BIG DATA ANALYTICS AND DATA MINING TECHNIQUES IN THE PUBLIC HEALTH INFORMATICS

Dr.Rakesh Kumar Giri,
Assistant Professor,
Department of Computer Science & Engineering,
Bharath University, Chennai, India

Abstract: The field of clinical medicine, along with the administration of healthcare, is currently going through a period in which there is a rapid increase of the methods and procedures that are related with Big Data. This expansion is simultaneously taking place in a number of different components of the healthcare system. The use of healthcare analytics has the ability to lower the financial burden of obtaining treatment, foresee the beginning of epidemics, aid in the prevention of diseases that may be averted, and, in general, increase the quality of life. In this article, we would like to discuss the relevance of big data with regard to the fields of medicine and healthcare as the background for our discussion. The idea that is being alluded to by the phrase "Big Data" is one that is one that is regarded as being one that is rather abstract, and the meaning of this notion is not generally acknowledged within the scientific community. On the other hand, when it comes to the research conducted in the field of health informatics, enormous data sets of this kind are not something that are seen very frequently. The discipline of public health informatics makes use of a number of different methodologies, two of which are data mining and analytics. In order to get a more in-depth comprehension of a variety of medical conditions, these methodologies are used to data gleaned from populations. This is done so that a greater quality of service may be provided to the general community as a whole.

***IndexTerms*: Big Data, Data Mining, Public Health Informatics**

I. INTRODUCTION

Big data refers to an enormous quantity of data that is stored in a variety of formats (both clinical and financial), as well as a combination of various formats, which can be unsettling for professionals in the medical field. This is due to the fact that big data comprises clinical data formats in addition to financial data formats. It is not out of the question that enormous amounts of data could end up being an incredibly useful resource for healthcare organisations. This could be the case in terms of improving the level of care that is given to patients as well as reducing the costs that are associated with providing medical services. For a variety of reasons, including the management of financial risk and regulatory compliance, amongst others, big data is becoming increasingly important to organisations operating in the pharmaceutical and biotechnology industries. It is anticipated that this pattern will carry on well into the foreseeable future. Big data in the medical field refers to the massive volumes of data that may be collected because to advancements in digital technology. The administration of hospitals' general operations is made easier thanks to the huge amounts of data that are collected, and the quality of treatment provided to patients also benefits from their use. Even while these digital technologies are capable of collecting patient data, the volumes of data that they acquire are, on the whole, too enormous and complex to be managed by ordinary technological platforms. There are a few essential obstacles that need to be conquered in order to successfully provide digital health care to mobile patients. One of these challenges is the necessity of protecting and integrating information on patients in order to be in a position to make proactive decisions and diagnoses pertaining to patient care. Rasid Z. Big data infographic (2013) If the delivery of healthcare is going to be improved, the focus of those who offer healthcare will need to shift from making healthcare available and effective inside a certain region or place to making healthcare available and effective independent of the individual's mobility. This is essential in order to realise the objective of strengthening the delivery of healthcare. Big data will be one of the most crucial factors to consider throughout this shift due to the fact that it will have a significant impact on this change and will be one of the most critical components during this transition. The application of prescriptive analytics and personalised medicine, clinical risk intervention, predictive analytics, waste and care variability reduction, automated external and internal reporting of patient data, standardisation of medical terms and registries, and fragmented point solutions are some of the challenges that need to be addressed, and "are some of the advantages that can be gained by using big data analytics in the healthcare sector".

Big data analytics have the potential to make a significant difference in the field of medicine, and there is even the possibility that it may save people's lives. Additionally, there is the possibility that it will lead to better outcomes for patients. The massive amounts of data that are generated as an immediate consequence of the broad adoption of digital technologies and that are afterwards analysed by those technologies are referred to as "big-style data," and the phrase "big data" is used synonymously with "big style data." Utilization in medicine has the potential to, among other things, contribute in the prevention of epidemics and pandemics, treat sickness, and reduce the expenses connected with seeking medical care.

It has been, and will continue to be, a painstaking endeavour that involves a large commitment of both time and money in order to collect huge volumes of data for the aim of applying it to medical research. The gathering of information and the writing of reports that are related to medical care are two instances of labor-intensive operations that have gotten easier as a direct result of improvements in technology. These are both related to the provision of medical treatment. These data may be analysed and put to use in order to enhance the standard of medical treatment that is delivered to patients. In order to maintain coherence throughout the entirety of this post, from this point forward, we are going to refer to the analysis of data relevant to healthcare as "predictive and prescriptive analytics." Our goal is to make sure that the text is clear and understandable at all times. Utilizing a variety of various technologies, it may also speed up the process of analysing treatments and procedures, maintain accurate records of stock and inventory, and enable patients to take responsibility for their own healthcare. These potential benefits might end up materialising.

II. BIG DATA IN HEALTH INFORMATICS

The concept that is referred to by the phrase "Big Data" is one that is considered to be somewhat abstract, and the meaning of this concept is not commonly accepted within the scientific community. According to a working definition of the phrase "massive data," any data with a size of about one petabyte (1015 bytes) or larger would be considered to fall under this category. On the other hand, in the field of health informatics research, massive data sets of this scale are not something that are often observed. As a direct result of this, we will be utilising a definition that is considerably more all-encompassing in order to accommodate a higher quantity of study. To be more explicit, we will be using a definition that characterises big data based on the following five characteristics: volume, velocity, diversity, veracity, and value. This will allow us to better understand what big data is. The terms "volume" and "velocity" refer to the rapid rate at which new data is generated; the term "variety" refers to the level of difficulty posed by the data; the term "veracity" determines whether or not the data are authentic; and the term "value" determines how advantageous the quality of the data's source is in relation to the outcomes that are desired. [CDATA[The terms "volume" and "velocity" refer to the rapid rate at which new data is generated The data that was collected for the investigation into health informatics does, in point of fact, illustrate some of these qualities. The term "high volume" refers to the copious amounts of patient medical histories that are meticulously documented for each and every one of the facility patients. For instance, in some datasets, each instance is extremely large (such as datasets that use MRI pictures or gene microarrays for each patient), whereas in other datasets, there is a large pool from which to collect data, and in still other datasets, there is no differentiation between the two types of instances (such as social media data gathered from a population). Big Velocity is when new data are coming in at a high speed, which can be seen when trying to monitor real-time events, such as monitoring a patient current condition through medical sensors or attempting to track an epidemic through a large number of incoming web posts. Big Velocity can be seen when trying to monitor real-time events, such as monitoring a patient current condition through medical sensors. When attempting to monitor real-time events, such as a patient present status being monitored by medical sensors, one can observe the effects of Big Velocity. When employing medical sensors to try to keep track of the patient's current condition, it's possible that a large quantity of velocity may be detected (such as from Twitter) .

Big Variety can refer to datasets that contain a large number of different types of independent attributes, datasets that are gathered from many sources (for example, search query data comes from many different age groups that use a search engine), or any dataset that is complex and, as a result, needs to be seen at many levels of data throughout Health Informatics. Big Variety can also refer to datasets that are gathered from many different sources (for example, search query data comes from many different age groups that use a search engine). The term Big Variety may also be used to refer to datasets that have been compiled from a wide variety of sources (for example, search query data comes from many different age groups that use a search engine). The phrase Big Range may also be used to refer to datasets that have been gathered from an extensive variety of diverse sources (for example, search query data comes from many different age groups that use a search engine)". Due to the fact that it is possible to deal with data that is noisy, incomplete, or wrong, high data veracity is a problem in Health Informatics, just as it is in any other sector that utilises analytics. This worry is shared by all other fields that use analytics (as could be seen from faulty clinical sensors, gene microarrays, or from patient information stored in databases). In order for this sort of information to be used successfully, it has to be carefully managed and examined. Because the fundamental goal of health informatics is to enhance HCO, a substantial amount of attention is placed on the value of data. This is because health informatics was developed for this reason. When contrasted with the value of data obtained through traditional methods (such as in a clinical setting), which is typically considered to have High Worth, the value of data gathered through social media (data supplied by anybody) may be called into question. Traditional methods of data acquisition are typically considered to have High Worth. Having said that, it was demonstrated in the preceding part of the chapter under the heading "Using population level data - Social media" that this may also have High Value. Not all of the research that is being covered in this article or in the field of health informatics meets all five of the criteria that Demchenko et al(2014)". listed as being characteristics of big data. This is an essential point to keep in mind. In the paper that was mentioned previously, these factors were discussed as being features of large amounts of data.

In spite of this, a sizeable fraction continue to impose severe computational restrictions that need to be addressed in one way or another. It can be difficult to preserve datasets offline, such as electronic health records, even when the data in question "does not exhibit Big Velocity or Variety (EHR). For high-throughput processing, which is obvious in real-time continuous data, methods are necessary that are competent and efficient even when individual data instances do not contain a Big Volume of data. These methods must be able to handle small volumes of data. Data that has Big Value but not Big Veracity may require complicated approaches in order to" reach a consensus among the multiple models, or it may require time-consuming alterations to the data, which may result in an increase in the size of the dataset. Both of these scenarios may be required if the data has Big Value but not Big Veracity. If the data has Big Value but not Big Veracity, then both of these possibilities might need to be considered. Both of these possible scenarios bring up a large amount of difficulties.

It is necessary to first go through all of this information and then carefully analyse it before the health care system can reap any benefits from it. It is possible for electronic health records (EHRs), which are able to store more than 44 petabytes of patient information, to be used to store health information. This is in addition to the many different presentation formats that are available for the data pertaining to a patient's medical history. This explosion of data has also been witnessed in the field of bioinformatics, which is related to the fact that genome sequencing may yield many terabytes of data. The reason for this is due to the fact that genome sequencing is becoming more and more common. One of the areas that is seeing an explosion of data is the field of health informatics. The community of people who work in health informatics has the responsibility of figuring out how to deal with the abundance of data, which originates from a wide variety of sources and can take on a variety of formats. This responsibility comes with the added complexity of the fact that these data can take on a variety of forms. It would appear that the practise of integrating and combining data from multiple sources, even across various subfields (for example, "translational bioinformatics), and even between Health Informatics and Bioinformatics" is gaining in popularity at a rate that is accelerating at an ever-increasing rate. This could be due to the fact that integrating and combining data from multiple sources can lead to the discovery of previously unknown relationships between data sets. Patients, who are the health care system's ultimate benefactors, stand to gain an incredible lot from the effective integration of the huge volumes of data that are involved in this endeavour. Desombre T, et al.(2014)

III. PUBLIC HEALTH INFORMATICS – SOCIAL MEDIA

In the discipline of "public health informatics, data mining and analytics are applied to population data in order to" achieve the goal of gaining a deeper comprehension of various medical concerns. The population is the source of the "data that is utilised in public health informatics, and this data can be gathered either through traditional means (such as from experts or hospitals) or directly from the people themselves (social media)". In any event, there is an abundance of demographic data, not to mention a high velocity and a wide variety of it. The amount of this data is particularly noteworthy. If the appropriate methods are used to extract the relevant "information from social media (for example, Twitter posts), then this line of data can also have a Big Value. However, it is possible that the data collected from the population through social media will have a low Veracity which will result in a low Value

IV. APPLICATIONS OF DATA MINING ON MEDICAL INFORMATICS IN INDIA

Clinical care, administration of health services, medical research, and education and training make up the majority of the subfields that fall under the umbrella of medical informatics . In this section, we will give the specifics of each sector with regard to India.

Clinical care

The vast majority of the time, the process by which doctors or nurses make recommendations for new drugs to patients is informed by the information that is contained in previous medical reports. However, because India does not have a centralised "electronic patient record database system, the vast majority of the country's physicians prescribe" drugs without first doing a thorough examination of the patient's medical history or family records. Even in the vast majority of situations, they only prescribe medicine to the patient based on the symptoms they are experiencing without first double-checking those symptoms utilising a variety of diagnostic procedures. If India's health care system uses a centralised database of patient records, then clinicians will be able to go beyond what is considered to be the most acceptable way to treat a patient by employing data mining techniques on that database. In other words, they will be able to treat patients in a manner that is more effective than what is currently considered to be the most acceptable treatment. A new doctor may suggest a more effective drug to a patient if the patient's circumstance is similar to a prior one by searching the centralised database for a decision. In this scenario, the patient would benefit from the new medication (Raghupathi, 2010).

The Clinical Decision Support System, more commonly referred to as CDSS, is a piece of software that is designed to aid medical practitioners in making decisions regarding the treatment of patients. This programmer is designed to operate on computers. The numerous implementations of CDSS include search capabilities for medical questions, the capacity to "monitor inputs and check them for specific triggers, reminders for periodic chores, advice based on medical information, and various prediction models such as diagnosis and prognosis (Raghupathi, 2010). Despite this, India does not yet have access to the CDSS. The terms knowledge based systems" and "data mining based systems" refer to the most popular types of CDSS". By merging these two distinct kinds of systems, it is possible to achieve a high degree of overall performance. In spite of the fact that it was developed with India's rural areas in mind, the hybrid Community Development Service Delivery System has not been put into use (Iqbal, 2012). provides a visual representation of how the CDSS works. A number of the CDSS's subsystems, including the "Health Evolution through Logic Processing (HELP) system, the Acute Physiology and Chronic Health Evolution (APACHE) series of models, and the pneumonia Severity of Illness Index, are currently being put to use in the healthcare facilities of developed countries (Raghupathi, 2010). It is now time for India to begin taking advantage of the various technical advances that have been made".

Administration of health services

Administrators at health care organisations are responsible for a vast array of tasks, all of which are performed with the intention of enhancing the population's health as a whole. These acts are totally contingent on the information that is relevant to the scenarios with regards to which judgements are going to be made. Administrators are going to make judgments on the arrangement of additional tools and equipment for a certain period of time, such as during an epidemic of a disease, and they are going to decide whether or not any additional services are required in terms of cost and benefit. For instance, during an epidemic of a disease, administrators are going to make decisions regarding the arrangement of additional tools and equipment. These types of decisions are often made with the assistance of a computer-based system in more developed nations. This kind of system is able to provide an accurate estimation of the needs of the nation (Raghupathi, 2010). However, health care facilities in India do not have this kind of organisation, and as a consequence, there are times when they are unable to handle the number of patients or treat them adequately owing to a lack of tools and beds at that time. POMDPs, which stands for "Partially Observable Markov Decision Processes, is a method that was created to identify instances of disease outbreaks by Izadi and Buckeridge(2016). This system is able to suggest a better solution to overcome disease outbreaks in terms of cost and effects, as well as predict about outbreaks; however, the amount of false detection of outbreaks by this system is not affordable for practical use. This system also has the ability to predict about outbreaks (Raghupathi, 2010). The administrators of healthcare facilities in India require such a system in order to improve the quality of treatment they provide to patients and to improve the quality of the facilities themselves. Despite the fact that such a system can only have a small margin of error, it is necessary for them to improve both the quality of treatment they provide to patients and the quality of the facilities themselves".

V.BIG DATA AND DATA MINING IN HEALTH INFORMATICS

"Data mining" refers to the process of collecting meaningful information from large data sets, which are sometimes referred to as "big data" at times. At this point in the procedure, an analysis of these massive raw datasets will be carried out utilising one or more of the available software programmes. Data mining is a relatively new concept that emerged in the middle of the 1990s along with a fresh notion and approach for the analysis of data as well as the finding of new information. The first annual ACM conference on knowledge discovery and data mining was held in the United States of America in 1995. The conference was held there since it was the host country. However, an application to register the term "Data Mining" with the 2010 Medical Subject Headings (MeSH1) was not first made until the latter half of the year 2009. (Yoo et al., 2012). Big data may be distinguished from other types of data by a few key characteristics. These characteristics are as follows: volume, velocity, diversity, truth, and value. The size of the collected data is referred to as the volume, the rate at which new data is generated is referred to as the velocity, the norm and nature of the data is referred to as the variety, the accuracy of the data is measured by veracity, and the value of the data is evaluated in terms of the result that is desired. Volume, Velocity, Variety, and Veracity all refer to the same thing (Hilbert, 2016; 2015). The data that is utilised for research in health informatics may have some of the characteristics that are associated with big data. Each individual patient generates a sizeable amount of data, "such as datasets that comprise MRI scans or gene microarrays for each unique patient". This data may then be used to draw conclusions about the patient's condition.

Big Velocity is characterised by the rapid creation of data that can be monitored in real time, for as when a medical device is used to monitor a patient's condition or when internet posts are used to monitor the spread of an epidemic. Big Velocity occurs when data is created quickly and may be examined alongside real-time events (such as from Facebook, Twitter etc.) Big Variety can refer to datasets that contain a large quantity of independent variables of a variety of types, datasets that were collected from a variety of sources, or any dataset that is complex and, as a result, needs to be evaluated at a number of different levels of data throughout the field of health informatics. Big Variety can also refer to datasets that were collected from a variety of sources. Big Variety can refer to datasets that were collected from a variety of sources. Big Variety can refer to datasets that The term "Big Variety" may also be used to refer to datasets that have been compiled from several different types of sources. The term "Big Variety" may also be used to refer to datasets that have been compiled from several different types of sources. In the realm of health informatics, a high level of data veracity can be accomplished by utilising defective clinical censors, gene microarrays, or the data of individual patients that are kept in a database. Other methods include employing gene microarrays. It is possible to obtain information of great value through the use of more conventional approaches, such as the collecting of data in clinical settings (Herland et al., 2014). Bioinformatics, image informatics (including neuroinformatics), clinical informatics, public health informatics, and translational bioinformatics are only few of the subfields that fall under the umbrella of health informatics (TBI). [10] Even though bioinformatics is not typically thought of as a part of health informatics, it is becoming an increasingly important source of health data. In order to offer a description of the data that was collected from biological research, this area makes use of a wide variety of various equipment and creates new procedures. In order to accomplish this, it applies the principles of computer science, statistics, biology, and engineering to the task of analysing and interpreting the biological data in the interest of enhancing the quality of the health care system as it exists at the moment (Lesk, 2011).

One notable example of what is known as "Big Volume" is the collection of data in the field of bioinformatics, such as the DNA sequences of thousands of different organisms. McDonald is responsible for the creation of khemr, which is a bioinformatics software suite with the goal of resolving a computational issue using hardware. This piece of software assists in the preprocessing of large volumes of genomic sequence by turning it into tiny fragments of sequence that can then be saved in a hash table that is based on the Bloom filter. This makes it possible for the data to be analysed in a way that is both useful and efficient (McDonald and Brown, 2013). In the subject of health informatics known as neuroinformatics, the primary focus is on the investigation of data pertaining to brain images. The purpose of this study is to acquire an understanding of the functioning of the brain, the connections that exist between the various sections of the body and the brain, as well as the connections that exist between data pertaining to medical occurrences and images of the brain. In the field of neuroinformatics, the primary areas of emphasis are research in the fields of neuroscience and informatics, as well as the creation of computer-based tools and their subsequent application. This is done with the intention of gaining a deeper comprehension of both the functional and structural components of the brain.

The creation of analytical and visualisation tools for the nervous system, as well as theoretical, mathematical, and simulation frameworks for describing the structure and function of the brain take up the majority of this field's attention. Clinical informatics has the potential to assist physicians in making choices regarding their patients that are both quicker and more accurate by analysing the data provided by patients. The use of predictive modelling enables this to be accomplished. One piece of research indicates that there is an approximate 152-year lag between the clinical research and the ramifications of that research in practise (Bennett and Doub, 2011).

VI. CONCLUSIONS

Recognizing "Big Data," gaining an understanding of it, and putting it to use in the context of scientific research and medical practise are all absolutely necessary steps to take right now if we are going to be successful in producing the most compelling evidence in a world where the volume of data is constantly growing. The convergence of Big Data interpretations is expected to continue as a result of the proliferation of data resulting from scientifically led endeavours, the acceleration of innovations in healthcare, and the rise of Big Data in higher education as a result of embedding technologies and the proliferation of e-Learning in higher education. All of these factors contribute to the rapid advancement of healthcare innovations. It is vital that healthcare professionals make use of big data analytics in order to save time, money, and energy; by doing so, we may minimise the number of readmissions, adverse events, and treatment optimizations for illnesses that impact several organ systems. Big Data is Saving Lives. During this interim period, the fields of medicine and healthcare are slipping farther and further behind in their implementation of big data methods. A greater emphasis will be placed, as the use of big data analytics becomes more widespread, on concerns such as the assurance of confidentiality, the maintenance of safety, the formulation of guidelines and policies, and the ongoing development of appropriate instruments and methods. Big Data Analytics, and if so, where areas of the country make use of this form of company operation. if Big Data Analytics is used, which regions of the country use it. It was feasible to draw the inferences and conclusions that are described below as a direct result of the results that were obtained. These findings may be found below. The medical organisation is engaged in the process of analysing and making use of both structured and unstructured data, the likes of which are produced by databases, transactions, the unstructured content of emails and documents, devices, and sensors, respectively.

REFERENCES

1. Rasid Z. Big data infographic(2013) What is big data? [Internet] 2013. [Cited March 17, 2014]. Available from: http://www.asigra.com/blog/big-data-infographicwhat-big-data.
2. de Lusignan S, Cashman J, Poh N, Michalakidis G, Mason A, Desombre T, et al.(2014)Conducting requirements analyses for research using routinely collected health data: a model driven approach. Stud Health Technol Inform 2012;180:1105–7. [PubMed]
3. Henke N, McKinsey, ‘ Data analytics: Changing the practice of medicine’ https://www.mckinsey.com/industries/healthcare-systems-and-services/ourinsights/the-big-data-revolution-in-us-health-care
4. Laney D. (2015) 3D data management: controlling data volume, velocity and variety. Available at: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3DData-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf. Retrieved: May 2014.
5. Laney, Douglas. “3D Data Management: Controlling Data Volume, Velocity and Variety”. Gartner.
6. IHTT . Transforming Health Care through Big Data Strategies for leveraging big data in the health care industry. 2013.
7. Bellazzi R. Big data and biomedical informatics: a challenging opportunity. Yearb Med Inform. 2014;9:8–13. doi: 10.15265/IY-2014-0024.
8. Frost & Sullivan: Drowning in Big Data? Reducing Information Technology Complexities and Costs for Healthcare Organizations. http://www.emc.com /collateral/analyst-reports/frost-sullivan-reducing-information-technologycomplexities-ar.pdf
9. Newman HB, Ellisman MH, Orcutt JA. Data-intensive e-science frontier research. Communications of the ACM 2003:46(11):68–77.
10. Müller H, Hanbury A, Al Shorbaji N. Health information search to deal with the exploding amount of health information produced.