2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

Importance of Data Mining in Healthcare: A Survey

Mohammad Hossein Tekieh[1], Bijan Raahemi[1]
[1] Telfer School of Management
University of Ottawa
Ottawa, Canada
(mteki040@uottawa.ca, braahemi@uottawa.ca)

Abstract—**In this survey, we collect the related information that demonstrate the importance of data mining in healthcare. As the amount of collected health data is increasing significantly every day, it is believed that a strong analysis tool that is capable of handling and analyzing large health data is essential. Analyzing the health datasets gathered by electronic health record (EHR) systems, insurance claims, health surveys, and other sources, using data mining techniques is very complex and is faced with very specific challenges, including data quality and privacy issues. However, the applications of data mining in healthcare, advantages of data mining techniques over traditional methods, special characteristics of health data, and new health condition mysteries have made data mining very necessary for health data analysis.**

Keywords—data mining; health data analysis; data quality; predictive modelling; health big data; data mining applications

I. Introduction

During the past decades, many organizations have transformed their paper-based systems to electronic systems. Electronic systems provide significant benefits for both the employers and employees to increase the efficiency of work and outcomes. As a matter of this transformation, large amount of data is being gathered via these electronic systems on a daily basis. While data holders were getting concerned about the feasibility of maintaining this daily increasing amount of data, they got aware of the invaluable information hidden in their datasets. In other words, the collected data in any organization is now seen as a new source of very important information that can directly affect the operational efficiency of that organization, provide higher quality outcomes, and even cut the unnecessary expenditures that waste the budget. Therefore, data mining has received a lot of attention due to its strong ability of extracting information from data. As a result of this attention, data mining algorithms are also getting improved and even new versions being developed to provide more robust and reliable information.

Among the organizations that are using electronic systems, healthcare institutes and organizations are also actively gathering electronic health data using various methods, including computer-based surveys, online insurance claims, and eventually, Electronic Health Record (EHR) or Electronic Medical Records (EMR) systems. Particularly, as a result of implementing EHR systems, a significant amount of data is being collected by hospitals, clinics and other health providers. However, most of these datasets are not very well structured and appropriate for analytical purposes. In addition, health data are usually very complex and not easy to analyze [1]. Therefore, health sector is still a very demanding domain in applying data mining techniques to find trends and hidden information to make the health organizations more cost efficient, and provide complementary clinical solutions.

As an instance, massive healthcare cost in Canada is causing severe challenges for the government. According to an article in the Policy magazine [2], Canada will be spending a quarter of a trillion dollars on healthcare by 2020. In 2011, more than 50% of the total government budget in both provinces of Ontario and Quebec has been consumed for provincial health spending, and this number is still increasing. Same thing will happen to the other provinces – British Columbia, Alberta, Saskatchewan, and New Brunswick – by 2017. Surprisingly, this enormous amount of expenditure has not also provided the required quality in health services and deliveries. Among the 11 developed countries, Canada has the worst access times to a doctor or a nurse when in need of care, the most extreme delays for specialist appointments, and the highest use of emergency rooms. In addition, in terms of medical error incident rate, Canada has been ranked 7th among 15 peer countries. On the other hand, the good news is that Canada has been able to collect huge amount of health data due to the existence of its public health plans from most segments of its population. Adding that to the data collected by EHR systems that have been implemented in the public health providers (e.g. hospitals) across the country, and also linking these health data to other sources of data, such as the databases of Immigration Canada, could provide rich sets of data that contain very useful trends and can be used for analysis. The outcomes of these analyses can be saviors for financial, clinical, and administrative issues of Canada's healthcare.

In this paper, we present the advantages of data mining and special characteristics of health data that makes data mining very important to be considered in health data analysis. In this section, we demonstrated some evidences and explanations about why data mining has received attention, especially in the health domain. In the next section, data mining techniques and concepts are briefly presented. In continue, the applications, necessities, and challenges of data mining in healthcare have been described and summarized. Finally, we conclude about the influences of applying data mining in healthcare.

ASONAM '15, August 25-28, 2015, Paris, France

ACM ISBN 978-1-4503-3854-7/15/08 $15.00
DOI: http://dx.doi.org/10.1145/2808797.2809367

II. Data Mining

In the middle of 1990s, data mining came into existence as a strong tool to extract useful information from large datasets and find the relationship between the attributes of the data [3]. Data mining originally came from statistics and machine learning as an interdisciplinary field [4], but then it was grown a lot that in 2001 it was considered as one of the top 10 leading technologies which will change the world [5].

A. *Main Techniques*

Data mining techniques are divided into two main categories: descriptive (or unsupervised learning) and predictive (or supervised learning) [6][7][8][9]. Descriptive data mining is an exploratory analysis that attempts to measure the similarity of records, and discover the patterns and relationships. The most important techniques in descriptive data mining are clustering and association. On the other hand, predictive data mining tries to generate predictive rules as a model to classify the records based on a specific target (or label). Classification is the most widely used technique in predictive data mining. As there is no standard framework for a data mining process [1], it is important that the analyst himself holds good amount of skills in this area and understands the techniques very well to design an appropriate framework and achieve high quality and reliable outcomes. The main techniques have been introduced briefly in continue.

1) Classification

This technique is used when the data is required to be classified into different groups based on a target attribute – e.g. patient cost. Therefore, the classifiers predict the target label for each record using the input attributes. Some of the famous classification techniques are: decision trees, neural networks, K-nearest neighbors, support vector machines, bayesian methods. According to a survey [4], decision tree algorithms are the most popular ones among all other classification techniques in the applications of data mining in healthcare.

Classification techniques are widely used in health data analyses, including: analyzing microarray data [10], diagnosing skin diseases [11], performance of different classifiers on cancer datasets [12], predicting cost of healthcare services [13][14], identifying significant factors in healthcare coverage and predicting the status [15].

2) Clustering

This technique is used when we do not have much information about the different types of data objects involved in a population. As it is an unsupervised learning, it tries to find the cluster of data objects that have similarities to each other without considering any specific target label. Therefore, there are no predefined classes in contrast to classification. Different clustering techniques are: partitioned clustering, hierarchical clustering, and density based clustering.

3) Association

This technique is used when the relationship of attributes in a dataset needs to be identified – e.g. the association among the purchased items of a customer's basket. This technique is mainly used in healthcare to detect the relationship between diseases. In addition, association techniques can also be combined with classification techniques to increase the capability of analysis. For instance, the rules in a database or relationship of attributes in a dataset are detected, and then an efficient classifier is built by just considering the identified rules and including just the main attributes.

B. *Advantages Over the Traditional Statistics*

In the recent history, traditional statistics has been considered as the main data analysis method and is still actively contributing in most analysis studies. Although statistics is viewed as the *primary* data analysis in the current era, data mining is considered as the *secondary* data analysis due to its strengths and rapid developments [16]. While the fundamental of both analysis methods is mathematics, data mining actually includes statistics as part of its process. However, as an interdisciplinary field which benefits from the advantages of other fields, such as machine learning, artificial intelligence, and visualization, it has some important gains over the traditional statistics.

First, statistics prefers to use more conservative strategies in the first phases of analysis, and in general, employ concrete mathematical methods to run analysis. On the other hand, data mining is open to consider various approaches in regards to mine the data in different orders [17]. Due to this flexibility, data mining uses heuristics as well when facing with real-world issues, so that categorical (discrete) attributes are included in the analysis too [4].

Second, statistics runs analysis only on a sample of data, as this was probably the approach to handle large datasets for analysis in past, and it has retained in this method's nature. In contrast, data mining has the ability to consider the whole dataset for analysis which in return provides more reliable results by considering all details of the population.

Third, statistical methods can only work with numeric data [17]. However, there are a lot of categorical (discrete) attributes – e.g. race, gender, diagnosis code – in addition to numeric and even other types of data in the current databases. Most data mining techniques are capable of handling these types of data in addition to numeric data.

Finally, in statistics, a hypothesis is first created and then the data gets analyzed to prove or reject the hypothesis (hypothetico-deductive analysis). On the other hand, data mining does not consider any clear hypotheses. It starts exploring the data and tries finding knowledge out of the data (inductive analysis) [17]. This can be very useful when studying the prevalence of new diseases that their causing factors are unknown.

III. Application of Data Mining in Healthcare

Data mining can be applied on health data for many different purposes and investigations. These applications can roughly be grouped into the four main categories which have been discussed in continue.

Note that, text mining is a broader field of science that includes data mining as well. "Text mining in healthcare" has also received many attention to analyze unstructured texts – e.g. doctor notes, prescriptions. It has its own specific applications which is out of the scope of this survey.

A. Clinical Decision Making

Patients visiting clinics or hospitals will be examined by clinicians to diagnose their issue or disease. Although the medical experts do all their best in identifying the causes of every symptom in the patient, the nature of this examination is experimental and sometimes the diagnoses might go wrong. Data mining techniques can help the experts in the field to receive a second opinion for most diagnoses, especially to make sure the disease is not under-estimated during diagnosis [18]. This information can help the clinicians to make more accurate and reliable decisions, the providers to deliver higher quality services, and even insurance companies to save money as a result of avoiding claim re-submissions for just one diagnosis.

Moreover, high-risk patients can be identified in an earlier stage of their disease progression using predictive models, which again will provide higher quality of service for the patient and also reduce healthcare cost through intervention and prevention plans [19]. Some new studies are also focusing on data stream mining to support real-time clinical decision making which requires very high performance data mining algorithms [20].

B. Biomedicine and Genetics

Some specific diseases are studied in the biomedical and molecular level, in addition to the clinical level. As the amount of extracted biomedical and molecular data are increasing, it can help researchers investigate the effects of genetics on different diseases in micro-level (molecular analysis); therefore, we have separated this section from the population health section which looks at the trend of diseases in macro-level. Having said that, in microarray data analysis, clustering techniques have received more attention comparing to classification and association as there are not a lot of information available about genes, in contrast to health conditions and disease symptoms that a lot of information are known [4].

Researchers can classify the microarray data of diseases [10][21] for different purposes such as predicting whether a patient will have a health condition [22], and differentiating similar diseases based on their DNA expression microarray results extracted from the infected sample cells or tissues [23]. Also, some researchers claim that clustering microarray data related to a disease (in their case, it was extracted from breast cancer tumors) to find the vulnerable groups and then classify the records based on the initial findings can provide much better outcomes with higher accuracy in identifying specific patients involved with this disease, comparing to predicting using clinical factors [24].

C. Population Health

Epidemiologists and other health analysts focused on the prevalence of diseases are interested in identifying the patterns, trends, and causes of spreading a specific disease across a population. For these studies, they consider different risk factors and health determinant, including early-life, lifestyle, and socio-demographic. Fig. 1 demonstrates the level of these risk factors that can contribute on causing and spreading a disease.

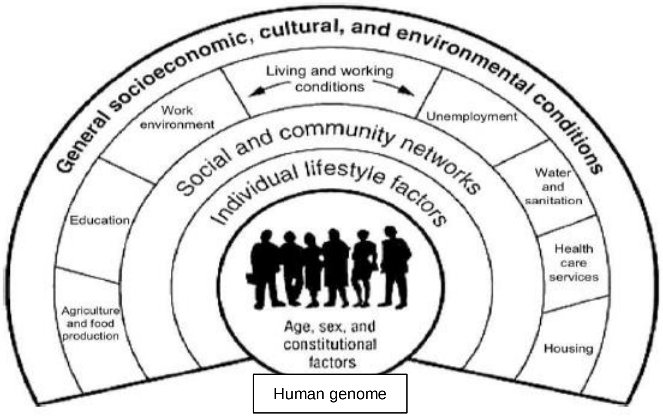
General socioeconomic, cultural, and environmental conditions
Living and working conditions
Work environment
Unemployment
Social and community networks
Individual lifestyle factors
Education
Water and sanitation
Health care services
Agriculture and food production
Age, sex, and constitutional factors
Housing
Human genome

Fig. 1. Factors responsible for diseases [25]

Many different studies have focused on classifying and predicting type of cancers, such as breast cancer which its rate has increased globally [12]. In addition, data mining techniques can help to analyze the survivability and identify the significant factors affecting survival from specific diseases, including breast cancer and kidney dialysis patients [26][27].

D. Health Administration and Policies

One of the big challenges in the area of health administration is handling insurance plans and their mechanism. Usually each country or state has its own specific mechanism. However, one issue that all health insurance companies (no matter it is public or private) deal with is insurance fraud. Data mining has been applied to detect insurance fraud in which the patients, doctors, or hospitals claim drugs that were not necessary or procedures that partially or totally did not actually happen [28]. This can put a massive amount of financial load on the government's shoulder (if is public coverage) or lead the insurance companies (if is private coverage) toward bankruptcy. In this case, the built predictive models in an almost real-time fraud detection systems can help to identify what type of drug or procedure is necessary for every diagnosis, and where this has not been respected.

As the healthcare cost is increasing globally, some health administrative data analyses are mainly focused on identifying high-cost patients, and solutions for preventing significant costs by providing early stage care if possible [14]. Moreover, finding the relationship between the contributing factors in a health condition using association rules can help to provide more efficient policies, at the level of government (such as a government policy on hypertension management) or even care providers, in controlling that health issue [29].

IV. NECESSITY OF DATA MINING IN HEALTHCARE

In addition to the items mentioned in the previous section (applications of data mining in healthcare) such as fraud detection, causing factors discovery in unknown diseases, microarray data classification, and high-cost or high-risk patient prediction, that makes data mining techniques necessary in healthcare, health data analysis is going toward a direction that a tool like data mining is becoming more important and essential to provide successful analyses. The data being collected in the health sector has its own specific characteristics that makes any analysis very challenging. Some of the main concerns in this regard has been described in the next section. However, there are other specifications tied to health data that prevents traditional methods to evolve and provide new information.

A. Health Big Data

"Big Data" has become a known term referring to a massive collection of data to the point that the traditional data analysis techniques do not function efficiently. In addition to the volume of data, velocity (streaming data) and variety (semi-structured and un-structured) are major concerns in "big data" too. As a result of implementing EHR and similar systems to collect electronic health and medical data, in addition to the data being collected via online insurance claims and health surveys, a large amount of data has been collected by health organizations which is increasing on a daily basis. Data mining techniques have the ability to consider the whole population into the analysis which can provide information about specifics, details, and also minorities.

Although data mining algorithms also need to be improved in terms of performance and scalability, at least current data mining algorithms have the capability of working with large datasets. Moreover, new frameworks are being studied and developed to handle health big data analytics, in terms of better understanding the data, overcoming with the challenges, and providing higher quality results. Normally, the steps that need to be taken to support health big data analysis are: data aggregation, data maintenance, data integration, data analysis, and pattern interpretation [30].

B. Health Condition Mysteries

As discussed in the previous section, not a lot of information is available about genes' impacts on health disorders. Therefore, exploratory methods can significantly help the researchers to identify the relationships between the attributes and also cluster the groups of contributing genes. Descriptive data mining has started evolving in this area with the attempt of going from specific (data) to general (knowledge) about the effects of genes on health conditions. In the next steps, classification and prediction models can also be built for the discovered groups to be able to predict the health disorders using gene expression patterns observed in patient's related organs.

In addition, by the changes introduced in human's lifestyle and living conditions in the modern era, newly discovered diseases are being introduced and spread globally, such as Inflammatory Bowel Disease (IBD) and Ebola, that are still mysteries. The main causes of these diseases and dysfunctions are not clearly identified yet, and based on the previous observations, maybe more of these new diseases will be discovered in future as well. In this case, data mining techniques can also be used to discover the trends related to these diseases considering clinical, biomedical, and even environmental factors.

V. Challenges of Data Mining in Healthcare

There have been some studies conducted to examine the issues and challenges of applying data mining techniques in healthcare [31][32]. In overall, the top important challenges that make data mining in healthcare (and biomedicine) applications differ from other application are as following:

A. Data Quality

To achieve useful and reliable information from a data mining process, it requires data with good quality. Health data is usually collected from different sources with totally different set-ups and database designs which makes the data complex, dirty, with a lot of missing data, and different coding standards for the same fields. For instance, although problematic handwritings are no more applicable in EHR systems, the data collected via these systems are not mainly gathered for analytical purposes and contain many issues – missing data, incorrectness, miscoding – due to clinicians' workloads, not user friendly user interfaces, and no validity checks by humans. In addition, as EHR data are observational and not experimental, they might not represent all cases involved in an issue and face with different types of biases (selection, confounding, measurement) which makes it challenging to generalize the outcomes to the whole population [33]. Dealing with this type of data is not easy and requires significant efforts in cleaning and pre-processing before building the actual data mining models. Moreover, there have been some studies to compare and examine the difference of EHR data with other types of data collection methods, such as health surveys, in analyzing specific health conditions [34].

If a health service provider – e.g. a hospital – wants to gather data so that it can be used for secondary purposes, especially analytics, to obtain useful information for improving their service quality and making their costs efficient, it is essential to define frameworks and methods that data with acceptable quality be collected from the first place. This requires very flexible and scalable database designs for every unit that can easily be linked – whenever required – while using the same standards in entering the data. For instance, when the patient has to fill out a lot of questions at the check-in time when visiting a hospital, he/she is sick and might not even have enough concentration in answering the questions; therefore, this will lead to collecting a data with lots of missing and even incorrect entries. Other considerations which can make a gathered dataset – e.g. via EHR systems – appropriate for secondary uses is to have interoperable data collecting systems with improved infrastructure, and also using proper data analysis techniques [33].

B. Data Sharing and Privacy

Since the health data contains personal health information (PHI), there will be legal difficulties in accessing the data due to the risk of invading the privacy. This issue puts a big gap between the collected data and the data analyst, and connecting these two is sometimes not very easy. Health providers are not usually comfortable with sharing their data with the analysts to avoid any risk that threats the privacy of patients. On the other hand, preparing a secure infrastructure to gather data from different sources is very time consuming and expensive [1]. No access to data essentially means no input into the data mining techniques, and subsequently, no analysis and outcome information.

To be able to access the required health data for analysis, there should be sufficient security protocols implemented in the data warehouse site, so that analysts can reach the data to run the required analysis. Although implementing these security protocols might be annoying, it is important to have them in place to both protect the privacy of patients as much as possible, and allow the analysts to access data to be able to conduct analysis. In addition, health data can be anonymized using masking and de-identification techniques, and be disclosed to the researchers based on a legal data sharing agreement [35]. However, the amount of anonymization is very crucial. If the data gets anonymized so much with the aim of protecting the privacy, on the other hand it will lose its quality and would not be useful for analysis anymore. Therefore, coming up with a balance between the privacy-protection elements (anonymization, sharing agreement, and security controls) is essential to be able to access a data that is usable for analytics.

C. Relying on Predictive Models

There should not be unrealistic expectations from the constructed data mining models. Every model has an accuracy. When a predictive model for diagnosing a health issue – e.g. diagnosing the type of thyroid gland – is built, it usually does not have an accuracy of 100%. Depending on the amount of model's accuracy and the importance of the decision being made using that model, we should decide how much we can rely on the outcomes of this data mining model. Especially when we are dealing with clinical medicine decisions based on the outcomes of one or multi data mining studies, it is important to consider that it would be dangerous to only rely on the predictive models when making critical decisions that directly affects the patient's life, and this should not even be expected from the predictive model.

However, data mining models and specifically predictive models, can be very helpful when they are used as a second opinion for the physician's decision on a treatment. In some cases, the doctor might have under-estimated the severity of a patient's disease, and the constructed model will alert the doctor to consider some aspects that might have been missed and makes the patient to be at risk. On the other hand, if the doctor is not too much sure if for instance a diagnostic surgery is required for this patient, then predictive model will help the doctor to decide – either with going for the diagnostic surgery or not – with more confidence by considering the previous cases via the constructed model.

D. Variety of Methods and Complex Maths

As the underlying math of almost all data mining techniques is complex and not very easily understandable for non-technical fellows, thus, clinicians and epidemiologists have usually preferred to continue working with traditional statistics methods. They understand the *p* value much better comparing to data mining's measurement methods such as correctness, sensitivity, specificity, and ROC curve. In addition, as there are many different data mining techniques and methods, it is difficult for clinicians to get familiar with all the different methods and easily select the proper one.

It needs to be considered that usually there is no one best technique that works for different datasets. It is essential for the data analyst to be familiar with the different techniques, and also the different accuracy measurements to apply multiple techniques when analyzing a specific dataset. This way, the chance of getting better results and outcomes will be higher. In this case, it would be necessary to get help from analysts that are more familiar with the different data mining techniques to be able to apply different methods and take advantage of data mining's strength in dealing with large datasets and accepting more input attributes comparing to traditional statistics.

VI. CONCLUSION

Some countries or states are spending a lot of money on their health sector; however, they do not achieve the expected quality of services in outcome. It is believed that by analyzing the related data and extracting the hidden information out of that data, many useful and applicable solutions can be developed. With the increase of implementing electronic systems, such as electronic health records (EHR) systems, in the health sector, there are massive amount of data being captured every day. Health data holders have shown more attention to data mining techniques in the past decades, as these techniques can help them to extract very useful information from their gathered data. These information can be used to improve the health services and deliveries, find the unknown relationship between diseases, and make the organization cost efficient.

There are different fields of application in the health area for data mining. But what makes data mining more in the spotlight, is the necessities of using data mining techniques in healthcare due to its specific properties that makes it more suitable when dealing with current health data. Discovering unknown issues through large amounts of health data that traditional statistics can no more handle this volume and consider all the records at once, makes data mining more essential for conducting health data analysis in the current era. In addition, data mining is also capable of detecting the patterns that cause over-budgeting, classifying microarray data for unknown health issues, predicting insurance frauds and high-risk patients, which each cause severe problems for any healthcare industry. However, we need to consider that there will be some challenges and issues when applying data mining techniques in healthcare, such as algorithm performance, information reliability, data quality, and variety of methods. There are some solutions to mitigate the concerns around these challenges. But yet, what is important is that even the current data mining algorithms can perform much better in analyzing health data comparing to other data analysis methods.

REFERENCES

[1] D. Tomar and S. Agarwal, 'A survey on Data Mining approaches for Healthcare', *Int. J. Bio-Sci. Bio-Technol.*, vol. 5, no. 5, pp. 241–266, 2013.

[2] J. Natale, 'Leveraging Technology to Revolutionize Canadian Health Care', *Policy: Canadian Politics and Public Policy*, vol. 2, no. 6, pp. 27–30, Dec-2014.

[3] D. J. Hand, H. Mannila, and P. Smyth, *Principles of Data Mining*. MIT Press, 2001.

[4] I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J.-F. Chang, and L. Hua, 'Data Mining in Healthcare and Biomedicine: A Survey of the Literature', *J. Med. Syst.*, vol. 36, no. 4, pp. 2431–2448, May 2011.

[5] 'The Technology Review Ten', *MIT Technology Review*, Feb-2001.

[6] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, 'The KDD Process for Extracting Useful Knowledge from Volumes of Data', *Commun ACM*, vol. 39, no. 11, pp. 27–34, Nov. 1996.

[7] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2006.

[8] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, 'From Data Mining to Knowledge Discovery in Databases', *Commun ACM*, vol. 39, no. 11, pp. 24–26, 1996.

[9] S. Velickov and D. Solomatine, 'Predictive Data Mining: Practical Examples', in *2nd Joint Workshop on Applied AI in Civil Engineering*, Cottbus, Germany, 2000.

[10] H. Hu, J. Li, A. Plank, H. Wang, and G. Daggard, 'A Comparative Study of Classification Methods for Microarray Data Analysis', in *Proceedings of the Fifth Australasian Conference on Data Mining and Analystics*, Darlinghurst, Australia, Australia, 2006, vol. 61, pp. 33–37.

[11] H. Cataloluk and M. Kesler, 'A diagnostic software tool for skin diseases with basic and weighted K-NN', in *2012 International Symposium on Innovations in Intelligent Systems and Applications (INISTA)*, 2012, pp. 1–4.

[12] R. Potter, 'Comparison of classification algorithms applied to breast cancer diagnosis and prognosis', presented at the 7th Industrial Conference on Data Mining, ICDM 2007, Leipzig, Germany, 2007, pp. 40–49.

[13] G. A. Beller, 'The rising cost of health care in the United States: Is it making the United States globally noncompetitive?', *J. Nucl. Cardiol.*, vol. 15, no. 4, pp. 481–482, Jul. 2008.

[14] D. Bertsimas, M. V. Bjarnadóttir, M. A. Kane, J. C. Kryder, R. Pandey, S. Vempala, and G. Wang, 'Algorithmic Prediction of Health-Care Costs', *Oper. Res.*, vol. 56, no. 6, pp. 1382–1392, Dec. 2008.

[15] M. H. Tekieh, B. Raahemi, and S. A. Izad Shenas, 'Analysing healthcare coverage with data mining techniques', *Int. J. Soc. Syst. Sci.*, vol. 7, no. 3, pp. 198–221, 2015.

[16] D. J. Hand, 'Data Mining: Statistics and More?', *Am. Stat.*, vol. 52, no. 2, pp. 112–118, May 1998.

[17] D. J. Hand, 'Statistics and Data Mining: Intersecting Disciplines', *SIGKDD Explor Newsl*, vol. 1, no. 1, pp. 16–19, Jun. 1999.

[18] 'Highmark maximizes Medicare revenues with SAS.' SAS, 2006.

[19] 'Healthways Heads Off Increased Costs with SAS.' SAS, 2009.

[20] Y. Zhang, S. Fong, S. Fiaidhi, and S. Mohammed, 'Real-time clinical decision support system with data stream mining', *J. Biomed. Biotechnol.*, vol. 2012, p. 8, 2012.

[21] T. Haferlach, A. Kohlmann, L. Wieczorek, G. Basso, G. T. Kronnie, M.-C. Bene, J. De Vos, J. M. Hernandez, W.-K. Hofmann, K. I. Mills, A. Gilkes, S. Chiaretti, S. A. Shurtleff, T. J. Kipps, L. Z. Rassenti, A. E. Yeoh, P. R. Papenhausen, W. -m. Liu, P. M. Williams, and R. Fo, 'Clinical utility of microarray-based gene expression profiling in the diagnosis and subclassification of leukemia: Report from the international microarray innovations in leukemia study group', *J. Clin. Oncol.*, vol. 28, no. 15, pp. 2529–2537, 2010.

[22] R. Salazar, P. Roepman, G. Capella, V. Moreno, I. Simon, C. Dreezen, A. Lopez-Doriga, C. Santos, C. Marijnen, J. Westerga, S. Bruin, D. Kerr, P. Kuppen, C. van de Velde, H. Morreau, L. Van Velthuysen, A. M. Glas, and R. Tollenaar, 'Gene expression signature to improve prognosis prediction of stage ii and iii colorectal cancer', *J. Clin. Oncol.*, vol. 29, no. 1, pp. 17–24, 2011.

[23] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, 'Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring', *Science*, vol. 286, no. 5439, pp. 531–537, Oct. 1999.

[24] L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend, 'Gene expression profiling predicts clinical outcome of breast cancer', *Nature*, vol. 415, no. 6871, pp. 530–536, Jan. 2002.

[25] R. Kandwal, P. K. Garg, and R. D. Garg, 'Health GIS and HIV/AIDS studies: Perspective and retrospective', *J. Biomed. Inform.*, vol. 42, no. 4, pp. 748–755, Aug. 2009.

[26] D. Delen, G. Walker, and A. Kadam, 'Predicting breast cancer survivability: a comparison of three data mining methods', *Artif. Intell. Med.*, vol. 34, no. 2, pp. 113–127, Jun. 2005.

[27] S. Shah, A. Kusiak, and B. Dixon, 'Data mining in predicting survival of kidney dialysis patients', in *Proceedings of Photonics West—Bios 2003*, Belingham, 2003, vol. 4949, pp. 73–79.

[28] 'First Things First—Highmark makes healthcare-fraud prevention top priority with SAS.' SAS, 2006.

[29] Y. M. Chae, S. H. Ho, K. W. Cho, D. H. Lee, and S. H. Ji, 'Data mining approach to policy analysis in a health insurance domain', *Int. J. Med. Inf.*, vol. 62, no. 2–3, pp. 103–111, Jul. 2001.

[30] M.-H. Kuo, T. Sahama, A. W. Kushniruk, E. M. Borycki, and D. K. Grunwell, 'Health big data analytics: current perspectives, challenges and potential solutions', *Int. J. Big Data Intell.*, vol. 1, no. 1/2, pp. 114–126, 2014.

[31] R. D. Canlas Jr, 'Data mining in healthcare: Current Applications and Issues', Carnegie Mellon University, Australia, 2009.

[32] F. Hosseinkhah, H. Ashktorab, R. Veen, and M. M. Owrang O., 'Challenges in Data Mining on Medical Databases', *IGI Glob.*, pp. 502–511, 2009.

[33] S. Hoffman and A. Podgurski, 'Big Bad Data: Law, Public Health, and Biomedical Databases', *J. Law. Med. Ethics*, vol. 41, pp. 56–60, Mar. 2013.

[34] C. Violán, Q. Foguet-Boreu, E. Hermosilla-Pérez, J. M. Valderas, B. Bolíbar, M. Fàbregas-Escurriola, P. Brugulat-Guiteras, and M. Á. Muñoz-Pérez, 'Comparison of the information provided by electronic health records data and a population health survey to estimate prevalence of selected health conditions and multimorbidity', *BMC Public Health*, vol. 13, no. 1, p. 251, Mar. 2013.

[35] K. El Emam, *Guide to the De-Identification of Personal Health Information*. CRC Press, 2013.