The role of data mining techniques and tools in big data management in healthcare field

Adel Al-Zahrani1*, Abdullah Safhi2
1 Department of Safety Management, National Guard Health Affairs, Jeddah, Kingdom of Saudi Arabia
2 Department of Information Science, King Abdul-Aziz University, Jeddah, Kingdom of Saudi Arabia

Abstract
Data mining is one of the most important modern techniques used to achieve high output standards at all levels. The twenty-first century saw the advent of a new trend to improve medical services in the healthcare sector. To bridge the gap between previous studies and the practical applications of data mining, this study aimed to review the theoretical literature and previous studies related to the demonstration of data mining techniques and tools and their role in big data management. To achieve the objectives of the study, the researchers used a descriptive, analytical, documentary method. The study concluded many results including that in the era of the knowledge and technology revolution, data mining is one of the important issues, that requires everyone to take into account its achievements in our current era, as well as the existence of a correlation between big data and the provision of a separate health service in the field of healthcare, and work to address epidemics and discover vaccines for them. In the healthcare industry, data mining plays a vital role, especially in predicting various types of diseases. In detecting diseases, diagnosis is the main tool. The study recommended the need to conduct more experimental and exploratory studies dealing with healthcare data mining techniques and tools and their effect on the management of big data volumes, especially in our Arab countries and the need for the development of models and action plans and the development of processes and methods from which data in the healthcare sector can be explored.

Keywords: Techniques; Tools; Data Mining; Healthcare; Big Data

1. Introduction

In various fields, data mining is one of the points of interest for organizations and institutions that want to get accurate information by the shortest route because the amount of data in certain sectors can be very large. The spread of technology directly affects the increase and inflation of knowledge in a way that has not been known in history, and it is known that new scientific discoveries are being developed every day, which make it impossible to collect data if there are no resources and available methods.

The extraction of knowledge is one of the most critical elements sought by data mining, and because of its importance, the major countries are very involved in this issue at the leader’s level. The best proof of this is the involvement of former US President Barack Obama in this topic during his two president’s terms (2009-2016), as he held several meetings with major companies that have a tremendous amount of data, to find out the most successful ways in which data can be mined. Many interested in health issues have recently been concerned with data mining issues in the healthcare sector, especially with the spread of the new Coronavirus (COVID-19). The health sector is one of the sectors with tremendous data that needs to improve methods and techniques from which creative systems can collect it.

This work is licensed under a Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/) that allows others to share and adapt the material for any purpose (even commercially), in any medium with an acknowledgement of the work’s authorship and initial publication in this journal.

58
The growing gap between the cost of health care and its results is a significant problem, and many attempts to bridge this gap are made in developing countries. This gap was examined as a result of inadequate management of research-based insights, abuse of available data, and poor access to care experience. They all resulted in lost opportunities, misused money, and potential harm for the patient.

It has also been suggested that the gap between health care costs and outcomes can be determined by implementing a "continuous learning healthcare system" where a virtuous circle is formed between health research and organizational forces and where data can be used effectively. The initiations of the big data era in healthcare are in a critical need to boost the quality of healthcare and patient outcomes, increase the availability of data, and increase analytical capabilities.

Several obstacles need to be addressed before big data technology can dramatically increase healthcare costs, status, and outcomes. This confirms that the volume of big data in the health sector makes a large amount of data of no great importance, which confirms that the process of data mining may be useless unless the correct techniques and tools are available that achieve this purpose. However, if a large amount of data were available in the health sector and has been reconstructed per the correct research principles and regulations, this will have a huge impact on developing the healthcare sector, handling epidemics, delivering outstanding healthcare services to patients and improving production rates. After that, the health sector may be one of the active sectors, and the statistics show that a large percentage of people in the countries of the world want to share their health data to access better treatment. A large percentage of British citizens, estimated at 60%, are not bothered by the use of their medical data in medical research and vaccine discoveries due to the numerous epidemics that have frighteningly occurred in our world recently. Several trends have emerged that aim to benefit from data mining in health care, and this plays a major role in the growth of big data in which, if appropriate mining rules are used, they can be valuable for improving health services.

It is worth noting that the relationship between data mining techniques and the increase in big data in the healthcare sector has not been closely discussed in previous studies, as some studies have indicated the value of data mining in the healthcare field without referring to the relationship between mining techniques and the increase in big data. There are some distinctive characteristics of medical big data that vary from other big data disciplines. Medical Big Data is also difficult to access, and for reasons such as the possibility of data abuse by third parties and the shortage of incentives to share data, most medical investigators are hesitant to practice open data science. Medical Big Data is mostly collected based on protocols (i.e., static models) and is reasonably standardized, partially because of the initial data extraction process. Another significant aspect is that medicine is performed in a vital sense of critical safety, where explanations would help decision-making practices. Medical big data, due to employee engagement, the use of expensive devices, and potential discomfort for the patients involved can be costly. Medical big data is relatively small compared to data from other disciplines and can be gathered from a non-replicable situation. Compared to data from other fields, medical big data is comparatively limited and can be collected from a non-replicable circumstance. Many sources of inconsistency, such as measuring errors, missing data, or encoding information errors in text files, can affect large medical data. Therefore, in both data analysis and outcome interpretation, the position of domain knowledge may be dominant. Such defining features of medical big data in empirical ways include the various types of patient descriptions, which can often involve weighting; the sequential structure that might be of an additional dimension; treatment details and the timing of treatment decisions and changes; (i.e., time-dependent confusion).

Based on the aforementioned and the importance of data mining and its impact on health and increasing data, the idea of this scientific study came with the intention to review previous research on this topic and identify the extent of the effects of data mining in health care in the process of increasing large data that can be useful for improving the health services sector.

This study consists of seven parts: introduction, study problem, study methodology, previous studies, results and analysis of results, conclusion and recommendations.

2. Study problem

There has been a great development of data related to the healthcare sector in recent years, and it has become important to make use of it. At a time, and when infectious diseases and epidemics are spreading, the focus has been put on the need to pay attention to and improve the field of health care to solve these epidemics. Big data may also be an important element in the detection of diseases and epidemics and the pace of development and treatment of vaccines for them.
The significant value of medical big data is shown by 1) providing customized medicine; 2) using clinical decision support systems such as instrumental analysis of medical imagery and retrieval of the medical literature; 3) adjusting diagnosis and care choices to support preferred patient actions using mobile devices; 4) population health analyzes limited to big data display correlations that would have been missing when smaller collections of randomly formatted data were analyzed instead; and 5) fraud identification and prevention. Diagnosis based on high-resolution measurements such as microarray analysis or next-generation sequencing, monitoring of molecular properties for use in diagnosis and treatment decisions during treatment, and continuous monitoring of patients' health care among possible uses of medical big data.

There are eight fields of which big data analytics are used to enhance healthcare: 1) statistical risk modeling and resource use; 2) population management; 3) safety tracking of drugs and medical devices; 4) disease and heterogeneity of treatment; 5) personalized medicine and clinical decision-making support; 6) quality of care and success measurement; 7) environmental health; 8) search technologies. Predictive analytics using big data technologies is a methodology that learns from experience (data) to forecast the potential behavior of individuals to make smarter choices, i.e., future observations, based on a comprehensive overview of correlations, for example, overtime or across a wide geographic region or experienced in a large portion of the overall population. Big data is necessary, but it is not sufficient. In other words, a vast data set is of no benefit if data cannot be understood such that potential information can be produced and the importance of big medical data can be assessed for this possibility.

With the great development and changes the world is undergoing, the large increase in population numbers, the large spread of diseases, as well as the availability of technology, all this confirms the need to pay attention to data mining and confirms as well that there is a problem related to data mining in the field of healthcare scientifically, and a problem in the size of big data. The study problem can be developed from the above through the following main question: How do data mining techniques and tools in healthcare increase the volume of big data?

Sub-questions were extracted from this main question as follows:
1. How important is data mining in healthcare?
2. What are the main techniques and tools for data mining in the field of healthcare?
3. What are the obstacles facing data mining in healthcare?

3. Study methodology

This study adopted the descriptive, analytical, documentary method, which depends on reviewing documents and literature such as research, articles, books, and dealing with them descriptively and analytically to extract conclusions and indications which are relevant to answering the study's questions. From this standpoint, the current study will critique and analyze previous studies on the subject of techniques and tools for data mining in the field of healthcare and their role in increasing the volume of big data.

Big data analysis uses different algorithms to retrieve data that can be described as the automated extraction of useful data that is often previously unknown from large datasets or data sets using sophisticated search techniques and algorithms to discover trends and similarities in pre-existing large databases. Data mining tasks can be summed up as explaining and identifying similarities and associations that can be understood by humans and predicting any interesting response. Clinical data mining can be characterized as the application of data mining to a clinical problem.

Data mining algorithms are known as supervised, unsupervised, and semi-supervised learning. Supervised learning involves predicting known outcomes of the target, using a training set that contains already-classified data to extract inference or predict possible test data. Although there are no predetermined outputs of unsupervised learning, researchers aim to identify naturally occurring correlations or combinations within unmarked data. Finally, semi-supervised learning requires combining efficiency and precision using limited sets of labeled or annotated data and a much greater set of unmarked data.

4. Previous studies

The analytical objectives of medical big data are prediction, simulation and inference, while classification, aggregation and regression are typical approaches that are used in these contexts. Classification is used as a form of supervised learning and can be used as predictive processing in which the performance vector or the predictive variable is critical. Classification is the development of a rule for assigning objects to a set of predefined classes (prediction variable) centered on a vector from the measurements taken on these items.
Classification techniques involve logistic regression, methods of Naive Bayes, decision trees, neural networks, Bayesian networks and Vector Machines Support. Score performance may be measured by various performance metrics tested in an individual evaluation or validity category. Such techniques can be used to build a decision support framework that determines a diagnosis with several possible diagnoses or to create predictive models based on data from the analysis of many biomarkers. Clustering is unsupervised learning that is used to identify clusters of data by using distance scales. Clustering methods entail k-mean clusters, component-based clustering, and self-organizing maps (SOM). The performance of the cluster can be measured by its performance in a corresponding supervised teaching assignment. Clustering is also used in the study or analysis of the development and growth of micro-matrix data. It may also be used to reshape disease through pathophysiology pathways to offer more precise treatment choices. Regression is subject to supervised learning when the output variable is constant and is a statistical analysis tool that describes the interaction between a dependent variable and one or more independent variables to demonstrate data patterns. Linear regression is the most widely used method in this classification. Examples of its uses are the longitudinal analysis of medical data or the decision support system. This section aims to review and critically review several Arab and foreign studies related to the topic of artificial intelligence and its impact on the quality of decision-making in organizations. They will be arranged in descending chronological order from the newest to the oldest, which falls within the time period 2010-2020, noting that the total number of studies that will be reviewed is 8 studies distributed as follows:

Study (5): Big data and its analytics: concept, characteristics and applications [2]
Big data in the digital era and the implications of mining are based on modern foundations, laws and strategies. This study adopted the descriptive survey method to answer the study's questions. The study concluded that the health sector and other sectors are in dire need to provide the largest amount of big data that, in turn, contribute to the development of all fields. The study recommended the need to foresee the future about providing databases in the health sector to ensure its progress and development.

Study (6): Transforming healthcare institutions into institutions managed by big data [3]
This study aimed to identify the importance of big data in upgrading the health system and develop a proposal by which healthcare institutions would be transformed into institutions managed by big data. The study adopted the inductive analytical approach to reach its objectives. The study reached several conclusions including that big data has a major role in improving the health service and upgrading the health system. It also concluded that there are several steps through which health institutions can be transformed into big data-managed institutions through identifying and processing data assets, establishing data partnership networks, managing data as the longest important strategy, and using data to achieve institutional innovation. The study recommended the necessity of paying attention to databases in our Arab countries to improve the health system.

Study (3): The use of data mining techniques in the health field for tuberculosis in Khartoum State [7]
This study aimed to identify the role of data mining techniques in overcoming tuberculosis in Khartoum State. The study applied for a computer program to achieve its objectives. The study adopted the inductive method to reach its objectives and answer its questions. This study concluded that data mining greatly contributes to outbreak prevention and epidemic planning and infectious diseases by pre-prepared action plans. It found that there is a deficiency in the issue of data mining in the health sectors. It recommended that attention be paid to health data mining to boost health institutions’ efficiency.

Study (9): Efficient algorithms for mining healthcare data [8]
This study aimed to develop algorithms through which data can be explored in the health field and benefit from the increase in the number of big data for the health sector. It also aimed to identify the role of data mining in providing health care to patients and develop scientific research and discoveries of vaccines and drugs. The study relied on the inductive descriptive-analytical method to reach the results of the study. The study reached several conclusions, including that the last period was a remarkable leap in the implementation of healthcare data to overcome many obstacles related to the health sector. The study has discovered a new algorithm through which data mining can be improved in the health sector. It concluded that the algorithms used in data mining in the health sector are still in need of great improvement and consistent follow-up, and recommended that data transparency studies discuss algorithms that, in turn, lead to accessing any data that can be used in the different areas of the health sector.

Study (11): A systematic review on healthcare analytics: Application and theoretical perspective of data mining [10]
This study aimed to outline the major role that data mining plays in the healthcare sector in raising the size of data and making great use of the identification of skin diseases and to identify the role that big data plays in the treatment of skin diseases. The study adopted a descriptive survey method to reach the answers to its questions. The study reached several conclusions, including that the techniques of disease mining have a prominent role in treating skin diseases, and that the availability of big data and the availability of mechanisms through which data are searched helps to treat many diseases and reach solutions to the problems facing the health sector.

Study (8): Prototype for a tele-dermatology system in Oman: Disease detection using data mining techniques

This study aimed to outline the major role that data mining plays in the healthcare sector in raising the size of data and making great use of the identification of skin diseases and to identify the role that big data plays in the treatment of skin diseases. The study adopted a descriptive survey method to reach the answers to its questions. The study reached several conclusions, including that the techniques of disease mining have a prominent role in treating skin diseases, and that the availability of big data and the availability of mechanisms through which data are searched helps to treat many diseases and reach solutions to the problems facing the health sector.

Study (10): Healthcare data mining using in-database analytics to predict diagnosis of inflammatory bowel disease

This study aimed to identify the role of data mining in improving the services provided to patients and their role in the classification of patients and studies of stomach diseases. The studies also aimed to identify some frameworks and rules in which attention is paid to analyzing health sector data and develop an application that works on Bismarck database analyzes in a healthcare environment. The study concluded that open-source databases on which statistics can be applied produce reliable results in the field of healthcare. The study recommended the necessity of activating the mechanism of data disclosure in the health sector by adopting and innovating new methods.

Study (12): Optimization and data mining in healthcare: Patient’s classification and epileptic brain state transition study using dynamic measures, pattern recognition and network modeling

This study aimed to identify the key role that data mining plays in the healthcare sector in raising the size of data and making great use of the identification of skin diseases and to identify the role that big data plays in the treatment of skin diseases. The study relied on several models and dynamic and electrical programs to achieve the objectives of the study. The study found that studies in the field of data mining are still somewhat scarce and that the future predicts a great deal in terms of data mining. The study recommended the necessity of working on devising techniques through them; data can be explored in the field of treatment for epilepsy.

5. Results

A large volume of data is generated in the medical field, including patient personal data, patient history, genetic information and clinical information. Such an immense volume of medical data provides useful data that could allow, when analyzed legally and thoroughly, to clarify the principles of illness and well-being and thus make quantum leaps forward in the medical field, especially in the areas of disease detection and prevention. There have already been many practical implementations and many other possible applications in this area, such as disease diagnosis, disease detection, infection control, telemedicine, fraud prevention, etc.

Through reviewing, inspecting and analyzing previous researches, it became clear that they addressed and focused on the following points:

1. Previous studies proved that the interest in data mining in the healthcare field is one of the basic trends reinforced by administrators in charge in this large sector, which confirms the importance of paying attention to this process due to its significant role in providing big data. In this regard, study (11) [10] proved that research literature in health care filed pays attention to data mining, study (10) also proved that data mining in healthcare field contributes, to a great extent, to improve the service provided to patients in the health sector, and study (3) [7] concluded that mining big data can contribute to treating people infected with tuberculosis in Sudan.

2. Some studies have shown that big data plays a major role in all sectors, particularly health care, which has shown the immediate need for the availability of big data and the sharing of such data between countries and major health institutions. In this regard, study (6) [3] demonstrated that big data could contribute to boosting the economy and improving data in all fields and that the big gap caused by lack of attention to
data limits the effectiveness of solutions, in particular concerning disasters, epidemics and diseases afflicting humanity till the recent times.

3. Some studies showed that depending on data mining in healthcare has been restricted to the patient clinical data for a long time. In this regard, Study (11) [10] has shown that there is a recent dominant trend which seeks to leverage and benefit from the enormous scale of data on websites and social media websites for mining purposes. Study (8) [4] also mentioned that the internet and electronic databases are crucial to ensuring the availability of in-depth databases from which data could be mined and the health system could be strengthened.

4. Some studies demonstrated that there are still deficiencies in the data mining field, as the study (11) [10] indicated that the indicative aspects left so much to be desired, while study (12) [12] mentioned that data mining in the field of treatment discovery still in need for development to take advantage of the enormous amounts of the data available in the health sector. Also, study (3) [7] mentioned that lack of data mining contributed to a great extent to the failure in defeating infectious diseases immediately after they are detected in most Arab countries.

5. Some studies have shown that the health systems implemented in Arab countries still lack mechanisms and methods that could allow mining data to strengthen the health system. In this regard, study (3) [7] revealed that there is a severe lack of electronic records, while Study (5) [2] demonstrated that Arab countries need a specific plan for handling health institutions across big databases.

6. Some studies suggested mechanisms and methods that could help mining data in the health sector, as study (9) [8] suggested an algorithm for improving the mechanism of data mining in health sectors, and study (6) [3] also suggested many steps through which healthcare institutions could be transformed into institutions managed by big data.

6. Discussion

Big data analysis refers to the use of structured and unstructured data tools. The transformation into an integrated data environment is still an insurmountable challenge to be solved. Interestingly enough, the main part of big data relies mostly on the assumption that the greater the knowledge, the more insights to be extracted from it and therefore, the greater the prediction of future events.

Besides, several reputable research firms and healthcare companies are projected to see an unprecedented growth rate in the Big Data healthcare market. Conversely, within a brief span of time, we have seen a range of analytics, which are currently, used that have a tremendous effect on the health industry's decision-making and efficiency.

In different fields, the spectacular increase of medical data compelled computer experts to develop groundbreaking methods to analyze and classify an immense amount of data during a particular time frame. Also, the integration of data processing computing systems has seen the growth of medical scientists and practitioners. Accordingly, by incorporating physiological data, create a comprehensive human body model and Wi MAX (Worldwide Interoperability for Microwave Access) techniques could be the next great goal. Such a great idea could enhance our knowledge of the pathogenesis and could contribute to developing new diagnostics tools. The persistent growth of available genomic data, including deeply rooted errors from the previous experiments and analytical data requires more interest. However, opportunities for systematic improvements in healthcare research exist in each step of this inclusive process.

The results of the previous studies showed the following:

1. The previous studies covered a large number of the institutions working in the healthcare field in Arab and forging countries, also these previous studies were conducted as exploratory studies for some diseases and the role of data mining in overcoming such diseases and improving healthcare services.

2. Previous studies have shown that there is a great deal of analysis and study needed in healthcare data mining because the appropriate evaluation of data mining has not been adequately addressed in previous studies and theoretical literature.

3. Earlier research confirmed that the data mining strategy needs to be applied immediately in the medical field due to challenges in several countries in the world in the health sector.

4. Some countries came up with suggestions to transform the healthcare sector into a sector managed by huge data. However, this will require great efforts and capabilities at all levels, and it is not expected to be fully achieved shortly.
5. We can find a range of gaps by evaluating previous data, of which studies investigating the issue, which are techniques and tools in the field of healthcare used for data mining and their impact on big data development, are very few. Also, some of the previous studies did not have a clear methodology, and some of them did not depend on large and sufficient samples during the experiment or did not demonstrate the nature and characteristics of these samples, such as the study (5) [2], and the study (10) [9].

6. All the previous studies have the same purpose as the current research, as they aimed at cognizing the data mining techniques and tools used in the healthcare sector and their effect on the growth of big data. The previous studies, however, diverged in some ways from this current research, as the preceding studies concerned data mining at certain institutions or application for certain diseases while the current study aimed at defining the large impact of Big Data mining.

7. With regard to the methodology, the current research adopted the descriptive-analytical documentary approach, so it agreed with the study (9) [8] and differed from the other studies, which adopted different approaches.

8. The studies (12) [12] and 10 [9] proved that there is a significant correlation between the development of the methods and tools used for detecting data in the healthcare field, contributes to great extent to the growth of big data, which could be utilized in the health sector and for discovering treatments and vaccines for epidemics and diseases afflicted by humanity.

7. Conclusion

This research discussed a variety of recent studies discussing the most common sub-branches of health informatics, using big data at all levels of human existence that could be accessible, to answer the question at all levels. Big data analysis within the scope has only recently become feasible thanks to the increased computing capacity and the resources used by algorithms.

It is fair to say that the analysis of the use of such tools and techniques in the area of health information systems is particularly critical, as this field requires so much testing and verification before the implementation of modern decision-making techniques in the real world at all levels. The fact that computational power has made it easier to work with big data through successful algorithms (alongside hardware advancements) paved the way for data mining and managing large volumes, speed, diversity, validity and data value produced by health informatics (whether traditional or otherwise).

The use of big data provides benefits for health informatics by allowing additional experiments or more testing advantages, thereby facilitating rapid verification of the quality of the studies and the opportunity to obtain sufficient cases for training if there are only a few cases in the positive category. Moving forward with health informatics definitely requires the exploitation of big data generated throughout all different levels of medical data and discovering the best way for analyzing, mining, and answering the largest possible number of medical questions.

After the research, we have reached several important results as follows:

1. Data mining is of great significance in the age of knowledge revolution and technology, and in our recent period, everyone should make use of the achievements in this area.

2. The research concluded that data mining in the healthcare field contributes to a great extent to the improvement of healthcare, treatment of diseases, and the development of strategies through which pandemics and epidemics could be overcome.

8. Recommendations

In light of the results achieved by this research, some recommendation can be offered:

1. It is necessary to conduct more experimental and exploratory studies to address data mining techniques and tools used in the healthcare field, and their effect on the growth of big data, in particular in our Arab countries.

2. It is necessary to develop models, action plans, mechanisms and methods, through which data could be mined in the healthcare field.

3. It is necessary to conduct experimental studies covering data mining techniques and tools used in the healthcare field, and their effect on the growth of big data, in particular in all countries and healthcare institutions.
4. It is necessary to hold local, Arab and international conferences and seminars sufficiently and regularly, to present the recent updates in this field.

5. It is necessary to take the required actions at the local level and conduct studies and scientific researches in the field of data mining in the healthcare sector, in line with the kingdom vision 2030.

References


