ORIGINAL ARTICLE

An Exploratory Study to Find the Early Trend and Pattern Recognition of COVID-19 Infection in India: A Severity Model-Based Prediction

Afreen Khan[1], Swaleha Zubair[2], Najam Khalique[3], Samreen Khan[4]

[1]Senior Research Fellow, Department of Computer Science, Aligarh Muslim University, Aligarh, India; [2]Associate Professor, Department of Computer Science, Aligarh Muslim University, Aligarh, India; [3]Professor, Department of Community Medicine, Jawaharlal Nehru Medical College, Aligarh Muslim University, Aligarh, India; [4]Assistant Professor, Department of Community Medicine, Integral Institute of Medical Sciences and Research, Integral University, Lucknow, India

Abstract | Introduction | Methodology | Results | Conclusion | References | Citation | Tables / Figures

Corresponding Author

Dr Samreen Khan, Assistant Professor, Department of Community Medicine, Integral Institute of Medical Sciences and Research, Integral University, Lucknow, India
E Mail ID: drsamreen2k4@gmail.com

Citation

Khan A, Zubair S, Khalique N, Khan S. An Exploratory Study to Find the Early Trend and Pattern Recognition of COVID-19 Infection in India: A Severity Model-Based Prediction. Indian J Comm Health. 2021;33(2):299-303. https://doi.org/10.47203/IJCH.2021.v33i02.013

Source of Funding: Nil **Conflict of Interest:** None declared

Article Cycle

Received: 06/02/2021; **Revision:** 12/04/2021; **Accepted:** 02/06/2021; **Published:** 30/06/2021

Abstract

Background: Recent Coronavirus Disease 2019 (COVID-19) pandemic has inflicted the whole world critically. Although India has been listed amongst the top ten highly affected countries to date, one cannot rule out COVID-19 associated complications in the near future. **Aim & Objective**: We aim to build the COVID-19 severity model employing logistic function which determines the inflection point and help in the prediction of the future number of confirmed cases. **Methods and Material**: An empirical study was performed on the COVID-19 patient status in India. We performed the study commencing from 30 January 2020 to 12 July 2020 for the analysis. Exploratory data analysis (EDA) tools and techniques were applied to establish a correlation amongst the various features. The acute stage of the disease was mapped in order to build a robust model. We collected five different datasets to execute the study. **Results**: We found that men were more prone to get infected with the coronavirus disease as compared to women. On 165-days based analysis, we found a trending pattern of confirmed, recovered, deceased and active cases of COVID-19 in India. The as-developed growth model provided an inflection point of 72.0 days. It also predicted the number of confirmed cases as 17,80,000.0 in the future i.e. after 12th July. A growth rate of 32.0 percent was obtained. We achieved statistically significant correlations amongst growth rate and predicted COVID-19 confirmed cases. **Conclusions**: This study demonstrated the effective application of EDA and analytical modeling in building a mathematical severity model for COVID-19 in India.

Keywords

Coronavirus Disease 2019; COVID-19; Pandemic; Modeling; Data Analysis; Exploratory Data Analysis; India

Introduction

Coronavirus disease 2019 (COVID-19) emerged as a serious health threat to the lives of many people around the globe. According to WHO situation report-175, as of 10:00 CEST, 13th July 2020, the virus had inflicted 12,768,307 people with around 566,654 deaths globally (1). Whereas the South-East region (India belongs to this region) had around 1,163,556 infection cases with 29,258 deaths as of 12th July 2020 (1). In the context of India, 35 states and union territories had reported coronavirus infection on July 12th, 2020. There has been a rise in the number of confirmed cases and deaths in India. Among the 1,33,40,516 subjects reported for the coronavirus symptoms, 8,49,553 were detected as positive while 5,34,621 patients recovered and 23,139 patients had died from this disease in India (2).

The early and timely analysis of COVID-19 is essential for complete monitoring of the growth and spread of the disease. Thus, in the present paper, we carried out a comprehensive study to learn the effect of coronavirus disease on the Indian population. We worked on statistical data analysis, exploratory data analysis (EDA), and logistic modeling to carry out the present study.

Aims & Objectives

We aim to build the COVID-19 severity model employing logistic function which determines the inflection point and help in the prediction of the future number of confirmed cases.

Material & Methods

1. Subjects and Data Collection

The COVID-19 patients from all the 36 states and union territories of India were enrolled in this retrospective study. We acquired data between 30th January to 12th July 2020, a total of 165-days (i.e. 5 months 13 days) period for the analysis. The samples used in the study were obtained and integrated from various data consortium websites.

2. Experimental Frame

We divided our study into two sections i.e. data analysis and building a severity model for COVID-19. The data analysis section further consisted of statistical analysis and exploratory data analysis. Statistical analysis is a type of analysis that includes collecting, uncovering and presenting the big data to find the underlying patterns (3). It is necessary for formulating data-centric decisions. Whereas, EDA is another type of analysis that examine and evaluate the datasets and thus, summarizes their essential properties (4).

Results

The results of the executed analysis are discussed in the below sub-sections.

a. Statistical Analysis

In the gathered data, age and gender features consisted of many missing values. Several patients' age and gender values were not found in the dataset, hence these values are termed as missing values. Out of the total details present in the dataset, the demographics of the subjects used in the study is shown in (Table 1). Moreover, it can be inferred that amongst the COVID-19 patients, men were found to be greater in number as compared to women and transgenders. The range of age was found to be between 0 years to 98 years.

b. Exploratory Data Analysis

Along with the patient details, the major feature that the dataset contained was the confirmed, deceased and recovered COVID-19 patients. Using these details effectively, we analyzed to arrive at better results.

b.1 Age by Gender Distribution of the Confirmed subjects

In the age distribution of the infected and confirmed patients concerning the number of male, female and transgender subjects, it was found that males are more likely to be under the effect of coronavirus than females and transgenders. Within this, the men's age group of 30-50 years were more likely to be infected.

b.2 Total Number of Infected cases

As of 12th July 2020, the number of confirmed, deceased and recovered coronavirus patients belonging to all the Indian states and union territories is plotted in (Figure 1). For all the 165 days, starting from 30th January to 12th July 2020, (Figure 1) displays a rising graph. Beginning from a single confirmed patient (on 30th January) mounting to 33,330.0 confirmed patients (30th April) and 8,49,553.0 cases by the end of 12th July. We got to see an extreme rise of infected patients after 30th March. The total number of deceased patients were found to be 22,674.0 while recovered patients as 5,34,621.0. The plot for all the three goes in parallel, with a rise in all of the three.

b.3 Overall Cumulative Scale

The cumulative recovery rate and mortality rate of India are reported in (Table 2). Up until 2nd March, the recovery rate was 0%. And, till 11th March, the mortality rate was found to be 0%.

b.4 Weekly Trend

An outline of the weekly trend has been shown in Figure 2. It provides a general notion of the week-wise trend in India. The data show the number of confirmed, deceased and recovered cases on seven-time points i.e. 30th January, 29th February, 30th March, 30th April, 31st May, 30th June and 12th July respectively.

c. Severity Model for COVID-19: Logistic Modeling

The scatterness of any contagious disease can be demonstrated by a logistic function. Herein, the progression of disease starts exponentially but after some time, it slows down. The position where it slows down, that particular point is called an inflection point. The main task of this study is to gain insights and look for the inflection point using two growth metrics i.e. growth ratio and growth factor.

(1) Growth Ratio:

It is calculated by dividing the total number of confirmed cases on the nth day by the total number of confirmed cases on the (n-1)th day. The plotted graph for equation (1) is illustrated in (Figure 3). It simply signifies the progression of confirmed cases of a particular day with respect to the previous day. The peak point we got to see is between 29th February and 10th April. After that, it starts declining till the end date.

(2) Growth Factor:

It is calculated by dividing [the total number of confirmed cases on the nth day – the total number of confirmed cases on the (n-1)th day] by [the total number of confirmed cases on the (n-1)th day – the total number of confirmed cases on the (n-2)th day]. It can be seen from (Figure 4), that the peak point is reached around the 5th of April. The growth factor tells whether or not the inflection point is reached. If it becomes stable nearly to 1.0, then it has reached a state of inflection; if not stabilized at around 1.0, then it has not reached an inflection point. From (Figure 3) and (Figure 4), we can see that there is a remarkable reduction in both growth ratio and growth factor. The growth factor stabilizes near 1.0 while the growth ratio is close to 1.0 but did not exactly reach this point. Using these values, we built a logistic model which defines the severeness of this deadly disease. The results of this model are exemplified in (Figure 5). The number of confirmed coronavirus cases in India were fitted by the logistic curve, and further predicted the new cases in the coming days.

From the statistical report, it was concluded that the inflection point hit around 72.0 days. The inflection date was 3rd April 2020. The growth rate is 32.0 percent. And the number of confirmed cases will maximize at around 17,80,000 cases in the coming days (after 12th April). This means that the given model predicts the new confirmed cases in the upcoming days. The more the number of data, the better are the predictions. Also, a strong correlation has been observed between these two i.e. a value of 0.991. (Table 3) shows the combined results of the complete growth analysis performed. From (Table 3) it can be seen that we have emphasized on April. It is because the inflection point was reached this month. Then the heavy upsurge of the spread of the pandemic was observed in April. This is the month from where certain major changes were observed.

Discussion

EDA is a powerful and crucial step in the field of predictive modeling. It provides several insights which help in creating strong correlations amongst various features and modeling a complete prediction system. In the context of health-related data, it becomes necessary to recognize the patterns within the dataset to build a robust working model (5). This study had manifold objectives. Firstly, to establish and look for the positive and negative correlations amongst the various features from the several integrated datasets. Secondly, it also helped in gaining the maximum perception to understand the data structure effectively. Thirdly, how COVID-19 has affected the Indian population up until now and how it can affect in the future. Lastly, it also helped in determining the COVID-19 growth factor and growth ratio, obtaining the inflection point and finally building a severity predictive model using a logistic function.

Both qualitatively and quantitatively, the cases of the COVID-19 are diverse as compared to the earlier epidemics. The pandemic pattern shows that it is an extremely severe virus spread. Recently, many studies have been performed employing machine learning and deep learning. Alimadadi A. et al. in their paper talked about how machine learning and deep learning can be used to fight with COVID-19 (6). They presented the application of both these technologies pictorially. Furthermore, Allam Z. et al. performed a survey of earlier viral outbreaks and explored the use of Artificial Intelligence can aid in early detection of COVID-19 in China (7). A study accomplished by Yang Z. et al. showed the COVID-19 epidemic trend in China by using Artificial Intelligence and SEIR (Susceptible-Exposed-Infectious-Removed) modeling (8). Mavragani A. applied the infodemiology (information epidemiology) approach which employs web-based data to notify public health and policymaking (9). This paper talked about tracking the coronavirus disease in Europe using the infodemiology technique. A related study was performed by Ayyoubzadeh S. M. et al. In their paper, they communicated the findings of COVID-19 applicable to Iran using data mining and deep learning techniques (10).

The present study is slightly different from other reported studies. It provides the initial description of critically infected COVID-19 patients in India. We achieved noteworthy results on around 14,06,848.0 diseased subjects. We observed that men were more prone to the infection with this virus than women. In both men and women, a protein called ACE2 (Angiotensin Converting Enzyme 2) is present in the lungs, heart and gastrointestinal tract in a larger amount. But not all tissues get affected by this virus. The ovarian tissue does not produce ACE2 protein while testicular tissue does that at a higher rate. This may form a possible correlate of the infection with male subjects.

We looked for an overall cumulative trend and then the weekly trend. All these analytical evaluations uncovered the hidden relationships amongst the various features present in the dataset. At a broader level, we worked on time-series data along with the other data. This helped us in understanding the underlying structure and function that generated the results. Through this, it was possible to explain the data in a manner that further assisted in predicting the growth rate. As of 12th July, the growth rate of confirmed cases was 3.49 percent, deceased cases as 2.49 percent while for recovered cases, the increment rate came out to be equal to 3.73 percent respectively. From this, we calculated the recovery rate and fatality rate. The mean value of recovery and mortality rate were found to be 38.74 and 1.81. When compared to the growth rate of April (30th April) in India, it was found that the growth rate of confirmed cases was 5.44 percent, deceased cases was 6.95 percent and recovered cases as 7.48 percent. Also, the mean value of recovery and mortality rate for April were found to be 18.59 and 2.14. From our results and the given data, we can say that a balance was maintained between these two. As compared to the other countries likewise Italy, Iran, USA, China, there was a huge rise in deaths along with the confirmed cases. These figures tell us that a huge number of individuals reported corona-like symptoms. Many Indians, after the imposition of lockdown, took heavy precautions. This is one of the main reasons for not the huge growth of COVID-19 in India whereas many countries failed to abide by this.

The next part of our study was building a severity model. This model was built employing ML and later a mathematical model was presented. The growth ratio and growth factor were calculated, which showed that a stabilization point was reached near to 1.0 in the growth factor. But growth ratio reaches closely to 1.0. We noticed a significant decrease in both of them. Because the growth factor stabilized at 1.0, the inflection point was reached. Using a logistic growth curve, the number of confirmed cases in the future was predicted. Along with this, this model gave an entire statistical report. The report gave the best parameter values for the predicted model. It gave 72.0 days as an inflection point, 32.0 percent as growth rate while the number of predicted confirmed cases in the future after 12th July 2020 was found to be 17,80,000.0. A strong positive correlation was witnessed between the growth factor and the predicted count of confirmed cases

Conclusion

We explored the role of empirical, analytical and mathematical modeling in the prediction of early trends and pattern recognition of the COVID-19. It demonstrates how all these applied tools and technologies can extract more information to make the required predictions. Even though certain aspects need to be accomplished and bolstered, the empirical analysis and predictive modelling reported in the present study is noteworthy. It is legitimate and effective to predict the pandemic state when the number of infectious subjects is bound to increase. This study delivers certain preliminary understanding and practices concerning the related features in COVID-19 infected subjects in India. Our study aids in building such a model where strong relationships were observed amongst the confirmed, recovered, active and deceased coronavirus cases in India.

Recommendation

Although, there is further scope of improvement in the model, nevertheless, our approach opens many possibilities for more advanced research related to the COVID-19 disease especially in the context of the diagnosis and formulating certain protocols to combat this infectious disease at the earliest possible.

Limitation of the study

With this modeling, there was still statistical uncertainty due to many reasons. The accomplished study has limitations. Because the pandemic is still in its growth state and has not attained any peak point, the sample size was limited. Secondly, a lot of missing data were present. The perceived correlations are centred on limited observations. These results are not static. They are subject to change as we move ahead and the days pass until a peak point is not reached or the pandemic completely stops spreading. At this stage, we can't settle for the predictions attained. But it will surely and positively aid once the spread of this virus finally comes to an end so that we could assemble more data and thus, we will be able to build more robust models.

Relevance of the study

The technological revolution has made an increasing use and application of computational processes. With time, we will achieve more accurate and significant results which will aid in better management decisions on the COVID-19 spread.

Authors Contribution

All authors contributed equally.

References

1. WHO Report, Coronavirus disease 2019 (COVID-19) Situation Report – 175 (2020). https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200713-covid-19-sitrep-175.pdf?sfvrsn=d6acef25_2 (Last accessed on 25.06.2021).
2. COVID-19 INDIA, Ministry of Health and Family Welfare Government of India https://www.mohfw.gov.in/.
3. "NIST/SEMATECH e-Handbook of Statistical Methods," 2012.
4. A. Khan and S. Zubair, "Longitudinal Magnetic Resonance Imaging as a Potential Correlate in the Diagnosis of Alzheimer Disease: Exploratory Data Analysis," JMIR Biomed. Eng.2020;5(1):1–13
5. D. C. Hoaglin et al., Understanding robust and exploratory data analysis. New York: John Wiley & Sons, 2000
6. A. Alimadadi, S. Aryal, I. Manandhar, P. B. Munroe, B. Joe, and X. Cheng, "Artificial Intelligence and Machine Learning to Fight COVID-19," Physiol. Genomics, 2020;52(4):200–202
7. Z. Allam, G. Dey, and D. S. Jones, "Artificial Intelligence (AI) Provided Early Detection of the Coronavirus (COVID-19) in China and Will Influence Future Urban Health Policy Internationally," Ai, 2020;1(2):156–165
8. Z. Yang et al., "Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions," J. Thorac. Dis., 2020;12(3):165–174
9. A. Mavragani, "Tracking COVID-19 in Europe: An Infodemiology Study," JMIR Public Heal. Surveill. 2020;6:1–13
10. S. M. Ayyoubzadeh, S. M. Ayyoubzadeh, H. Zahedi, M. Ahmadi, and S. R Niakan Kalhori, "Predicting COVID-19 Incidence Through Analysis of Google Trends Data in Iran: Data Mining and Deep Learning Pilot Study," JMIR Public Heal. Surveill., 2020;6(2):18828,.

Tables

TABLE 1 SUBJECTS' DEMOGRAPHIC VALUES

Factors	Values
Gender n (%)	
Men	27165 (67.0)
Women	13376 (33.0)
Transgender	7 (0.0)
Age (in years)	
Mean (SD)	35.40 (16.53)
Median	34.0

TABLE 2 CUMULATIVE REPORT FOR RECOVERY AND MORTALITY RATE

	Recovery Rate		Mortality Rate	
	Highest	Lowest	Highest	Lowest
Date	12th July	13th March	17th June	13th March
%	63.04	3.70	3.44	1.33
12th July	63.04		2.68	
Mean	38.74		1.81	

TABLE 3 CONFIRMED CASES: PREDICTED VS. ACTUAL

	Predicted (Confirmed Cases)	Actual (Confirmed Cases)
72.0 Days (3rd April)	4,874.0	2,653.0
85·0 Days (23rd April)	23,077.0	21,700.0
92.0 Days (30th April)	34,863.0	33,062.0
After 12th July	17,80,000.0	

Figures

FIGURE 1 TOTAL NUMBER OF COVID-19 INFECTED CASES REPORTED TILL DATE IN INDIA

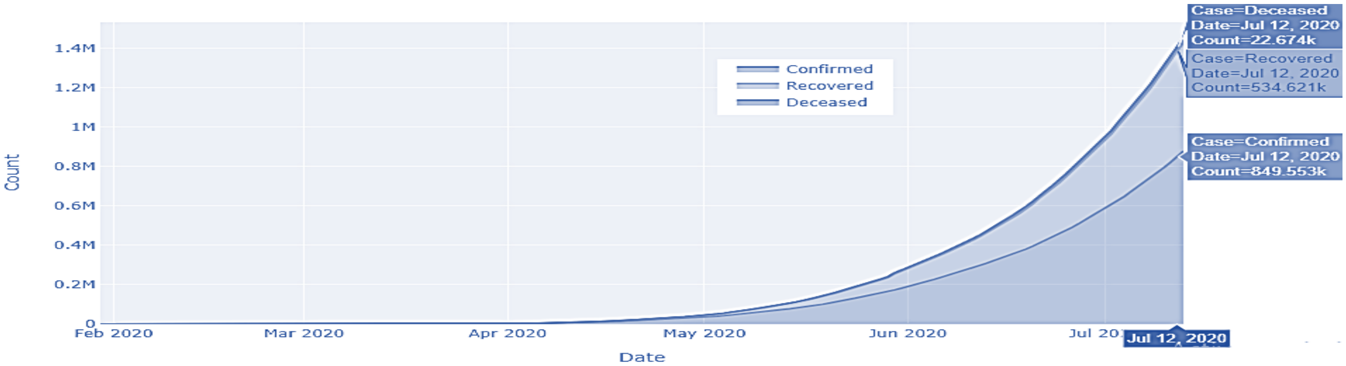


FIGURE 2 WEEKLY TREND IN INDIA FOR COVID-19 CONFIRMED, DECEASED AND RECOVERED CASES

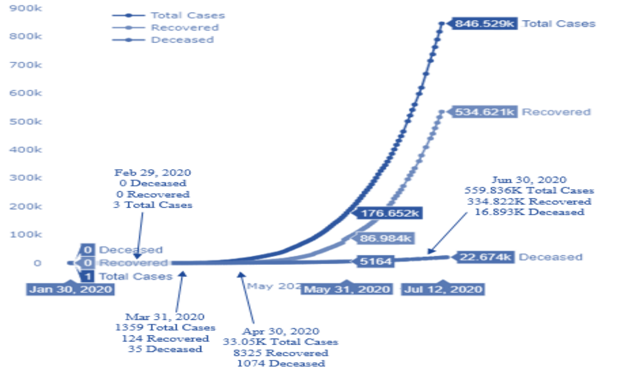


FIGURE 3 GROWTH RATIO

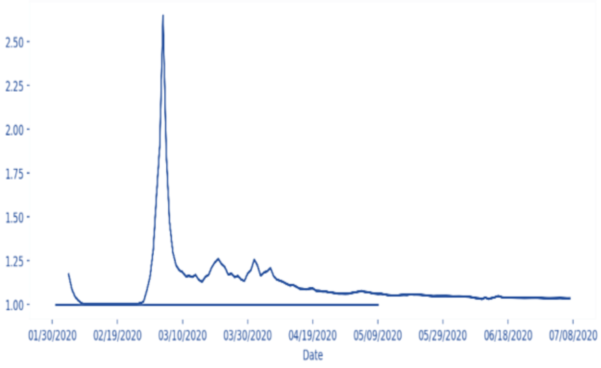


FIGURE 4 GROWTH FACTOR

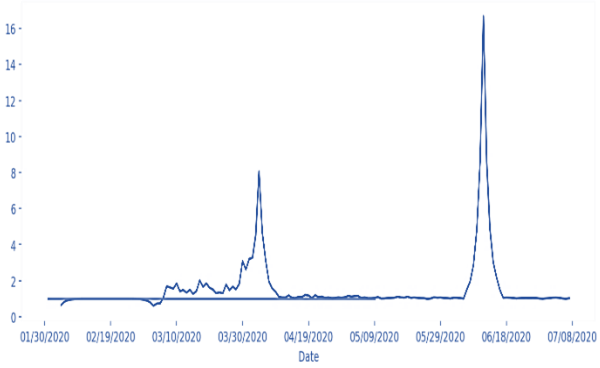
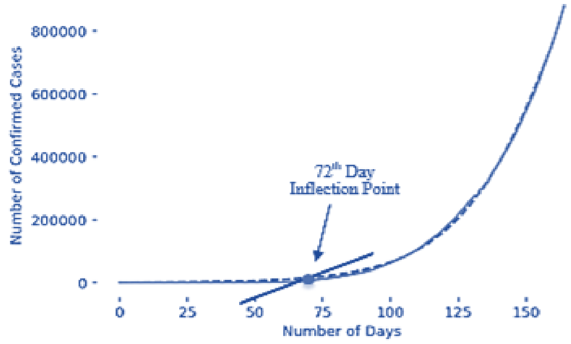


FIGURE 5 LOGISTIC GROWTH CURVE


Number of Confirmed Cases
72th Day
Inflection Point
Number of Days