IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.114208

Application of the data mining model in the field of health

Salunke Aniket Vikram, Guide Sheetal Wadhai

Department of Computer Science, Universal College of Engineering & Research, pune

Abstract: This study looks at the benefits of data mining in everyday life, particularly in healthcare. Statistical analysis has gotten a boost thanks to the prevalence of computing technologies. Data mining uses and improves existing statistical approaches to forecast human behaviour in a variety of fields, from supermarket purchases to cancer vaccine production.

The paper begins with a quick overview of data mining, including examples of common and daily retail uses. The technology and methods used in data mining are briefly discussed. A brief conversation with a company that uses data mining to its advantage is mentioned.

The publication then goes on to describe a number of research studies that have employed data mining to answer important health problems. What age group is the most vulnerable tocardiovascular disease? Which cancer vaccine trial is the most popular? How many of these experiments were successful? What is an effective treatment for a rare paediatric disease? How can data mining be utilised to solve challenges in medical applications in different countries? How can one reliably calculate life expectancy? This document answers themajority of these questions.

The legal and ethical implications of data mining are then discussed. Finally, we end on a positive note about this intriguing technology's future possibilities.

INTRODUCTION

In today's society, it may appear that planning without data mining is tough, but picture waking up one day and discovering that you have no way of accessing any information that isimportant to you. Assume you are a doctor who has discovered that there is no way to look up the patient's habits and activities in the computer. There was no method to look for effective therapies and best practises, and there was also no means to analyse the data and avoid some of the industry's pitfalls. We all know how valuable and powerful data is. But how?

With the evolution of data mining, we can now answer critical questions like "What kind of surgery resulted in patients being in hospitals for more than five days?" and "What were themost prevalent pre-surgery symptoms in patients who stayed in the hospital for a longer amount of time?" Data mining is useful not only in the healthcare field, but also in enhancing customer happiness, better targeting marketing campaigns, detecting high-risk clients, and improving manufacturing processes in all industries. Our focus will be on healthcare becausethis article is centred on data mining applications in healthcare.

DEFINATION AND USAGE

Data mining is a powerful new technique that has the ability to help businesses focus on the most critical information in the data they've gathered about their customers' and potential customers' behaviour. You may learn and study a lot about trends and behaviours by using data mining. This can assist in making important business decisions. Data mining can be usedfor a variety of purposes, including:

1) Fraud Detection: Big companies like Macy's or J. C. Penny, as well as smaller small firms, can keep track of which customers purchase items and then return them after using them. If the transactions are done with a single credit card, this type of information can be traced. During one of the author's job interviews, she interacted with Mr. Shane Johnson, a business analyst at Buckle, Inc., stated that many customerswill purchase a certain item, such as child apparel or a women's outfit, and then returnit within a few days. These dresses are commonly worn, and the store discovered thatthe customers who were doing this were mostly ladies between the ages of 18 and 29, and of Hispanic heritage, after capturing credit card information and delving deeper. However, there is nothing we can do to solve the problem. However, we can only tell them that they have a good return track record. As a result, this group of customers will believe that the store is knowledgeable.(Jhonson,2014)

2) Can recognise complementary goods for a specific product type:

a) Amazon provides a good illustration of how descriptive data can be utilised to make predictions. Amazon

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.114208

discovered a link between cocktail shaker and martini glass purchases by looking at the user's buying history (The Atlantic, 2012).

Another example could be:

b) Target assigns each customer a Guest ID number, which is linked to their credit card, name, or e-mail address and serves as a bucket for storing a history of everything they've purchased as well as any demographic data Target has gathered from them or purchased from other sources (Hill, 2012)

Application of the data mining model in the field of health

According to the WHO, the two deadliest killers in the world are cardiovascular disease and cancer, in that order (Mathers CD, 2009). Better understanding of causes and symptoms can undoubtedly prevent or delay mortality. Data on patients can be found in international hospital databases. However, there appears to be no consistency in the data's format or accessibility. Even if all or most of the data could be delivered in a mutually understandable format, human conclusions from hidden patterns are impossible. The majority of the hidden information or pattern would go unnoticed, limiting the usefulness of the valuable data to a small group of isolated patients. Physicians in advanced technical nations such as the United States and the United Kingdom would be unable to effectively research such data and discover new ground-breaking remedies for the entire human race.

DISEASES OF THE CARDIOVASCULAR SYSTEM

Peyman Rezaei Hachesu1 (2013) employed traditional data mining strategies such as Decision Trees, Artificial Neural Networks (ANNs), and Support Vector Machine (SVM) to try to predict the early start of Coronary Artery Disease (CAD). Although the study was limited to a single location, and the beginning of CAD is also race-related, it provides vital information into CAD prediction. The three algorithms above were applied to a cohort of roughly 5000 CAD patients.

To ensure the research's validity and sanctity, the following actions were taken:

1) The sample population was carefully selected with expert medical help, comprising patients from a specific heart health institution in Teheran, Iran, meeting the study'scriteria.

2) All of the patients in the available pool lacked consistent or comprehensive information. Data was pre-processed to eliminate noise, and missing data weremostly replaced with average values. Outliers were also deleted. Values outside the first and third quartiles were considered as outliers. Minitab14 was used to explore the data distribution in greater depth.

3) Only about 2000 data points were found to be complete and legitimate after the cleanup. Because data mining requires the separation of a training and testing set,80 percent of the data was used for training and 20% for testing.

The average age of beginning of CAD was 58, with the 54-64 year old age group being themost vulnerable. The SVM method was determined to be the most accurate overall.

The beginning of CAD in various nations, including the United States, can be predicted usingidentical data sets and the same analysis algorithm (SVM). The American Heart Association predicts that the cost of treating heart disease in the United States will quadruple by 2030. (American Heart Association, 2011). Further investigation into the variables that cause CAD could greatly minimise this cost.

CANCER

Cancer is not far behind cardiovascular disease as the leading cause of death. Cancer is catching up as the leading cause of death, with global cancer deaths expected to rise from
7.1 million in 2002 to 11.5 million by 2030. (World Health Organization, 2007)

4. The world's leading pharmaceutical corporations are (literally) racing to develop new cancer-curing drugs and chemicals. Clinical trials are an important aspect of the introductionof every new medicine or vaccination. Clinical trials (National Institutes of Health, 2014) are research investigations that examine whether a medical approach, treatment, or equipmentis safe and effective in humans. As a result, clinical studies generate large amounts of data; yet, gathering data is meaningless if it cannot be mined or evaluated effectively. ClinicalTrials.gov, a US government website, has a variety of publicly available clinical data.

Three US researchers attempted to synthesise and illustrate cancer vaccine clinical trials using data available on the disease (Xiaohong Cao*1, 2008). The researchers concluded that,despite the vast amount of data accessible, only basic

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.114208

querying strategies had been applied thus far. The researchers were able to answer important issues such as when the trials started and if they were successful, the vaccination platforms utilised, and the trial phase using advanced data mining and bioinformatics. The most crucial question answered was whether any of the cancer kinds were overlooked in research and trials. The researchers discovered that several types of cancer that are similarly fatal, such as bladder, liver, pancreatic, stomach, and esophageal, were overlooked.

Other major findings from publicly available cancer clinical trial data utilising data miningtools include:

1) Despite the fact that the first cancer vaccination trial (lung) was undertaken in 1971, theprogressive prevalence of trials did not begin until the early 2000s. Since then, the numberof trails has constantly increased.

2) Melanoma (skin cancer), cervical cancer, prostate cancer, breast cancer, and leukaemia are the top five cancers targeted by vaccine treatment in clinical studies. Melanoma has themost trial candidates, followed by cervical cancer.

3) In terms of the institutions conducting the trials, the National Cancer Institute was shown to be the overwhelming leader, followed by GSK (GlaxoSmithKline). All other pharmaceutical companies contributed roughly equally to cancer vaccine trials and research.

4) The effectiveness of cancer vaccination studies can also be determined by the vaccineapproach employed. The bulk of the trials used an antigen-based vaccine followed by a cellular-based vaccine, according to the researchers. Antigen- and cellular-based vaccinations account for more than 80% of the studies.

5) An intriguing scatter-plot depicting current cancer prevalence and survival rates with existing treatment on the X-axis and five-year survival rates on the Y-axis. Prostate, melanoma, breast, and cervical cancers all have high clinical trial rates (dark red circles).Prostate cancer, too, has a good survival rate. Figure 1 in the appendix.

PEDIATRICS

Pediatrics is garnering more attention in the medical community. With new specialised hospitals like St. Jude Children's Research Hospital in Memphis, TN, mining all accessible inpatient data is more crucial than ever. The appropriately called 'KID,' or Kids' Inpatient Database,' is a one-stop shop for all paediatric clinical data (Bliss-Holtz, 2012). The KID is part of the HCUP (Healthcare Costs and Utilization Project) family, which was developed in collaboration with the Agency for Healthcare Research and Quality (AHRQ), a federal agency, in a federal-state-industry partnership. Because the data sets are big, disorders thatare relatively uncommon in children, such as prune belly syndrome, can be easily studied.
Primary and secondary diagnoses; primary and secondary procedures; admission and discharge status; patient demographics, including gender, age, race, and median income (by ZIP code data); total charges; duration of stay; and hospital characteristics are among the variables included in the KID (e.g., ownership, size, teaching status). The KID is thus a genuine gold mine that, if correctly mined, can help clinicians answer numerous pediatrics- related problems.

HEALTHCARE FOR OUTPATIENTS

Most outpatients are not treated as well as inpatients in a typical hospital – likely because they cost less – but outpatient illnesses can be complex, and having proper understanding about diseases, ailments, and treatments can save both the patient and the care provider money. A study released in 2013 (Huang, 2013) used a medical database from a Taiwanese hospital to find the optimum algorithm for analysing such a data set. Between abnormal health examination results and outpatient illnesses, association rules can be built. After that, a disease preventive knowledge database can be created to aid healthcare providers with follow-up treatment and prevention. The author also presents a novel algorithm for more successfully analysing such a data set. The strength of data mining and the possibility for additional research is easily proved, even if it is a candidate for more thorough testing.

A few observations about the study's data mining methods and research methodology:

1) To illustrate association rules as required by this study, apriori algorithms are commonly used. Apriori algorithms were initially proposed in 1993 and have been popular ever since (Huang, 2013). Apriori, on the other hand, necessitates frequent database scans, which are inefficient. As medical research advances, the need to link multiple diseases and causes has become more important.
2) Because the study was done in Taiwan, the data consisted of two parts: health examination results and

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.114208

outpatient medical records from a Taiwanese hospital. There was no differentiation established in the medical department. The data frompatient health checkups was classified into three categories: normal (01), below normal (02), and above normal (03). (03). Because the link sought was between aberrant health findings and outpatient illness (about 100,000 data points), normalhealth data was filtered out.

3) Six months before and after the clinical data, outpatient illness records were gathered. In addition, data that was incomplete, prenatal, or dental was eliminated from the dataset. A flowchart of the data integration process can be seen in Figure 2in the appendix.

4) Due to the limitations of the Apriori method, a new algorithm DCSM – Data Cutting and Sorting Method was proposed.

The DCSM is a seven-step process that includes:

a. Converting data to a Boolean matrix.
b. For high frequency data, create huge item sets.
c. Create a reductions matrix to eliminate unpaired data.
d. Repeat step (b).
e. Repeat step (c).
f. Repeat step (d).
g. Repeat steps (c) and (d) to return to step (b) (f).

5) Empirical study demonstrated that association rules discovered using DCSM and Apriori were identical, proving the new algorithm's validity. DCSM, on the other hand, was shown to be ten times faster than Apriori.

6) Medical doctors and independent research backed up the association guidelines.

ASPECTS OF DATA MINING LAW

In the legal sector, data mining of health-care databases has two broad applications. The first is in non-healthcare legal cases, where data mining can be utilised as reliable evidence during depositions. It's worth noting that Federal Rule of Evidence 404(b) makes no distinction between prior acts discovered by humans and prior acts discovered by computer utilising data mining. As a result, a plaintiff with a claims-related case can employ acceptabledata mining techniques to defend himself in court. The healthcare data itself is the second legal issue of data mining. The US Supreme Court judgement in Sorrell versus IMS Health Inc. in June 2011 concluded that Vermont's legislation preventing pharmacists from selling prescription data to "data-mining businesses" violated the First Amendment's Free Speech Clause (Cohen, 2012). HIPPA (Health Insurance Portability and Accountability Act of 1996) plays a key role when it comes to healthcare data. Because the Federal Privacy Rule, which administers HIPPA, bars any illegal use or disclosure of protected health information for marketing reasons, the Supreme Court's decision is somewhat surprising. However, because regulations are normally construed "in context" (and this was a marketing case, not a research case), the Supreme Court decision presents numerous hurdles to data mining evangelists who aim to make all healthcare research data worldwide. It must be carefully established where marketing ends and profitable research begins. However, privacy laws vary by country, and what is legal in one area may be prohibited in another. Due diligence must be performed prior to any large monetary or time commitment, especially when building multinational healthcare systems.

PROPOSED SOLUTIONS AND FINDINGS

Technology is seductive (and lucrative), but there must be legal safeguards in place to avoidabuse. There is currently no international legislation in place. There are only a few laws in afew advanced countries. The formation of consortia of major countries (including emerging markets) to deliberate and regulate on transnational and ethical aspects of data mining is required. To prevent abuse, laws must benefit the poorer economies. In a difficult field like data mining, education is essential. Many prominent colleges have begun to offer Data Mining courses, but much more has to be done to reach the general public. Data mining isn't just for the wealthy; it will be employed in everyday lives in the near future.

CONCLUSION AND SUMMARY

Data mining is a relatively new technique that is still in its early stages. There are few applications, and just a small portion of the pie has been discovered so far. Applications are now limited to more experimental regions. Every day,

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

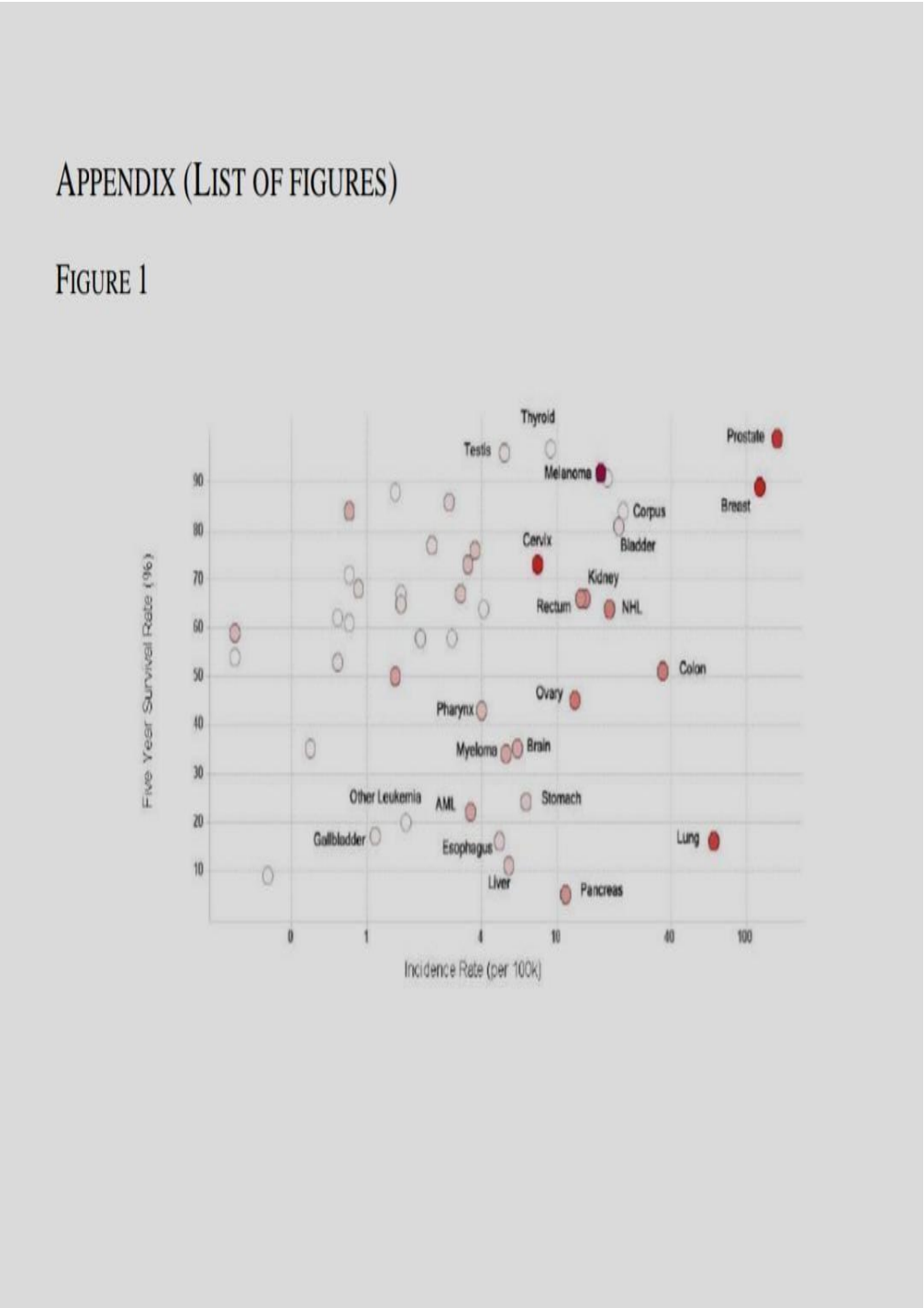
Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.114208

data mining should become easier and more prevalent. However, data mining algorithms should be able to'self-tune' in the near future, assisting researchers, particularly in healthcare, in eradicating terrible diseases like cancer. Furthermore, most generated data mining patterns are now more mathematicalthan practical, and are considered "rocket science" by most people who are not skilled in the field.

More technical abstraction layers (by built application software) should be used in the future to make use of and comprehend data mining technologies, similar to how e-mail isused today (Borgwardt, 2007).

APPENDIX (LIST OF FIGURES)



IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.114208

FIGURE 2

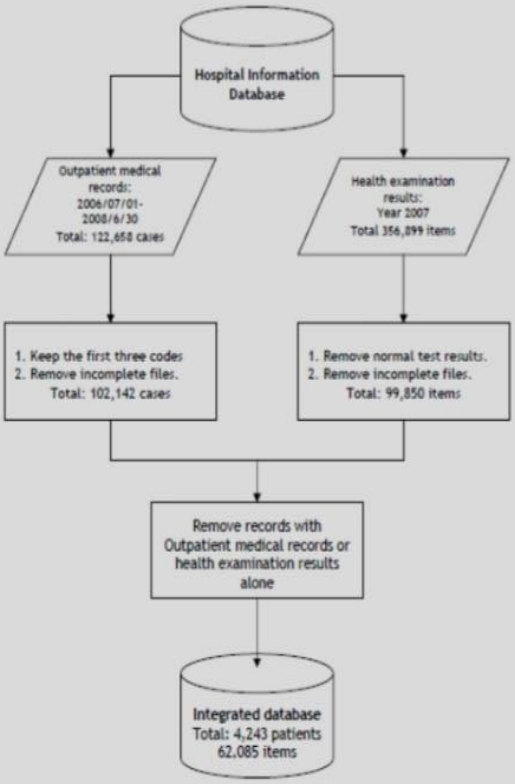
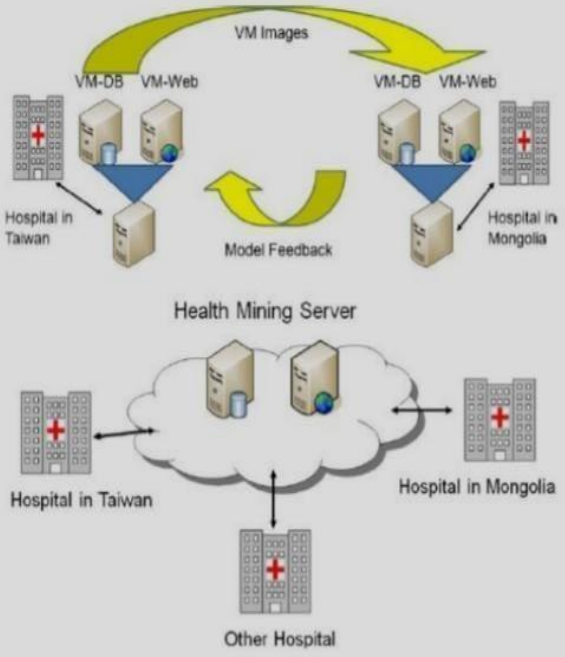


FIGURE 3



DOI: 10.17148/IJARCCE.2022.114208

REFERENCES

1. (2012). Retrieved from The Atlantic: http://www.theatlantic.com/technology/archive/2012/04/everything-youwanted-to-know-about- data-mining-but-were-afraid-to-ask/255388/
2. (2014). Retrieved from Microsoft Technet: http://technet.microsoft.com/en-us/library/ms175595.aspx
3. American Heart Association. (2011). Retrieved from Cost to treat heart disease in United States willtriple by 2030: www.sciencedaily.com/releases/2011/01/110124121545.htm
4. Bliss-Holtz, J. (2012). THE KIDS' INPATIENT DATABASE (KID) AND DATA MINING. Informa Healthcare
5. USA, Inc.
6. Borgwardt, H.-P. K. (2007). Future trends in data mining. Springer Science+Business Media.
7. Cohen, B. (2012). REGULATING DATA MINING FOST-SORRELL: USING HIPAA TO RESTRICT MARKETING USES OF PATIENTS' PRIVATE MEDICAL INFORMATION. Wake Forest Law Review.
8. Hernandez, D. (2014). Doctors monitor patients remotely via smartphones and fitness trackers. Retrieved from http://www.pbs.org/newshour/updates/doctors-monitor-patients-vitals-via- smartphones-fitness-trackers
9. Hian, C. K. (n.d.). Data mining applications in healthcare. Retrieved from Journal of Healthcare Information Management: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.92.3184&rep=rep1&type=pdf
10. Hill, K. (2012). How target figured out a teen girl was pregnant before her father did. Retrieved from http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was- pregnantbefore-her-father-did/
11. Huang, Y. C. (2013). Mining association rules between abnormal health examination results andoutpatient medical records. Health Information Management Journal.
12. Jason Scott Mathias, 1. A. (2013). Development of a 5 year life expectancy index in older adults using predictive mining of electronic health record data. Journal of the American Medical InformaticsAssociation.
13. Jigjidsuren, C.-P. S. (2011). A Data-Mining Framework for Transnational Healthcare System. Journal of Medical Systems.
14. Johnson, S. (2014). (K. K, Interviewer)
15. Koh HC1, T. G. (n.d.). US National Library of Medicine National Institutes of Health. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15869215
16. Mathers CD, L. D. (2009). Projections of global mortality and burden of disease from 2002 to 2030.
17. National Institutes of Health. (2014). Retrieved from https://www.nhlbi.nih.gov/health/healthtopics/topics/clinicaltrials/
18. Peyman Rezaei Hachesu1, M. A. (2013). Cardiac diseases prediction and rule extract with data mining - Classification techniques. HealthMed.
19. Wikipedia. (2014). Retrieved from http://en.wikipedia.org/wiki/Data_mining