomenon necessitates multiple simplified experiments. This suggests that the more data we have, the better we will be able to understand biological processes. Modern techniques have advanced at a rapid rate as a result of this concept. This suggests that the more data we have, the better we will be able to understand biological processes. Modern techniques have advanced at a rapid rate as a result of this concept. A large amount of data is generated in an effort to decode human genetics. Complex and efficient technologies such as next-generation sequencing (NGS) and genome wide association studies (GWAS) are being used to decode human genetics. The ‘-omics' (The omics branches of science are various disciplines in biology with names that end in the suffix -omics, such as genomics, proteomics, metabolomics, metagenomics, and transcriptomics. The goal of omics is to characterise and quantify pools of biological molecules that translate into the structure, function, and dynamics of an organism or organisms.) Through big data analytics, scientists can now study the entire ‘genome' of an organism instead of studying a single ‘gene’ in ‘genomics' studies in a limited period of time. Same is the case with ‘transcriptomics’ studies. Each of the experiments conducted on omics data generates a significant amount of data with more depth than ever. However, in order to fully know a particular system or mechanism, this level of detail may not be sufficient. So, it is necessary to analyse huge volumes of data gathered from several studies in order to understand various biological phenomena. In-depth analysis of such data from biomedical studies and healthcare systems can be extremely useful in developing new healthcare technologies. The most recent technological advancements in data generation, collection, and analysis would lead to a personalised medicine revolution in the near future.

1.5 Big Data in IOT based healthcare devices

In the field of healthcare, the Internet of Things (IoT) has become a growing movement. IOT devices generate data continuously while keeping track of people's health status. So, IOT devices are the healthcare big data’s major contributors. The data generated by IOT devices can provide elderly and chronic illness patients with an effective diagnosis and treatment services. Through IOT devices, doctors can keep track of their patients’ by measuring and monitoring various parameters in their place of stay itself. As there is early diagnosis and treatment, the patient may not require hospitalization. As a result most of the healthcare cost is saved. Fitness or health-tracking wearable devices, biosensors, clinical devices for monitoring vital signs, and other types of devices or clinical instruments are examples of IoT devices used in healthcare. IoT devices like these generate a lot of health-related data. If the data from these kinds of devices is combined with other healthcare information, such as the data from EHRs or PHRs, the health status of a patient along with progress in his/her recovery can be predicted. Big Data generated from IOT devices has proven to be beneficial in a number of areas, allowing for better diagnosis, predictions and treatment. Figure 4 shows how data from EHRS, PHRS, EMRs, IOT devices and omics data can be integrated and analysed can help in making smart and cost effective healthcare decisions. Huge amounts of data generated from different kinds of sources are stored in data warehouses. This data is processed using analytic pipelines to produce more intelligent and cost-effective healthcare opportunities.

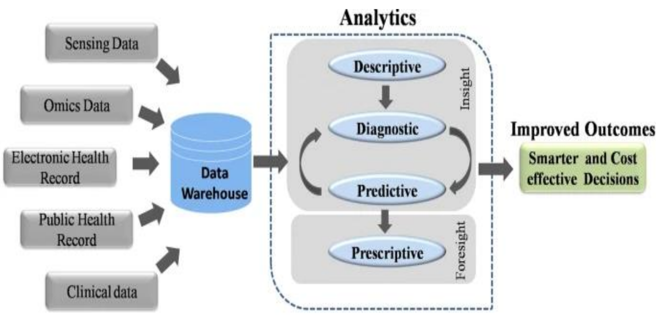

Fig.4. Workflow of Big Data

2. Big Data Analysis and Management in Healthcare

Big data refers to massive amounts of diverse data produced at a fast rate.The information collected from different sources is mainly used to enhance the services provided to customers rather than to increase consumption. Same is the case for healthcare data and data from biomedical studies.The main challenge is figuring out how such a large and homogeneous amount of data(big data) can be analysed. The data must be stored in a format that is readable for analysis before it is made available for the scientific community. There is another challenge for healthcare data, to implement high-end computing protocols, software and hardware. Hadoop and Apache Spark are two of the most widely used platforms for working with big data. These platforms are briefly described below.

2.0.1. Hadoop

The logical way to analyse massive amounts of data is to distribute across multiple nodes and process. But this method needs several computing machines as the data is so large. As a result several issues such as how to distribute data, how to handle errors and failures etc must be addressed.

Hadoop is an open-source application which is very popular. It is mainly designed to address the issues mentioned above that arise due to parallelized analyses of data. Hadoop basically uses the MapReduce algorithm to process large datasets. In this algorithm, map and reduce primitives are used. Each logical record in the input data is mapped into a set of key/value pairs, and the reduce primitive is used to combine all the values that have the same key. It schedules communication among various machines across various clusters of machines and also parallalizes computations. HDFS (Hadoop Distributed File System) is a file system component that allows for scalable, efficient, and replica-based data storage across multiple nodes in a cluster. Hadoop is being used in many large projects, such as drug development using “-omics” data, correlation between bronchial asthma and air quality etc. As a result of the Hadoop system's implementation, efficient use of healthcare data is possible.

2.0.2 Apache Spark

Apache Spark is also an open source platform used for big data analysis. It has high-level libraries such as Spark SQL for SQL query support, Spark streaming for data streaming, GraphX for processing of graphs, and MLlib for machine learning algorithms. As less coding effort involved in the programming interface is used, these libraries help developers to be more productive.

2.1 Commercial Platforms for big data analysis in healthcare

Several firms have used AI to analyse published findings, data in the form of text and image in order to generate relevant findings and handle big data difficulties.

2.1.1 Linguamatics

It is a natural language processing (NLP) algorithm that is based on a text mining algorithm named “I2E" which is interactive. This algorithm can extrapolate and analyse a wide range of data. This technique produces results ten times faster than other tools. It also does not require any expertise in interpretation of data. It is used for analysing and extracting information from unstructured data as well. In order to produce clean and filtered results, traditional ML requires input data to be systematic and carefully chosen. However, when NLP is integrated into healthcare records in general it makes it easier to retrieve clean and organised information that is often concealed in unstructured input data.

2.1.2 IBM Watson

It is IBM's original idea for big data analytics in practically every industry. To extract large amounts of data from the least amount of data, this platform makes substantial use of Machine Learning and Artificial Intelligence based algorithms. In order to offer useful and organised data, IBM Watson maintains a rigid routine of integrating a wide range of healthcare areas. With the purpose of speeding up the discovery of new combinations of immunity and oncology in an effort to make new drugs to treat. The researchers can interpret complex genomic data sets with Artificial Intelligence technologies and deep learning modules of IBM Watson. Based on gene expression patterns acquired from a variety of huge data, it has been used to forecast certain forms of cancer, suggesting the presence of many druggable targets. By integrating carefully chosen literature and constructing network architecture, IBM Watson is also utilised in drug development programmes to offer a complete perspective of the molecular landscape in a given disease model.

2.1.3 Ayasdi

Ayasdi is a large framework which uses Machine Learning based methodologies to provide an application framework and Machine Intelligence platform. It offers a variety of healthcare analytics applications, such as understanding and managing healthcare data variations and transforming medical and clinical costs. Also, various other data such as conversations among physicians, how clinics and hospitals are organised, decision making in risk-based treatments and the care given to the patients can be analysed and managed. It has an application which is used for population health management and assessment, a proactive approach to risk management that goes beyond typical risk assessment. It employs machine learning intelligence to forecast risks in the future, identify risk factors, and provide solutions to achieve the best potential results.

3. Applications of Big Data in Healthcare

3.1 Real time alerts

Medical data is analysed in real time by Clinical Decision Support (CDS) software in hospitals, providing guidance to doctors when they make prescriptive decisions. Doctors, on the other hand, prefer that patients leave hospitals in order to avoid expensive procedures performed in the hospitals. The buzzword of business intelligence in 2019 is "analytics", which has the ability to become a booming topic in the modern world. IOT based wearables capture patient health data. The data collected is transferred into the cloud. Furthermore, this data will be linked to a database on the general public's well being, allowing the comparison of data in a socio-economic sense and accordingly distributing resources. This can be tracked by the institutions and care managers with advanced tools and react when the results seem to be alarming. For example, if there is a dangerous raise in a patient's blood pressure, the machine will send a real-time warning to the physician or caretaker, who will contact the patient and prescribe pressure-lowering measures. Asthmapolis is another example, which has begun to use trackers which are GPS-enabled in inhalers to detect asthma patterns on a personal level as well as in large populations. In order to create improved asthma care plans, this information is being combined with the data obtained from CDS.

3.2 Enhancing the Engagement of Patient

Many customers – and thus potential patients – are already interested in smart devices that track their every move, heart rate, sleeping patterns, and other data continuously. This data can be used to uncover hidden health risks when combined with other data such as EHRs. For example, chronic insomnia as well as an increased heart rate may indicate a potential risk of heart attack.

Patients are directly interested in their own health monitoring, and health insurance incentives will encourage them to live a healthier lifestyle (e.g.: giving money back to people using smartwatches). New wearables in development are another way to do so, monitoring individual health patterns and then transmitting them to the cloud, from where doctors can control them. Patients with asthma or high blood pressure would benefit from it, being more self-reliant and reducing unnecessary doctor visits.

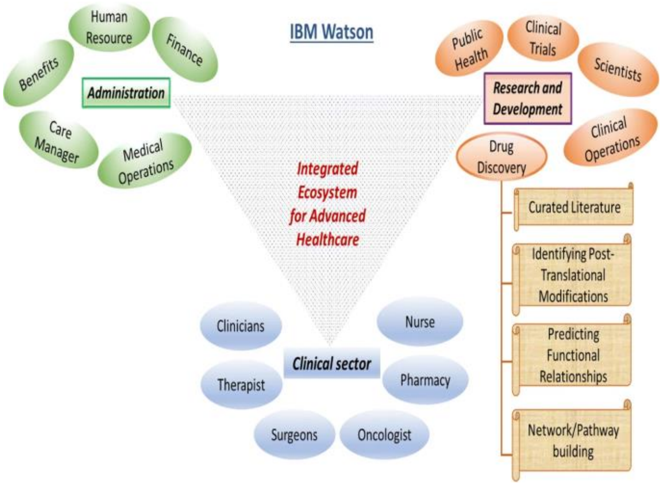

Fig.5. Schematic diagram of multiple models of big data in Healthcare Package of IBM Watson.

3.3 Curing Cancer

The Cancer Moonshot program is another intriguing example to show the usage of big data in healthcare. President Obama devised this program, prior to the conclusion of his second term with the goal of making progress towards cancer curing.

Large volumes of data about cancer patients' diagnosis, treatment methods and recovery rates can be used by medical researchers to find trends and therapies with the best success rates in the real world.Researchers can look at tumour samples in biobanks that are connected to patient treatment data, for example. Researchers may use this information to evaluate how particular mutations and cancer proteins interact with various therapies, as well as identify trends that will lead to better patient outcomes. This information can also lead to unanticipated advantages, such as discovering that the antidepressant Desipramine can help treat some forms of lung cancer. However, patient datasets from other organisations, including hospitals, colleges, and organisations, must be connected in order to make these sorts of insights more accessible.

3.4 Integrating with Medical Imaging

Medical imaging is critical, with approximately 600 million imaging procedures performed in the United States each year. The most common types of medical imaging are X-Rays, CT scans and MRI(Magnetic Resonance Imaging). Manually analysing and storing these images is time-consuming and costly, as radiologists must look at each image separately, and hospitals must keep them for so many years.

Curemetrix, a medical imaging company, demonstrates how healthcare is digitized rapidly through big data analytics. It may transform the way images are read: algorithms built via the analysis of lakhs of photos might discover certain patterns in the pixels and transform them into a number to aid the physician in diagnosis. They even go so far as to argue that radiologists may no longer need to look at images, instead analysing the results of algorithms that will undoubtedly study and memorise more pictures than they could ever recall in a lifetime. This would surely have an influence on radiologists' roles, education, and skill set.

3.5 Prevent unnecessary ER (Emergency Room) visits

There is an example which shows how big data can prevent unnecessary ER visits.

PreManage ED is a programme developed by Alameda County hospitals in which patient information is shared across emergency rooms. This system provides information to the emergency room staff such as:

- If the patient being treated has previously undergone specific tests and the results of those tests at other departments or facilities.
- If the patient being treated has a manager for his case in some other hospital, avoiding unnecessary assignments.
- What advice has previously been given to the patient, so that professionals may keep their message consistent.

This shows that big data plays an important role in reducing unnecessary ER visits.

3.6 Prior Risk Management

Healthcare data and big data analytics are critical for reducing the likelihood of hospitalisation for certain chronic disease patients. Through the combination of healthcare data and big data analytics, degeneration can also be prevented.

Healthcare facilities can give correct preventative treatment and, as a result, reduce hospital admissions by diving down into details like the type of medicine, symptoms, vital signs and the number of medical appointments, to name a few. Not only will this type of risk assessment save money on treatments given in hospitals but also ensure that resources such as ICU beds, equipment, medicines etc are available for those who need them most.

3.7 Suicide Prevention

Every year, almost 800,000 individuals die by suicide around the world. In addition, almost 17% of the population would harm themselves at some point in their lives, in the world. These figures are concerning. While this is a very uncertain topic to address, big data applications in healthcare are assisting in the prevention of suicide and self-harm. Healthcare organisations can utilise data analysis to identify individuals who are prone to kill themselves because they see a large number of patients every day.

A combination of data from big data components in healthcare such as EHRs, PHRs etc and the data of persons who are identified as persons with an increased risk of suicide attempt through a questionnaire containing standard depression questions with excellent accuracy in a 2018 study by KP and the Mental Health Research Network. The researchers revealed that among the top 1% of patients who are identified based on certain statistics, using a prediction algorithm., suicide attempts and successes were 200 times more probable. This shows that big data analysis can also be used in the suicide and self-harm prevention.

3.8 Big Data Applications in COVID-19 Pandemic

COVID-19's rapid global spread has brought powerful big data analytics techniques to the forefront, with entities from all sectors of the healthcare business attempting to monitor and mitigate the virus's impact.

The big data analytics are used in multiple areas like diagnosis, estimation of risk, healthcare decision making and pharmaceuticals during this COVID-19 pandemic.

3.8.1 Diagnosis

The Reverse Transcription-Polymerase Chain Reaction (RT-PCR) test is used to diagnose suspected COVID-19 patients. Depending on the various variables, this test can take anywhere from 24 hours to several days. Many nations saw an upsurge in demand for identifying suspected COVID-19 cases, which outstripped local testing capabilities. As a result, various researchers have proposed alternate solutions for the COVID-19 RT-PCR diagnostics test, such as the ones listed below.

1. A model to differentiate COVID19 from four other viral chest diseases was proposed by researchers. The model collects data and monitors the patient's health using a variety of body sensors, including temperature, blood pressure, heart rate, respiration monitoring, glucose detection, and others. The collected data is saved on a cloud database with AI-enabled expert systems that aid in the diagnosis of symptoms in patients infected with or suspected of having COVID-19 in order to select the best course of action.

2. The design of the medical device which can be used to identify and track COVID-19 symptoms was flexible and cheap cost, offered by the researchers. It uses a mobile phone and headphones to identify respiratory issues. Signals are captured and saved via the Mobile App in an audio file format, after which the signals are processed using the application MATLAB to detect COVID-19 respiratory symptoms.

3. Researchers also created a tool to remotely monitor COVID19 patients who have been released. A pulse oximeter and Thermometer are provided to report daily symptoms and O2 saturation and temperature to each patient registered in the app. A group of nurses is responsible for evaluating the anomalous sensors of vital signs and symptoms. The patient may be readmitted to the Emergency Department according to the evaluation results (ED). The programme reduces the use of EDs and gives the person who is discharged from the hospital scalable remote monitoring capabilities.

4. The researchers also invented smartwatches that are used to detect COVID-19 pre-symptomatically. They examined the physiological and activity data acquired from the infected COVID-19 cases' smartwatches. They concluded that a two-level warning system based on substantial increase in resting heart rate relative to individual baseline may detect 63 percent of COVID-19 cases before symptoms appeared. Furthermore, they discovered that adopting wearable devices for activity tracking and health monitoring can aid in the early detection of respiratory infections.

3.8.2 Estimation of Risk

Estimating the risk score aids in evaluating the level of care and priority for each patient, as well as providing insight into the necessary proactive steps. The researches and inventions in this area that used big data are listed below.

1. The researchers sought to validate a notion that COVID-19 infection could result in significant cardiovascular disease or worse. They made use of statistics. To investigate COVID-19, a multi-factorial logistic regression model was used. cause and effect The study included 54 patients of various ages, genders, and ethnicities and vital signs, of which 39 were classified as serious COVID-19 cases, 15 of which are critical. The information was gathered clinically from patients who had attached vital sign measurement devices that were updated every four hours. The findings revealed that older men, diabetes patients, and hypotension patients are more likely to have a significant heart problem and require more care.

2. A risk score to predict adverse outcomes in COVID-19 patients is developed. They conducted a retrospective cohort study of adult emergency department visits. The primary outcome of the trial was death or no respiratory decompensation within 7 days. They employed the Least Absolute Shrinkage and Selection (LASSO) and Logistic Regression methods to calculate the risk score. They determined that the COVID-19 Acuity Score (COVAS) can help with patient discharge decisions during the COVID-19 pandemic. They also reported on the development and validation of measures for pneumonia or COVID-19 diagnosis in cohorts and subgroups.

3. They also proposed an Internet of Things (IoT) based system to discover unregistered COVID-19 patients, as well as infectious places. This would help the responsible authorities to disinfect contaminated public places and quarantine the infected persons and their contacts even if they did not have any symptoms. The newly confirmed and recovered cases would be recorded in the system by the healthcare staff, while the geolocation data will be collected automatically by Global Positioning System(GPS) technology in the IoT devices.

4. A model that anticipates the course of the outbreak to aid in the planning of an effective preventative technique was proposed. SIDURTHE (susceptible, infected, diagnosed, unwell, recognised, threatening, healed, and extinct) is a model stage. It differentiates infected people based on whether or not they have been diagnosed and the severity of their symptoms. The simulation results generated by merging the model with the existing data on the COVID-19 pandemic in Italy indicate that it is a critical need.

3.8.3 Healthcare Decision Making

During the COVID-19 epidemic, there was a surge in the demand for emergency rooms and medical equipment such as ventilators. As a result, several studies have tried to create monitoring tools and models that aid in making various medical decisions using big data to reduce potential hazards, and these solutions include the following.

1. A patient monitoring platform that enables for daily electronic checking of symptoms, advice and reminders via text messages, and phone care is developed. Patients who have registered in the system complete a daily questionnaire in which they rate 10 symptoms on a scale of 0 to 4. Questionnaire responses are used to categorise patients and specify the care needed, in addition to assessing how much the infection affects them, the amount of analgesic/antipyretic tablets they take, and the temperature monitored. The platform depends on the following data: the number of patients tracked over time, the daily symptoms score, and the number of ED visits per day.

2. A COVID-19 app was developed which has user satisfaction and data usability gathered to assist decision-makers and healthcare providers. The app collects information from patients on a daily basis, such as symptoms, vital signs, and an assessment of their health. The obtained data is disseminated on an interactive map for each user based on their postal code, which aids in their understanding of the area’s healthcare consumption along with regional distribution of infection dissemination.

3. The COVID-19 pandemic, an analytical model for predicting patient census and determining ventilation needs for a given hospital is developed. This model is trained on data such as length of the hospital stay and the number of days spent on a ventilator of several patients to predict the number of ventilators and patients. It is found that there was no correlation between the age of hospitalized patients and their chance of requiring ventilators or between the gender of the inpatient and the length of the stay. It is suggested that hospitals rely on internal data for accurate resource planning.

4. Researchers also made an attempt to summarise the clinical characteristics of COVID-19 patients and discovered criteria that predict ICU admission. They discovered that the necessity for a COVID-19 patient to be admitted to the ICU may be anticipated by examining a set of easily acquired medical parameters: age, fever, and tachypnea with/without respiratory crackles. To extract information from the medical data, they employed the EHRead technology To classify the retrieved data, deep learning convolutional neural network classification methods are also used.

4. Challenges

In this section, we will go through some of the challenges of big data in healthcare

4.1 Storage

One of the key challenges is storing massive amounts of data, although many firms are interested in storing data within their own organization. This has a number of benefits such as less risk of security breaches, access controls etc. Using a server network with the firm, on the other hand, can be costly to scale and complex to operate. With falling costs and improved reliability, it appears that storage in a cloud employing IT infrastructure is a better alternative, which has been by most healthcare organizations. Organizations must partner with cloud providers who understand the importance of healthcare related conformity and security issues. Cloud storage has also resulted in lower startup costs, faster catastrophe recovery, and simpler expansion. Organizations can also employ a hybrid data storage approach, which is the adaptive and practical choice for service providers with varying data storage and access requirements.

4.2 Data processing before Analysis

After acquisition, the data must be pre-processed using certain techniques to maintain accuracy, uniformity, relevance, and clarity. To maintain high levels of correctness and integrity, pre-processing techniques such as data scaling, data integration, data reduction etc can be performed manually or automatically using certain logic rules. Cleaning large volumes of data manually is a very tedious task. Machine Learning algorithms are used in more complex and precise technologies to save time and money and to keep poor data from impeding big data initiatives.

4.3 Consistency in data format

Healthcare systems generate a large amount of data in different forms and formats that is difficult to gather because data in various formats is complex and difficult to handle. It is quite tough to manage massive data, specifically to healthcare providers who lack a perfect data organisation. There was a need to maintain a certain format in all the clinical and medical data information for the purposes of assertions, billing and medical analytics. To capture the essential clinical ideas, some coding systems such as Current Procedural Terminology (CPT) and International Classification of Diseases (ICD) were developed and deployed in the field of medicine. These code sets, however, have their own set of constraints.

4.4 Maintaining Accuracy

Some studies have found that reporting the data of patients into EHRs or PHRs is not totally precise yet, most likely due to Poor EHR functionality, convoluted procedures, and a lack of awareness of why big data gathering is so important. These elements might put up to big data quality challenges throughout its lifecycle. Although reports suggest disparities in these situations, EHRs promise to improve data quality and communication in healthcare workflows. The use of self-report surveys from patients for their symptoms may improve documentation quality.

4.5 Security Issues

Because of the numerous breaches of security, phishing and hacking attempts, and malware occurrences, healthcare businesses have made data protection a top concern. Following the discovery of a number of possibilities of security breach, many technical protections for Protected Health Information(PHI) were produced. These standards, known as Health Information Portability and accountability Act(HIPAA) security rules and regulations, assist enterprises in storing, transmitting, and authenticating data, as well as controlling unity, accessibility and scrutiny. Common security precautions such as employing anti-virus software that is up to date, firewalls, making sensitive data difficult for cryptanalytic attacks by encrypting sensitive data using the best ciphers, and strong authentication using several factors may all help save time and effort.

4.6 Meta-Data Storage

To implement an effective data governance strategy, every stored data must have updated, complete and precise meta-data. It includes details such as the moment of creation, the goal and person in charge of the data, and by whom, how, when and where the data is used for academics and data scientists. This would enable analysts to reproduce earlier questions, which would aid in future scientific investigations and precise specifications. This improves the quality of data usage and puts a stop to the formation of "data trash" with little use.

4.7 Extracting information

Through metadata, extracting data(querying) and getting responses is easy for organizations. However, because of the incompatibility of datasets, data extraction tools may not be able to obtain and maintain the entire storehouse of data. Furthermore, diverse dataset components must be properly integrated, connected, and conveniently available, or else a comprehensive picture of a patient's health cannot be generated. Medical coding systems such as SNOMED-CT, ICD-10 or LOINC must be used to transform free-form ideas into a standard ontology. Structured Query Language (SQL) may be used to query enormous datasets and relational databases if the data's correctness, completeness, and standards are not in question.

4.8 Data sharing

Patients may not receive treatment in more than one location. In this situation, data sharing with other healthcare groups would be critical. If the data is not compatible during such exchange, data transfer across various entities may be substantially hampered. Causes for this can be technical or organisational constraints. Because of this, doctors may not receive critical data needed to make decisions about patient diagnosis and treatments methods. A non-profit organization “Commonwell”, a standard exchange platform built on consensus “Carequality” and Fast Healthcare Interoperability Resource (FHIR) are making data sharing simple and protected from security breaches. The idea of data as a commodity that may give a competitive advantage is the most major hindrance to data sharing. As a result, suppliers and customers may purposely disrupt the flow of data in order to hinder the transfer of data across the EHR system.

5. Future of Big Data in Healthcare

Through big data analytics, both structured and unstructured data can be processed and information can be extracted. The transition to a data environment that is well integrated is the biggest challenge. Surprisingly, the big data notion is based on the notion that the more data available, the more perceptions and projections for future occurrences may be gained. The big data systems in healthcare business is expected to increase at an enormous rate, according to many credible consulting organisations and health care companies. We've seen a wide range of analytics now in use, all of which have shown to have major ramifications on healthcare decision-making and performance.The enormous rise of clinical data from diverse fields has compelled professionals to devise novel methods for analysing and interpreting such massive amounts of data in a certain time limit. Several computational systems are integrated by researchers and practical medical practitioners for processing of signals. As a result, creating a detailed model of the biological system by the combination of functional data of various body parts with “-omics" approaches would be the upcoming goal. This revolutionary concept has the potential to improve our understanding of illness conditions and aid in the creation of newer tools for diagnosis. The continual increase in genetic data availability, as well as inherent concealed mistakes from experimentation and analytical procedures, necessitates additional investigation. However, there are chances to implement systemic changes in healthcare research at each stage of this lengthy process. In terms of cautious integration and application, the large size of clinical data acquired from disparate sources has been a challenge for data analysts. As a result, it is claimed that a new and drastic change in healthcare is required to bring together biomedical information, health information, and analytics to encourage tailored and more efficient therapies. In order to extract valuable information, new techniques and technologies must be created to comprehend the nature of the data (structured, semi-structured, unstructured), heterogeneity (dimensions and characteristics), and amount of data.The main advantage of large data is its endless potential. Within the last few years, the emergence and integration of big data has resulted in significant advances in the healthcare systems and organizations, spanning from medical data management to drug development programmes for severe human illnesses including cancer and neurological illnesses. To offer a basic example, advances in the EHR system in terms of data acquisition, administration, and usability have been seen in the healthcare market since the late 2000s. It is believed that, Big data will complement and boost the existing way of healthcare breakthroughs, rather than replacing qualified people, subject knowledge specialists, and data scientists, as many have stated. The shifts in the healthcare market from a broad capacity base to custom domain are clearly visible. As a result, technologists and professionals must be aware of this changing situation. Big data analytics is expected to progress toward a predictive system in the coming year. This would imply predicting future outcomes in a person's health based on currently existing or historical information (like EHRs data and Omics-based data). In the same way, it's possible that structured data gathered from a specific location will lead to the generation of people’s health data. Big data is going to improve healthcare systems by allowing for the early detection of illness conditions, the identification of novel biomarkers, and the development of therapeutic intervention using intelligence techniques for a better living.

Conclusions

This paper explains big data in healthcare, management and analysis of big data in healthcare, big data solutions in healthcare, use of big data during COVID-19 pandemic, challenges being faced and future prospects. In recent years, big data has gained a wide popularity. The paper explains various forms of big data in healthcare (EHRs, EMRs, MPMs and PHRs), big data in biomedical research(-omics data) and big data in IOT based healthcare devices. The paper explains how big data is managed and analysed in healthcare and also mentions platforms for big data analysis (Apache Hadoop and Apache Spark). Furthermore, the paper explains how machine learning is used in information extraction and analysis of data in healthcare. Then various approaches of using big data in healthcare are mentioned along with the utilization of big data in healthcare during COVID-19 pandemic. Then various challenges faced for big data analysis in healthcare are explained along with future prospects of big data in healthcare. We believe this paper will prove to be helpful to anyone who wants to know various advantages of big data in healthcare systems, its analyses, management and problems being faced. A subsequent work with respect to efficient analysis and management of big data in healthcare can be undertaken. That can serve as a basis for tools and technologies to be developed to make big data analytics an asset to doctors, healthcare workers and also patients.

References

[1]. Doyle-Lindrud S. "The evolution of the electronic health record," Clin J Oncol Nurs. 2015;19(2):153–4.

[2]. Yu XT, Zeng T. "Integrative Analysis of Omics Big Data," Methods Mol Biol. 2018;1754:109-135, doi: 10.1007/978-1-4939-7717-8_7, PMID: 29536440.

[3]. Dash S, Shakyawarm S.K, Sharma M. "Big data in healthcare: management, analysis and future prospects."J Big Data 6, 54 (2019).

[4]. Al Sunaidi S.J, Almuhaideb A.M, Ibrahim N.M, Shaikh F.S, Alqudaihi K.S, Alhaidari F.A, Khan I.U, Aslam N, Alshahrani M.S "Applications of Big Data Analytics to Control COVID-19 Pandemic," Sensors 2021, 21, 2282.

[5]. A.Oussous, Z.Benjelloun, A.A.Lahcen, S.Belfkih "Big Data technologies: A survey" J King Saud Univ- Comput Inf Sci, 18 (2017).

[6]. Aurelle Tchagna Kouanou, Daniel Tchiotsop, Romanic Kengne, Djoufack Tansaa Zephirin, Ngo Mouelas Adele Armele, René Tchinda, "An optimal big data workflow for biomedical image analysis", "Informatics in Medicine Unlocked", Volume 11, 2018, Pages 68-74, ISSN 2352-9148.

[7]. Yin Y, et al. "The internet of things in healthcare: an overview." J Ind Inf Integr. 2016;1:3–13."

[8]. 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.

[9]. Blagoj Ristevski, Ming ChenJ - "Big data Analytics in Medicine and Healthcare" Integr Bioinform. 2018 Sep; 15(3): 20170030. Published online 2018 May 10.