Scientific Journal of Biology

Research Article

Significance of Mutation Rate of Structural and Non-Structural Proteins of SARS-Cov-2 Showed with Lower Death Rate of COVID-19

Saakshi Jalali[1], Bhaskar Bhadra[1]* and Santanu Dasgupta[1]

[1]*Synthetic Biology Group, Reliance Corporate Park, Reliance Industries Limited, Navi Mumbai 400701, India*

***Address for Correspondence:** Bhaskar Bhadra, Synthetic Biology Group, Reliance Corporate Park, Reliance Industries Limited, Navi Mumbai 400701, India, Tel: 91-22-4475-0885; E-mail: Bhaskar.Bhadra@ril.com

Submitted: 29 August 2020; **Approved:** 02 September 2020; **Published:** 04 September 2020

Cite this article: Jalali S, Bhadra B, Dasgupta S. *Significance of Mutation Rate of Structural and Non-Structural Proteins of SARS-Cov-2 Showed with Lower Death Rate of COVID-19*. Sci J Biol. 2020;3(1): 017-022. https://dx.doi.org/10.37871/sjb.id18

Abstract

The outbreak of the infectious and rapidly expanding coronavirus disease 19 (COVID-19) caused by the SARS-CoV-2 virus has led to a devastating effect on public health and the global economy. The daily country-wise updates from the World Health Organization on the number of infected cases and death rates show diverse statistics. In this study, the mutation rate of six proteins of SARS-CoV-2 over the last four months (April – July 2020) was correlated with the death rate utilizing 7200 genome sequences from various countries. From our findings, we suggest a significant correlation between the mutation rates of NSP6 and Surface glycoprotein with the death rate. Additionally, a cumulative mutation rate of these two proteins with the death rate of three major cluster countries enabled us to hypothesize that mutations of these 2 proteins would grow consistently while the death rate would drop below 0.5% by end of 2020 in cluster I countries. Hence, we propose that with the current mutation rate trend, the COVID-19 death rate would significantly weaken by the end of this year. In addition, decrease in death rate and infectivity is also depends on the prevention measures and therapeutics approaches adopted in different regions.

Keywords: COVID-19; SARS-CoV-2; Virus mutation; NSP6; S glycoprotein; Mutation rate

INTRODUCTION

The COVID-19 pandemic caused by Severe Acute Respiratory Syndrome Virus (SARS-CoV-2) is a beta-coronavirus containing 29,903 nucleotides-long single-stranded RNA genome. After the first report at Wuhan, China, in December 2019, the disease was spread over 230 countries within a couple of months. It was believed that, like other beta-coronaviruses, SARS-CoV-2 is a human zoonotic pathogen sharing almost 96% similarity with BatCoV RaTG13 which was found in horseshoe bats [1]. Over the last couple of decades, viral zoonotic infection [2] caused by avian influenza, chikungunya virus, nipa virus, hantaviruses, SARS-CoV-1, MERS (middle east respiratory syndrome), etc have significantly impacted human health and economy. It is to be believed that zoonosis viruses evolve through mutations in the viral genome, altering the structure of surface proteins thereby leading to a change in host specificity.

In the past century, diseases which were acquired from animals wreaked havoc, e.g. 'Spanish flu' and 'Hong Kong flu' causing 50-60 million deaths all over the world. In current times, urbanization, increased global travel, changes in land use, and changes in the environment [3,4] has led to the likelihood and spread of infectious diseases. From the past, it can be determined that pandemics like influenza in 1918, 1957, and 1968 years slowly subsided either via acquiring heard-immunity or by loss of infectivity of the virions, by accumulating a cluster of mutations after several months of replications in the population [5]. Viruses have smaller genome size and therefore, are prone to accumulate mutations at a much higher rate when compared to other organisms with larger genome sizes. It has been shown that spontaneous mutation rate varies among viruses compared to DNA viruses, single-stranded RNA viruses (ssRNA) mutate faster. The mutation rates of ssRNA viruses are influenced by polymerase fidelity, proofreading, secondary structure, replication mechanisms, sequence, and access to post-replicative repair systems [6].

Comparative analysis based on more than 7000 genomes of SARS-CoV-2 indicated that the virus has a very high mutation rate for various structural genes [7]. An accurate estimation of the mutation rate of the viral genome is often considered complex [8]. The viral genome mutations are represented as substitutions per nucleotide per cell infection (s/n/c), and the rate of mutation ranges from 10^{-8} – 10^{-6} s/n/c for DNA viruses whereas the mutation rate for RNA viruses ranges from 10^{-6} – 10^{-4} s/n/c. Therefore, it could be argued that after several rounds of multiplication across a wide range of population zoonotic RNA viruses, such as SARS-CoV-2 may accumulate a wide range of mutations. Optimistically, we could believe that some of those mutations could result in loss of infectivity. In this study, we have calculated the mutation rate of specific proteins of SARS-CoV-2 over the last four months (April-July 2020) and correlated with the death rate caused by COVID-19 across various countries. We were able to show a significant correlation between the mutation rates of a couple of proteins with the death rate. We have also proposed that with the calculated rate of mutation, the COVID-19 pandemic may be significantly weakened by the end of this year. However, the factors such as epidemic management strategies to curb community spread and discovery of novel drugs and treatment protocol could significantly impact this prediction.

MATERIALS AND METHODS

Genes and amino acid sequences

For the current study, four structural and two nonstructural proteins of the SARS-CoV-2 were selected. The full-length protein sequences of the six SARS-CoV-2 proteins were retrieved from the NCBI GenBank and the downloaded data were arranged month-wise for each of the proteins. A total of 41,304 SARS-CoV-2 protein sequences isolated from 49 different countries, deposited in the NCBI as of 31st July 2020 were utilized. Table 1 lists the number of protein sequences of the SARS-CoV-2 virus used in the current study.

Calculation of mutation rate

By comparing the protein sequences of six SARS-CoV-2 structural and nonstructural proteins base pair differences were identified. Briefly, Multiple Sequence Alignment (MSA) of each of the amino acid sequences of SARS-CoV-2 proteins with their respective reference sequences of SARS-CoV-2_Wuhan (accession numbers YP_009725302, YP_009724396, YP_009724390, YP_009724393, YP_009724397, and YP_009724392) was performed using CLUSTALW with default parameters. The multiple sequence

Table 1: Total number of SARS-Cov2 genomes sequenced from March-July 2020 which were used as a resource for the studies.

Month	Country names	Total number of sequences
March	Australia, China, India, USA + 36 countries	3357
April	Australia, Bangladesh, China, India, USA +12 countries	2181
May	Australia, Bangladesh, China, India, USA + 6 countries	635
June	Australia, Bangladesh, China, India, USA + 3 countries	668
July	Australia, Bangladesh, USA	43

alignment helped in constructing an aligned view of query sequences based on an evolutionary relationship. Using the recorded variations corresponding to the reference amino acid residues a mutation profile was generated. Hence, an elaborative month-wise mutation profile for all the six proteins of SARS-CoV-2 was created. The mutation rate was calculated using the formulae mentioned below.

Mutation Rate = (Number of unique mutations / Number of SARS-Cov2 sequences) X 1000

Calculation of death rate

To analyze the death rate month-wise data for total cases and deaths over the period from March to July 2020 was derived from a web platform [9]. The month-wise death rate was calculated by simply dividing the number of deaths in a month by the number of cases reported for the month. The complete calculated death rate is presented in table 2.

Table 2: Covid-19 cases and deaths rate tabulated for the months March-July 2020.

Months (2020)	Cumulative data of 49 countries of		
	Number of cases	Number of Deaths	Death Rate/ 1000 cases
March	687327	34604	50.35
April	1490791	99428	66.69
May	2507683	173926	69.36
June	3434107	149320	43.48
July	4746206	155342	32.73

Death Rate = (Number of deaths/ Number of COVID-19 cases reported). X 1000.

Statistical analysis

A correlation coefficient between the mutation rate and the global death rate for the six proteins was calculated using Pearson's correlation method. Also, regression statistical analysis was performed on selected proteins using the ANNOVA test.

RESULTS AND DISCUSSIONS

Selection of proteins for the study

The viral RNA once inside the host cell encodes for both structural and nonstructural proteins. The structural proteins play an important role in host recognition and infection, whereas several non-structural proteins aid the process of replication and assembly of the virion to enhance infection efficiency. In the present study, four structural and two non-structural proteins were selected from 7195 genomes sequenced across different geographical regions and analyzed for mutational changes for over the last four months (March-June 2020).

NSP6: The Non-Structural Protein 6 (NSP6) is a putative transmembrane protein translated from the largest gene (ORF1ab polyprotein) of SARS-CoV-2, which is predicted to act as membrane anchor during assembly of the viral replication complexes [9]. The 290 amino acid residues long sequence of NSP6 protein, produced by both pp1a (polyproteins short) and pp1ab (polyproteins long) was downloaded from NCBI (YP_009725302). The mutation rate of NSP6 protein was observed to be 11.6 in March followed with a gradual dip in May (4.75) and then rise in June (14.97). This was one of the two proteins where we noticed a higher mutation rate of 46.5 in July when compared to the previous month. The number of unique mutations identified for the last 4 months was as follows April (23), May (3), June (10), and July (2). Previous reports have stated a significant role of L37F missense mutations in viral infection [10], in our analysis frequency of L37F mutation across all months was observed to be higher.

ORF8 protein (ORF8): The non-structural protein 8 (121 amino acid residues) is known to have a possible role in host-virus interactions (UniProtKB-P0DTC8; NCBI-YP_009724396). This protein showed variation in mutation rate ranging from 6.25 in March to 9.49 in May. The two prominent mutations on position 24 with a change from serine to leucine and position 84 leucine to serine were noticed across all months in ORF8 protein.

Surface glycoprotein (S): The longest of the four proteins, S protein (1273 amino acids; UniProtKB-P0DTC2; NCBI-YP_009724390) is involved in initiating the infection by interacting with the host's Angiotensin-Converting Enzyme 2 (ACE2) receptor and therefore playing an important role in rapid human to human transmission [11]. It was noted from the line graph that S protein showed a gradual increase in the rate of mutation from 38.72 in March to 50.39 in June. This was the second protein wherein we observed that the July month mutation rate was higher (68.86) than the previous month. Unique mutations encountered in S protein were 85 in April, 32 in May, 46 in June, and 6 in July. The mutation frequency of D614G was highest across all months.

Membrane glycoprotein (M): The M protein (222 aa; UniProtKB - P0DTC5; NCBI-YP_009724393) is a component of the viral envelope which plays a central role in virus morphogenesis and assembly via its interactions with other viral proteins. In figure 1, M protein had the highest mutation rate (9.03) in June.

Nucleocapsid phosphoprotein or Nucleoprotein (N): The N protein (419 aa; UniProtKB-P0DTC9; NCBI-YP_009724397) holds a fundamental role during virion assembly through its interactions with the viral genome and M protein, in enhancing the efficiency of sub-genomic viral RNA transcription as well as viral replication. For N protein the mutation rate in May was 31.57 which gradually decreased to 26.77 in June. In total, 212 mutations were identified in N protein, with 53 in April, 17 in May, and 20 in June with S194L mutation occurring dominantly across all months.

Envelope small membrane protein (E): The E protein (UniProtKB-P0DTC4; NCBI-YP_009724392) similar to M protein plays a vital role in virus morphogenesis and assembly. It also acts as a viroporin and self-assembles inside host membranes forming pentameric protein-lipid pores that allow ion transport. The mutation rate in the E protein was highest in June (5.98) and least in May (1.5)

3.2. The total number of unique mutations identified in all six SARS-CoV-2 proteins is represented using a mosaic plot. Figure 2, represents the unique mutations per kb of the genome in different months. The highest number of mutations was observed in S and N proteins. These 2 proteins being the longest out of the six proteins, could be one of the possible reasons for accumulating more mutations, however in order to normalize the mutation rate we have expressed the mutation as per KB of genome.

SARS-CoV-2 genome data collection

The amino acid sequences for the six viral proteins were retrieved from the NCBI (www.ncbi.nlm.nih.gov) and were segregated as per their genome sequencing date (March-July 2020). Table 1 lists the

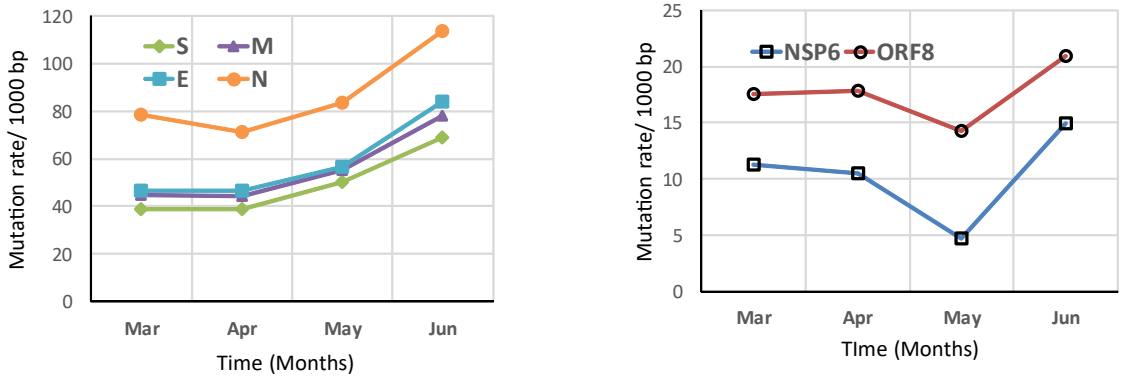

Figure 1: Analysis of the amino acid sequences of structural (A) and non-structural proteins (B) of SARS-CoV-2 is plotted in different months showing increase in mutation rate of the viral proteins from March – Jun 2020.

number of protein sequences of the SARS-CoV-2 virus used in the current study.

Mutation analysis

Protein sequences of SARS-CoV-2_Wuhan (accessions: YP_009725302, YP_009724396, YP_009724390, YP_009724393, YP_009724397, and YP_009724392) were used as reference in this study. All these genomes were sequenced in China in December 2019/ January 2020. Mutations in the six SARS-CoV-2 structural and nonstructural proteins were scored manually from multiple sequence alignment. A mutation profile that records all the variations corresponding to the reference amino acid residues was prepared for each of the proteins and tabulated for all months. An elaborative month-wise tabulation of mutation profile for all proteins NSP6, S, N, E, M, and ORF8 shows a consistent increase over the last three months (Figure 1).

COVID-19 cases and death rate

Month-wise data for total COVID-19 cases (infection) and deaths over the period from March to July 2020 were collected from a web platform [12]. The death rate per 1000 cases was derived and tabulated in table 2. It was observed that the death rate was at its peak during May and showed a downward trend in June and July. This motivated us to compare the mutation rate with the death rate to draw our initial hypothesis that the death rate and mutation rate are inversely proportional.

Comparative analysis of COVID-19 cases and death rate

A comparative analysis of the global death rate and cumulative mutation rates of all six proteins used in the study indicated that the death rate is inversely proportional to the mutation rate. The slope for the death rate (1.29) and cumulative mutation rate (0.79) was derived from the data of April, May, and June. The slope value is used to calculate the mutation rate and death rate, and the projected comparison is presented in figure 3. The data analysis clearly illustrated that by November 2020 the cumulative mutation rate of the six protein will be close to 500/ kb of the genome which will lower the global death rate to 1.1%.

Regression analysis was done to score the impact of mutation of each gene on the death rate (Table 3). Among the six proteins used in the study, a significant positive relationship between the death rate (global) and mutation rate was noted in the case of NSP6 and S proteins [r (5) = 0.081-0.83 $p < 0.1$]. The p-value for ORF8 and

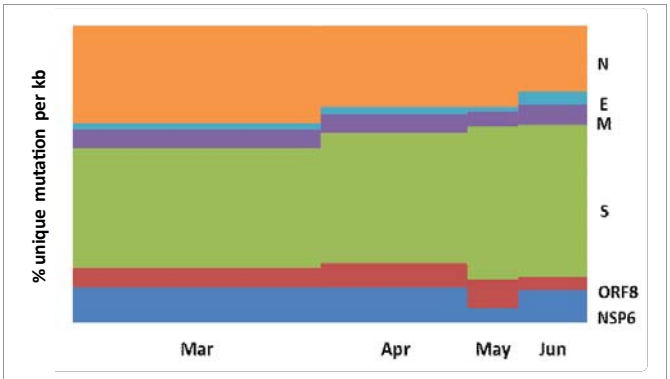

Figure 2: Month-wise unique mutations of six proteins of SARS-CoV-2 showing the highest amount of mutations in S and N gene per 1000 bp of genome.

M proteins are marginally higher ($p > 0.13$) than the p-value of NSP6 and S (Table 3). As the p-value of NSP6 and S proteins were comparatively lower than other proteins (M, N, E, and ORF8), the cumulative mutation rate of these two proteins was calculated with the death rate of three major clusters Cluster I (genomes sequenced from India and Bangladesh), Cluster II (genomes sequenced from United States) and Cluster III (genomes sequenced from Australia and New Zealand).

Analysis of the mutation rate and death rate was performed for these three clusters which is represented in figure 4. The cumulative mutation rate of NSP6 and S proteins is highest in cluster I, followed by cluster II and cluster III. Over the last three months, a sharp decline in death rate is also observed for cluster I and cluster III (Figure 4). Independent analysis of the calculated death rate of these three clusters indicated that by the end of 2020 the death rate of cluster I will decline below 0.5%, whereas the death rate of cluster III will be below 1%, and the death rate of cluster II will be around 1.5% (Figure 5). Therefore, it could be predicted from the analysis that in the cluster I countries (India and Bangladesh) the virion of SARS-CoV-2 will accumulate more mutations for NSP6 and S proteins and the death rate will be significantly reduced to >0.5% by the end of 2020. As the rate of mutations of these two proteins is slower in cluster II country (USA) the calculated death rate of COVID-19 will be 1-1.5% by the end of 2020. It could be predicted that high mutation of NSP6 and S will result (a) defective virion assembly, (b) which may inhibit subsequent infection, (c) negatively impact the geometric progression of viral titer, and (d) lower death rate. Lower viral titer

will provide a competitive advantage to the host immune system to fight the infection efficiently.

CONCLUSION

There have been pandemic outbreaks in the past which affected larger populations of the world such as the century-old Influenza pandemic or 1780s smallpox or other outbreaks from cholera in the 1830s to HIV-AIDS in the 1980s. But, on a positive note, there are lessons to be learned from the past. First, the naming of the virus is important. For instance, naming coronavirus initially as 'Chinese flu' gave the wrong indication of the limited spread of this disease to China and thereby delayed the response to the pandemic threat. Second, social distancing is presently the only available preventive measure for this disease, lately, people have started behaving responsibly because of awareness and knowledge hence we can see a slight downward trend in death rate. Third, one should learn that global cooperation and knowledge sharing can help in times of outbreaks. Immunization is the best solution one can have for such deadly infections and we are already on the track of developing it sooner. At last, one should believe that "this shall too pass" because we are in the era of owning advanced public health systems, scientific tools, and medical supplies, far superior to what we had in past.

The viral genome has undergone a series of mutational changes over the period since its parent genome has been released. Among the structural and non-structural proteins, NSP6 and S showed a high degree of correlation with the death rate. Cumulative mutation rates of these two proteins showed a high degree of correlation with the death rate in three focused clusters (I, II, III). All these studies have helped us to propose that, the mutation rate of NSP6 and S proteins are lower in cluster II country (USA) than in cluster I and cluster III countries. Therefore, we propose that the death rate of the COVID-19 pandemic in cluster II country (USA) will weaken later than cluster I and III countries (India, Bangladesh, Australia, and New Zealand). Implementation of strict measures to reduce infection rate, public awareness, advanced treatment strategies, novel drugs, and mass vaccination could be some of the factors which can influence this prediction.

We propose that a similar strategy of correlating country-wise mutation and death rates could be estimated in a big-data platform and AI models could be developed to eradicate COVID-19 pandemic in a much precise manner. In the model, integration of efficient treatment, epidemic management, and vaccination strategies could also be added to the mutation rate of structural and non-structural proteins (viz., NSP6 and S proteins). Country-wise mapping of unique mutations in structural genes and docking studies could open windows on host-pathogen interactions, infectivity, and mortality. This data could be used for developing personalized treatment strategies. The ongoing trials in vaccine development across the globe and progressive treatment methodologies would certainly contribute in the next quarter of the year to significantly weaken this pandemic. In this scenario along with the frontline workers and researchers, people could also contribute by social distancing, using a face mask, and following hygiene guidelines to help break the chain and reduce the burden on the healthcare system. Sequencing Monte Carlo sampling method was used to study the structural changes due to ongoing mutations of SARS-CoV-2 for developing efficient management of the epidemic [13].

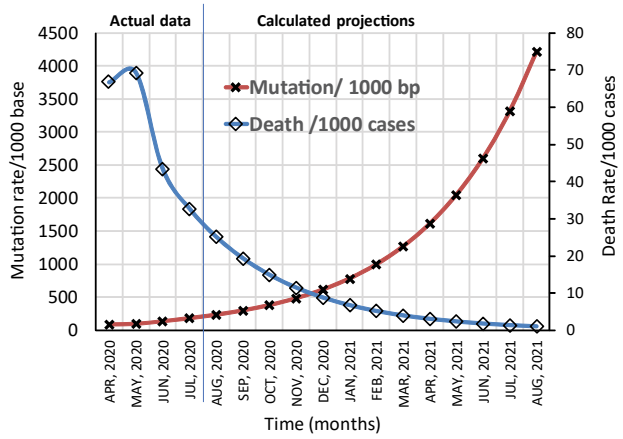

Figure 3: Cumulative mutation rate of the six proteins of the SARS-CoV-2 and death rate of COVID-19 were plotted for April-June 2020 (actual data). Calculated mutation rate and death rate derived from slope were also plotted for Aug 2020-Aug 2021.

Table 3: Regression analysis of mutation rate of structural and non-structural proteins and global death rate was scored using ANNOVA.

Regression Statistics calculated with death rate vs mutation rates						
	NSP6	**ORF8**	**S**	**M**	**E**	**N**
Multiple R	0.83	0.87	0.81	0.87	0.77	0.83
R Square	0.69	0.76	0.62	0.76	0.59	0.69
Standard Error	10.04	7.61	10.09	7.59	9.80	8.62
Observations	5	4	5	4	4	4
P-value	0.083	0.131	0.099	0.130	0.229	0.172

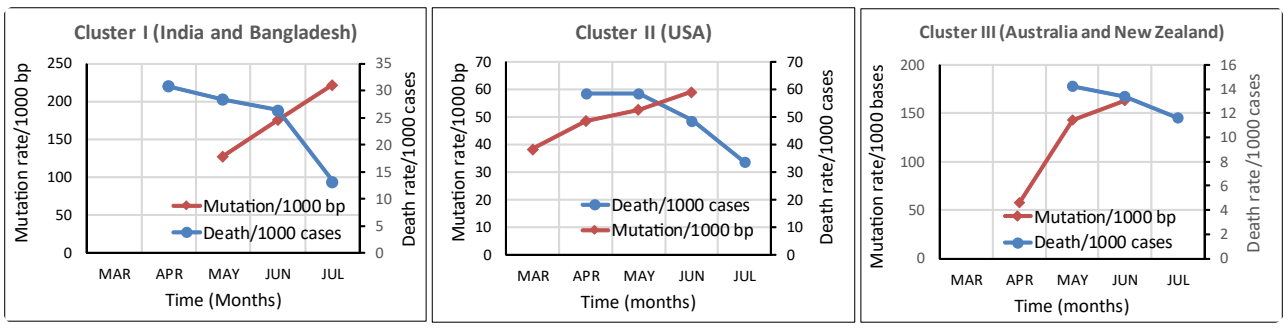

Figure 4: Cumulative mutation rate of the NSP6 and S proteins of SARS-CoV-2 and death rate of COVID-19 cases were plotted for three clusters (Cluster I, Cluster II and Cluster III).

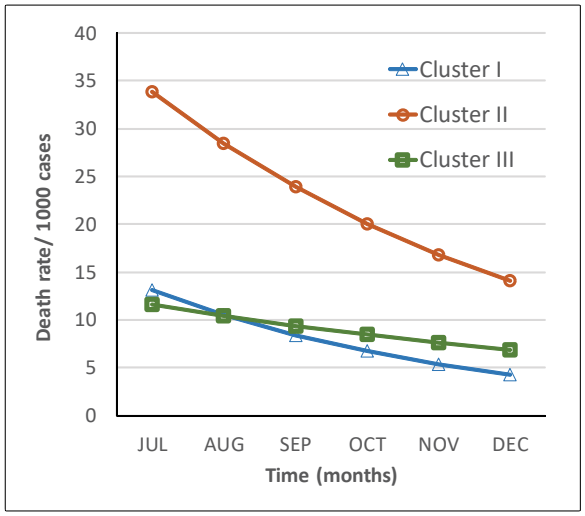

Figure 5: Calculated death rate (/1000 cases) were plotted for Cluster I, Cluster II and Cluster III countries. Cluster I= India and Bangladesh; Cluster II = USA; Cluster III= Australia and New Zealand.

REFERENCES

1. Peng Zhou, Xing-Lou Yang, Xian-Guang Wang, Ben Hu, Lei Zhang, Wei Zhang, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020; 579: 270-273. DOI: 10.1038/s41586-020-2012-7
2. Yu H, Wu JT, Cowling BJ, Liao Q, Fang VJ, Zhou S, et al. Effect of closure of live poultry markets on poultry-to-person transmission of avian influenza A H7N9 virus: an ecological study. Lancet (London, England). 2013; 383: 541-548. DOI: 10.1016/s0140-6736(13)61904-2
3. Kate E Jones, Nikkita G Patel, Marc A Levy, Adam Storeygard, Deborah Balk, John L Gittleman, et al. Global trends in emerging infectious diseases. Nature. 2008; 451: 990-993. DOI: 10.1038/nature06536
4. Morse SS. Factors in the emergence of infectious diseases. Emerg Infect Dis. 1995; 1: 7-15. DOI: 10.3201/eid0101.950102
5. World Health Organization (WHO), IHR core capacity monitoring framework: Checklist and indicators for monitoring progress in the development of IHR core capacities in states parties 2013. International Health Regulations (2005) document; WHO/HSE/GCR/2013.2
6. Sanjuán, R, Domingo-Calap P. Mechanisms of viral mutation. Cell Mol Life Sci. 2016; 73: 4433-4448. DOI: 10.1007/s00018-016-2299-6
7. Wang R, Hozumi Y, Yin C, Wei GW. Mutations on COVID-19 diagnostic targets. Cornell University. 2020. https://tinyurl.com/yyg7h53v
8. Sanjuán R, Nebot MR, Chirico N, Mansky LM, Belshaw R. Viral mutation rates. Journal of virology. 2020; 84: 9733-9748. DOI: 10.1128/JVI.00694-10
9. Hannah Ritchie, Esteban Ortiz-Ospina, Diana Beltekian, Edouard Mathieu, Joe Hasell, Bobbie Macdonald, et al. Coronavirus pandemic (COVID-19). 2020. https://tinyurl.com/t9b3bs3
10. Baliji S, Cammer SA, Sobral B, Baker SC. Detection of nonstructural protein 6 in murine coronavirus-infected cells and analysis of the transmembrane topology by using bioinformatics and molecular approaches. J Virol. 2020; 83: 6957-6962. DOI: 10.1128/JVI.00254-09
11. Zhang H, Penninger JM, Li Y, Zhong N, Slutsky AS. Angiotensin-converting enzyme 2 (ACE2) as a SARS-CoV-2 receptor: Molecular mechanisms and potential therapeutic target. Intensive Care Med. 2020; 46: 586-590. DOI: 10.1007/s00134-020-05985-9
12. Wang R, Hozumi Y, Yin C, Wei G. Decoding asymptomatic COVID-19 infection and transmission. Cornell University. 2020. https://tinyurl.com/y492xh4c
13. Wong SWK. Assessing the impacts of mutations to the structure of COVID-19 spike protein via sequential Monte Carlo. Cornell University. 2020. https://tinyurl.com/y36mx5yb