Recent Developments in Electronics and Communication Systems
KVS Ramachandra Murthy et al. (Eds.)

doi:10.3233/ATDE221334

Forecasting of COVID-19 Outbreak in India: A Time Series Analysis

Rajni[a,1] and Tushant Kumar[b]
[a] *Jindal Global Business School, O. P. Jindal Global University, India*
[b] *Department of Radiodiagnosis, Dr RMLIMS Lucknow, India*
ORCiD ID: Rajni https://orcid.org/0000-0002-7187-6363
Tushant Kumar https://orcid.org/0000-0002-2196-1404

Abstract. The World Health Organization (WHO) declared the status of coronavirus disease 2019 (COVID-19) to a global pandemic on March 11, 2020. Since then, numerous statistical, epidemiological and mathematical models have been used and investigated by researchers across the world to predict the spread of this pandemic in different geographical locations. The data for COVID-19 outbreak in India has been collated on daily new confirmed cases from March 12, 2020 to April 10, 2021. A time series analysis using Auto Regressive Integrated Moving Average (ARIMA) model was used to investigate the dataset and then forecast for the next 30-day time-period from April 11, 2021, to May 10, 2021. The selected model predicts a surge in the number of daily new cases and number of deaths. An investigation into the daily infection rate for India has also been done.

Keywords. ARIMA, stationarity, COVID-19, Daily Infection Rate

1. Introduction

World is passing through a very distressful stage due to novel coronavirus also known as COVID-19. As of April 10, 2021 the number of confirmed cases reported worldwide were 134,342,276 and deaths were 2,909,082. It is a highly infectious disease and was declared a global pandemic by the World Health Organization (WHO) on 11th March 2020. It originated in Wuhan, Hubei Province, People's Republic of China (PRC) in late December 2019, when a case of unidentified pneumonia was reported [1]. The disease was renamed as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) by the International Committee on Taxonomy of Viruses (ICTV). It has been found that this is a class of β-coronavirus. It has many potential natural hosts as it can be from animal to human transmission and human to human transmission. Thus, these characteristics, pose a great challenge for prevention and treatment of the virus infection. Despite many cases worldwide [2] and low mortality rate [3] compared to SARS and the middle east respiratory syndrome (MERS), this virus has high infectivity and transmissibility among humans.

WHO issued guidelines of preventive measures for COVID-19 including maintaining social distancing, washing of hands frequently, avoiding touching the mouth, nose, and face [4]. India reported its first case of COVID-19 on 30th January 2020 with

[1] Corresponding Author: Rajni, rajni@jgu.edu.in.

origin from China [5]. Since then, it has spread to almost all the districts of the country. As on 10th April 2021 the total confirmed cases reported in India were 13,205,926 and 168,436 deaths. Currently, India is going through the second wave of the COVID-19 [6]. Under the first wave, however, the rate of infection was lower as compared to other countries. But, recently due to different variant of COVID-19, the rate of infection is higher than the first wave.

As COVID-19 does not have specific set of treatments, and it is spreading quickly, it is crucial to make healthcare services prepared for the future scenario. The increasing number of cases puts a lot of stress on the part of administration and healthcare officials for accommodating patients with possible symptoms of COVID-19. In this situation, some prediction tools help to know about the number of cases in coming days for preparation at the administrative level. Machine learning algorithms and statistical analysis such as time series analysis has been used to make projections for COVID-19 with context to different scenarios and regions [7-17]. The machine learning techniques have also been used for prediction in many different areas [18].

In this paper, we propose the data-driven ARIMA method for the prediction of number of patients (daily new cases) to be accommodated in the subsequent days based on the data available. The proposed model can approximately predict the number of daily new COVID-19 cases, and the number of deaths so, the administration can prepare accordingly to accommodate them.

This paper has been organized as follows. In Section 2, ARIMA technique for the prediction of COVID-19 has been explained in detail. In Sections 3 and 4, the results and discussion followed by conclusion of the work is presented, respectively.

2. Methodology

Auto Regressive Integrated Moving Average model, also known as ARIMA model [19-20], has been used for analysis of daily new cases, active cases and deaths in India. The Box-Jenkins approach has been applied to this dataset for model identification, estimation, diagnostic checking and forecasting. ARIMA model is specified by three orderly parameters: (p, d, q); where 'p' is the order of the auto aggressive order part referring to the past values of the variable, 'd' is the order of the differencing also called as degree of integrated I(d) component, and 'q' is the order of the moving average part also referred to as model error which is combination of past forecast error terms.

The ARIMA model can be modified to perform the function of an ARMA model as well as a simple AR, I or MA model. AR (p) model refers to the current value of the time series Yt linearly in terms of its previous values Y_{t-1}, Y_{t-2}, …, Y_{t-p} and the current residuals ε_t. MA (q) model refers to the current value of the time series Y_t linearly in terms of its current and previous residual series ε_{t-1}, ε_{t-2}, …, ε_{t-q}. Combining the equations, the final forecast equation is written as

$$Y_t=\alpha + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q} + \varepsilon_t \quad (1)$$

where, α is a constant, ϕ and θ are the autoregressive and moving average parameters, respectively. Y_t is the observed value at time t and ε_t is the value of the random shock at time t. In the current scenario we are dealing with non-seasonal ARIMA as there is absence of seasonal component.

The first step of the ARIMA model is to check whether the time series is stationary. A time series is considered as stationary if its statistical properties such as mean, variance, autocorrelation are constant over time. The stationarity of a timeseries observation is

important as it makes it easier to get accurate prediction estimates. After converting all non-stationary time series into stationary using differencing, different ARIMA models were developed. Based on the ARIMA model accuracy evolution of COVID-19 Indian data on mentioned time period, we considered Akaike Information Criterion (AIC) parameter. We also give an account of Daily Infection Rate (DIR) for India.

3. Results and Discussion

3.1. ARIMA Modeling for India

The daily new COVID-19 epidemic cases, and deaths in India from March 12, 2020 to April 10, 2021 are available from the John Hopkins Database [2]. The data was analyzed using the statistical software R. The dataset consists of 395 values. 90% of the dataset was used for training and 10% for testing. Figure 1 shows the time series plot of the daily new cases in India.

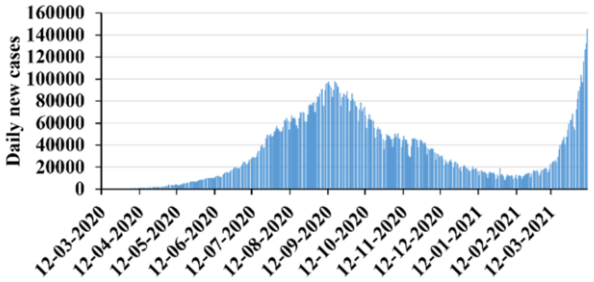

Figure 1. Time series plot of daily new cases in India as of 10th April 2021.

It is be observed that the original time series is nonstationary and follows an increasing trend. The differencing transformation was used to obtain stationarity on the original nonstationary time series. The Augmented Dickey-Fuller (ADF) and KPSS unit-root test were used to identify whether the time series is stationary or not. In addition, the R package "tseries", "urca", "forecast" and "fpp2" was used to obtain the output for ARIMA [21, 22]. Thus, it requires differencing to proceed further, and which will help to stabilize the original data. In Figure 2, the second-order differenced time series is shown, and it looks stationary when compared to the original time series. From the original data, using ADF test and KPSS test, stationarity was checked, and the original data was differenced two times until the ADF and KPSS test confirmed stationarity. After confirming the stationarity, the ACF and PACF plots of the second-order differenced data for daily new cases were obtained which shows significant spikes at lags 1, 3, 5 and 7. From PACF plot, it is observed to have significant spikes at lags 1, 2, 3, 4, 5.

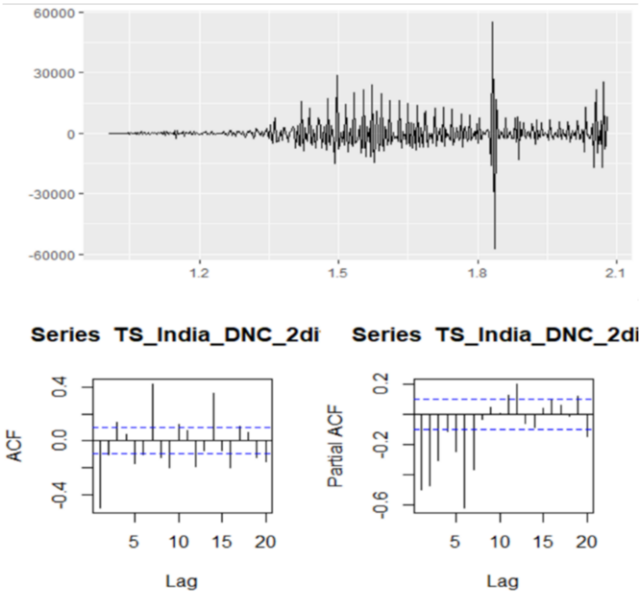

Figure 2. Plot of second-order differenced series of daily new cases and second row showing the autocorrelation and partial auto-correlation function of second-order differenced daily new cases time series.

Based on the ACF and PACF plots as shown in Figure 2, 12 initial models for daily new cases, and by observing correlogram for death and confirmed cases data 7 models for deaths cases and 10 models for number of total confirmed cases were checked for ARIMA modeling. Their orders and AIC values are mentioned in Table 1.

Table 1. Summary of ARIMA modeling for daily new cases, active cases and number of deaths in India.

	Order of ARIMA Model	**AIC values**
Daily New Cases	(2,2,2)	7652.303
	(0,2,0)	8038.13
	(1,2,0)	7928.298
	(0,2,1)	7719.549
	(1,2,2)	7655.797
	(2,2,1)	7686.52
	(3,2,2)	7654.81
	(1,2,1)	7708.688
	(1,2,3)	7651.616
	(0,2,3)	7648.854
	(0,2,2)	7690.525
	(0,2,4)	7650.617
Daily Number of Deaths	**(1,1,3)**	4843.481
	(0,1,3)	4857.835

	(2,1,3)	4846.486
	(1,1,4)	4845.383
	(0,1,2)	4856.88
	(0,1,4)	4853.613
	(2,1,2)	4844.425
Total Confirmed Cases	(2,2,2)	7605.553
	(0,2,0)	7616.991
	(1,2,0)	7619.242
	(0,2,1)	7617.807
	(1,2,2)	7602.676
	(0,2,2)	7606.935
	(1,2,1)	7604.161
	(1,2,3)	7604.306
	(0,2,3)	7605.826
	(2,2,1)	7604.162

The residuals were also checked for each model and Box.test were also performed. Combining the parsimony principle and AIC value, ARIMA (0,2,3) was selected as the best model for forecasting and the results are shown in Figure 3. The model predicts that by 10 May 2021, India will see around 396101 new cases per day. The estimated results have been provided with interval values in the supplementary data. The diagnostic check on the final ARIMA model was performed, which followed the plot of residuals, ACF of the residuals and p-values of the Ljung-Box test, which show the residuals in good agreement with the white noise. Thus, we can say that the final ARIMA (0,2,3) model captures the structure of the daily new cases in the best way. In a similar manner, the best fitting model for the number of deaths was obtained as ARIMA (1,1,3) and the forecasting is shown in Figure 4. Also, the best model for the number of total confirmed cases is ARIMA (1,2,2) and the forecasting is shown in Figure 5.

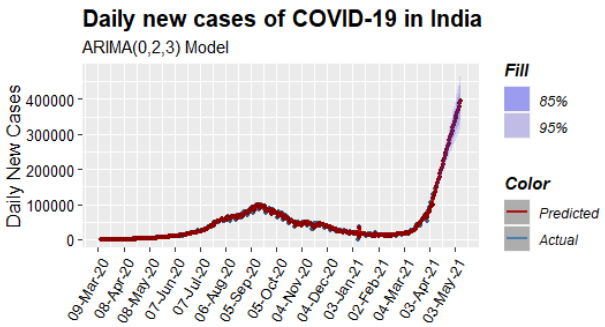

Figure 3. Predicted and Actual plot of daily new cases and forecasting for the next 30 day time period.

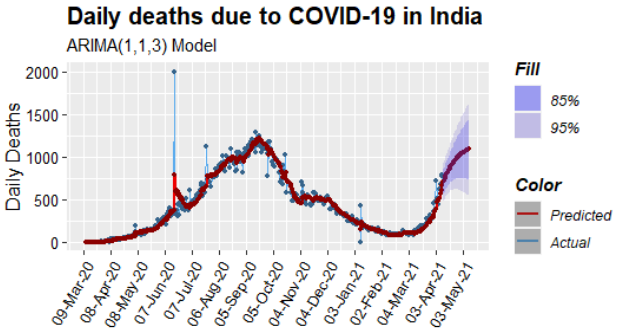

Figure 4. Predicted and Actual plot of daily death cases and forecasting for the next 30 day time period.

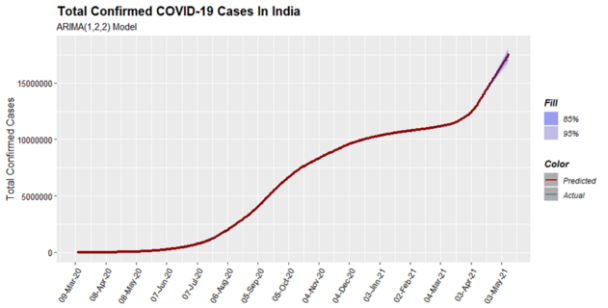

Figure 5. Predicted and Actual plot of total confirmed cases and forecasting for the next 30-day time period.

The predicted values for number of daily new cases, deaths and total confirmed cases are shown in Table 2.

Table 2. Summary of predicted values of daily new cases, number of daily deaths and total confirmed cases in India.

Date	Daily new cases	Number of Deaths	Total confirmed cases
11-04-2021	147514	709	13348985
12-04-2021	152571	727	13492090
13-04-2021	161269	753	13635167
14-04-2021	169966	779	13778261
15-04-2021	178664	802	13921345
16-04-2021	187361	824	14064435
17-04-2021	196059	845	14207521
18-04-2021	204756	865	14350609
19-04-2021	213454	883	14493696
20-04-2021	222151	901	14636784
21-04-2021	230849	917	14779871
22-04-2021	239546	932	14922959
23-04-2021	248244	947	15066047

24-04-2021	256941	960	15209134
25-04-2021	265639	973	15352222
26-04-2021	274336	985	15495309
27-04-2021	283034	997	15638397
28-04-2021	291731	1007	15781485
29-04-2021	300429	1017	15924572
30-04-2021	309126	1027	16067660
01-05-2021	317824	1036	16210747
02-05-2021	326521	1044	16353835
03-05-2021	335219	1052	16496922
04-05-2021	343916	1059	16640010
05-05-2021	352614	1066	16783098
06-05-2021	361311	1073	16926185
07-05-2021	370009	1079	17069273
08-05-2021	378706	1085	17212360
09-05-2021	387404	1090	17355448
10-05-2021	396101	1095	17498535

Figure 6 shows the difference between the actual and predicted values from 11th April to 30th April 2021. Huge difference is seen towards the end of the month as many new cases emerged due to rapid spread of new strain of COVID-19 and lack of healthcare facilities led to a greater number of deaths within a 20-day span.

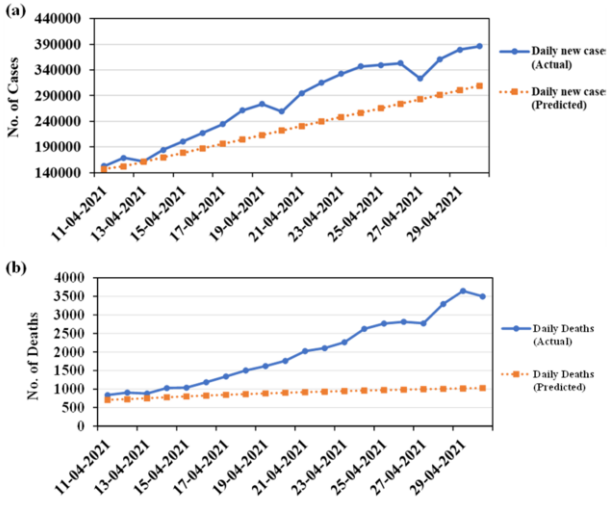

Figure 6. Predicted and Actual plot of daily new cases from 11th April to 30th April 2021.

3.2. Daily Infection Rate for India

In this section, we investigate the daily infection rate (DIR) for Indian region. obtained by using the following formula:

Daily Infection Rate (DIR)=$(T_1-T_2)/T_2$,

where T_1 is total number of active cases on a given day and T_2 is total number of active cases on a previous day. The value of DIR can be positive, zero or negative. If the number of active cases on the given day increases from the previous day, it will be positive. If the number of active cases on the given day decreases from the previous day, it will be negative. And if there is no change in the number of active cases, then DIR will be zero. Figure 7 shows the plot of DIR for India. The rise in DIR in March 2021, indicates the onset of second wave of COVID-19. It indicates that the health administration needs to be well prepared to handle the large number of projected cases in the next coming month.

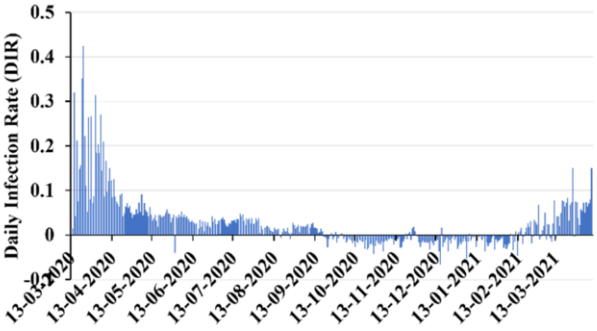

Figure 7. Daily Infection Rate for India from 13th March 2020 to 10th April 2021.

4. Conclusion

In this paper, a data-driven forecasting method has been used to predict the possible number of daily positive cases of COVID-19 in India for the next 30 days. The number of daily positive cases, active cases and deceased cases has also been estimated by using ARIMA. The preventive measures like social distancing and lockdown had helped contain the outbreak situation to a large extent till January 31st, 2021. Entry of new variants of COVID-19 which is highly infectious than the previous strain has led to a surge in the number of cases recently in February and March 2021. To contain the situation, there is a need to follow the preventive measures in the coming months to contain the outbreak. The start of vaccination has also helped to cut down the risk in vaccinated patients compared to those who are not vaccinated [23, 24]. As India is a very big country, the next challenge is also to speed-up and vaccinate many people. But, due to some misinformation about vaccines spreading among masses it will not be an easy task [25, 26].

Acknowledgements

The research work was supported by the Jindal Global Business School, O. P. Jindal Global University, India.

References

[1] Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, Zhang L, Fan G, Xu J, Gu X, Cheng Z, Yu T, Xia J, Wei Y, Wu W, Xie X, Yin W, Li H, Liu M, Xiao Y, Gao H, Guo L, Xie J, Wang G, Jiang R, Gao Z, Jin Q, Wang J, Cao B, Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 2020 395 (10223), 497–506.

[2] The John Hopkins Database. https://github.com/CSSEGISandData/COVID-19

[3] Liu Y, Gayle AA, Wilder-Smith A, Rocklöv J. The reproductive number of COVID- 19 is higher compared to SARS coronavirus. J. Travel Med. 2020 Vol. 27 (2), doi: https://doi.org/10.1093/jtm/taaa021 .

[4] WHO, Covid-19 Dashboard: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public .

[5] "One positive case reported in Kerala", https://pib.gov.in/PressReleasePage.aspx?PRID=1601095, (last accessed June 20, 2020)

[6] Samarasekera U. India grapples with second wave of COVID-19. The Lancet Microbe, 2021 Volume 2, Issue 6, e238. doi: https://doi.org/10.1016/S2666-5247(21)00123-3

[7] Rauf HT, Lali MIU, Khan MA, Kadry S, Alolaiyan H, Razaq A and Irfan R. Time series forecasting of COVID-19 transmission in Asia Pacific countries using deep neural networks. Pers Ubiquit Comput (2021), doi: https://doi.org/10.1007/s00779-020-01494-0

[8] Chimmula VKR, Zhang L. Time series forecasting of COVID-19 transmission in Canada using LSTM networks. Chaos, Solitons & Fractals 2020 Volume 135, 109864, doi: https://doi.org/10.1016/j.chaos.2020.109864

[9] Salgotra R, Gandomi M, and Gandomi AH. Time Series Analysis and Forecast of the COVID-19 Pandemic in India using Genetic Programming. Chaos, Solitons & Fractals 2020 Volume 138, 109945, doi: https://doi.org/10.1016/j.chaos.2020.109945

[10] Yadav M, Perumal M, Srinivas M. Analysis on novel coronavirus (COVID-19) using machine learning methods. Chaos, Solitons & Fractals 2020 Volume 139, 110050, doi: https://doi.org/10.1016/j.chaos.2020.110050

[11] Das RC. Forecasting incidences of COVID-19 using Box-Jenkins method for the period July 12-Septembert 11, 2020: A study on highly affected countries. Chaos, Solitons & Fractals 2020 Volume 140, 110248, doi: https://doi.org/10.1016/j.chaos.2020.110248

[12] Ballı S. Data analysis of Covid-19 pandemic and short-term cumulative case forecasting using machine learning time series methods. Chaos, Solitons & Fractals 2021 Volume 142, 110512, doi: https://doi.org/10.1016/j.chaos.2020.110512

[13] Gecili E, Ziady A, Szczesniak RD. Forecasting COVID-19 confirmed cases, deaths and recoveries: Revisiting established time series modeling through novel applications for the USA and Italy. PLoS ONE 2021 16(1): e0244173, doi: https://doi.org/10.1371/journal.pone.0244173

[14] ArunKumar KE, Kalaga DV, Kumar Ch.MS, Chilkoor G, Kawaji M, Brenza TM. Forecasting the dynamics of cumulative COVID-19 cases (confirmed, recovered and deaths) for top-16 countries using statistical machine learning models: Auto-Regressive Integrated Moving Average (ARIMA) and Seasonal Auto-Regressive Integrated Moving Average (SARIMA). Applied Soft Computing 2021 Volume 103, 107161, doi: https://doi.org/10.1016/j.asoc.2021.107161

[15] Yadav VK, Yadav VK and Yadav JP. Cognizance on Pandemic Corona Virus Infectious Disease (COVID-19) by using Statistical Technique: A Study and Analysis. Evergreen 2020 7 (3), pp.329-335, doi: https://doi.org/10.5109/4068611

[16] Bhatnagar P, Kaura S, Rajan S. Predictive Models and Analysis of Peak and Flatten Curve Values of CoVID-19 Cases in India. Evergreen 2020 7 (4), pp.458-467, doi: https://doi.org/10.5109/4150465

[17] Prabakaran G, Vaithiyanathan D and Kumar H. Fuzzy Decision Support System for the Outbreak of COVID-19 and Improving the People Livelihood. Evergreen 2021 8 (1), pp.36-43, doi: https://doi.org/10.5109/4372258

[18] Pariaman H, Luciana GM, Wisyaldin MK, Hisjam M. Anomaly Detection Using LSTM-Autoencoder to Predict Coal Pulverizer Condition on Coal-Fired Power Plant. Evergreen 2021 8 (1), pp.89-97, doi: https://doi.org/10.5109/4372264

[19] Box GEP, and Jenkins GM. Time Series Analysis, Forecasting and Control. 1976 San Francisco: Holden-Day.

[20] Haines LM, Munoz WP, and Van Gelderen CJ. ARIMA modelling of birth data. Journal of Applied Statistics. 1989 Vol. 16, 55–67, doi: https://doi.org/10.1080/02664768900000007

[21] Shumway RH, and Stoffer DS. Time Series Analysis and Its Applications with R Examples. Fourth ed, Springer (2017).

[22] Hyndman RJ, and Athanasopoulos G. Forecasting: principles and practice. 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2 (2018).

[23] Mahase E. Covid-19: One dose of vaccine cuts risk of passing on infection by as much as 50%, research shows. BMJ 2021Vol. 373: n1112, doi: https://doi.org/10.1136/bmj.n1112
[24] Moghadas SM, Vilches TN, Zhang K, Wells CR, Shoukat A, Singer BH, Meyers LA, Neuzil KM, Langley JM, Fitzpatrick MC, and Galvani, AP. The Impact of Vaccination on Coronavirus Disease 2019 (COVID-19) Outbreaks in the United States. Clinical Infectious Diseases 2021 Vol. 73 (12), pp: 2257-2264, doi: https://doi.org/10.1093/cid/ciab079
[25] Editorial. COVID-19 vaccines: the pandemic will not end overnight. The Lancet Microbe 2021 Volume 2, Issue 1, doi: https://doi.org/10.1016/S2666-5247(20)30226-3
[26] Padma TV. India's COVID-vaccine woes—by the numbers. Nature, 2021 Vol. 592 (7855) pp. 500-501, doi: https://doi.org/10.1038/d41586-021-00996-y