Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of recent Epidemic and Pandemic Viruses

With the advancement of technology in data science and network technology, the world has stepped into the era of Big Data and medical field is rich in data suitable for analysis. Hence, in the recent years there has been much research in medical big data, mainly targeting data collection, data analysis and visualization. Big Data Analytics is a systematic approach for analyzing and identifying different patterns, relations and trends within a large volume of data. In this paper, we will apply the Big Data Analytics to recent Epidemic and Pandemic disease data such as - COVID19, SARS, EBOLA, H1N1 and MERS. We will be using data mining techniques for exploratory data analysis to generate visualization and trends prediction. We will be comparing each of these diseases based on death toll, countries affected, number of cases and Mortality rates. From the results which are going to be obtained, we will be able to predict the likelihood of these diseases, effectively take precautionary measures in the future and also optimize the decision making process.

SECTION - I

Introduction

Continuous developments in medical field, along with the advent of big data Era, has crated rapid and expensive growth of medical information.As we know, Big data consists of 4 V's: Volume, Velocity, Variety and Veracity. In a statistic cited by DELL EMC, it is estimated that 2.5 Quintillion Bytes of data is generated each day[1]. The velocity of medical data is incredible, with health monitoring data generating every second. In terms of Variety, the medical field generates from many different sources such as, magnetic resonance imaging, health monitoring data, genome, and X-Ray data. For Veracity, the medical data maybe incomplete, biased or even filled with noise, preventing the necessary insights. In our example, large amount of datasets for the diseases have been collected, giving report from each region for each day. From the visualizations generated we can see how the diseases spread worldwide in such a short time.

To analyze the data, data preprocessing, data modeling, data visualization, and are needed. Ambiguous information, repetition, noise, and ultrahigh dimensions influence the medical data. Therefore, it is necessary to preprocess the data. In medical big data preprocessing, one must extra, transform, and load (ETL) the data, integrate multi-data sources; and unify the data model. Typical algorithms and tools based on the existing big data platform are mainly designed to make data analysis more convenient and effective, thus making data models accessible and repeatable. Medical big data visualization is the most effective tool when dealing with complex medical data and growing medical needs, and includes information visualization, interaction techniques and

architectures, modeling techniques, multi-resolution methods, visualization algorithms and techniques, and volume visualization, as well as other customized analytics. Medical big data is mainly used in clinical data, monitoring and early warning of chronic diseases, daily activities, and physical characteristic index detection and collection.

As one of the fundamental techniques of BDA, data mining is an innovative, interdisciplinary, and growing research area, which can build paradigms and techniques across various fields for deducing useful information and hidden patterns from data [2]. Data mining is useful in not only the discovery of new knowledge or phenomena but also for enhancing our understanding of known ones. With the support of such techniques, BDA can help us easily identify crime patterns which occur in a particular area and how they are related with time. The implications of machine learning and statistical techniques on crime or other big data applications such as traffic accidents or time series data, will enable the analysis, extraction and understanding of associated patterns and trends, ultimately assisting in crime prevention and management.

In this paper, big data analytics algorithms are utilized for mining of data from the 5 wide spread diseases. After preprocessing, including data filtering and normalization, various visualizations are implemented to produce statistical results. Major contribution of this paper can be summarized as follows:

1. A series of investigative explorations are conducted to explore and explain the data of diseases worldwide.
2. visual representation which is capable of handling large datasets and enables users to explore, compare, and analyze evolutionary trends and patterns of diseases.

SECTION - II

A. Related Work

Big data analytics (BDA) has been extensively applied and studied in the fields of data science and computer science for quite some time. While machine learning provide both opportunities and challenges when meets large amount of big data [9]. Raghupathi and Raghupathi [10] described the promise and potential of BDA in healthcare and the government. Archana and Anita [11] conducted a survey on applications of BDA in healthcare and the government. Lodhe and Rao [12] presented various software frameworks available for BDA and discussed some widely used data mining algorithms. Chinmayee, Manjusha and Siddarth [13] have used BDA approach to determine the dynamic of infectious diseases. They have even proposed a model which is used to predict the intensity of the disease and also the preventive measures. Fisher et

al. [14] explained the conception of big data in BDA, its analytics and the associated challenges when interacting among them.

B. ***Healthcare Data Mining, Visualization & Trends Forecasting***

Big data related to healthcare have the characteristics of variety, volume, velocity and in particular veracity. “Big data” in the healthcare industry comprises of data related to patient healthcare and well-being. A wide category of clinical data is involved. Few of them are the data from Computerized Physician Order Entry, clinical decision support systems like physician’s prescriptions, pharmaceuticals, laboratory, radiology, insurance and other organizational data, Patients’ records stored electronically, tweets, blogs, facebook status of social media web posts, publishing case studies in medical journals and so on. Data analytics in health care is broadly classified into Bioinformatics, Neuroinformatics, Clinical Informatics, Public Health Informatics, and Translational Bioinformatics [15]. Each of the field generates voluminous data which pronounces the necessity of big data management technology.

A report published by the McKinsey Global Institute [16], identified that big data analytics can be greatly used to improvise the healthcare sector by effective cost management, improved, timely and effective treatment. The three broad categories in healthcare that require data analytics include: 1) Clinical operations: Comparative Effectiveness Research can be used to provide optimal treatment to specific patients. Clinical decision support systems could be used to alert errors in events or reactions and to suggest treatment options to physicians based on image and signal analysis. Publishing results of data analytics on clinical operations can help patients to take informed decisions on choice of healthcare providers for effective service. It also enables remote patient monitoring and proactive care by analysing patient profiles. 2) Research & development: Predictive modelling enables greater and faster production of drugs and increases the rate of therapeutic success. Statistical tools and algorithms improve clinical trial design reducing the risk of trial failures and accelerating new products into the market. Analysing clinical trials data identifies adverse effects of products before they reach the market. R&D has greatly helped in personalized medicine offering early detection, effective diagnosis of disease and effective treatment. 3) Public health: Big data analytics can help predict/detect and prevent the outbreak of infectious diseases. Healthcare providers would be better prepared with required vaccines and medicines for any such diseases or outbreaks. This also leads to awareness among the public on health risks involved in these diseases and hence would be less probable to contracting infections.

C. ***Background***

The different ways a disease might be described based on the disease prevalence, incidence, and the known or unknown disease pathways are as follows[3].

1. *Sporadic* - Diseases which occurs infrequently or irregularly.
2. *Cluster* - Diseases that occurs in larger numbers even though the actual number or cause may be uncertain.
3. *Endemic* - Constant presence and/or usual prevalence of a disease in a geographic population
4. *Hyper-Endemic* - Persistent, high levels of disease well above what is seen in other populations.
5. *Epidemic* - Sudden increase in the number of cases of a disease above what is normally expected.
6. *Outbreak* - Same as Epidemic but is limited to a particular geographic event.
7. *Pandemic* - Epidemic that has spread over several countries or continents, usually affecting a large number of people.

In our study, we only consider the epidemic and pandemic diseases which is caused due to virus. Most common symptoms include high fever, dry cough, sore throat and tiredness. The viral infection can last from several days to two weeks. EBOLA, MERS and SARS are classified as epidemic diseases since they had only affected a particular geographical area. Whereas, H1N1 and COVID-19 have been classified as Pandemic by the World Health Organization(WHO) as they affected everyone around the world, killing thousands of them.

Severe Acute Respiratory Syndrome coronavirus (SARS-CoV) – virus identified in 2003. SARS-CoV is thought to be an animal virus from an as-yet-uncertain animal reservoir, perhaps bats, that spread to other animals (civet cats) and first infected humans in the Guangdong province of southern China in 2002. An epidemic of SARS affected 26 countries and resulted in more than 8000 cases in 2003. Since then, a small number of cases have occurred as a result of laboratory accidents or, possibly, through animal-to-human transmission. Transmission of SARS-CoV is primarily from person to person. It appears to have occurred mainly during the second week of illness, which corresponds to the peak of virus excretion in respiratory secretions and stool, and when cases with severe disease start to deteriorate clinically. Most cases of human-to-human transmission occurred in the health care setting, in the absence of adequate infection control precautions. Implementation of appropriate infection control practices brought the global outbreak to an end.

In the spring of 2009, a novel influenza A (H1N1) virus emerged. It was detected first in the United States and spread quickly across the United States and the world. This new H1N1 virus contained a unique combination of influenza genes not previously identified in animals or people. This virus was designated as influenza A (H1N1)pdm09 virus.

Since 2012, Middle East respiratory syndrome (MERS) coronavirus has infected 2,442 persons worldwide. Case-based data analysis suggests that since 2016, as many as 1,465 cases and 293–520 deaths might have been averted. Efforts to reduce the global MERS threat are working, but countries must maintain vigilance to prevent further infections [17]. MERS-CoV is a zoonotic virus, which means it is a virus that is transmitted between animals and people. Studies have shown that humans are infected through direct or indirect contact with infected dromedary camels. MERS-CoV has been identified in dromedaries in several countries in the Middle East, Africa and South Asia.

The Ebola virus causes an acute, serious illness which is often fatal if untreated. EVD first appeared in 1976 in 2 simultaneous outbreaks, one in what is now Nzara, South Sudan, and the other in Yambuku, DRC. The latter occurred in a village near the Ebola River, from which the disease takes its name. The 2014–2016 outbreak in West Africa was the largest Ebola outbreak since the virus was first discovered in 1976. The outbreak started in Guinea and then moved across land borders to Sierra Leone and Liberia. The current 2018-2019 outbreak in eastern DRC is highly complex, with insecurity adversely affecting public health response activities.

Currently in 2020, Coronavirus is a family of viruses that can cause illness, which can vary from common cold and cough to sometimes more severe disease. Middle East Respiratory Syndrome (MERS-CoV) and Severe Acute Respiratory Syndrome (SARS-CoV) were such severe cases with the world already has faced.

SARS-CoV-2 (n-coronavirus) is the new virus of the coronavirus family, which first discovered in 2019, which has not been identified in humans before. It is a contiguous virus which started from Wuhan in December 2019. Which later declared as Pandemic by WHO due to high rate spreads throughout the world. Currently (on date 30 April 2020), this leads to a total of 120K+ Deaths across the globe, including 83+ deaths alone in Europe. This pandemic is spreading all over the world; it becomes more important to understand about this spread.

SECTION - III

Data Analysis and Visualization

The 5 datasets which we use are publicly available, which cover all the countries. The Ebola virus dataset contains 28,616 cases from 01/02/2014 to 12/31/2016 [4]. Data from MERS has 5670 cases dating back from 01/03/2019 to 05/07/2019 [5]. In the COVID-19 dataset, 6724149 cases from 01/31/2020 to 05/03/2020 [6]. H1N1 pandemic dataset contains 6724149 cases from 04/12/2009 to 04/10/2010 [7] and SARS dataset contains 8432 cases from 11/16/2002 to 07/31/2003 [8].

A. Feature Attributes

For each entry of confirmed cases in the datasets, following featured attributes are included:

1. *Date* - date and timestamp of each case.
2. *Country* - case belonging to a particular country.
3. *Cases* - Number of confirmed cases for a particular date
4. *Deaths* - Number of Deaths for the specified date
5. *Recovered* - Number of Recovered Cases.
6. *Lat* - Latitude of the location
7. *Long* - Longitude of the location

B. Data Preprocessing

Before implementing any algorithm on our datasets, a series of preprocessing steps are performed for data conditioning as presented below:

1. For some missing values in EBOLA dataset, we imputed random values sampled from the non-missing values, computed their mean, and then replaced the missing ones.
2. We also omit some features that unneeded like lat and long.
3. Time is discretized into a couple of columns to allow for time series forecasting for the overall trend within the data.

C. *Narrative Visualization*

Considering the geographic nature of virus spread, an interactive map called choropleth was used for data visualization, where the confirmed cases are represented according to their spatial

variation. As shown in Fig. 1, as the intensity of color red increases, it means more number of cases were detected from that country. Also, hovering on a particular country displays the number of confirmed cases. By looking at these maps, we can see that COVID-19 and H1N1 are

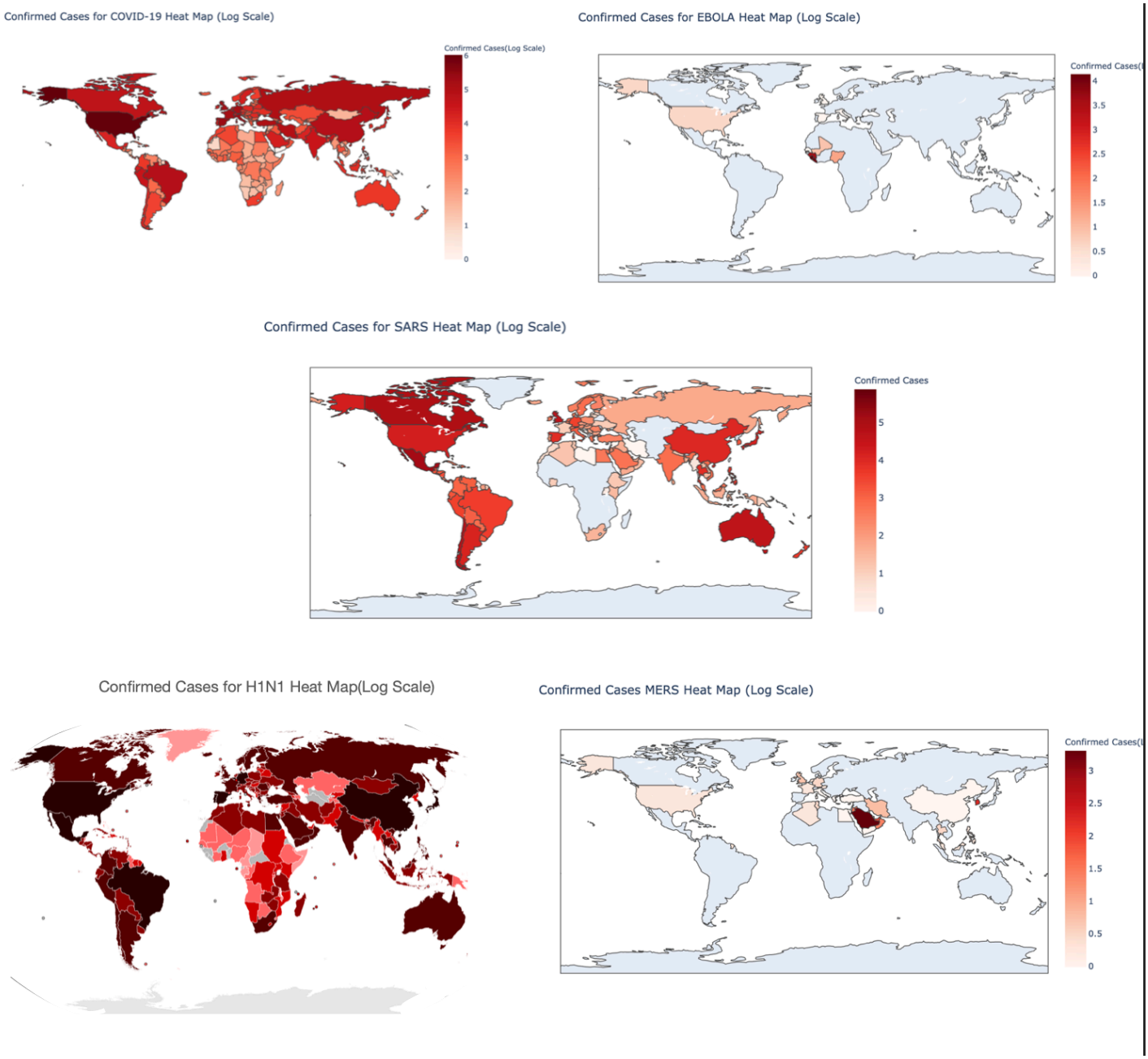

Fig. 1. Visualization of confirmed cases for virus spread world wide.

the two most widely spread diseases. whereas, EBOLA and MERS are limited only to limited countries.

Fig. 2 summarizes Number of cases, deaths and mortality rate of each disease. We see that though H1N1 is having more number of confirmed cases, its mortality rate is least. So, we can come to a conclusion that H1N1 is not very fatal and most of them can get cured. Although, about 7.14% of reported COVID-19 cases have died (as of May 2, 2020), it is having the highest number of deaths amongst all other diseases. In a contest that includes EBOLA and MERS, the COVID-19 death rates are closer to that of the seasonal flu. For EBOLA and MERS, out of 100 confirmed cases almost 40 and 34 people dies. So, this shows that though H1N1 and COVID-19 spreads vastly they are not very fatal, whereas EBOLA and MERS infected persons are more prone to death.

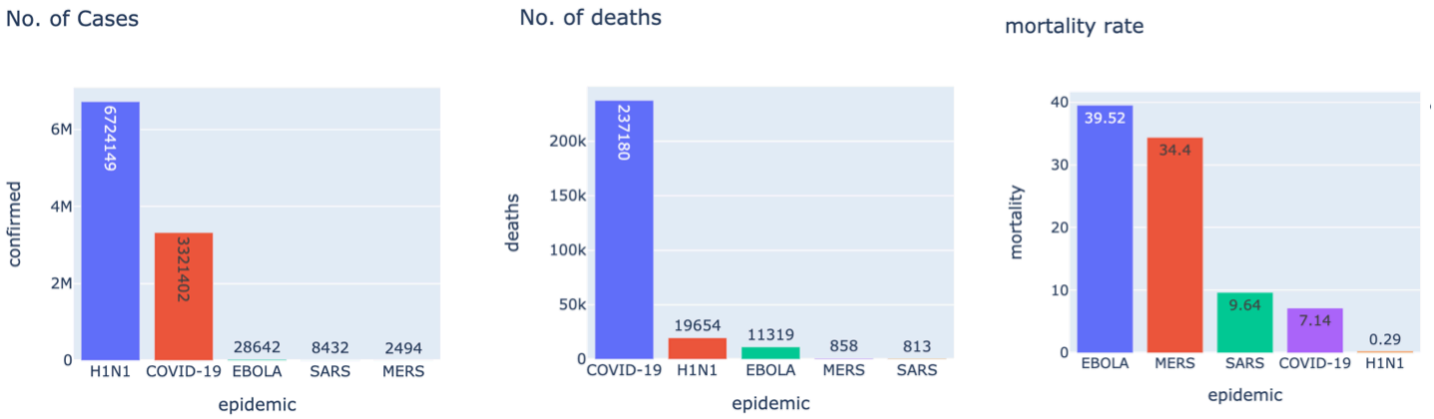

Fig. 2. Comparison of disease based on No. Of cases, deaths and mortality rate

In Fig. 3, we see the number of countries affected from each of the diseases. Almost 212 countries and territories have been affected so far and the count might also increase. H1N1 being second highest with 178 countries and EBOLA affected as least as 10 countries. The spread of virus is calculated based on a value called R0 (R Naught), it indicates how contagious an infectious disease is. R0 tells you the average number of people who will contract a contagious disease from one person with that disease. COVID is having an R0 of 5.6 where as H1N1 is having an R0 between 1.4 - 1.6, the existence of vaccines and antiviral drugs made the 2009 outbreak less deadly. For the SARS 2003 pandemic, scientists estimated the original R0 to be around 2.75. A month or two later, the effective R0 dropped below 1, this was due to tremendous effort that went into intervention strategies, including isolation and quarantine activities [18]. EBOLA and MERS are having an R0 of 1.75 and 0.6 respectively.

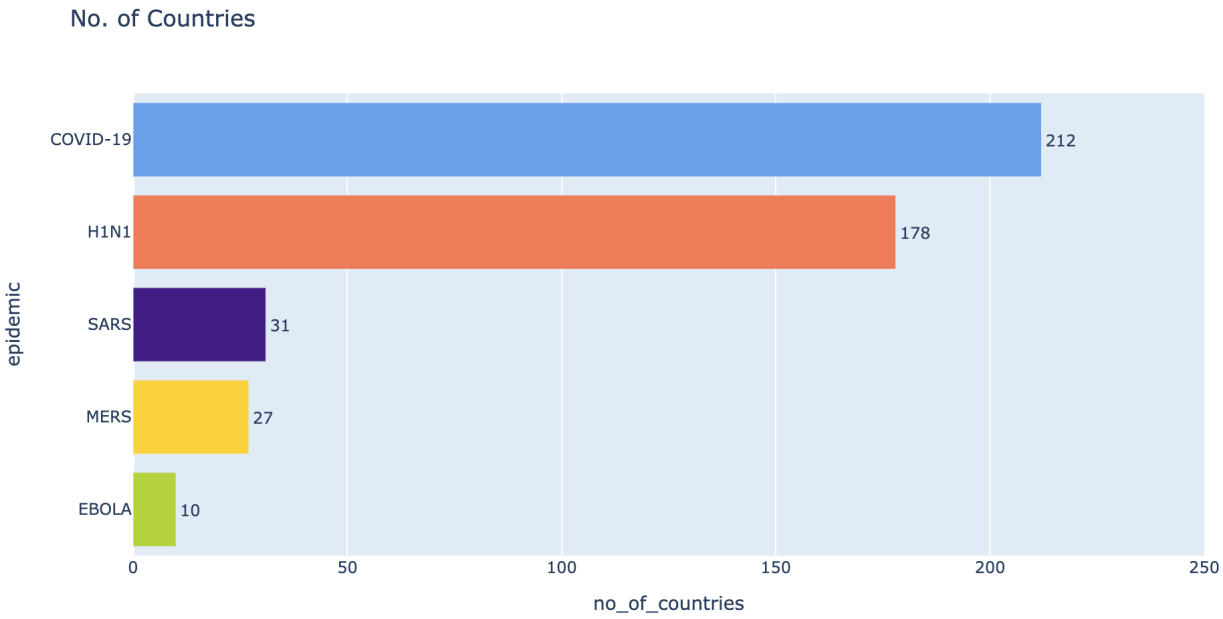

Fig. 3. Bar Graph depicting number of countries and territories affected worldwide

The table 1, gives the details of most affected age groups from CDC (Centre for Disease Control and Prevention). We see that young adults are most affected in H1N1 rather than older groups. The epidemiological data supports laboratory serology studies that indicate that older people may have pre-existing immunity to the novel H1N1 flu virus. This age distribution is very different from what is normally seen for seasonal flu, where older people are more heavily impacted [19]. Whereas, COVID -19 and SARS has affected the older population. This could be due to underlying chronic illness or younger immune system. Immunologists have identified some of the specific ways the immune system changes with age, allowing them to go beyond the simple assertion that it weakens. Janko Nikolich-Zugich an Immunologist from University of Arizona College of Medicine says "Older people are not as good at reacting to microorganisms they haven't encountered before" and he calls this " The Twilight of Immunity". Also, EBOLA and MERS have affected the medium aged groups.

Table 1. Most affected age groups

Epidemic	Age Groups (Yrs)
H1N1	5 - 24
COVID-19	> 80
SARS	> 64
EBOLA	25 - 49
MARS	50 - 64

SECTION - IV

Precautions and Cure

In an era of emerging and re-emerging communicable disease health threats, the importance of infection prevention and control measures in health-care settings should not be underestimated. Transmission of communicable disease/pathogen is an ever-evolving subject, and transmission of pathogens that cause acute respiratory diseases is no exception. The main mode of transmission of most ARDs are through droplets, but transmission through contact (including hand contamination followed by self-inoculation) and infectious respiratory aerosols of various sizes and at short range may also occur for some pathogens. Early recognition, isolation, reporting, and surveillance of episodes of ARD of potential concern are administrative control measures. From the two pandemics - H1N1 and COVID-19 and also for SARS, we can see that USA is the most affected country. One of the reasons are obesity, due to which the immunity decreases and therefore more prone to virus attacks. United States is currently topping the most countries with highest obesity rates. Other reasons as stated earlier can be because of weak immune system or also because of underlying chronic illnesses. Vaccines for EBOLA, H1N1 and SARS have been produced. whereas, vaccines for MERS and COVID-19 are still under development and testing phase. WHO recommends everybody to take the vaccines including 65 years and older [20].

SECTION - V

Conclusion

In this paper, data analytics and visualization techniques were utilized to analyze the spread of diseases worldwide. By using different visualization methods provided in the python libraries we were able to make many different conclusions. Though COVID-19 has spread enormously world wide, its not as fatal as EBOLA. We also saw the most affected country and also the precautions which have to be taken to minimize the spread of viruses. Additional results explained will help the users to compare the diseases and gain full understanding of the diseases.

SECTION - VI

References

[1] M. Zwolenski et al., “The digital universe: Rich data and the increasing value of the Internet of things,” Aust. J. Telecommun. Digital Econ., vo. 2, no 3, pp. 47.1–47.9, 2014.

[2] U. Thongsatapornwatana, ``A survey of data mining techniques for analyzing crime patterns," in Proc. 2nd Asian Conf. Defence Technol., Chiang Mai, Thailand, 2016, pp. 123128.

[3] “Difference Between an Epidemic and a Pandemic” https://www.verywellhealth.com/difference-between-epidemic-and-pandemic-2615168 Updated on March 17, 2020

[4] Ebola virus Data [Online]. Available: https://www.kaggle.com/imdevskp/ebola-outbreak-20142016-complete-dataset

[5] MERS Data [Online]. Available: https://www.kaggle.com/imdevskp/mers-outbreak-dataset-20122019

[6] COVID 19 data [Online]. Available: https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

[7] H1N1 Data [Online]. Available: https://www.kaggle.com/de5d5fe61fcaa6ad7a66/pandemic-2009-h1n1-swine-flu-influenza-a-dataset

[8] SARS Data [Online]. Available: https://www.kaggle.com/imdevskp/sars-outbreak-2003-complete-dataset

[9] L. Zhou, J. Wang, A. V. Vasilakos, and S. Pan, ``Machine learning on big data: Opportunities and challenges," Neurocomputing, vol. 237, pp. 350361, May 2017.

[10] W. Raghupathi and V. Raghupathi, ``Big data analytics in healthcare: Promise and potential," Health Inf. Sci. Syst., vol. 2, no. 1, pp. 110, Feb. 2014.

[11] J. Archenaa and E. A. M. Anita, ``A survey of big data analytics in healthcare and government," Procedia Comput. Sci., vol. 50, pp. 408413, Apr. 2015.

[12] A. Londhe and P. Rao, ``Platforms for big data analytics: Trend towards hybrid era," in Proc. Int. Conf. Energy, Commun., Data Anal. Soft Comput. (ICECDS), Chennai, 2017, pp. 32353238.

[13] Chinmayee Mohapatra, Manjusha pandey, Siddharth Swarup Rautray, “ Modeling and Dynamics of Infectious Disease: Big Data Analytics”, 2017 International Conference on Computer Communication and Informatics (ICCCI -2017), Jan. 05 – 07, 2017, Coimbatore, INDIA.

[14] D. Fisher, M. Czerwinski, S. Drucker, and R. DeLine, ``Interactions with big data analytics," Interactions, vol. 19, no. 3, pp. 5059, Jun. 2012.

[15] M. Herland, T. M. Khoshgoftaar and R. Wald, “A review of data mining using big data in health informatics”. Journal of Big Data, vol. 1:2, 2014.

[16] Big data: The next frontier for innovation, competition, and productivity, McKinsey Global Institute, June 2011.

[17] Donnelly, C. A., Malik, M. R., Elkholy, A., Cauchemez, S., & Van Kerkhove, M. D. (2019). Worldwide Reduction in MERS Cases and Deaths since 2016. *Emerging Infectious Diseases*, *25*(9), 1758-1760. https://dx.doi.org/10.3201/eid2509.190143.

[18] "How Scientists Quantify the Intensity of an Outbreak Like COVID-19” https://labblog.uofmhealth.org/rounds/how-scientists-quantify-intensity-of-an-outbreak-like-covid-19 updated on March 17, 2020 1:56 PM.

[19] “2009 H1N1 Early Outbreak and Disease Characteristics” https://www.cdc.gov/h1n1flu/surveillanceqa.htm Updated on October 27, 2009, 6:00 PM ET.

[20] “General Information About 2009 H1N1 Vaccines” https://www.cdc.gov/h1n1flu/vaccination/general.htm