MANAGING DATA ANALYSIS OF COVID-19

Syed Owais Bukhari
E-Mail Id: owais.bukhari@live.com
School of Engineering Science and Technology, Jamia Hamdard, New Delhi, India

Abstract- Covid 19 is a new animal in the zoo which has been commanded by homo sapiens for 200000 years. Data says that the battle against Covid 19 is going to be a long one. Intersectoral research is what is needed to amour sapiens to fight the invader. In this multifaceted research, the role of a data scientist is that of a defender. Innumerable reports, statistics and analysis of Covid 19 has generated a virtual data lake. The project lays down its prime focus on the various strategies to manage the Covid 19 data. In simple terms, machine learning tools like decision trees and clustering in addition to ensemble techniques can go a long way in creating a meaningful profile out of the repository of Covid 19 data.

Keywords: COVID-19, Machine learning, decision trees, clustering, K-means clustering, Supervised learning, training data, Data Lake, information gain, bootstrapping, random forest models.

1. INTRODUCTION

Covid 19 has waged a war against humanity. If every single individual commits a personalized fight against the disease, the end of the pandemic is not far. Being engaged in a niche of research which is directly related to data, the onus lies on us to generate a line of research which provides a lead in secondary research. With over 50 lakh cases and 3.5 lakh deaths, the numbers are enough to evoke the conscience of entire human race to fight Covid 19. An unending commitment to deeply probe the data analytics of Covid 19 is what can promise us a disease resilient future. This commitment serves as the prime motivation to drive the project.

2. TRANSACTIONS/ JOURNAL PAPER PREPARATION

Stage 1: This includes the collection of authentic and raw data from different sources which would range across various geographical regions around the globe. This would form the data repository which would serve as the basic infrastructure to work upon. After the herculean task of data collection has been completed, the task of data cleansing would begin. This process would involve a wide range of data tools so that redundant and less useful data is done away with and processed data is obtained. Techniques like K-means clustering would be of immense help. Supervised learning can be used at this stage which would learn from the training data of Covid 19 eg. Wuhan portfolio. This micro analysis can then be employed for comparative analysis across many geographic regions.

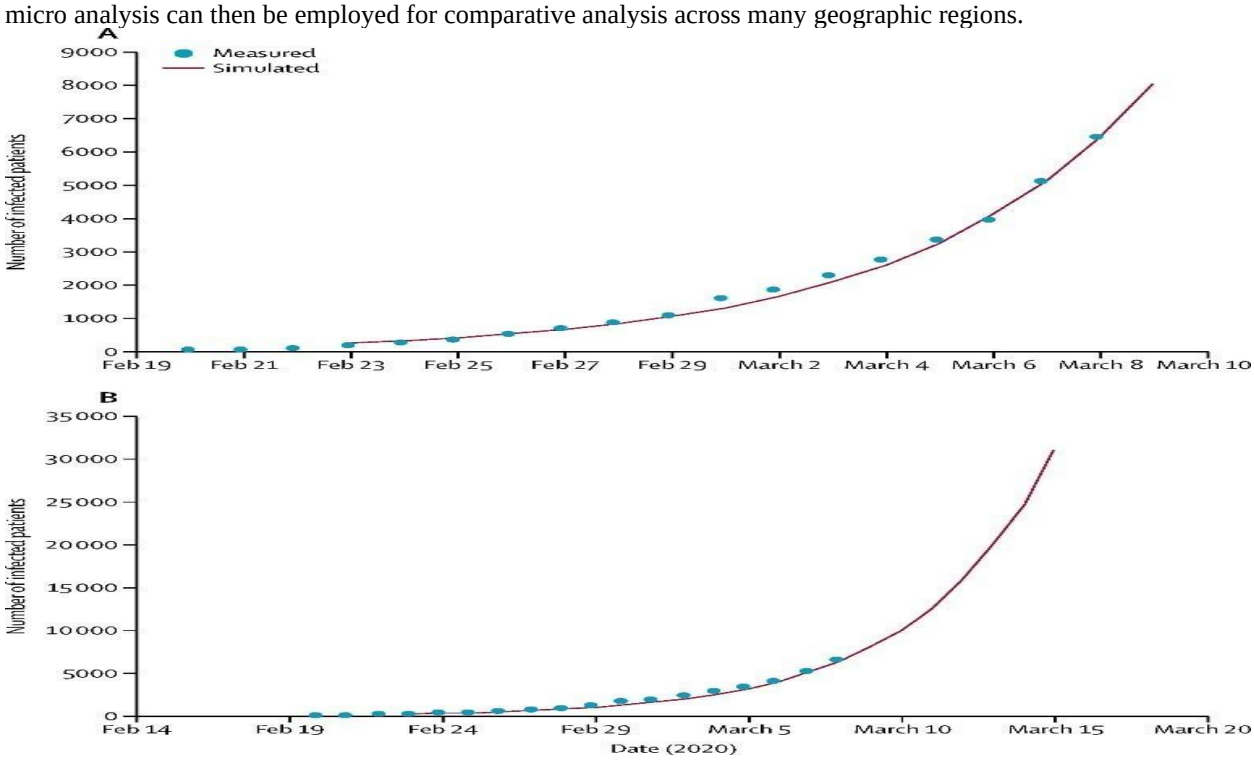

Fig. 2.1 Initial Sample of Covid Cases Used in Model

Stage 2: The data available at this stage would be in a more processed and clustered form. Similar data sets can now be taxonomically classified or segregated depending upon the attributes chosen. For instance, Grouping of infected

patients based on gender, age, climate, immunological response etc. To put it more specifically, this stage would form the 'Data Lake' on which different operations and algorithms can be employed. Algorithms like ID3 and C4.5 can prove handy on such processed data. The other goals of this stage include the reduction of entropy or disorder in the data sets and the maximization of overall information gain.

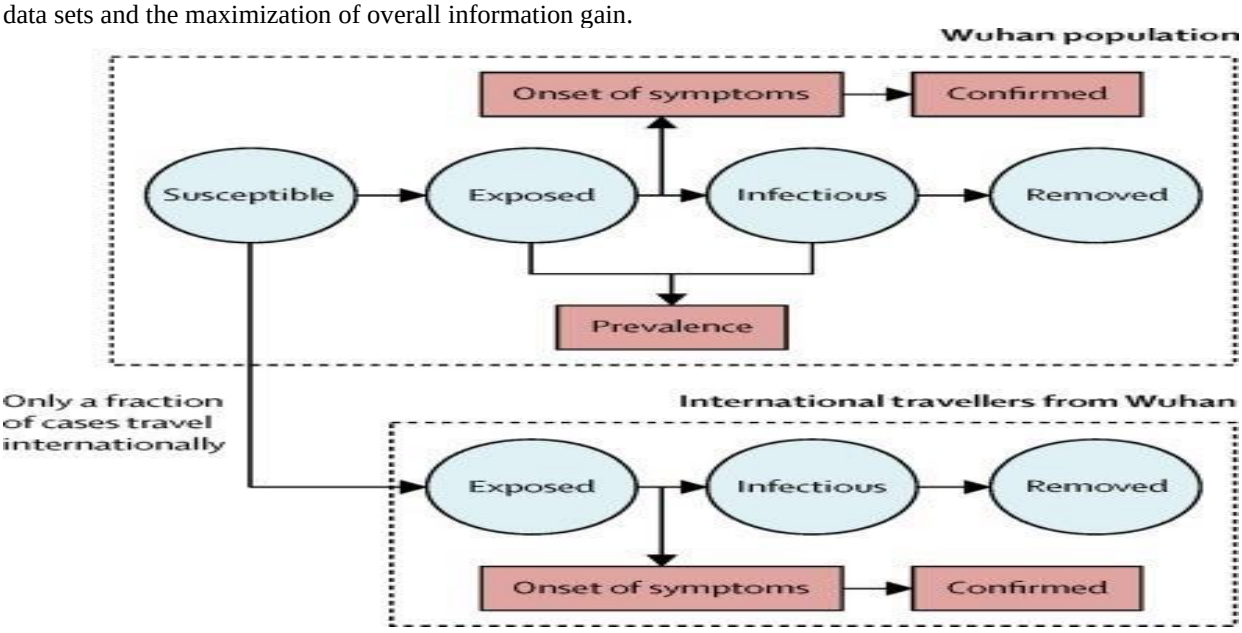

Fig. 2.2 Model Applied on Wuhan Portfolio

Stage 3: This is the most important stage as the research work here would culminate into visualization/output. The prime focus of this stage is to employ decision trees as well as ensemble modelling techniques like bootstrapping and random forest models so that we can deal with any amount of data which is generated relating to Covid 19. End products of this model may be summarized by visualization techniques like bar charts, graphs, pie charts, heat maps etc. In this way, a pool of data would be generated which can be easily accessed by various research agencies.

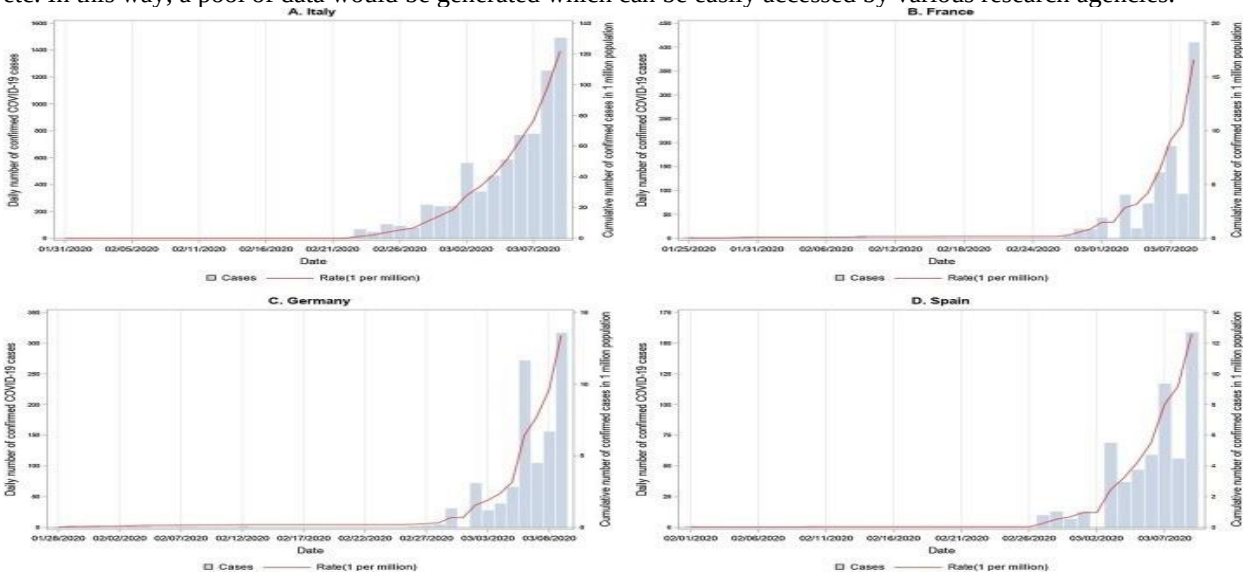

Fig. 2.3 Data Analysis Across Sample Nations

3. BENCHMARKS

3.1 Quantitative benchmarks used for analysis

- The Number of organizations that are relying on the data that we collected.
- Percentage of processed data that is being utilized by organizations.

4. RESEARCH OBJECTIVES/ GOALS

4.1 Short Term Goals

- Localized and region-specific prediction of infections and vulnerability of a population.
- Percentage of population that is responding to new drugs that are being administered.
- A repository of graphs based on various parameters to be formed which would retrieve data and predict trends.

4.2 Long Term Goals

- Accentuate the development of vaccine using data collected from clinical and human trials.
- Identify the lacuna in global health management systems to avert any impending pandemic in future.
- Boost the preparedness to fight diseases worldwide by harnessing the power of data.
- To encourage researchers to visualize the scope of data to deal with public health emergencies.

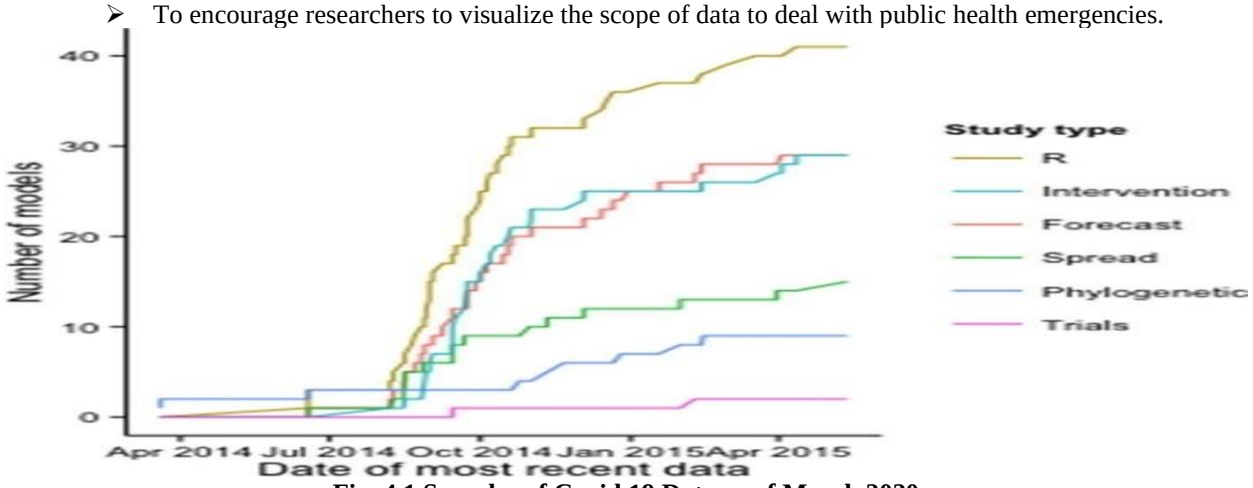

Fig. 4.1 Samples of Covid 19 Data as of March 2020

5. ANALYSIS AND FINDINGS

5.1 Measuring the Outcomes

The data processing model would be made available to various organizations involved in Covid 19 fight. An acknowledgment from various organizations that the data analysis is helping in vaccine development, clinical trials, public health system improvement etc. would push us to improve the model further. A feedback file would be developed to measure the outcomes in addition to model improvement.

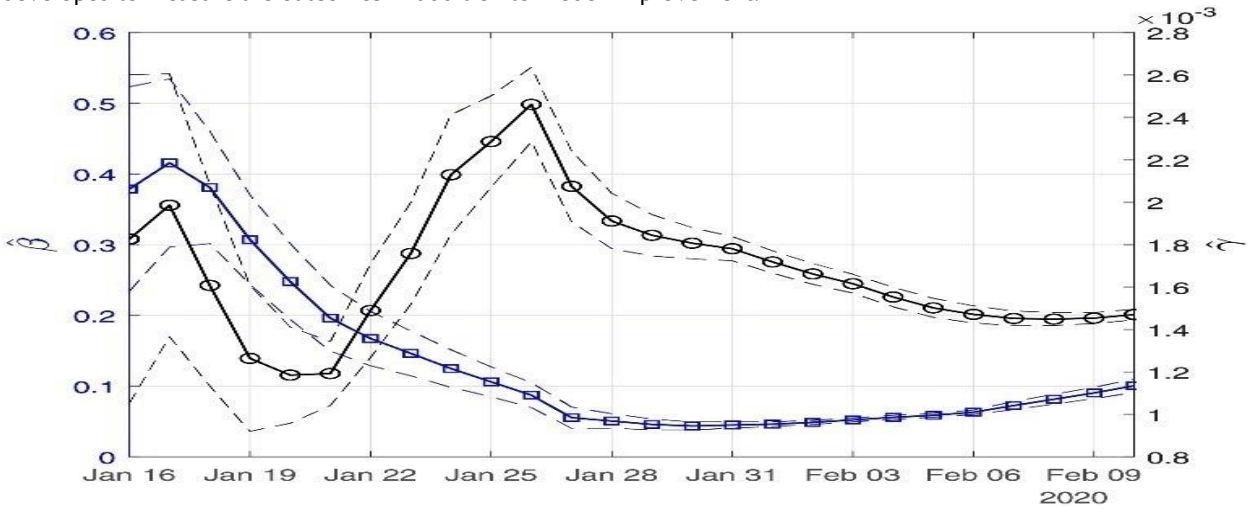

Fig. 5.1 Forecast Model Based on Sample Data

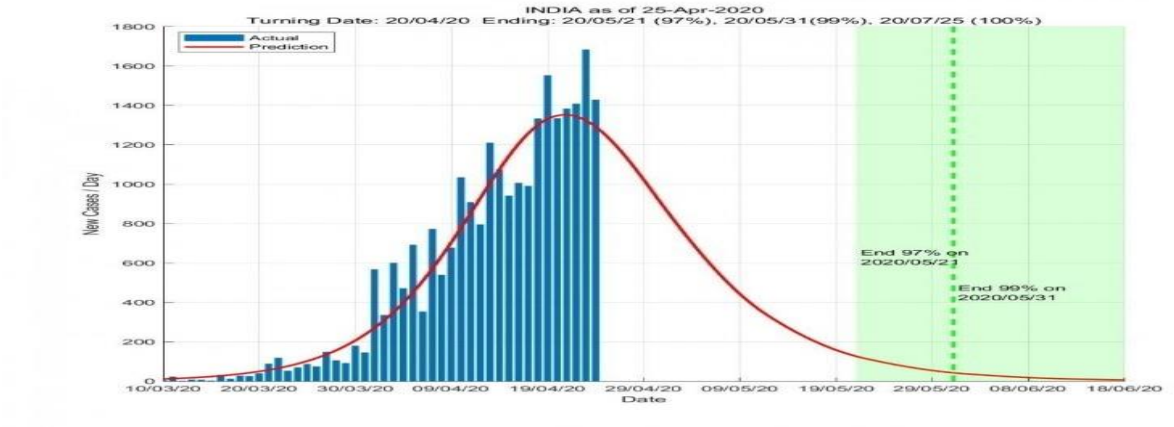

Fig. 5.2 Model Applied to Database Based on Indian Context

6. RECOMMENDATIONS AND DIRECTION FOR THE FURTHER RESEARCH

It Starting with a genre of localized research, the aim is to create an all-encompassing paradigm which would not only manage the current data streams but also those that may be generated in future.

No project can be deemed to have been completed unless it achieves the desired outcomes. The project "Managing data analysis of covid 19" is incomplete unless it achieves its short term and long-term goals. Let us first define the expected outcomes of the project: • Data mining of unexplored data which would provide path breaking interventions in the fight against Covid 19. • Development of an umbrella model which encompasses a vast majority of attributes and a major chunk of available databases. • Outcomes from data analysis would help in identifying the fault lines in our public health system and make us resilient to a future pandemic.

REFERENCES

[1] Woo PC, Huang Y, Lau SK, Yuen KY. Coronavirus genomics and bioinformatics analysis. Viruses,
[2] 2010; 2: 1804-20.
[3] Drexler, J.F., Gloza-Rausch, F., Glende, J., Corman, V.M., Muth,D., Goettsche, M., Seebens, A., Niedrig,
[4] M., Pfefferle, S., Yor-danov, S., Zhelyazkov, L., Hermanns, U., Vallo, P., Lukashev, A.,Muller, M.A., Deng, H., Herrler, G., Drosten, C., Genomic characterization of severe acute respiratory syndrome-related coronavirus in European bats and classification of coronaviruses based on partial RNA-dependent RNA polymerase gene sequences. J. Virol, 2010; 84: 11336–11349.
[5] Yin, Y., Wunderink, R. G. MERS, SARS and other coronaviruses as causes of pneumonia. Respirology,
[6] 2018; 23(2): 130-137.
[7] Peiris, J. S. M., Lai S. T., Poon L. et. al. Coronavirus as a possible cause of severe acute respiratory syndrome. The Lancet, 2003; 361(9366): 1319-1325.
[8] Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus AD,Fouchier RA. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. N. Engl. J. Med, 2012; 367: 1814–20.
[9] Seven days in medicine: 8-14 Jan 2020. BMJ, 2020;368-132.31948945.
[10] Imperial College London. Report 2: estimating the potential total number of novel coronavirus cases in Wuhan City, China. Jan. disease-analysis/news--wuhan-coronavirus, 2020.
[11] European Centre for Disease Prevention and Control data.Geographical distribution of 2019- nCov cases. Available online:
[12] (https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases) (accessed on 05 February 2020).
[13] World Helath Organization, nCoV Situation Report-22 on 12 February, 2020. source/coronaviruse/situation-reports/, 2019.
[14] Gralinski L.; Menachery V; Return of the Coronavirus: 2019- nCoV, Viruses, 2020; 12(2):135.
[15] Chen Z.; Zhang W.; Lu Y et. al.. From SARS-CoV to Wuhan 2019-nCoV Outbreak: Similarity of Early
[16] Epidemic and Prediction of Future Trends.: Cell Press, 2020.
[17] Luk H. K., Li X., Fung J., Lau S. K., Woo P. C. (Molecular epidemiology, evolution and phylogeny of SARS coronavirus. Infection, Genetics and Evolution, 2019; 71: 21-30.
[18] Coronavirinae in ViralZone. expasy.org/785 (accessed on 05 February 2019).
[19] Subissi, L.; Posthuma, C.C.; Collet, A.; Zevenhoven-Dobbe, J.C.; Gorbalenya, A.E.; Decroly, E.; Snijder, E.J.; Canard, B.; Imbert, I.One severe acute respiratory syndrome coronavirus protein complex integrates processive RNA polymerase and exonuclease activities. Proc. Natl. Acad. Sci. USA 2014, 111, E3900–E3909.
[20] Zhao L, Jha BK, Wu A, Elliott R, Ziebuhr J, Gorbalenya AE, Silverman RH, Weiss SR. Antagonism of the interferon-induced OAS-RNase L pathway by murine coronavirus ns2 protein is required for virus replication and liver pathology. Cell host & microbe, 2012; 11(6): 607–616.
[21] Randhawa GS, Soltysiak MP, El Roz H, deSouza CP, Hill KA, Kari L. Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study. Biorxiv; 2020.
[22] Dey SK, Rahman MM, Siddiqi UR,Howlader A. Analyzing the Epidemiological Outbreak of COVID-19: A Visual Exploratory Data Analysis (EDA) Approach. Journal of Medical Virology;2020.
[23] Xu Z, Shi L, Wang Y, Zhang J, Huang L,Zhang C, Tai Y. Pathological findings of COVID-19 associated with acute respiratory distress syndrome. The Lancet Respiratory Medicine; 2020.
[24] Sohrabi C, Alsafi Z, O'Neill N, Khan M, Kerwan A, Al-Jabir A, Agha R. World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19). International Journal of Surgery; 2020.
[25] Lu H, Stratton CW, Tang YW. Outbreak of Pneumonia of Unknown Etiology in Wuhan China: the Mystery and the Miracle. Journal of Medical Virology; 2020.
[26] Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, Yu T. Epidemiological and clinical
[27] characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China:A descriptive study. The Lancet. 2020; 395(10223):507-513.
[28] Bai Y, Yao L, Wei T, Tian F, Jin DY, Chen L, Wang M. Presumed asymptomatic carrier transmission of COVID-19. Jama; 2020.
[29] Santarpia JL, et al. Transmission Potential of SARS-CoV-2 in Viral Shedding Observed at the University of Nebraska Medical Center, med Rxiv 2020;03:

[30] https://doi.org/10.1101/2020.03.23.20 039446
[31] Velavan TP, Meyer CG. The COVID-19 epidemic. Trop Med Int Health. 2020;25(3): 278-280.
[32] https://www.cdc.gov/coronavirus/ 2019-ncov/lab/guidelines-clinical- specimens.htmlLAST [Accessed on 29 March 2020].
[33] COVID-19 Coronavirus Pandemic. Available:https://www.worldometers.info/coronavirus/ LAST [Accessed on 13 April 2020].
[34] Sajadi MM, Habibzadeh P, Vintzileos A, Shokouhi S, Miralles-Wilhelm F, Amoroso A. Temperature and latitude analysis to predict potential spread and seasonality for COVID-19; 2020. [ISSRN 3550308]
[35] Centers for Disease Control and Prevention. Frequently asked questions about SARS; 2005. Available:https://www.cdc.gov/sars/about/faq.html.
[36] World Health Organization. Middle East respiratory syndrome coronavirus (MERS-CoV). Available:https://www.who.int/emergencies/mers-cov/en.Google Scholar
[37] Mahase E. Coronavirus: Covid-19 has killed more people than SARS and MERS combined, despite lower case fatality rate; 2020.