Towards the Application of Big Data: A New Way to make Data Driven Healthcare Decision

Tanvir Ahammad
Department of CSE
Jagannath University, JnU
Dhaka, Bangladesh

Md. Sajib Al Mamun
Department of CSE
Jagannath University, JnU
Dhaka, Bangladesh

Mehnaz Tabassum
Ass. Prof., Dept. of CSE
Jagannath University, JnU
Dhaka, Bangladesh

ABSTRACT

In era of Big Data, large amounts of both structured and unstructured (also semi-structured) data are growing exponentially because of using social media, smart phones, archives and remote sensors. Most of these (approx. 80-90% data) are unstructured, commonly found in textual form and difficult to process using conventional database and software techniques. These large amounts of unstructured as well as complex data are giving new opportunities to the business organizations, governments, and researchers. Successful harnessing of Big Data provides valuable insights; improves data-driven decision making in healthcare, manufacturing, education, stock marketing, banking and insurance, agriculture, weather forecasting, travel and transportation, telecommunication, sports, and entertainment. This paper aims to emphasize on Big Data application and demonstrates that the analysis of healthcare data paves the way to make a better healthcare decision, reduce cost, and raise healthcare consciousness. The experimental analysis has been accomplished with a proposed methodology based on existing Text Mining and Natural Language Processing (NLP) techniques.

General Terms

Big Data, Text Mining, NLP, KDD, GATE.

Keywords

Knowledge Discovery from Healthcare Text, KDHT.

1. INTRODUCTION

Nowadays, the world is moving towards the information society, where large amount of data are needed to extract better knowledge; called Knowledge Discovery from Data, or KDD [1].The Internet represents a big space in Yottabyte (YB) level to accommodate this large amount of data. These data (called "Big Data") appear as a new power that changes everything it interacts with. Thus, it is considered to be a new weapon for the 21st century. It was in the early 21st century when the world first heard about the concept of big data and the attributes like too large, too unstructured and too fast-moving were used for describing the nature of the data [2]. These attributes are now familiarizing us about Big Data and its future. According to IBM, 2.5 quintillion bytes (2.5 billion GB or 2.5 Exabyte) of data was generated every day in 2012 and about 90% of the data in the world today originated in the last two years alone [3] [4]. In every second, 2.9 million emails are sent, 20 hours of video is uploaded in every minute and 50 million tweets are generated per day. So it is predicted that by 2020, the size of the data universe will reach 44 zettabytes, or 44 trillion gigabytes [5].

This rapid growth of data is not a matter of concern for us; rather bringing bigger opportunities [6]. Many business organizations as well as governments are now getting actionable insights with various key-decisions making in healthcare, manufacturing, education, banking, crime analysis, traffic management, stock marketing, insurance, sentiment analysis, etc. Healthcare industry is one of the leading sectors of these.

This paper has revealed how Big Data makes it possible to take better health decisions, to reduce cost, and to awaken healthcare consciousness. In the analysis, a methodology has proposed in the context of text mining and NLP, where healthcare information is first collected from different text data sources, e.g., healthcare news, blogs, social media, patient report and then various health issues such as disease name, cause, symptoms, outbreak period (season), and patient records are extracted; these are then stored into DB in a structured form for further tasks such as Data mining. The retrieved data was web based health sources, basically open source, where one patient can view others discussions who have the same disease or condition; track and share their own experiences and so on.

The structure of this paper is stated as follows: section 2 represents importance, sources, potentiality as well as challenges of Big Data in healthcare. In section 3 and 4, the proposed methodology and framework are described. Section 5 represents experimental results. Finally, in section 6 and 7, the key decisions are mentioned based on the experimental results; also mentioned limitations and suggesting as future works.

2. BIG DATA IN HEALTHCARE

Healthcare is one of the leading industries, generating Zettabyte of data every day. In Big Data era, data is growing and moving faster than healthcare organizations can consume it; 80% of medical data is unstructured and is clinically relevant [5]. Earlier, most of these were stored in hard copy form, while the current trend is toward rapid digitization [7]. That is why, healthcare is becoming digitized industry from multiple sources, classified as [5]: patient behavior and sentiment data, pharmaceutical and R&D data, healthcare data on the web, clinical data, claim, Cost & Administrative data, and streamed Data.

The impact of Big Data in healthcare system is based on the five new pathways [8]: right living, right care, right provider, right value, and right innovation.

Many healthcare organizations are leveraging Big Data technology to collect all data about a patient for getting more complete view of information. Successfully harnessing Big Data unleashes the possibility to achieve the following critical objectives for healthcare transformation [5]**:** reduced costs, access to complete information, better preventative care, personalized diagnoses, and improved evaluations.

However, the biggest challenge for data-driven care is to extract, or break down the information as cited in MIT Technology Review's report [9]. According to HealthCatalyst observation [10], the reasons that the healthcare data is unique and difficult to measure, including:

- Most of the data reside in multiple places and formats (e.g., text, numeric, and paper, digital, pictures, videos and multimedia);
- The data is both structured and unstructured;
- Inconsistent or variable definitions (e.g., one group of clinicians may define a group of asthmatic patients differently than another group of clinicians);
- The data is complex. For example, EMRs give a more complete picture of the patient's story, but difficult to process.

3. METHODOLOGY

This section demonstrates how textual information can be collected from multiple healthcare data sources, how information can be extracted and how text analyzer can be applied to process the extracted text. So it requires analysis to textual data as well as analysis with numerical data, related to the health. The textual data may be healthcare news, articles, report, prescription or sometimes Electronic Medical Record (EMR). On the other side, numerical data represents health cost, patient age, number of times to admit into hospital or number of diagnosis. Both types are equally important to extract desired information.

However, the challenge is to deal with textual data as these are unstructured; written in different formats; needed to be collected from different locations. That is why, these data are sometimes difficult to analyze and measure. To solve this problem, a methodology has been proposed that collects healthcare textual data from diverse sources; extracts information and filters to ensure better information extraction; uses text analyzer for other purposes. Besides, the information is stored at the same time to enrich the database for other type of analysis, e.g., more advanced data mining purposes. The Figure 1 shows a conceptual overview of the methodology.

4. FRAMEWORK

In this proposed methodology, data can be collected from various sources, e.g., healthcare blogs, social media, web-based consumer health information resource, etc. with different formats such as doc, pdf, txt, html, xml except non-textual (image, audio, video). Some of these sources are open access, while others are authenticated such as diagnostic center, hospital or clinical center data.

As data is collected from variety of formats, it will be difficult to process for extracting information. So, a common platform converting all of these into one format (i.e., html) has used. Besides, information extraction should be accomplished by processing multiple text documents at the same time rather than single. Hence the proposed methodology suggests a corpus based text system, which carries a collection of text document, used as the input for information extraction step.

Next step is the information extraction (IE), used for extracting healthcare information. The common steps involving in IE:

- Sentence splitting or segmentation,
- Tokenization,
- Parts-of-speech (POS) tagging,
- Name entity recognition,
- Semantic tagger,
- Coreference analysis.

In subsequent layer, a filtering process has applied on extracted text to determine the relevant text. After IE, various texts may be selected; some may have desired, while others may have desired, but redundant, or sometimes irrelevant output, e.g., to find a patient's age, has no clues for age. In this case, filtering can be applied, provided with healthcare keywords in a database (contains symptoms, disease names, causes, etc.). The most important part of this step is to maintain a system log, for storing filtering records. This is important because the logs can be analyzed to make a better IE process.

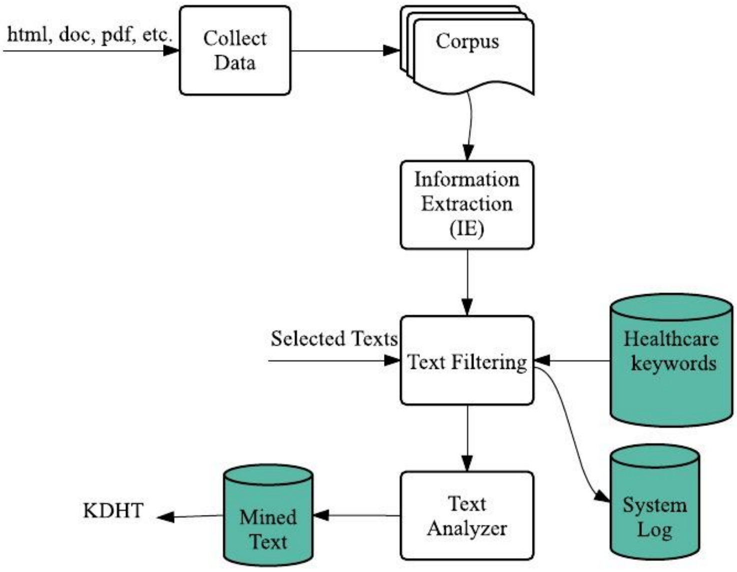
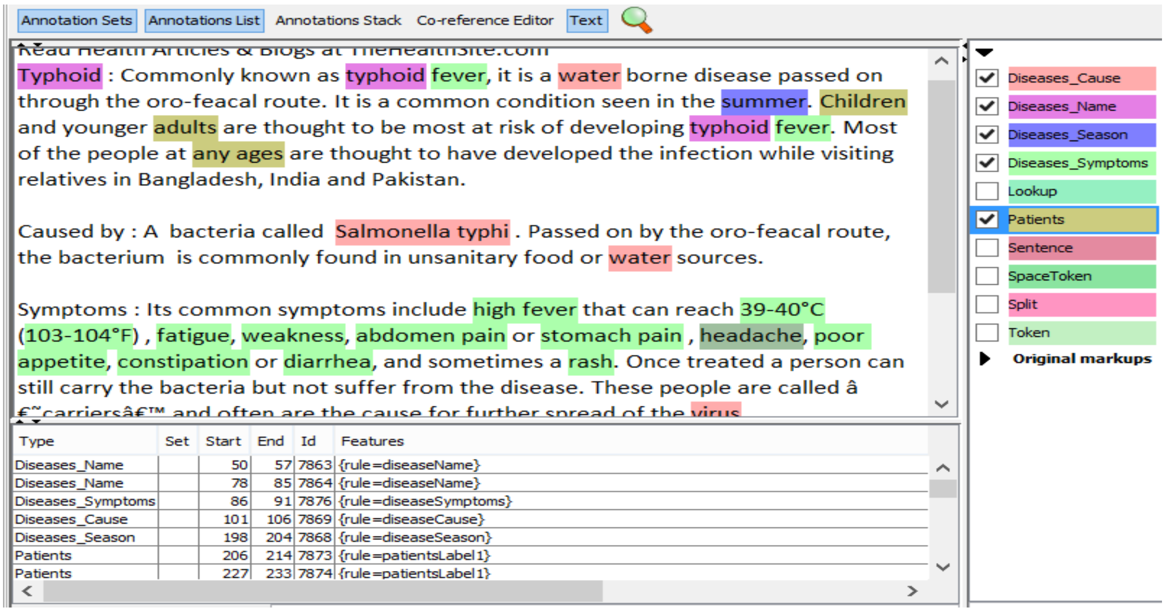


Fig 1: Overview of healthcare text data processing framework

5. EXPERIMENTAL RESULT

This section will demonstrate few examples of the analysis that provides some results of experiment. The experiment accomplished through an open source text mining tool, General Architecture for Text Engineering (GATE) capable of doing natural language processing tasks, including information extraction in many languages. It has also a defined and repeatable process for creating robust and maintainable text processing workflows [11].The data used in experiment mainly from web based healthcare sources [12] [13], which are directly accessible and reliable.

Document 1: The summary of information that has been extracted from the document 1 is given in the following figure:



Read Health Articles & Blogs at TheHealthSite.com
Typhoid : Commonly known as typhoid fever, it is a water borne disease passed on through the oro-feacal route. It is a common condition seen in the summer. Children and younger adults are thought to be most at risk of developing typhoid fever. Most of the people at any ages are thought to have developed the infection while visiting relatives in Bangladesh, India and Pakistan.

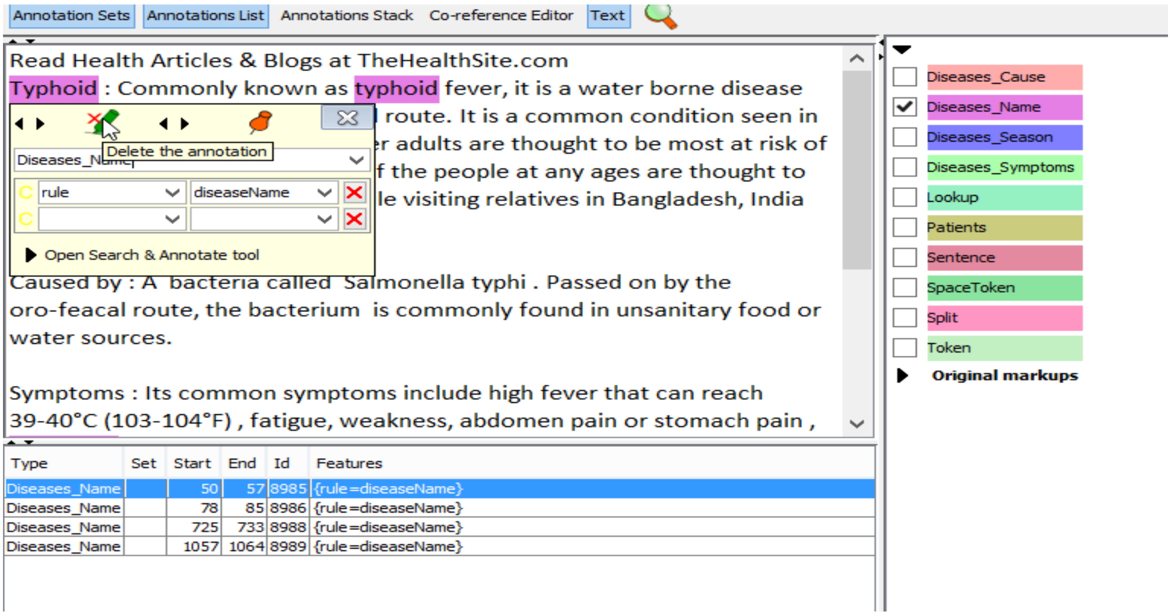
Caused by : A bacteria called Salmonella typhi . Passed on by the oro-feacal route, the bacterium is commonly found in unsanitary food or water sources.

Symptoms : Its common symptoms include high fever that can reach 39-40°C (103-104°F) , fatigue, weakness, abdomen pain or stomach pain , headache, poor appetite, constipation or diarrhea, and sometimes a rash. Once treated a person can still carry the bacteria but not suffer from the disease. These people are called â €˜carriersâ€™ and often are the cause for further spread of the virus

Type	Set	Start	End	Id	Features
Diseases_Name		50	57	7863	{rule=diseaseName}
Diseases_Name		78	85	7864	{rule=diseaseName}
Diseases_Symptoms		86	91	7876	{rule=diseaseSymptoms}
Diseases_Cause		101	106	7869	{rule=diseaseCause}
Diseases_Season		198	204	7868	{rule=diseaseSeason}
Patients		206	214	7873	{rule=patientsLabel1}
Patients		227	233	7874	{rule=patientsLabel1}

Fig 2: Disease name, cause, occurrence time, symptoms, and patient extraction from unstructured text

The above Figure 2 shows some redundant texts; i.e., typhoid is selected in three times and fever as a symptom is selected in two times. So filtering process was applied to eliminate the redundant terms. The following figure shows the result of this experiment:



Read Health Articles & Blogs at TheHealthSite.com
Typhoid : Commonly known as typhoid fever, it is a water borne disease
route. It is a common condition seen in
r adults are thought to be most at risk of
f the people at any ages are thought to
le visiting relatives in Bangladesh, India

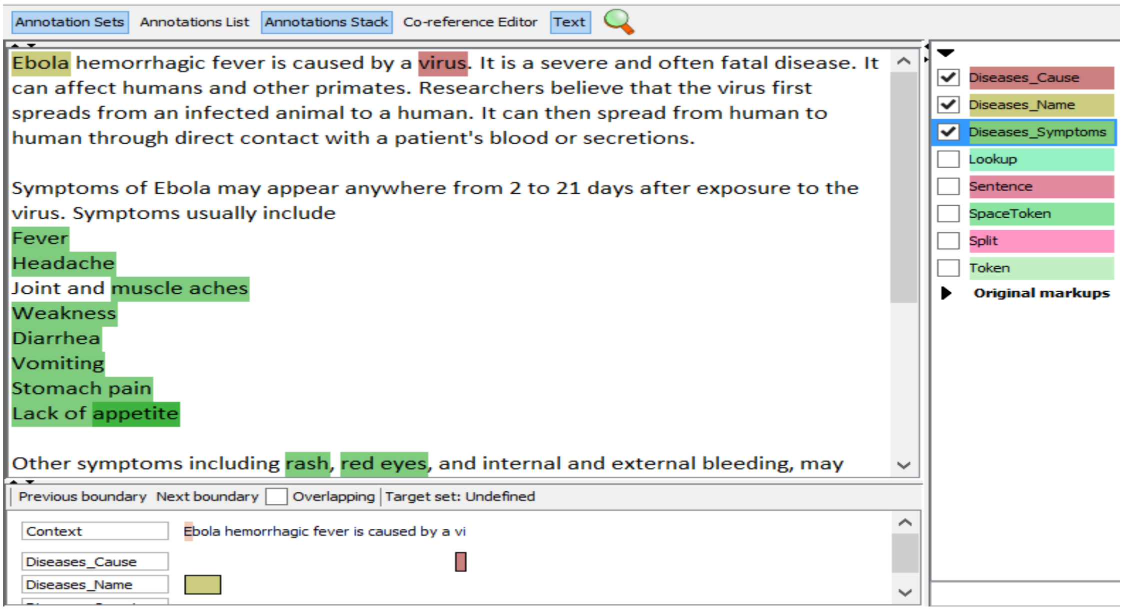
Caused by : A bacteria called Salmonella typhi . Passed on by the oro-feacal route, the bacterium is commonly found in unsanitary food or water sources.

Symptoms : Its common symptoms include high fever that can reach 39-40°C (103-104°F) , fatigue, weakness, abdomen pain or stomach pain ,

Type	Set	Start	End	Id	Features
Diseases_Name		50	57	8985	{rule=diseaseName}
Diseases_Name		78	85	8986	{rule=diseaseName}
Diseases_Name		725	733	8988	{rule=diseaseName}
Diseases_Name		1057	1064	8989	{rule=diseaseName}

Fig 3: Removal of redundant annotation as manual filtering process

Document 2: Consider another document for our analysis where some annotations may not be found. In this case NA (Not Available) needs to be inserted into DB for missing attribute, e.g., if the proposed system does not detect any patient or disease occurrence season, it needs to insert NA in the patient attribute of our DB. The Figure 4 shows a condition, where two types of information were not found.



Ebola hemorrhagic fever is caused by a virus. It is a severe and often fatal disease. It can affect humans and other primates. Researchers believe that the virus first spreads from an infected animal to a human. It can then spread from human to human through direct contact with a patient's blood or secretions.

Symptoms of Ebola may appear anywhere from 2 to 21 days after exposure to the virus. Symptoms usually include
Fever
Headache
Joint and muscle aches
Weakness
Diarrhea
Vomiting
Stomach pain
Lack of appetite

Other symptoms including rash, red eyes, and internal and external bleeding, may

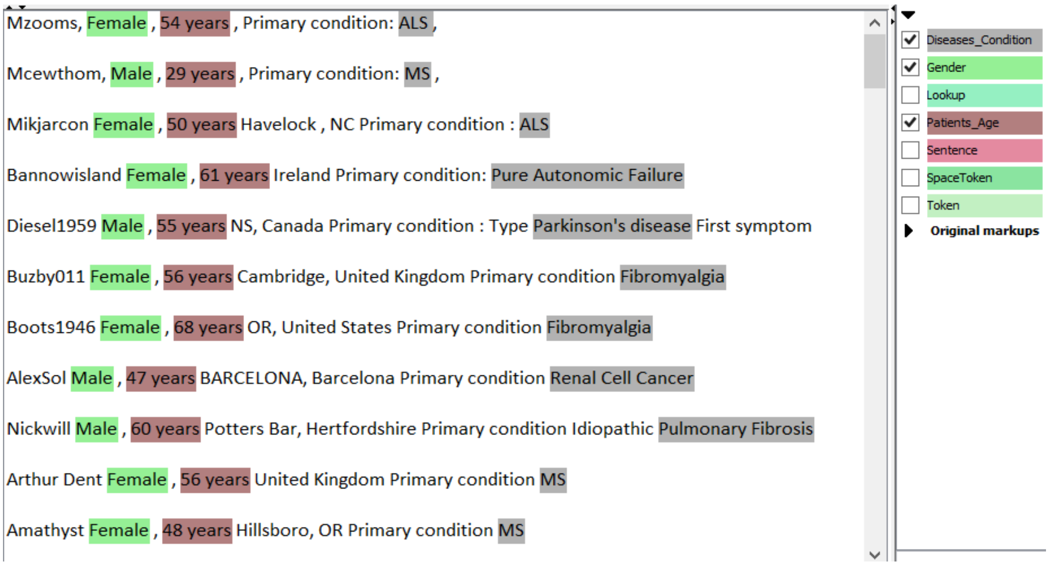
☑ Diseases_Cause
☑ Diseases_Name
☑ Diseases_Symptoms
□ Lookup
□ Sentence
□ SpaceToken
□ Split
□ Token
▶ Original markups

Previous boundary Next boundary □ Overlapping Target set: Undefined

Context: Ebola hemorrhagic fever is caused by a vi
Diseases_Cause
Diseases_Name

Fig 4: Two missing annotations: season, patient

Document 3: This document shows a report on 100 patient's profile where the information was collected from PatientsLikeMe [12], which is a patient powered research network; improves lives and a real-time research platform that advances medicine. Here all the patients share their health information including disease condition, symptoms, and diagnosis. The Figure 5 depicts the portion of result from 100 patient's profile.



Mzooms, Female , 54 years , Primary condition: ALS ,

Mcewthom, Male , 29 years , Primary condition: MS ,

Mikjarcon Female , 50 years Havelock , NC Primary condition : ALS

Bannowisland Female , 61 years Ireland Primary condition: Pure Autonomic Failure

Diesel1959 Male , 55 years NS, Canada Primary condition : Type Parkinson's disease First symptom

Buzby011 Female , 56 years Cambridge, United Kingdom Primary condition Fibromyalgia

Boots1946 Female , 68 years OR, United States Primary condition Fibromyalgia

AlexSol Male , 47 years BARCELONA, Barcelona Primary condition Renal Cell Cancer

Nickwill Male , 60 years Potters Bar, Hertfordshire Primary condition Idiopathic Pulmonary Fibrosis

Arthur Dent Female , 56 years United Kingdom Primary condition MS

Amathyst Female , 48 years Hillsboro, OR Primary condition MS

☑ Diseases_Condition
☑ Gender
□ Lookup
☑ Patients_Age
□ Sentence
□ SpaceToken
□ Token
▶ Original markups

Fig 5: Patient's information extraction based on the primary disease condition, gender, age

The highlighted texts in the above figure are the desired result; the primary condition describes the patient's own condition; i.e., a patient mainly suffers from this condition. However the Table 1 shows an overview of this experiment.

6. DATA DRIVEN DECISION

Harnessing the Big Data allows us to make better information based decision. That is, converting data into information, its meaning is determined, or discovered knowledge and then make a better decision from it. This section will describe how knowledge can be discovered from extracted healthcare text, called Knowledge Discovery from Healthcare Text, or KDHT; also demonstrate how to make important decisions based on KDHT. As the experiment accomplished on two types of information: one for disease information (diseases name, causes, symptoms, and patients), and another for 100 patient's profile from PatientsLikeMe [12], so the possible outcomes for both types will also describe here. In Figure 6 shows part of the database of the experimental result; that is extracted information.

Each row in the Table 1 and Figure 6 indicate a disease information, i.e., its name, why it occurs, when it generally spreads out, possible symptoms, and who can affect by it. So our insights from the above observation lead us to the following outcomes:

Table 1. Part of the extracted information from Document 3

Condition	Gender	Age
ALS	Female	54 years
MS	Male	29 years
ALS	Female	50 years
Pure Autonomic Failure	Female	61 years
Parkinson's disease	Male	55 years
Fibromyalgia	Female	68 years
Renal Cell Cancer	Male	47 years
Pulmonary Fibrosis	Male	60 years
MS	Female	56 years
MS	Female	48 years

- Government can alert mass people on a disease for any specific season through telehealth program.
- Healthcare organizations can notify people about necessary treatment based on their intelligent data analysis result. For example, common symptoms from diseases could be analyzed.
- Different age groups can be notified about diseases, which occur in any specific age group.
- For some diseases certain attributes can also be predicted, e.g., Ebola can spread in any season, and infect anyone. So awareness should be raised about it.

By observing the result from Table 1, it is possible to get more interesting information easily, i.e., to find the number of patients for a particular disease condition. The Figure 7 shows thirty one diseases; 100 patients usually suffer from these diseases condition. For disease MS (Multiple Sclerosis) there are 24 patients; 12 people were suffering from Parkinson's; 9 patients were suffering from Fibromyalgia and so on. Therefore, it can be said that MS (an autoimmune disorder characterized by destruction of myelin in the central nervous system) is more harmful than others, because maximum patients (24 out of 100) were suffering from it.

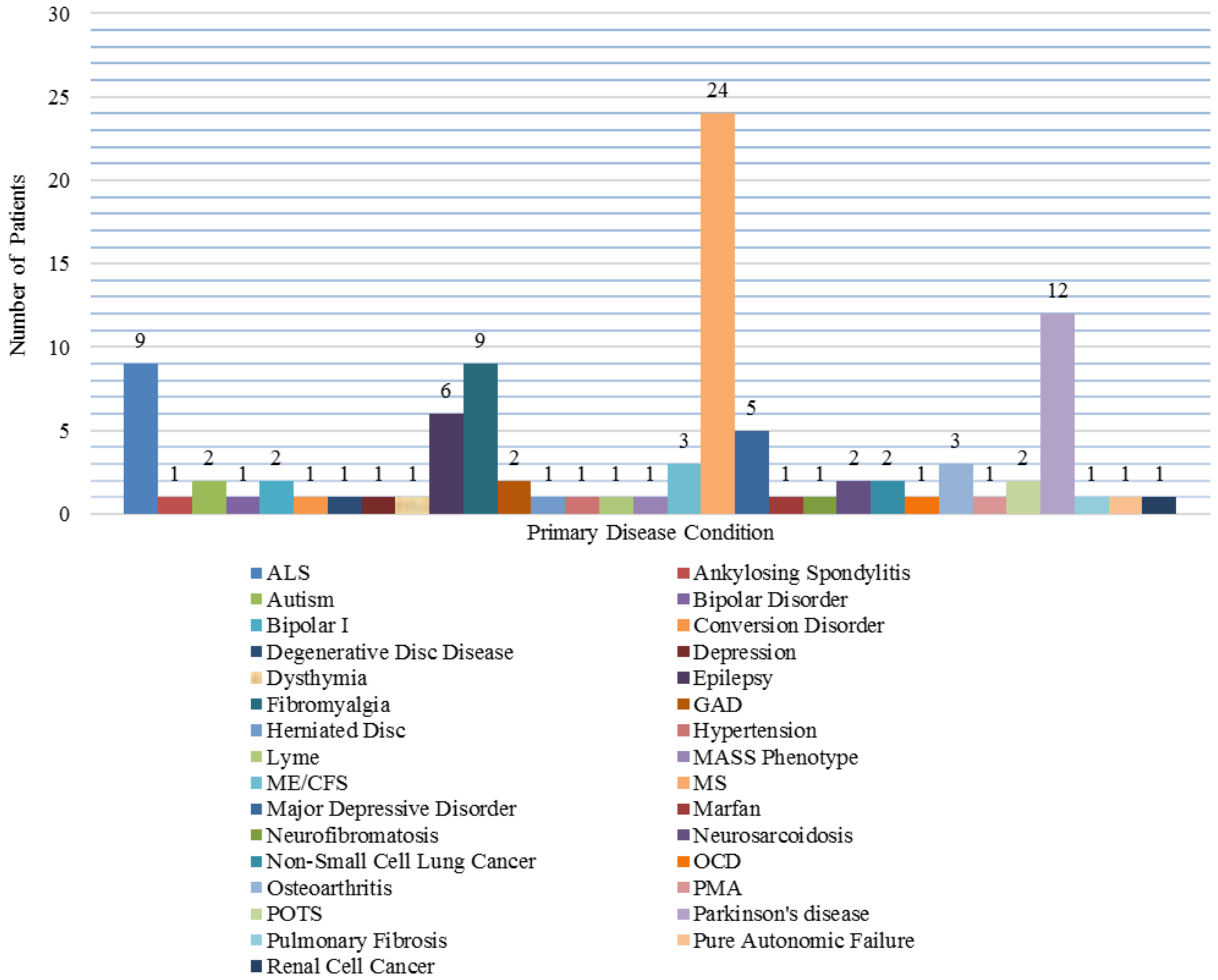


Fig 7: Number of patients suffer from a particular disease condition

There are many other decisions, can make from the result as shown in Table 1, i.e., the number of males and females for a certain disease condition. However, the Figure 8 shows this.

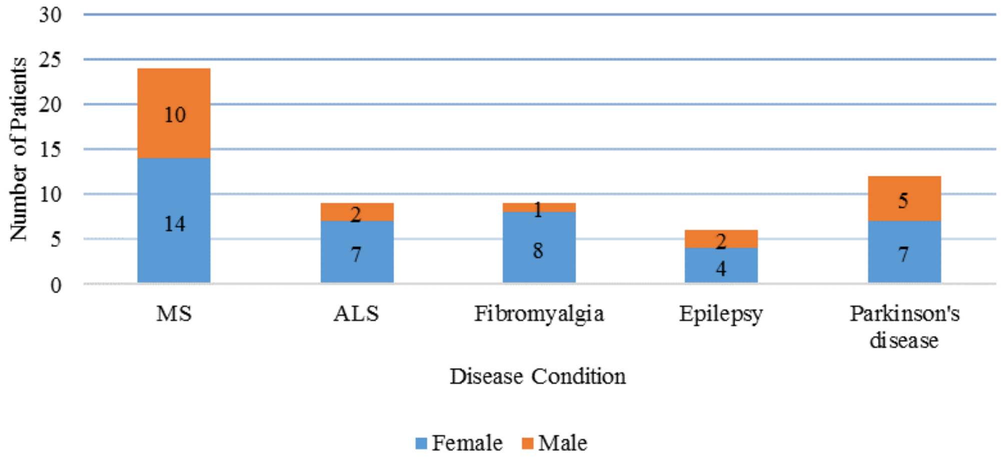


Fig 8: Number of males and females suffer from a particular disease condition

7. CONCLUSION AND FUTURE WORK

The more data will increase in Big Data universe, the more potentialities will be created; hence requires more analysis. In this paper, the most prominent applications that are making our life easier has been mentioned. Getting actionable insights from healthcare data is the most leading sectors of those.

Since doctors, pharmacists, researchers, and health conscious people are now sharing their opinions and experiences through blogs, social media, or journals. So we can find useful information from these sources. This helps to make realization on good health as well as reducing costs. Our approach in this paper showed this, which will create a new dimension on health.

The discussion is on the most possible outcomes of big data applications as well as how to apply these outcomes in health sector by making decisions, raising awareness and reducing costs. Although, the potentiality of healthcare is vast, so lot of scopes may arise to work in future in this topic, including:

- Cluster or classify diseases based on the similar symptoms and specify the possibility of occurring of one disease due to another, e.g., if a child is affected by common cold, there may be possibility to be affected by mumps, or fever.
- Analysis of large and complex health data (e.g., genomics, cancer, drugs, etc.) to extract more information for better health concern.
- Advanced filtering system can be introduced in place of existing one.
- Make efficient system log.
- Apply our methodology in other Big Data applications such as agriculture, weather forecasting, travel and transportation, etc.

8. ACKNOWLEDGMENTS

Special thanks to Dr. Mohammed Nasir Uddin, Chairman and Associate Professor and Dr. Uzzal Kumar Acharjee, Associate Professor, Department of CSE at Jagannath University for their continuous inspiration.

9. REFERENCES

[1] J. Han, M. Kamber and J. Pie,"Introduction," in Data Mining: Concepts and Techniques, 3rd ed., New York, Elsevier Inc., 2012, pp. 1-38.

[2] B. Nedelcu, "About Big Data and its Challenges and Benefits in Manufacturing," Database Syst. J., vol. IV, no. 3, pp. 10-19, 2013.

[3] S. Sagiroglu and D. Sinanc, "Big Data: A Review," in Int. Conf. on Collaboration Technologies and Syst., San Diego, CA, 2013, pp. 42-47.

[4] Data Growth, Business Opportunities, and the IT Imperatives, http://www.emc.com/leadership/digital-universe/,[Accessed 8 November 2014].

[5] Big Data and analytics, http://www-01.ibm.com, [Accessed 8 November 2014].

[6] J. Yan, "Big Data, Bigger Opportunities," MeriTalk: The Government IT Network, Washington, D.C., 2013.

[7] Y. Koumpouros, "Big Data in Healthcare," in Healthcare Administration: Concepts, Methodologies, Tools, and Applications, Pennsylvania, IGI Global, 2015, ch. 2, pp. 23-46.

[8] P. Groves, B. Kayyali, D. K. Knott and Stev, " The big data revolution in healthcare: Accelerating value and innovation," MGI, San Francisco, 2013.

[9] H. V. d. S. Jr., "MIT Technology Review: Data-Driven Healthcare," MIT Technology, Cambridge, 2014.

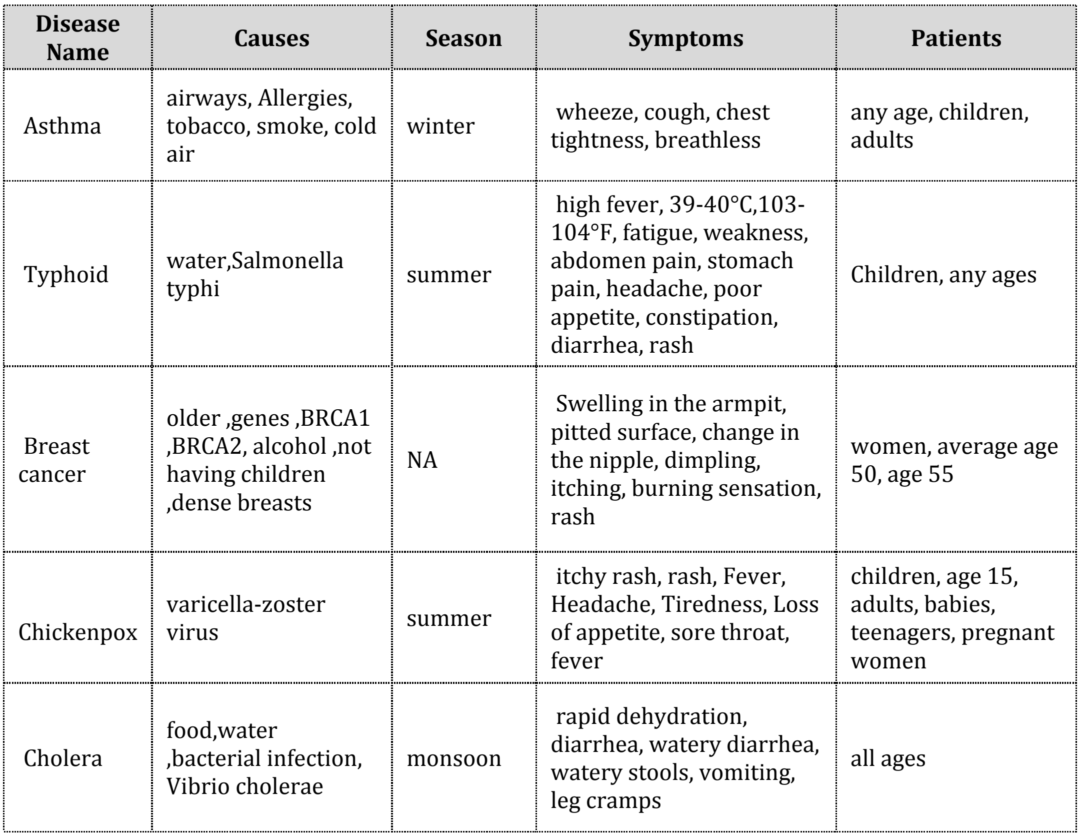
[10] Five Reasons Healthcare Data Is Unique and Difficult to Measure,HealthCatalyst, https://www.healthcatalyst.com, [Accessed 28 January 2015].

[11] Gate tools, https://gate.ac.uk/, [Accessed 28 January 2015].

[12] PatientsLikeMe, https://www.patientslikeme.com, [Accessed 5 May 2015].

[13] All diseases symptoms and medicines, http://www.nlm.nih.gov/medlineplus, [Accessed 5 May 2015].

10. APPENDIX



Disease Name	Causes	Season	Symptoms	Patients
Asthma	airways, Allergies, tobacco, smoke, cold air	winter	wheeze, cough, chest tightness, breathless	any age, children, adults
Typhoid	water,Salmonella typhi	summer	high fever, 39-40°C,103-104°F, fatigue, weakness, abdomen pain, stomach pain, headache, poor appetite, constipation, diarrhea, rash	Children, any ages
Breast cancer	older ,genes ,BRCA1 ,BRCA2, alcohol ,not having children ,dense breasts	NA	Swelling in the armpit, pitted surface, change in the nipple, dimpling, itching, burning sensation, rash	women, average age 50, age 55
Chickenpox	varicella-zoster virus	summer	itchy rash, rash, Fever, Headache, Tiredness, Loss of appetite, sore throat, fever	children, age 15, adults, babies, teenagers, pregnant women
Cholera	food,water ,bacterial infection, Vibrio cholerae	monsoon	rapid dehydration, diarrhea, watery diarrhea, watery stools, vomiting, leg cramps	all ages

Fig 6: Summary of information that has been extracted from experiment