JMIR Publications
The Leading eHealth Publisher

Severity Model Based Prediction of Early Trend and Pattern Recognition of the COVID-19 Infection in India: Exploratory Data Analysis and Machine Learning Study

Afreen Khan, Swaleha Zubair

Submitted to: Journal of Medical Internet Research
on: May 05, 2020

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Severity Model Based Prediction of Early Trend and Pattern Recognition of the COVID-19 Infection in India: Exploratory Data Analysis and Machine Learning Study

Afreen KhanMCA, ; Swaleha ZubairPhD,

Corresponding Author:
Swaleha ZubairPhD,
Phone: +919410059635
Email: swalehaowais123@gmail.com

Abstract

Objective: Recent Coronavirus Disease 2019 (COVID-19) pandemic has inflicted the whole world critically. Despite the fact that India has not been listed amongst the top ten highly affected countries, one cannot rule out COVID-19 associated complications in the near future. The accumulative testing facilities has resulted in exponential increase in COVID-19 infection cases. In figures, the number of positive cases have risen up to 33,614 as of 30 April, 2020. Keeping into consideration the serious consequences of pandemic, we aim to establish correlations between the numerous features which was acquired from the various Indian-based COVID datasets, and the impact of the containment of the pandemic on the current state of Indian population using machine learning approach. We aim to build the COVID-19 severity model employing logistic function which determines the inflection point and help in prediction of the future number of confirmed cases.

Methods: An empirical study was performed on the COVID-19 patient status in India. We performed the study commencing from 30 January, 2020 to 30 April, 2020 for the analysis. We applied the machine learning (ML) approach to gain the insights about COVID-19 incidences in India. Several diverse exploratory data analysis ML tools and techniques were applied to establish a correlation amongst the various features. Also, the acute stage of the disease was mapped in order to build a robust model.

Results: We collected five different datasets to execute the study. The data sets were integrated extract the essential details. We found that men were more prone to get infected of the coronavirus disease as compared to women. Also, the age group was the middle-young age of patients. On 92-days based analysis, we found a trending pattern of number of confirmed, recovered, deceased and active cases of COVID-19 in India. The as-developed growth model provided an inflection point of 85.0 days. It also predicted the number of confirmed cases as 48,958.0 in the future i.e. after 30th April. Growth rate of 13.06 percent was obtained. We achieved statistically significant correlations amongst growth rate and predicted COVID-19 confirmed cases.

Conclusion: This study demonstrated the effective application of exploratory data analysis and machine learning in building a mathematical severity model for COVID-19 in India.

(JMIR Preprints 05/05/2020:19872)
DOI: https://doi.org/10.2196/preprints.19872

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.
2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?
✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Severity Model Based Prediction of Early Trend and Pattern Recognition of the COVID-19 Infection in India: Exploratory Data Analysis and Machine Learning Study

Abstract

Objective: Recent Coronavirus Disease 2019 (COVID-19) pandemic has inflicted the whole world critically. Despite the fact that India has not been listed amongst the top ten highly affected countries, one cannot rule out COVID-19 associated complications in the near future. The accumulative testing facilities has resulted in exponential increase in COVID-19 infection cases. In figures, the number of positive cases have risen up to 33,614 as of 30 April, 2020. Keeping into consideration the serious consequences of pandemic, we aim to establish correlations between the numerous features which was acquired from the various Indian-based COVID datasets, and the impact of the containment of the pandemic on the current state of Indian population using machine learning approach. We aim to build the COVID-19 severity model employing logistic function which determines the inflection point and help in prediction of the future number of confirmed cases.

Methods: An empirical study was performed on the COVID-19 patient status in India. We performed the study commencing from 30 January, 2020 to 30 April, 2020 for the analysis. We applied the machine learning (ML) approach to gain the insights about COVID-19 incidences in India. Several diverse exploratory data analysis ML tools and techniques were applied to establish a correlation amongst the various features. Also, the acute stage of the disease was mapped in order to build a robust model.

Results: We collected five different datasets to execute the study. The data sets were integrated extract the essential details. We found that men were more prone to get infected of the coronavirus disease as compared to women. Also, the age group was the middle-young age of patients. On 92-days based analysis, we found a trending pattern of number of confirmed, recovered, deceased and active cases of COVID-19 in India. The as-developed growth model provided an inflection point of 85.0 days. It also predicted the number of confirmed cases as 48,958.0 in the future i.e. after 30th April. Growth rate of 13.06 percent was obtained. We achieved statistically significant correlations amongst growth rate and predicted COVID-19 confirmed cases.

Conclusion: This study demonstrated the effective application of exploratory data analysis and machine learning in building a mathematical severity model for COVID-19 in India.

Keywords: Coronavirus disease 2019; COVID-19; pandemic; modeling; data analysis; machine learning; India.

Introduction

Coronavirus disease 2019 (COVID-19) emerged as a serious health threat to the lives of many people around the globe. In 2020, this deadly disease has put the whole world on lock down forcing the people for social distancing and confinement inside their houses. The World Health Organization (WHO) announced the coronavirus outbreak as a state of Public Health Emergency of International Concern (PHEIC) on 30th January 2020. Further, the disease had been declared pandemic by the WHO in March 2020 [1]. In spite of the severity, the COVID-19 infection has only a 2 percent mortality rate. However, because of the lack of appropriate disease module, the onset of this disease has greater risk of the substantial injury to the alveolar and respiratory breakdown progressively that may eventually ensued in death of the concerned person [2].

According to WHO situation report-101, as of 30th April 2020, the virus had inflicted 30,90,445 people with around 2,17,769 deaths globally [3]. Whereas the South-East region (India belongs to this region) had around 54,021 infection cases with 2,088 deaths as of on 30th April 2020 [3]. In context of India, 32 states and union territories had reported coronavirus infection on April 30th, 2020. There has been a rise in the number of confirmed cases and deaths in India. From among the 9,02,654 subjects reported for the coronavirus symptoms, 33,614 were detected as positive while 9,074 patients recovered and 1,151 patients had died from this disease in India [4].

Since ages, new and emerging pathogens have caused a major health problem globally. On the positive note, the advancement in the technology has provided us potential for timely intervention [5]. Consecutively, this is true for virus-related infectious diseases that are swiftly transmitted and possess asymptomatic infection-period [6]. Thus, it becomes imperative that we build innovative models so as to regulate the fast spreading of the virus [7]. With increase in health-related data, the computer aided intervention in medical and other-related fields has been particularly transformed. This has emphasized on the ease to use numerous technologies like Machine Learning (ML), Artificial Intelligence, Big Data, etc. which can considerably help in analytical and predictive modeling of the disease [8].

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 chose a time period of 92-days for our analytical study, beginning from the first day when the coronavirus hit India i.e. from 30th January 2020 to 30th April 2020. We worked on statistical data analysis, exploratory data analysis, machine learning and mathematical modeling to carry out the present study. These tools and techniques have been applied for the empirical analysis and predictive modeling of COVID-19 for Indian cases. These techniques are much more robust and efficient than the traditional analysis methods. Subsequently, we built a prediction model using logistic function to establish the growth ratio in India to determine the inflection point within the gathered data.

We found very encouraging results that establish the effectiveness and potential of machine learning, more specifically, exploratory data analysis in analyzing the COVID-19 related cases. Although, there is further scope of improvement in the model, nevertheless, our approach open many possibilities for more advanced research related to the COVID-19 disease epically in context of the diagnosis and formulating certain protocols so as to combat with this infectious disease at the earliest possible.

Methods

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 17,656.0 subjects' data between 30th January to 30th April 2020, total of 92-days (i.e. 3 months 1 day or 13 weeks 1 day) period for the analysis.

The samples used in the study were obtained and integrated from various data consortium websites. The complete dataset description is illustrated in Table 1.

Table 1.

S.No.	Dataset	Number of	Total number	Description	Source

		features	of entries		
1	COVID-19 World Data	6	20,252	It includes worldwide data of number of affected, deceased and recovered cases day-wise. It is a time-series data.	https://github.com/CSSEGISandData/COVID-19
2	COVID-19 Indian Data	8	1,464	It comprises of daily data of number of confirmed, dead and recovered Indian cases. It is a time-series data.	https://www.mohfw.gov.in/
3	Patient-wise Details (India)	20	27,891	A complete subject-wise data of all the patients who reported for the symptoms of coronavirus.	https://www.covid19india.org/
4	ICMR[a] Testing Details	4	42	Total number of COVID-19 tests (sample tests, individual tests and positive tests) that have been performed at daily level.	https://icmr.nic.in/content/covid-19
5	ICMR[a] Testing Laboratories	6	267	Details of testing laboratories across Indian states and union territories.	https://icmr.nic.in/content/covid-19

[a]: Indian Council of Medical Research

Experimental Frame

Machine Learning (ML) is a mathematical description of real-life practices [9]. It is a mode of data analysis method which further employs different algorithms to learn from the data. With ML tools and techniques, valuable patterns are identified within the voluminous data. Thereby, by analyzing the data, from the perspective of ML modeling, we can look into the insights and discover the hidden trends and patterns, explore clusters within the data, if present any, establish the correlation amongst the different variables present in the dataset and can locate if they are linearly separable or not [10]. Thus, this predominant part of ML is known as Exploratory Data Analysis (EDA)**.**

We divided our study in 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. The statistical analysis is a type of analysis which includes collecting, uncovering and presenting the big data to find the underlying patterns [11]. It is necessary for formulating data-centric decisions. While, an EDA is another type of analysis which examine and evaluate the datasets and thus, summarizes their essential properties [10]. It is a graphical and visual approach and is not similar to statistical visualization. The prior one emphases only on single data characterization part while the subsequent one covers a larger aspect. The severity model was built using logistic function. Through this, inflection point and the future prediction of the number of confirmed cases was achieved positively.

In the present study, the empirical analysis performed on the coronavirus data was achieved using Python libraries operated on the Jupyter platform of Anaconda Navigator. This platform presents a well-defined framework for programmers to process and develop the models. The univariate and multivariate analysis performed was accomplished employing bunch of scikit-learn libraries.

Data Analysis and Major Findings

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

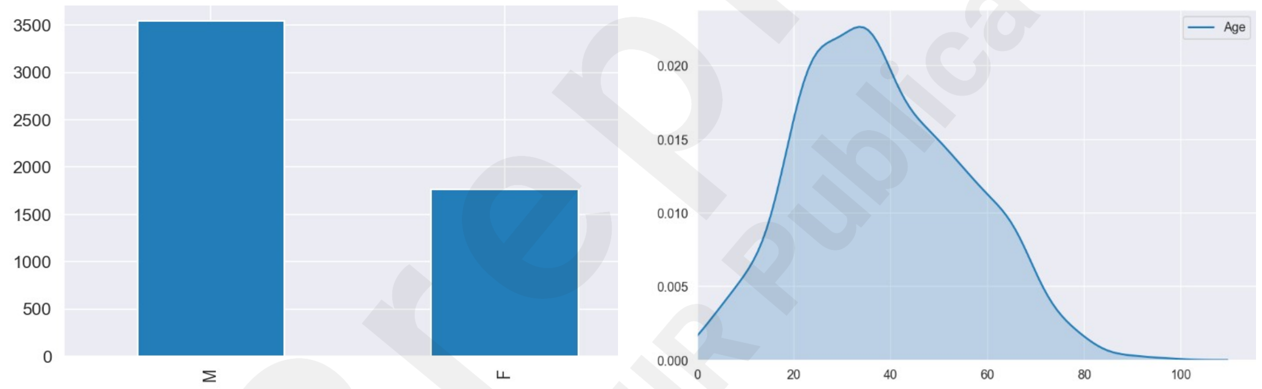
Statistical Analysis

In the gathered data, age and gender features consisted of certain 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 2.

Table 2.

Factors	Values
Gender n (%)	
Men	3547 (66.8)
Women	1766 (33.2)
Age (in years)	
Mean (SD)	38.48 (17.25)
Median	36.0

The graphs for both age and gender are plotted in the Figure 1.



(a) Gender (b) Age

Figure 1. Subject Distribution

From Table 1 and Figure 1, it can be inferred that amongst the COVID-19 patients, men were found to be greater in number as compared to women. The range of age was found to be between 0 years to 98 years.

Exploratory Data Analysis

Variety of data employed in this study was obtained by integrating several datasets, as mentioned above in Section 2. Before applying EDA, to establish correlation amongst the various features, it was necessary to understand the datasets correctly because of the integration of many datasets. As not all detail was present in one dataset, we sought to combine in such a way so that we could gain maximum insights and build a strong and robust severity model at the end. In the below sub-sections, we present the results of the applied EDA.

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 did the analysis so as to arrive at better results.

Age by Gender Distribution of the Confirmed subjects

Figure 2 shows the age distribution of the infected and confirmed patients with respect to the number of male and female subjects. It can be seen that the males are more likely to be under the effect of coronavirus than the females. Within this, the men's age group of 30-50 years are more likely to be infected.

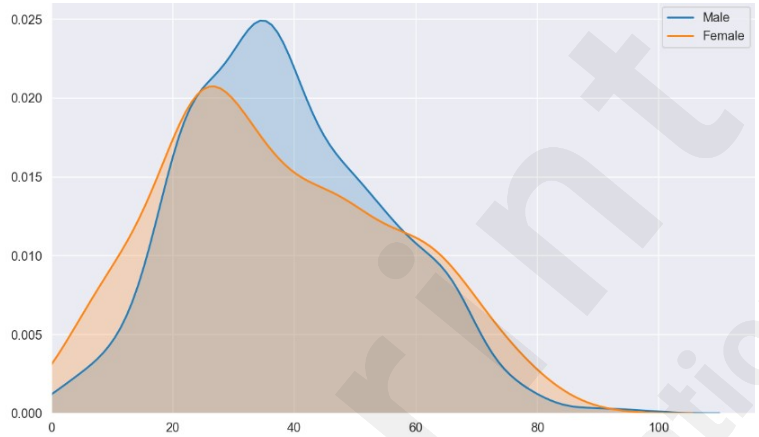


Figure 2. Age Distribution of Confirmed COVID-19 patients with respect to count of male and female subjects'

Total Number of Infected cases

As of 30th April 2020, the number of confirmed, deceased and recovered coronoavirus patients belonging to all the Indian states and union territories is plotted in Figure 3.

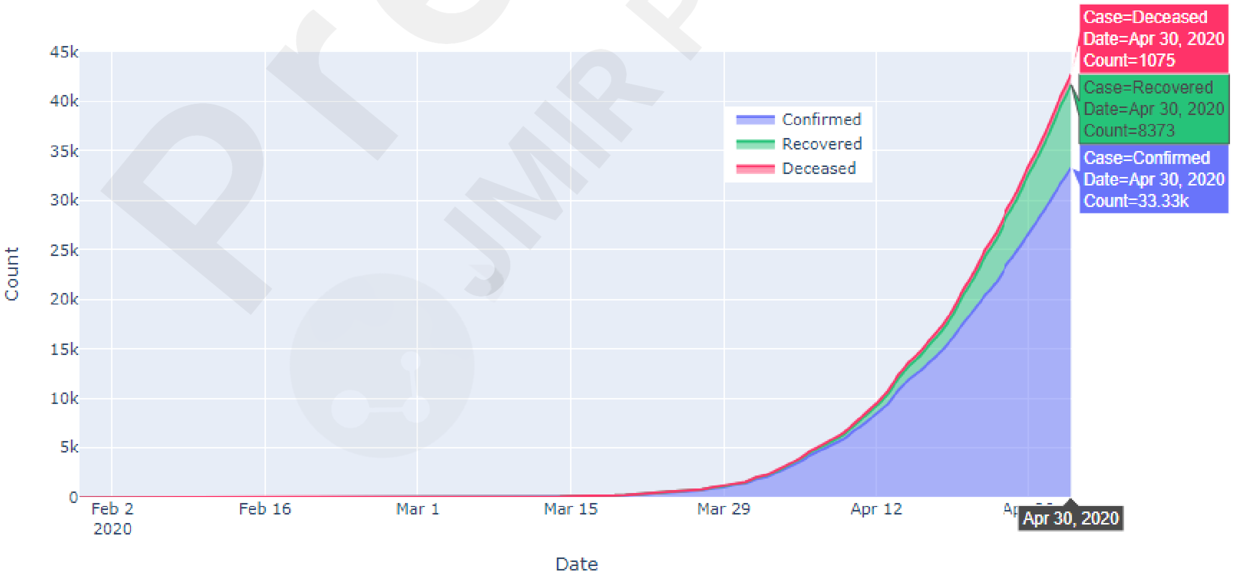


Figure 3. Total number of COVID-19 infected cases reported till date in India

For all the 92-days period, starting from 30th January to 30th April 2020, Figure 3 displays a rising graph. Beginning from a single confirmed patient (on 30th January 2020) mounting to a 33,330.0 confirmed patients (30th April 2020). We got to see an extreme rise of infected patients after 30th

March. The total number of deceased patients were found to be 1,075.0 while recovered patients as 8,373.0. The plot for all the three goes in parallel, with a rise in all of the three.
In India, complete lockdown was imposed from 24th of March till 15th April. Because of the huge escalation in daily number of infected patients, the Indian government extended the lockdown till 17th of May 2020.

Day-wise Count

Figure 4 gives an overview of the overall situation of India, starting from the day when the first coronavirus case was reported uptill 30th of April 2020. The zero deaths was uptill 12th March, it was when on 13th March the first death was reported. Furthermore, the first patient was recovered on 3rd March 2020; even now the patients are recovering on a daily basis, which shows a trending effect. Amongst the 92 days, the highest number of confirmed cases were noted on 30th April i.e. 1,801 in numbers. Also, 75 deceased and 631 recovered cases were reported on this day. This time-series shows the trends in counts of number of days over time in India.

If we look into the increment rate of confirmed, deceased and recovered cases, it was found as 5.44%, 6.95% and 7.48% respectively as on 30th April. While 10 days prior to this, a compartive analysis showed that the the increment rate of confirmed, deceased and recovered cases was 8.72%, 7.15% and 19.0% respectively (on 20th April).

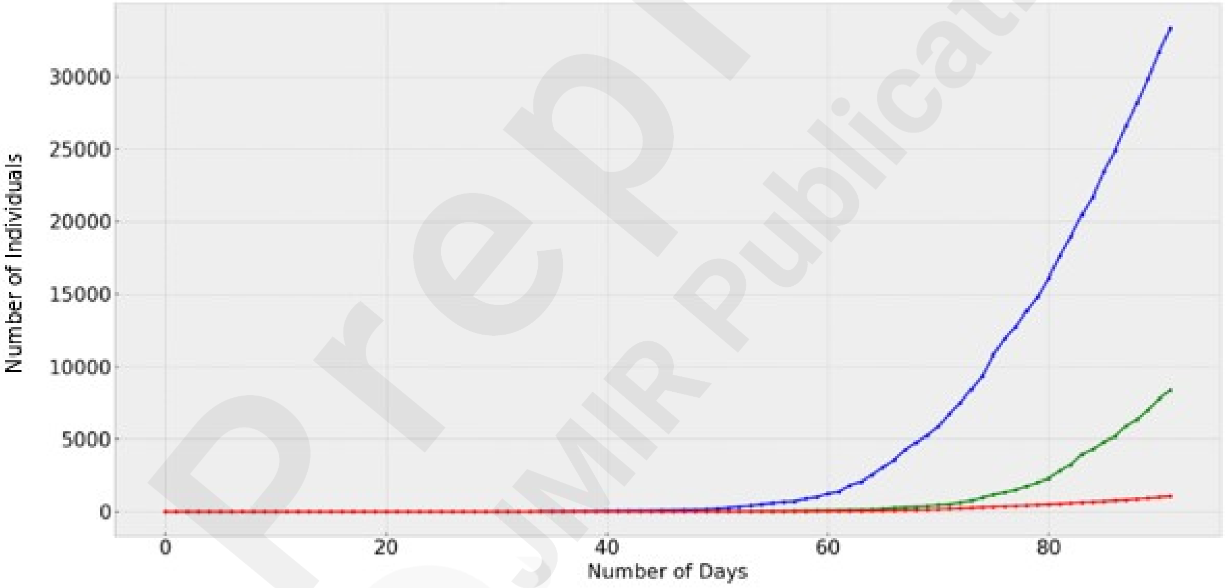


Figure 4. Time-series Plot for Day-wise Infections

Overall Cumulative Scale

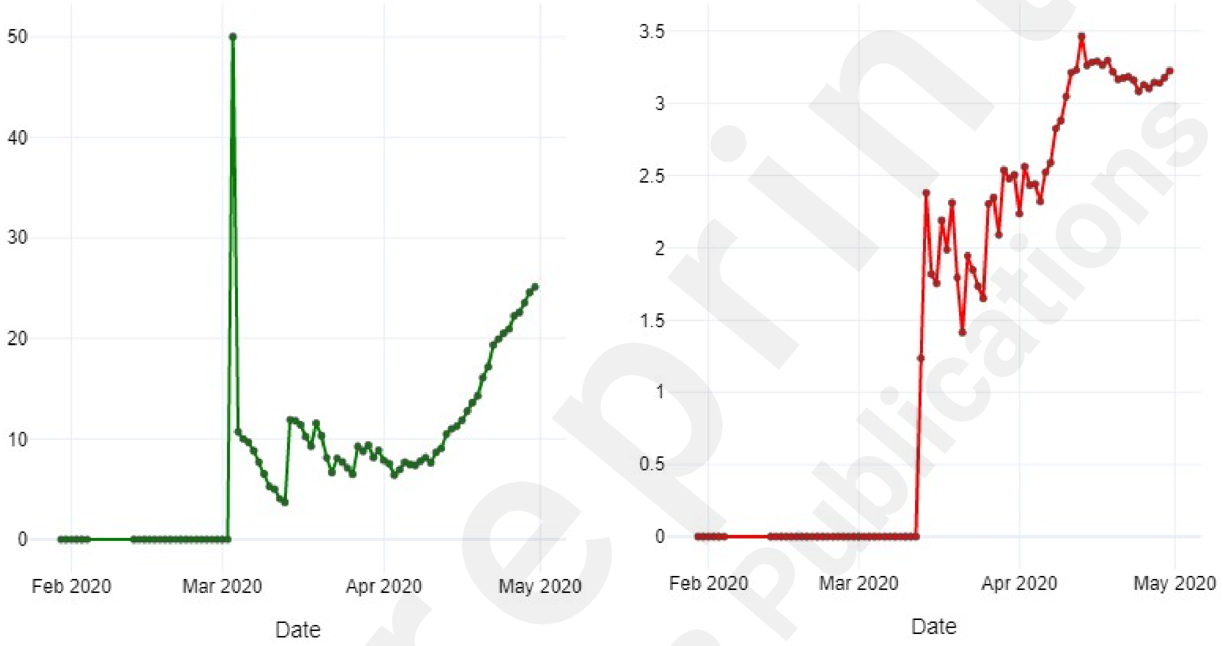
The cumulative recovery rate and mortality rate of India is reported in Table 3. Up until 2nd March, recovery rate was 0%. And, till 11th March, mortality rate was found to be 0%.

Table 3.

	Recovery Rate		**Mortality Rate**	
	Highest	**Lowest**	**Highest**	**Lowest**
Date	3rd March	13th March	13th April	13th March

%	50	3.70	3.46	1.23
30th April	25.12		3.22	
Mean	18.59		2.14	

The plotted graph for the above findings can be seen from the Figure 5. For both cumulative attributes, the scale (y-axis) is different. We got to see that mortality rate start to rise after 10th of March and a heavy drop in the recovery rate after 5th March. The reason behind this is that there was a sudden upsurge of cases in India after this period, more specifically in the beginning of April. The recovery rate again started to rise in the month of April and mortality rate started to fall which gave a notable indication pertaining to coronavirus.



(a) Recovery Rate (%) (b) Mortality Rate (%)

Figure 5. Recovery rate and Mortality rate

Weekly Trend

An outline of the weekly trend has been shown in Figure 6. It provides a general notion of the week-wise trend in India. The data show the number of confirmed, deceased and recovered cases on three time points i.e. 30th January, 29th Februaury, 30th March and 30th April respectively. As again, it gives a wholesome picture of the upsurge in cases. There were zero deaths and zero recovery of the patients up until 20th February.

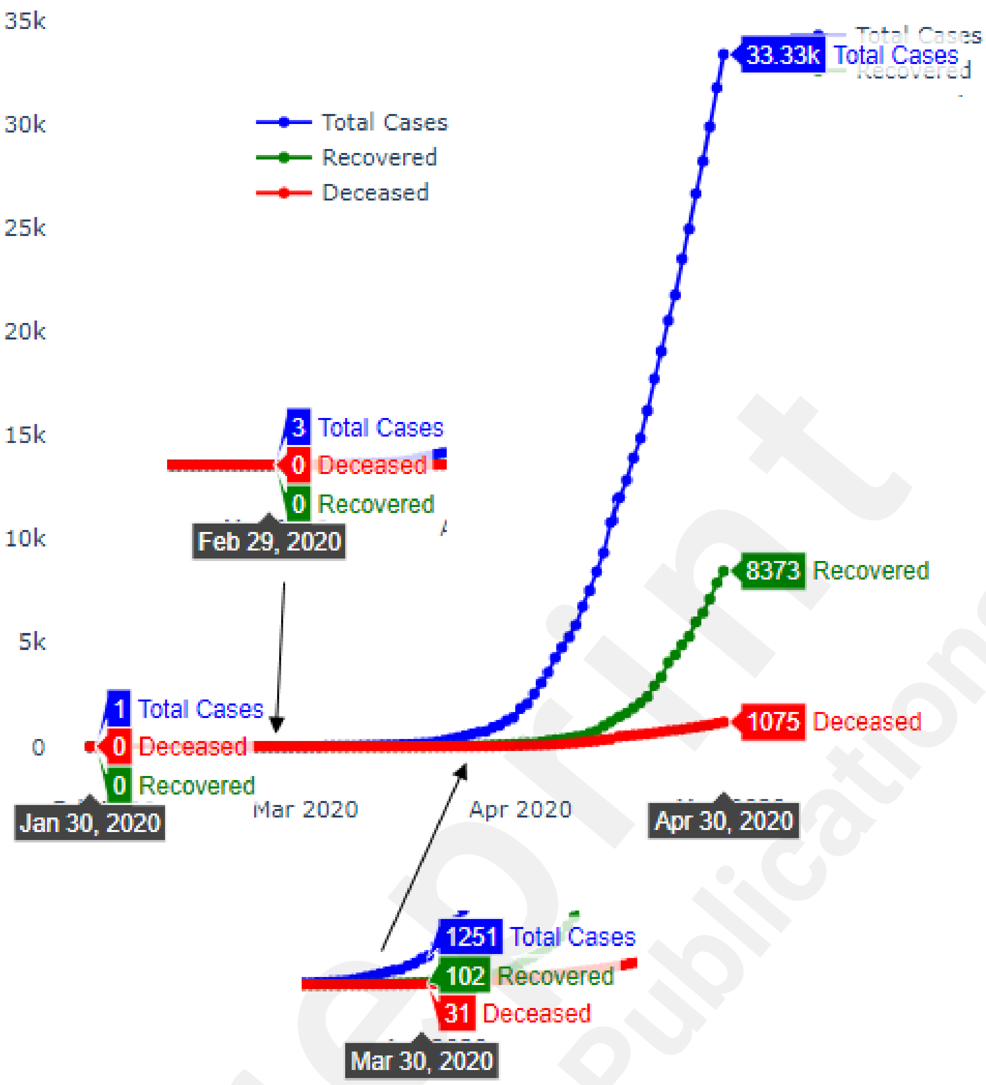


Figure 6. Pictorial representation of weekly trend in India for COVID-19 confirmed, deceased and recovered cases

The day-wise COVID-19 incident trend in India in lesser time. Even after the lockdown was imposed, the number of cases kept rising. Comparatively, this upsurge is still lesser as compared to the highly-hit countries like, USA, Italy and China etc.

COVID-19 subject-wise analysis

Table 4.

Infected with COVID-19	Minimum Age	Lower Fence	Quartiles			Upper Fence	Maximum Age
			25%	50%	75%		
Recovered	1	-	24.5	38	55	-	96
Hospitalized	1	-	25	36	50	85	98

Deceased	1	22	50	65	70	-	85

From the Table 4, it can be deduced that the majority of hospitalized patients belong to an age group of 25-50 years. The recovered patients were found to remain in the age-group of 24.5-55 years while the deceased patients were noticed to lie in a range of 50-70 years of age. For 30th April in particular, the data was explored and it was found that the count of hospitalized, recovered and deceased patients were found to be 1801, 631 and 75 respectively. Figure 7 depicts the count of hospitalized, recovered and deceased patients with respect to their age groups, as on 30th April. From the graph, it is inferred that the highest number of hospitalized patients that were confirmed with corona-positive belonged to an age group of 34-35.9.

The upper fence and lower fence are required for spotting extreme values present in the end of the data distribution. Table 4 and Figure 7 tells us about the existing upper fence and lower fence in the data, which in turn tells us about the outliers present. There is a lower fence of value 22 in the deceased plot which signifies it as a mild outlier. While an upper fence value of 85 in the hospitalized plot, stands out as an extreme outlier.

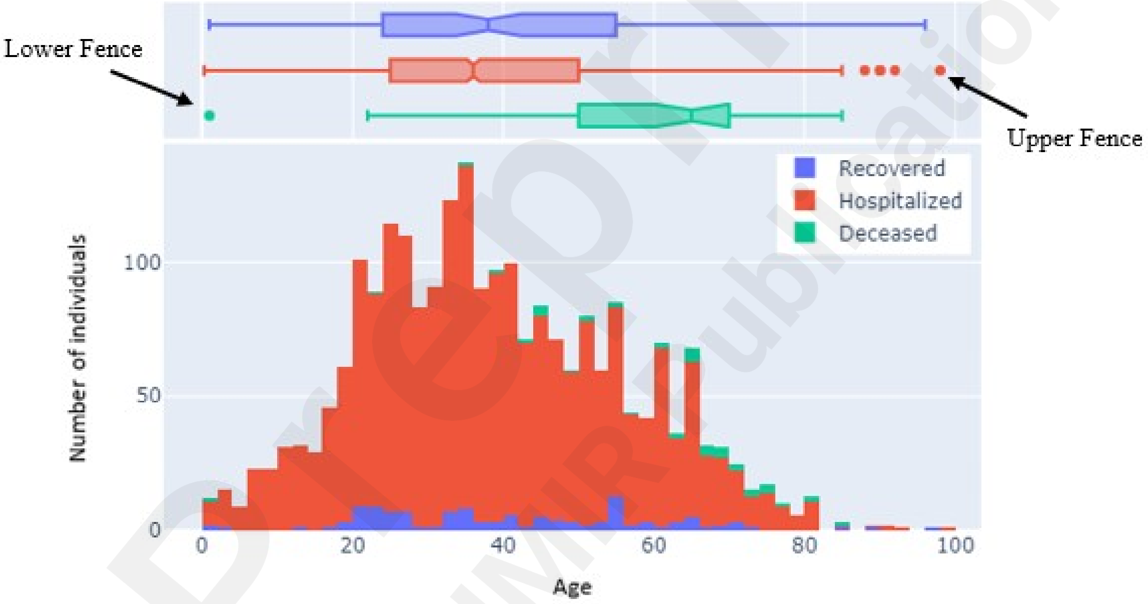


Figure 7. Number of reported individuals for recovered, hospitalized and deceased cases with respect to age class.

State-wise Analysis

Out of the 28 Indian states and 8 union territories (UTs), we found that only 33 of them were affected (from 15th April to 19th April). While the others remained unaffected from the coronavirus. From 30th January till 1st March, only one state (Kerala) remained affected. After this i.e. from 2nd March the COVID cases started to rise and reached to a total of 33,330 confirmed cases, affecting 32 Indian states and UTs by 30th April. The complete analytical graph is elucidated in Figure 8.

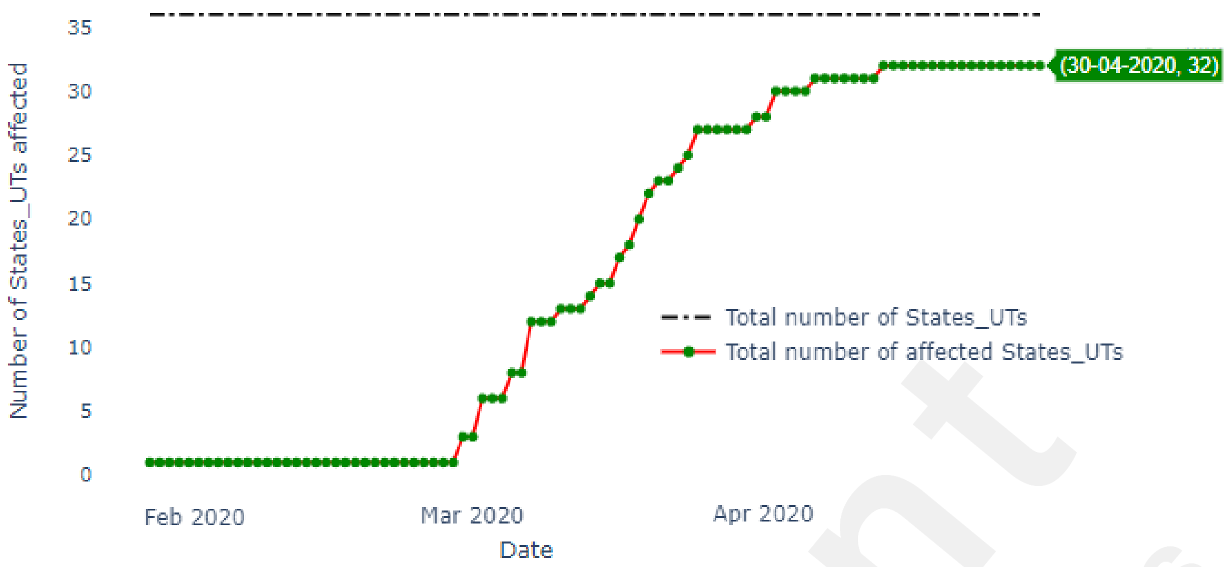


Figure 8. Count of Affected Indian states and Union Territories

Out of these affected states and UTs, we only show the latest trend of top ten which remain under influence of coronavirus as on 30th April (Table 5).

Table 5.

S.No.	State/Union_Territory	Confirmed (C)	Recovered (R)	Deceased (D)	Active (C-R-D)
1	Maharashtra	9915	1593	432	7890
2	Gujarat	4082	527	197	3358
3	Delhi	3439	1092	56	2291
4	Madhya Pradesh	2561	461	129	1971
5	Rajasthan	2438	768	51	1777
6	Tamil Nadu	2162	1210	27	1111
7	Uttar Pradesh	2134	510	39	1589
8	Andhra Pradesh	1332	287	31	1014
9	Telangana	1012	367	26	739
10	West Bengal	758	124	22	612

The graph was plotted for each of the confirmed, recovered, deceased and active cases for these ten states (Figure 9).

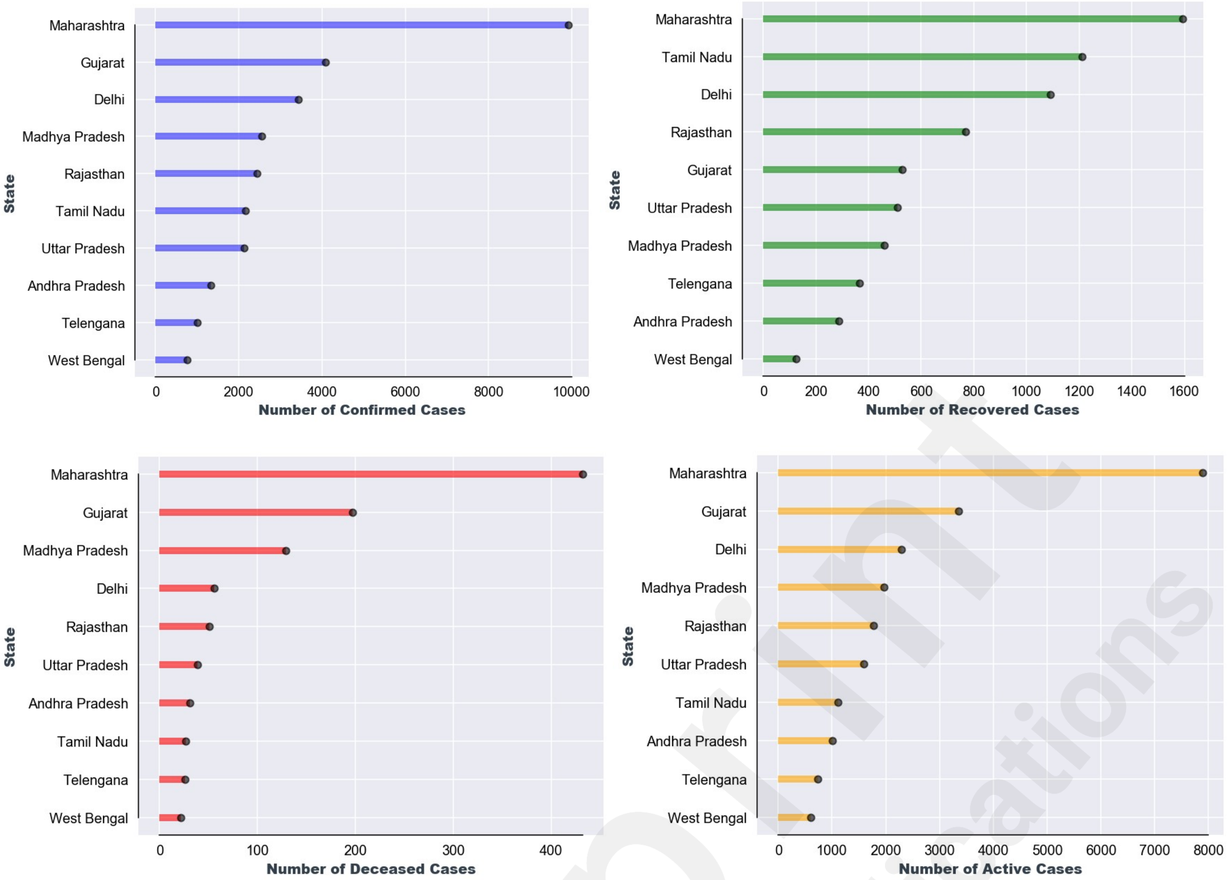


Figure 9. Top Ten Indian States affected by COVID-19

From the above plots, it can be deduced that Maharashtra remained highest in all of the four cases. The first case appeared in Maharashtra was on 9th March, with only 2 confirmed patients. Gradually, it emerged to 4,203 cases on 20th April. Comparably, we can see that West Bengal remained least in this list. Until 20th April, Kerala was at the 10th position in the top ten list. As on 30th April, Kerala was not amongst this list, which is a good indication despite of the fact that the first coronavirus case appeared in India was in the state of Kerala only. On 30th January, first confirmed case emerged and lasting to only 3 cases by the end of 29th February. The deceased rate is low and the recovery rate is high. Consequently, only 612 active cases was found at the end of 30th April, with a falling graph in number of confirmed cases in West Bengal.

COVID-19 Test Centre-wise analysis

ICMR has assigned test centres in the various cities of all the Indian states and UTs. The highest number of test centres is in Maharashtra, as can be seen from the Figure 10.

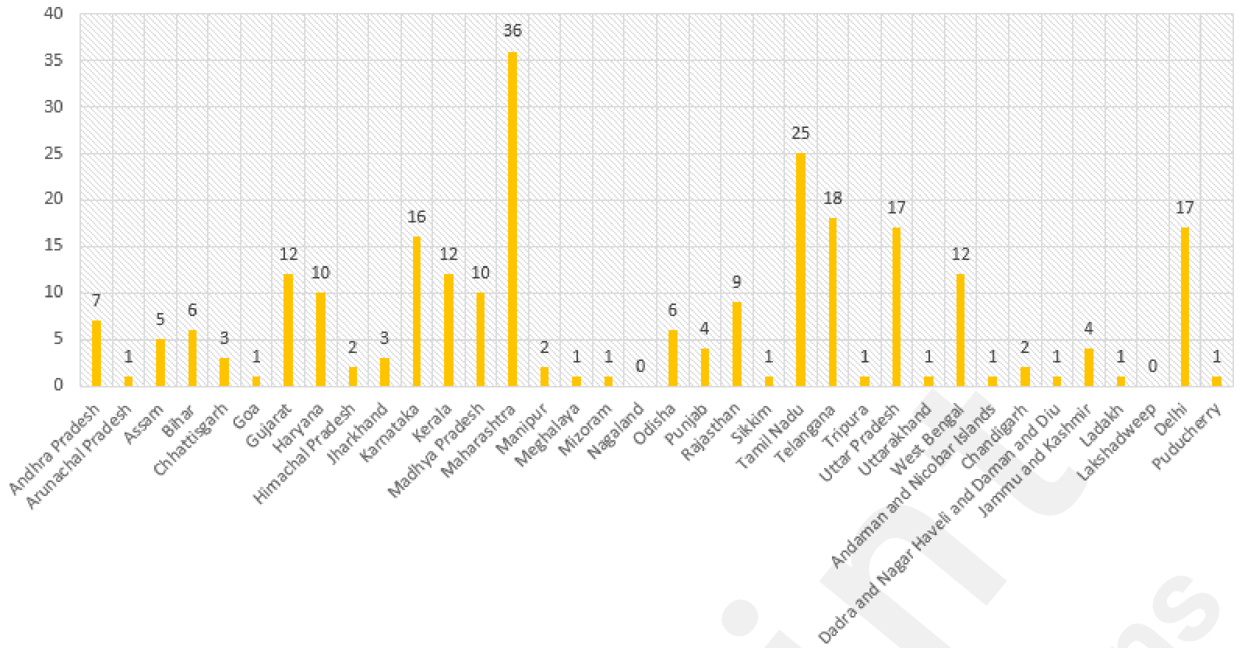


Figure 10. Count of COVID-19 Testing Laboratories within India

These labs tested a total of 9,02,654 of total samples by the end of 30th April country-wise. Also, the number of individual cases and total positive cases within the samples tested were discovered to be 8,690,40 and 33,614. The details provided by ICMR is from 13th of March, the details before this date were not available in the dataset. On 13th March, the total number of samples tested were 6,500. Out of this, the total number of individual cases and positive cases were only 5,900 and 78 respectively. After this a heavy increment in the number of confirmed cases appeared; the rising graph can be seen in the Figure 11. Figure 12 depicts the total positive cases ot of the total individuals tested. The graphs are plotted till the 27th April. The data was not available after this date.

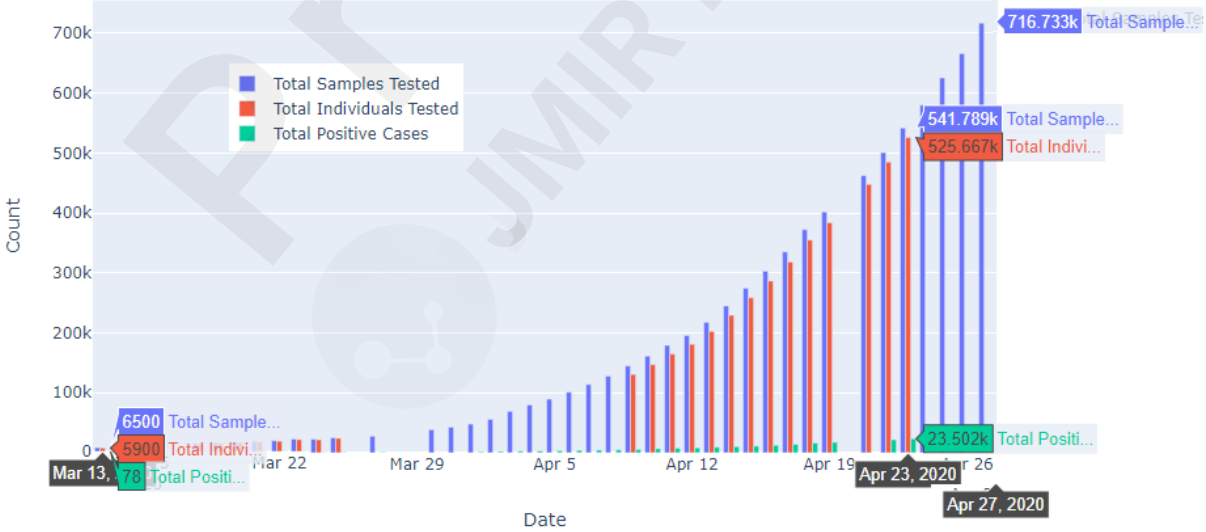


Figure 11. Total number of samples tests conducted for COVID-19

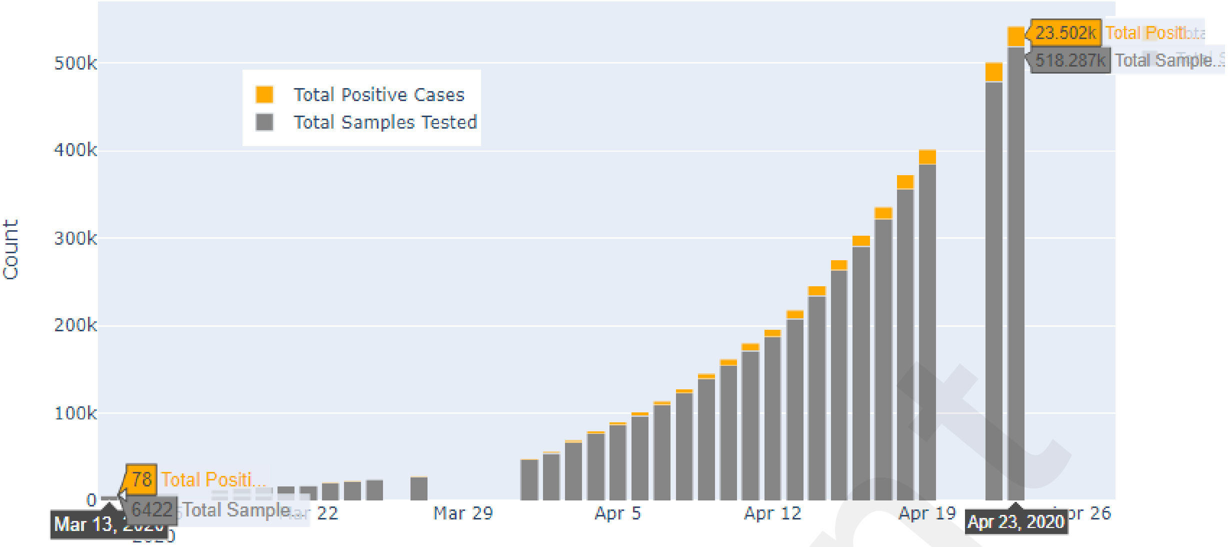


Figure 12. Total positive cases out of total samples tested

With the details provided, we discovered that the highest number of hospitals is in Uttar Pradesh amongst all the 36 Indian states and UTs. There are 3,277 primary health centres, 671 community health centres while 174 district hospitals. In this state, the hospital beds per 1000 person was found to be only 0.67. This figure is very low because of the huge population of Uttar Pradesh.

In contrast to this, we saw that Lakshadweep (a UT) has the highest number of hospitals per 1000 person i.e. 8.53. This is because it has less poulation with only 4 primary health centres.

Severity Model for COVID-19: Predicting with Logistic Function

With the escalating graph of all the active, confirmed, deceased and recovered coronavirus subjects, Figure 13 gives us a broader view of the 92-days analysis we performed on the Indian COVID data.

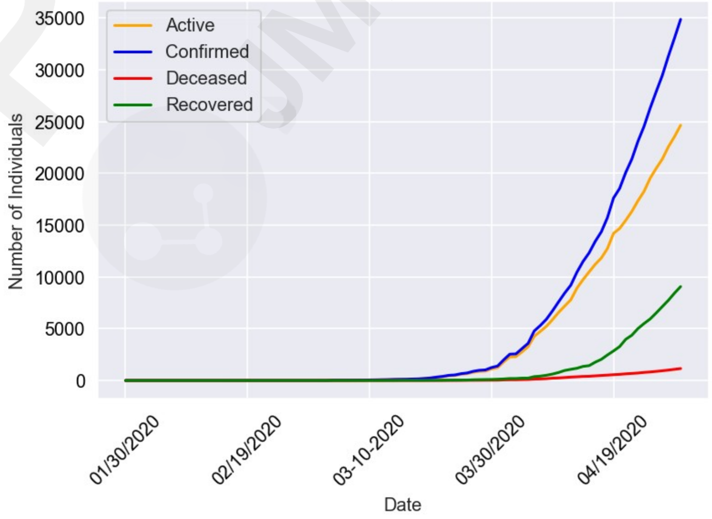


Figure 13. 92-days plot

Through this visualization, we further move towards building a severity model for COVID-19, a mathematical representation of all the empirical analysis done. As we have seen in the previous section that the state of art of Indian population got to see a sudden rise in the month of April despite of the fact that a complete lockdown was imposed from 24th of March, 2020 in India. Thus, in this section we dig deeper to look into the growth metrics with respect to the confirmed coronavirus subjects within the chosen timeperiod. These growth functions were plotted employing the logistic function. Later in this study, inflection point of India was predicted using machine learning approach.

The scatterness of any contagious disease can be demonstrated by making use of logistic function. Herein, the progression of a disease starts exponentially but after some time, it slows down. The position where it slows down, that particular point is called as an inflection point. The main task of this study is to gain the 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 the following formula:

$$Growth\,Ratio(d) = \frac{Total\,number\,of\,confirmed\,cases(d)}{Total\,number\,of\,confirmed\,cases(d-1)} \qquad (1)$$

where, d = nth day and d-1 = (n-1)th day

The plotted graph for equation (1) is illustrated in Figure 14. 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 in between 29th Februaury and 10th April. After that, it start declining till the end date.

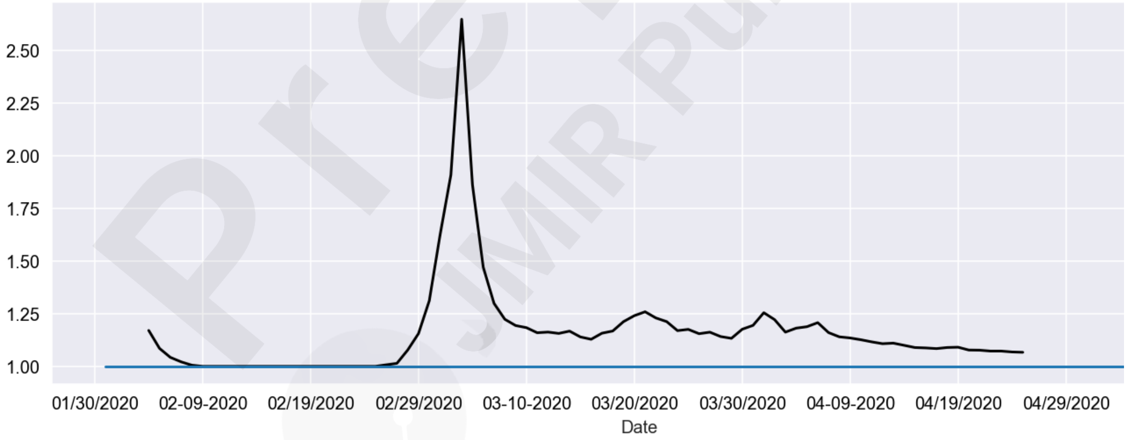


Figure 14. Growth Ratio

2. **Growth Factor:**

It is calculated by the following formula:

$$Growth\,Factor(d) = \frac{Total\,number\,of\,confirmed\,cases(d) - Total\,number\,of\,confirmed\,cases(d-1)}{Total\,number\,of\,confirmed\,cases(d-1) - Total\,number\,of\,confirmed\,cases(d-2)}$$

(2)

where, d = nth day

d-1 = n-1th day
d-2 = n-2th day

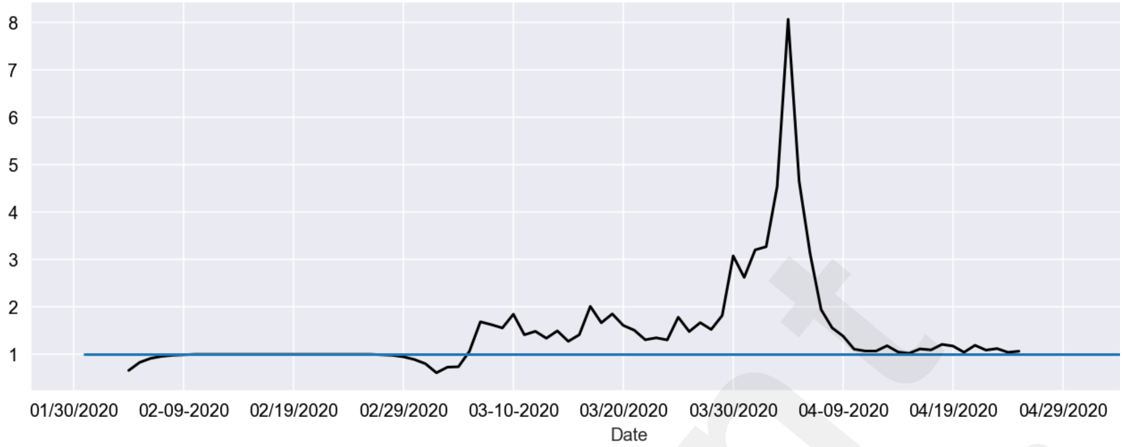


Figure 15. Growth Factor

It can be seen from Figure 15, that the peak point is reached around 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 inflection point.

From Figure 14 and 15, we can see that there is a remarkable reduction of both growth ratio and growth factor. The growth factor stabilizes near to 1.0 while 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 16.

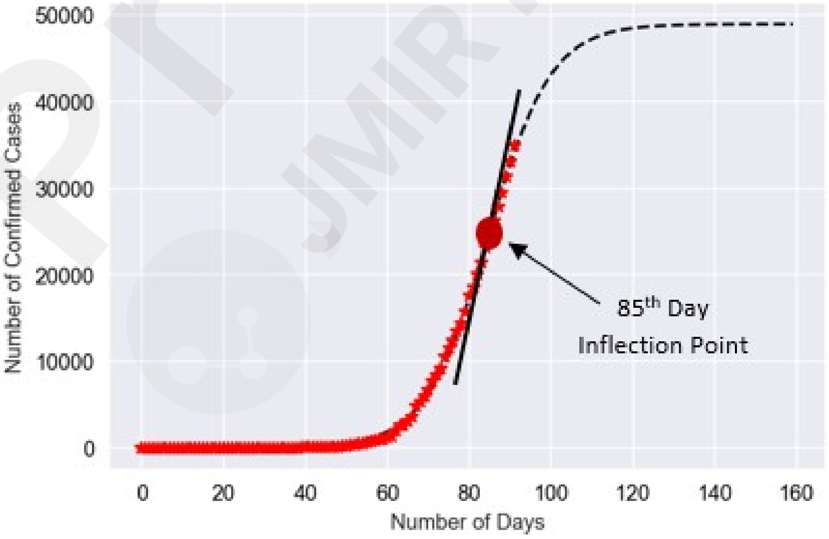


Figure 16. Logistic Growth Curve

(‘---’) represents Fitted line (‘***’) represents the Confirmed data.

The number of confirmed coronavirus cases in India were fitted by the logistic curve, as shown in

Figure 16 and further predicted the new cases in the coming days.

From the statistical report, it was concluded that the inflection point is hit around 85.0 days. The growth rate is 13.1 %. And the number of confirmed cases will maximize at around 48,958.0 cases in the coming days (after 30th April). This means that the given model predicts for the new confirmed cases in the upcoming days. The more the number of data, the better will be our predictions. Also, a strong correlation has been observed between these two i.e. a value of 0.991. Table 6 shows the combined results of the complete growth analysis performed.

Table 6. Confirmed Cases: Predicted vs. Actual

	Predicted (Confirmed Cases)	**Actual (Confirmed Cases)**
85·0 Days (23rd April)	23,077.0	21,700.0
92.0 Days (30th April)	34,863.0	33,062.0
After 20th April (Growth curve's maximum value)	48,958.0	

5. Discussion

EDA is a powerful and crucial step in the field of machine learning. It provides several insights which help in creating strong correlations amongst various features and modeling a complete prediction system. In context of health-related data, it becomes necessary to recognize the patterns within the dataset so as to build a robust working model [12]. In fact, EDA's chief objective is to design a parsimonious prototype i.e. a predictive model that could perform well with less number of features [13].

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 helps in gaining the maximum perception to understand the data-structure in effective manner. Thirdly, how COVID-19 has affected Indian population up until now and how it can affect in the near future. Lastly, it also helps 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 [14]. 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 [15]. 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 [16]. Mavragani A. applied the infodemiology (information epidemiology) approach which employs web-based data to notify public health and policymaking [17]. 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. 13. In their paper, they communicated the findings of COVID-19 applicable to Iran using the data mining and deep learning techniques [18].

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 33,330 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 (Angiostensin Converting Enzyme 2) is present in the lungs, heart and gastrointestinal tract in larger amount. But not all tissues get affected from 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.

For 92-days analysis, we found an escalating graph of the number of confirmed coronavirus cases in India. By the end of 30th April 2020, we discovered 33,330.0 confirmed cases, 8,373 recovered cases and 1075 deaths. Through this huge upsurge, we got to understand an informative trend and pattern in India. The systematic analysis may help us to take further precautions so as to combat with any disease or epidemic at the earliest.

We looked for an overall cumulative trend, day-wise trend, and then the weekly trend. All these analytical evaluation, 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 on 30th April, the growth rate of confirmed cases was 5.44 percent, deceased cases as 6.95 percent while for recovered cases, the increment rate came out to be equal to 7.48 percent respectively. From this, we calculated the recovery rate and fatality rate. The mean value of recovery and mortality rate 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.

A patient-wise analysis gave a complete picture of the hospitalized, deceased and recovered cases according to their age group. We noted that young to middle-aged men and women were more tested for coronavirus than the other age group. The COVID-19 dataset which contained the patient details included ‘travel history’ feature also. As already stated in the above sections, that many of the patient details were missing. Out of the details present, we discovered that 44.2 percent of subjects travelled in the past. While 26.5 percent revealed no travel history. These subjects were having the local transmission reasons other than that of travelling.

Among the 36 Indian states and UTs, we observed that only 32 states and UTs were affected from the virus spread. Moreover, Maharashtra was highly affected with 9,915 confirmed cases as on 30th April. In fact, Maharashtra remained at the highest position for all the four cases i.e. confirmed, recovered, deceased and active. Whereas Mizoram was at the least position with only 1 confirmed case.

We performed empirical tests on ICMR sample-wise data and state-wise laboratory data. As on 30th April, 9,02,654 samples have been tested for this virus. Out of these many samples, 34,863 individuals tested positive while the rest of them tested negative. Out of the details present, on 23rd April, the percentage of positive cases with respect to the total samples tested was 4.53%. While the percentage of positive cases with respect to the total samples tested was only 1.2% on 13th April. These figures tells us that the huge number of individuals reported for corona-like symptoms. Many Indians, after the imposition of lockdown, took heavy precautions. This is one of the main reasons of not a huge growth of COVID-19 in India whereas many countries failed to abide to 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 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, inflection point was reached. Using logistic growth curve, the number of confirmed cases in the future were predicted. Along with this, this model gave an entire statistical report. The report gave the best parameter values for the predicted model. It gave 85.0 days as inflection point, 13.1 percent as growth rate while number of predicted confirmed cases in the near future after 30th April 2020 were found to be 48,958.0. A strong positive correlation was witnessed between growth factor and the predicted count of confirmed cases.

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, lot of missing data was 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 to 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, therefore we will be able to build more precise machine learning and deep learning models.

6. Conclusion

We have explored the role of machine learning, exploratory data analysis and mathematical modeling in the prediction of early trend and pattern recognition of the COVID-19. It demonstrates how all ML and other 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. The technological revolution with machine learning 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 pertaining to the COVID-19 spread.

References

[1] Giovanetti M, Benvenuto D, Angeletti S, Ciccozzi M. The first two cases of 2019- nCoV in Italy: Where they come from? J Med Virol 2020 May;92(5):518-521. [doi: 10.1002/jmv.25699] [Medline: 32022275].

[2] Xu Z, Shi L, Wang Y et al (2020) Pathological findings of COVID-19 associated with acute respiratory distress syndrome. Lancet Respir Med.
https ://doi.org/10.1016/S2213 -2600(20)30076 –X.

[3] WHO Report, Coronavirus disease 2019 (COVID-19) Situation Report – 101 (2020 (accessed May 01, 2020)),
https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200430-sitrep-101-covid-19.pdf?sfvrsn=2ba4e093_2.

[4] COVID-19 INDIA, Ministry of Health and Family Welfare Government of India https://www.mohfw.gov.in/.

[5] A. Khan and S. Zubair, “Machine Learning Tools and Toolkits in the Exploration of Big Data,” *Int. J. Comput. Sci. Eng.*, vol. 6, no. 12, pp. 570–575, 2018.

[6] A. S. R. S. Rao and J. A. Vazquez, “Identification of COVID-19 Can be Quicker through Artificial Intelligence framework using a Mobile Phone-Based Survey in the Populations when Cities/Towns Are under Quarantine,” *Infect. Control Hosp. Epidemiol.*, vol. 1400, 2020.

[7] Expert: Better models, algorithms could help predict and prevent virus spread, The Augusta Chronicle Newspaper (accessed on February 11, 2020).

[8] A. Khan and S. Zubair, “An Improved Multi-Modal based Machine Learning Approach for the Prognosis of Alzheimer’s Disease,” *J. King Saud Univ. - Comput. Inf. Sci.*, 2020.

[9] A. Khan, S. Zubair, and M. Al Sabri, “An Improved Pre-processing Machine Learning Approach for Cross-Sectional MR Imaging of Demented Older Adults,” in *2019 First International Conference of Intelligent Computing and Engineering (ICOICE)*, 2019, pp. 1–7.

[10] 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.*, vol. 5, no. 1, pp. 1–13, 2020.

[11] “NIST/SEMATECH e-Handbook of Statistical Methods,” 2012.

[12] A. Khan and S. Zubair, “Usage Of Random Forest Ensemble Classifier Based Imputation And Its Potential In The Diagnosis Of Alzheimer’s Disease,” *Int. J. Sci. Technol. Res.*, vol. 8, no. 12, pp. 271–275, 2019.

[13] Komorowski *et al.*, " Exploratory Data Analysis," in Secondary Analysis of Electronic Health Records, MIT Critical Data, Springer, 2016, (doi:10.1007/978-3-319-43742-2_15).

[14] 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*, vol. 52, no. 4, pp. 200–202, 2020.

[15] 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*, vol. 1, no. 2, pp. 156–165, 2020.

[16] 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.*, vol. 12, no. 3, pp. 165–174, 2020.

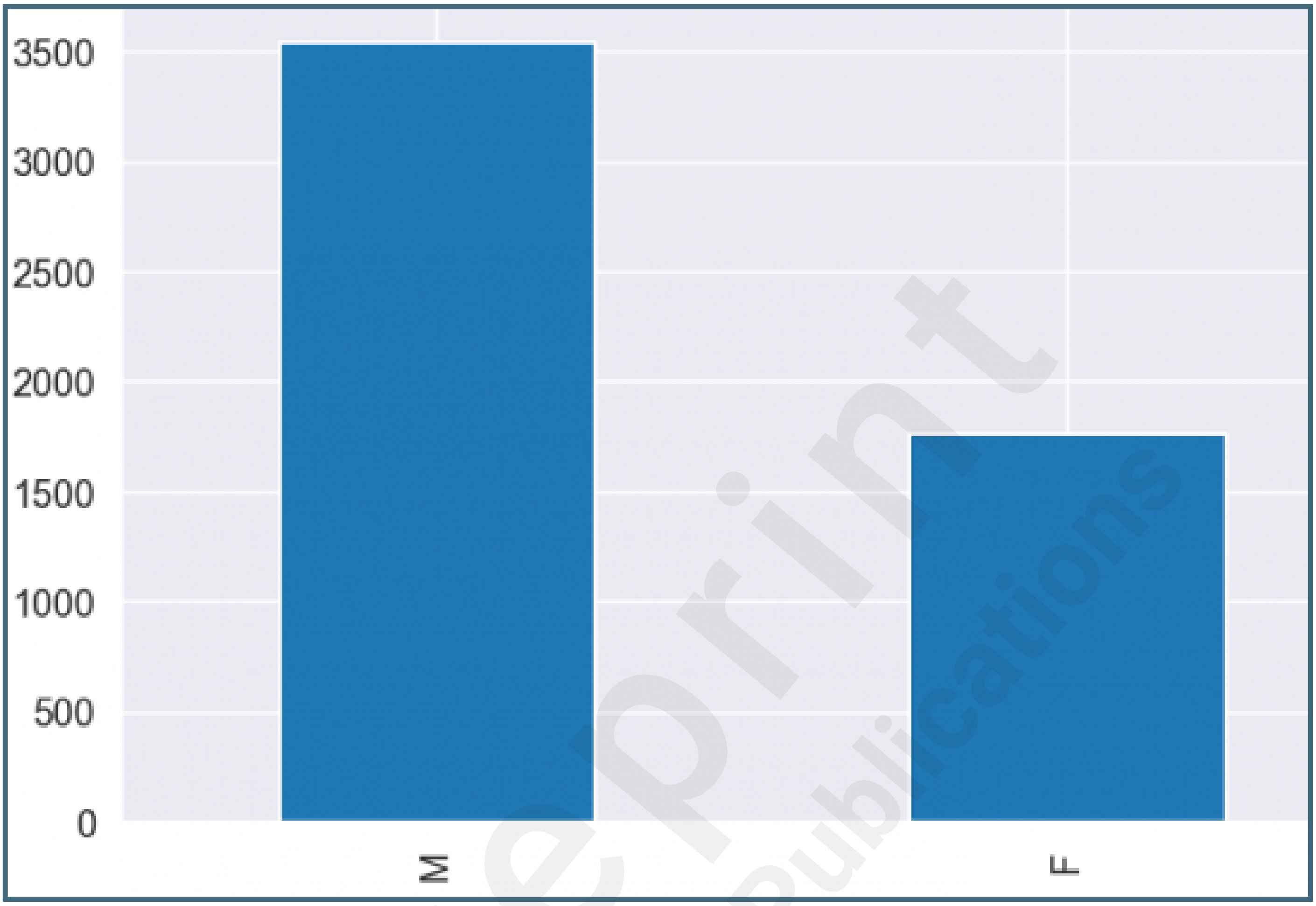
[17] A. Mavragani, “Tracking COVID-19 in Europe: An Infodemiology Study,” *JMIR Public Heal. Surveill.*, vol. 6, pp. 1–13, 2020.

[18] 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.*, vol. 6, no. 2, p. e18828, 2020.

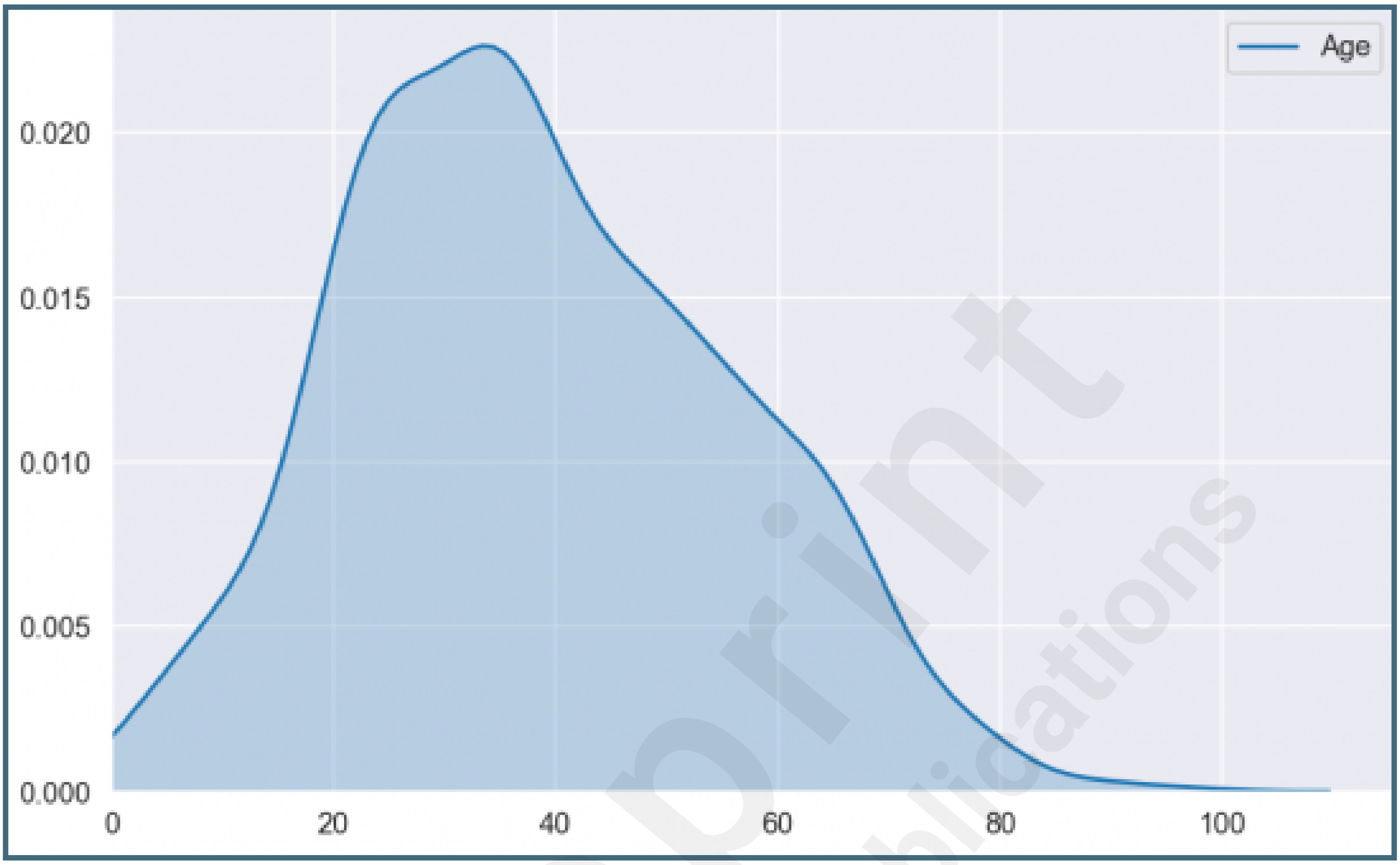
Supplementary Files

Figures

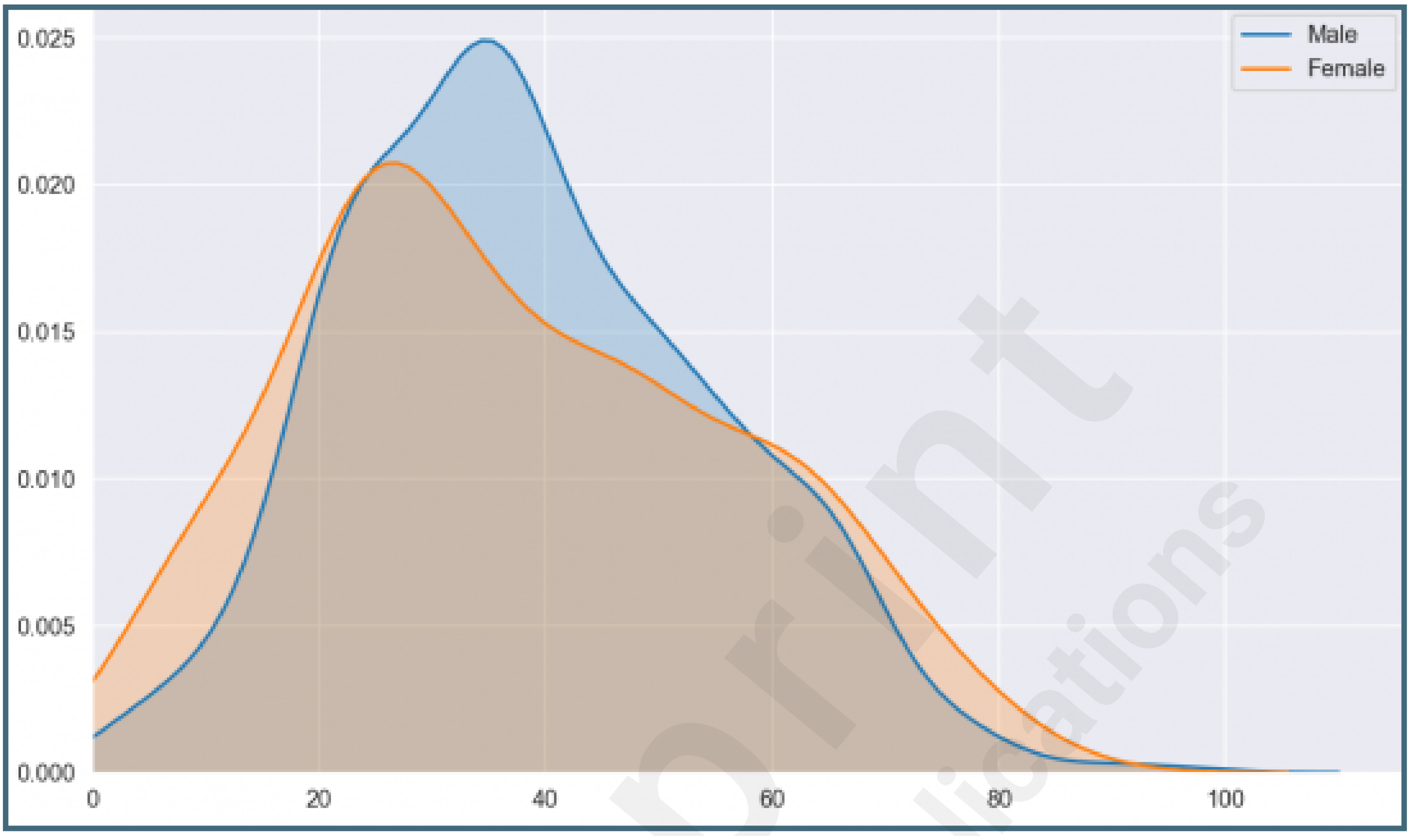
Subject distribution (Gender).



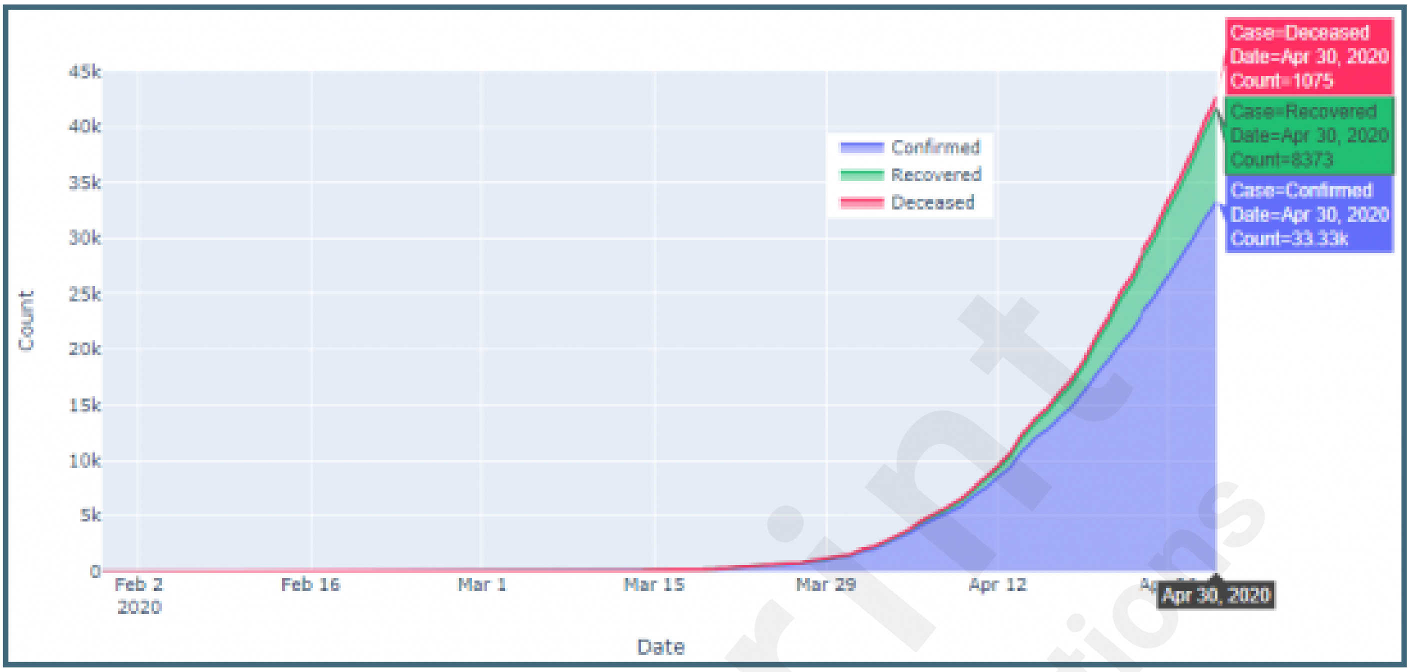
Subject distribution (Age).



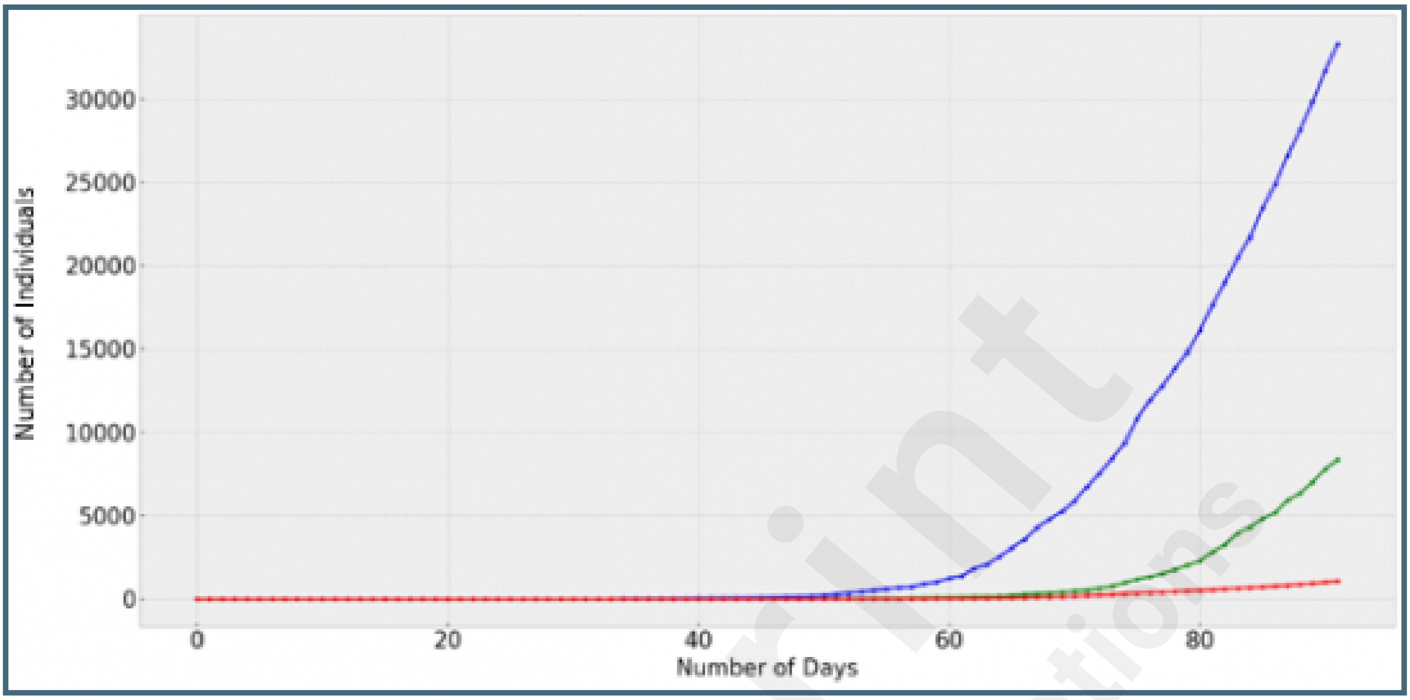
Age distribution of confirmed COVID-19 patients with respect to count of male and female subjects'.



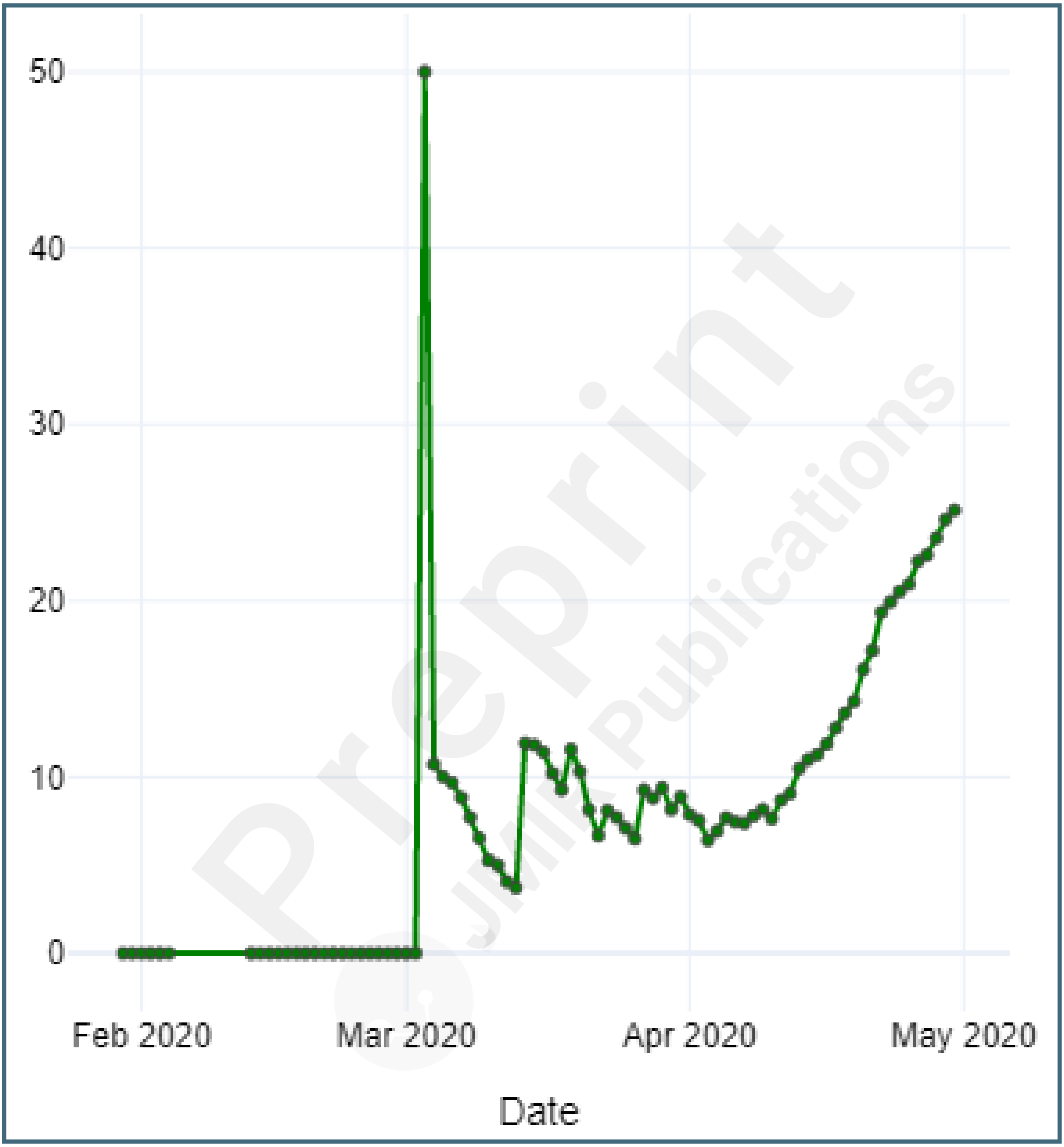
Total number of COVID-19 infected cases reported till date in India.



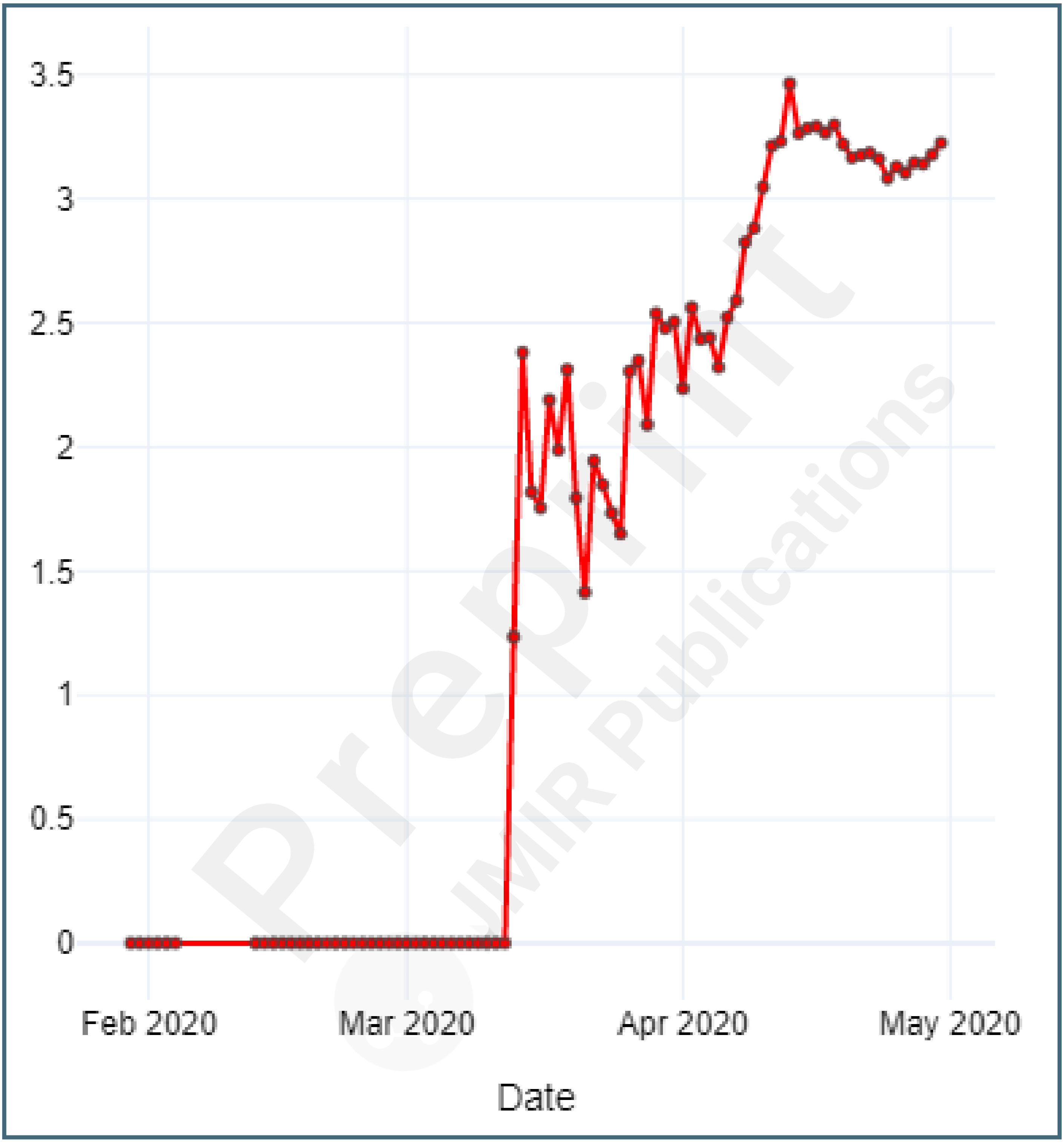
Time-series plot for day-wise infections.



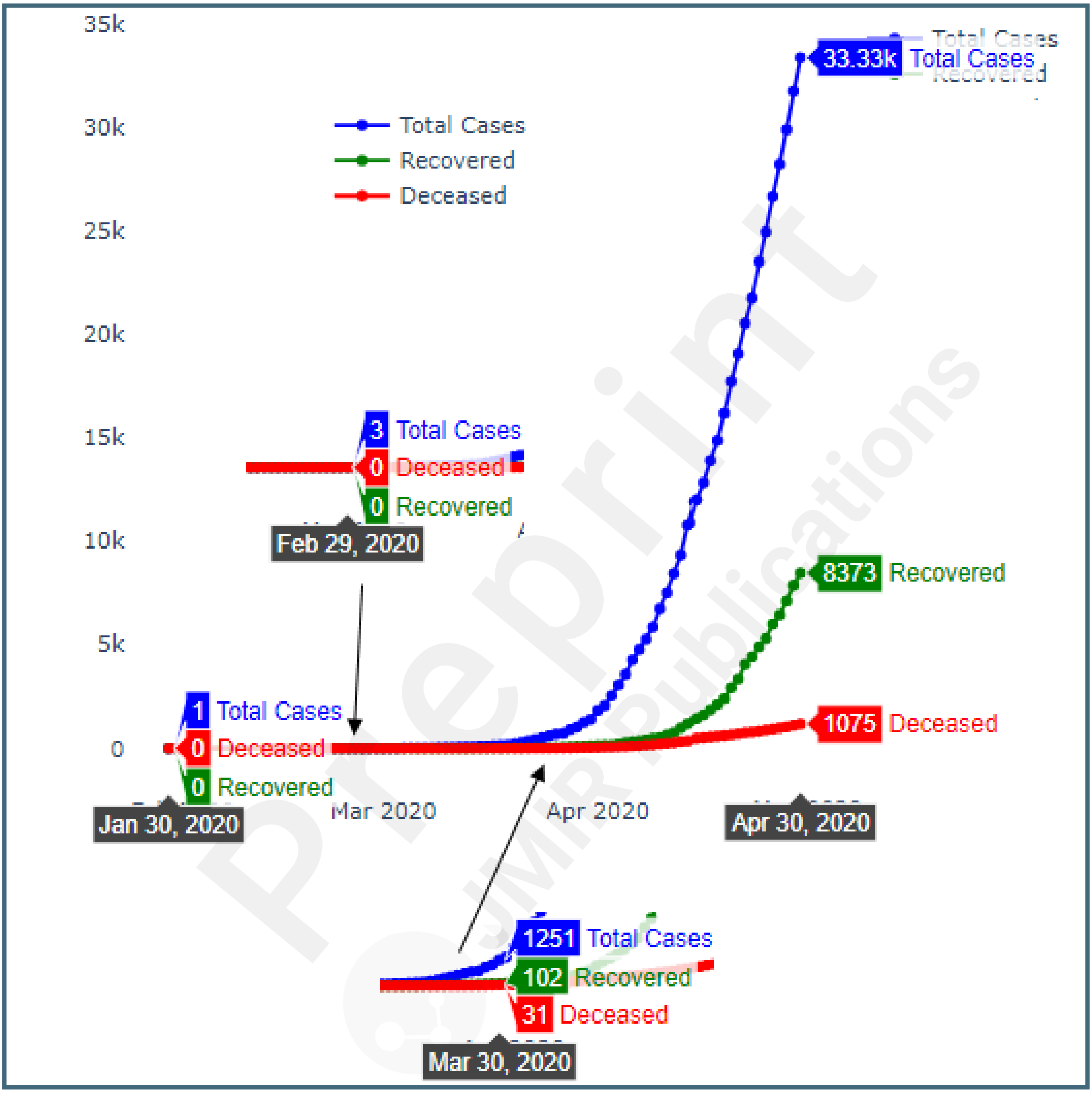
Recovery rate (%).



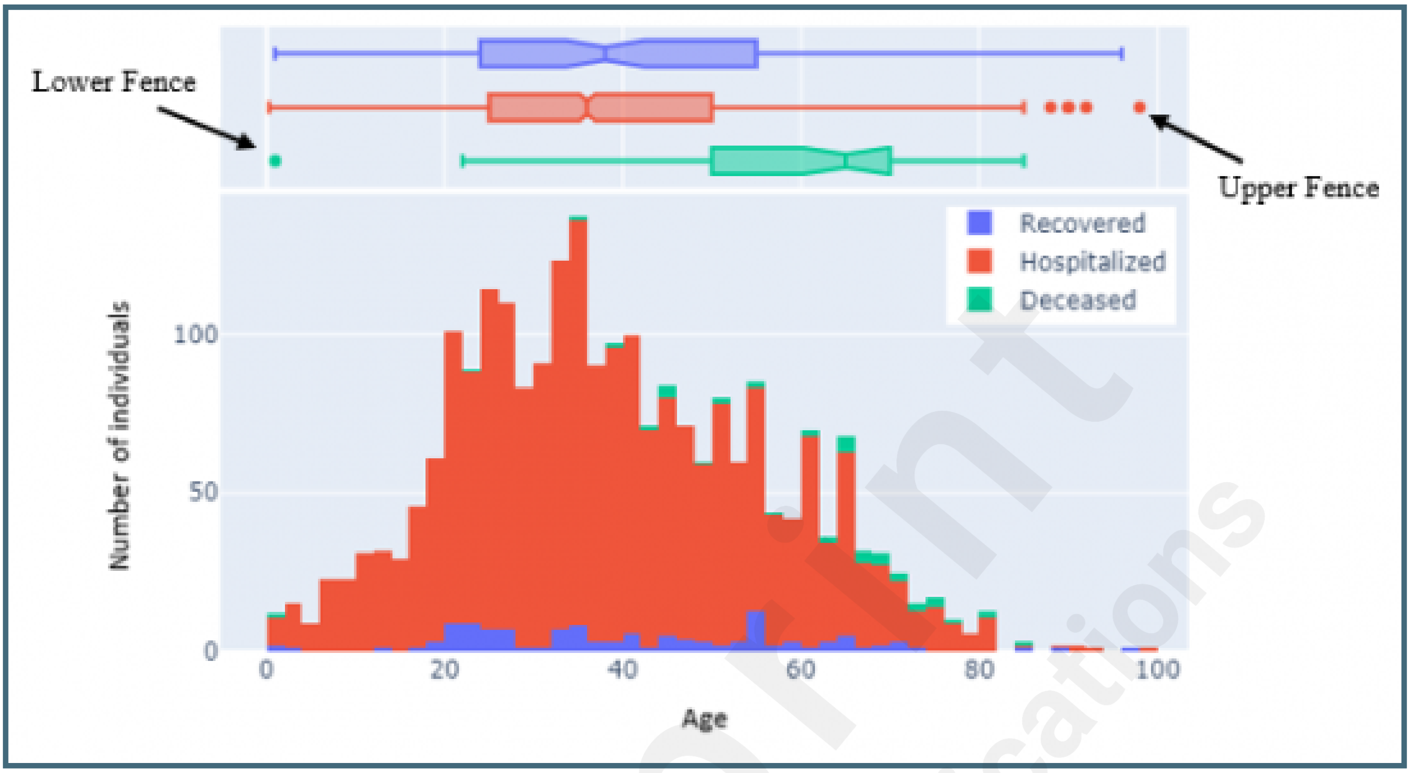
Mortality rate (%).



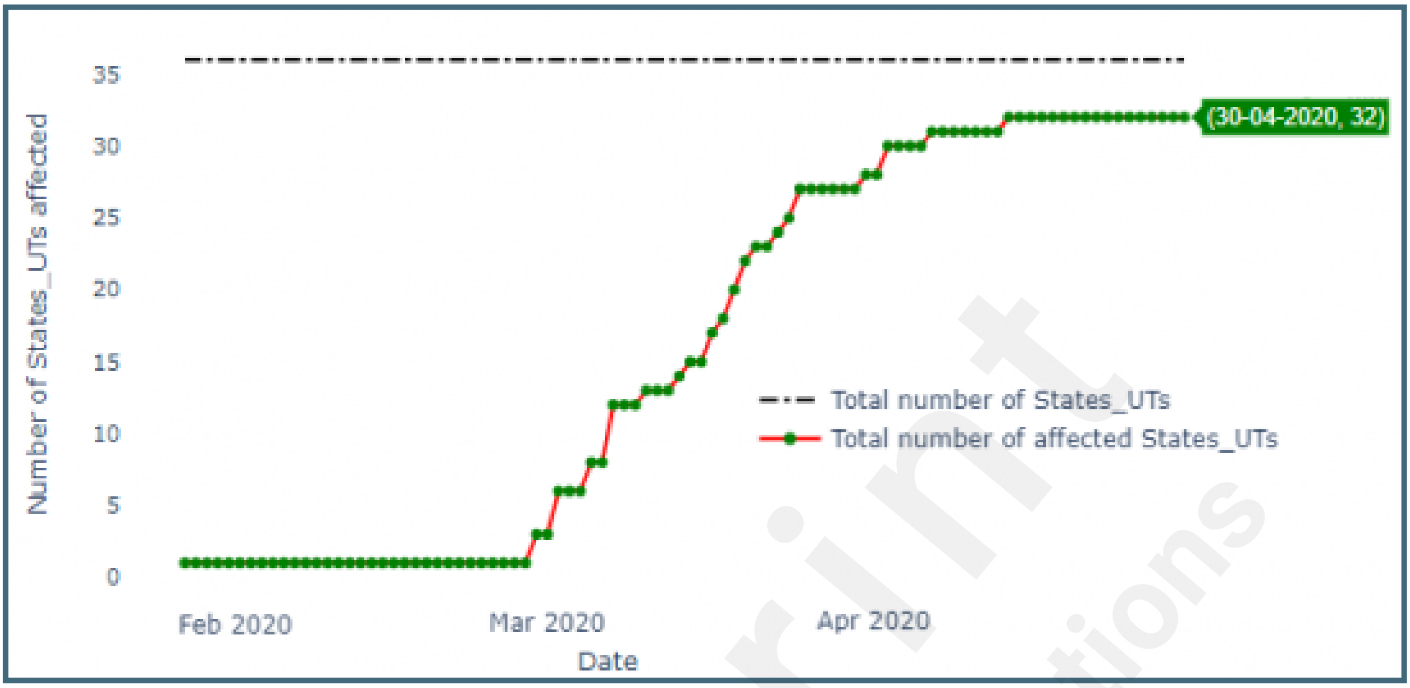
Pictorial representation of weekly trend in India for COVID-19 confirmed, deceased and recovered cases.



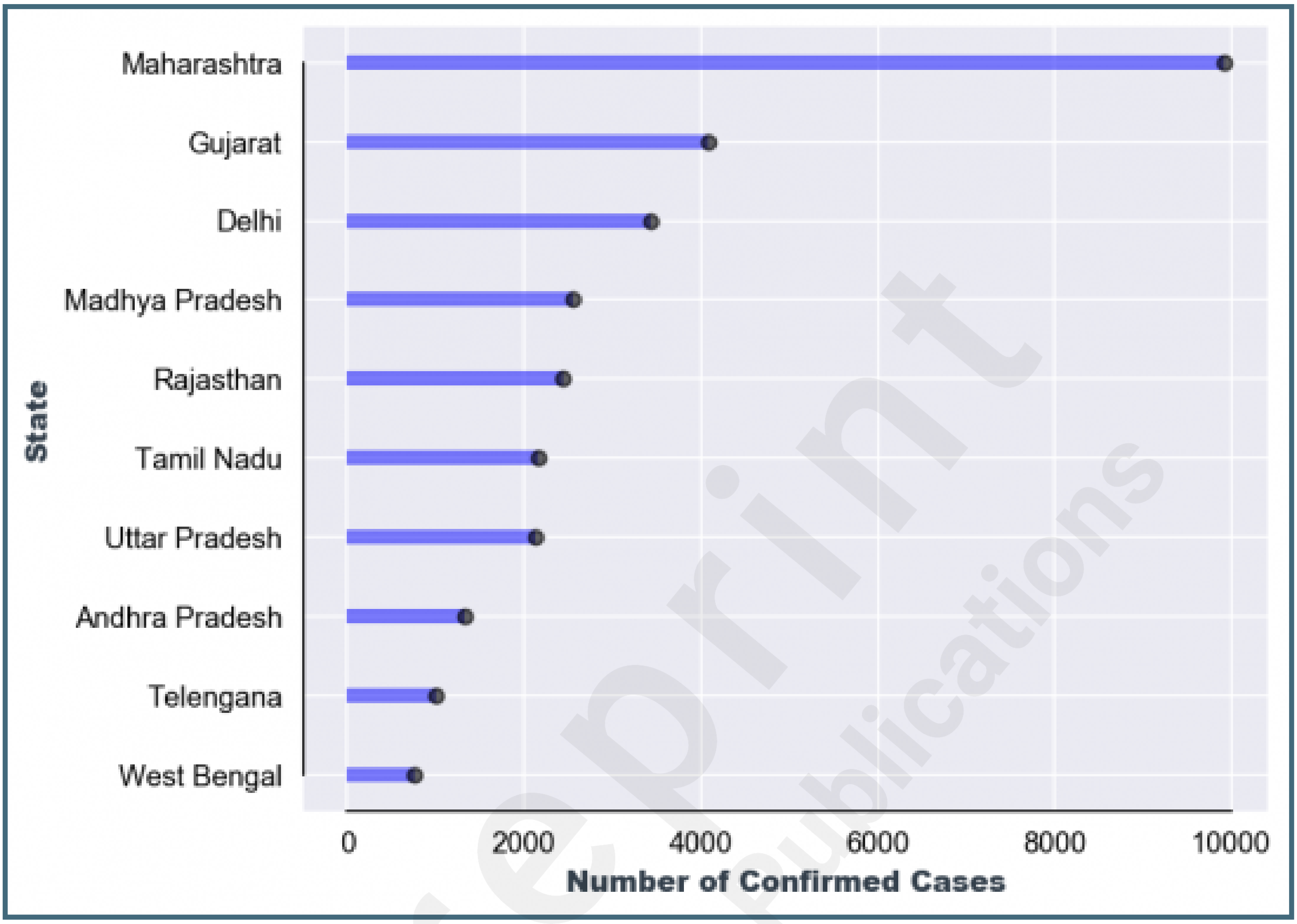
Number of reported individuals for recovered, hospitalized and deceased cases with respect to age class.



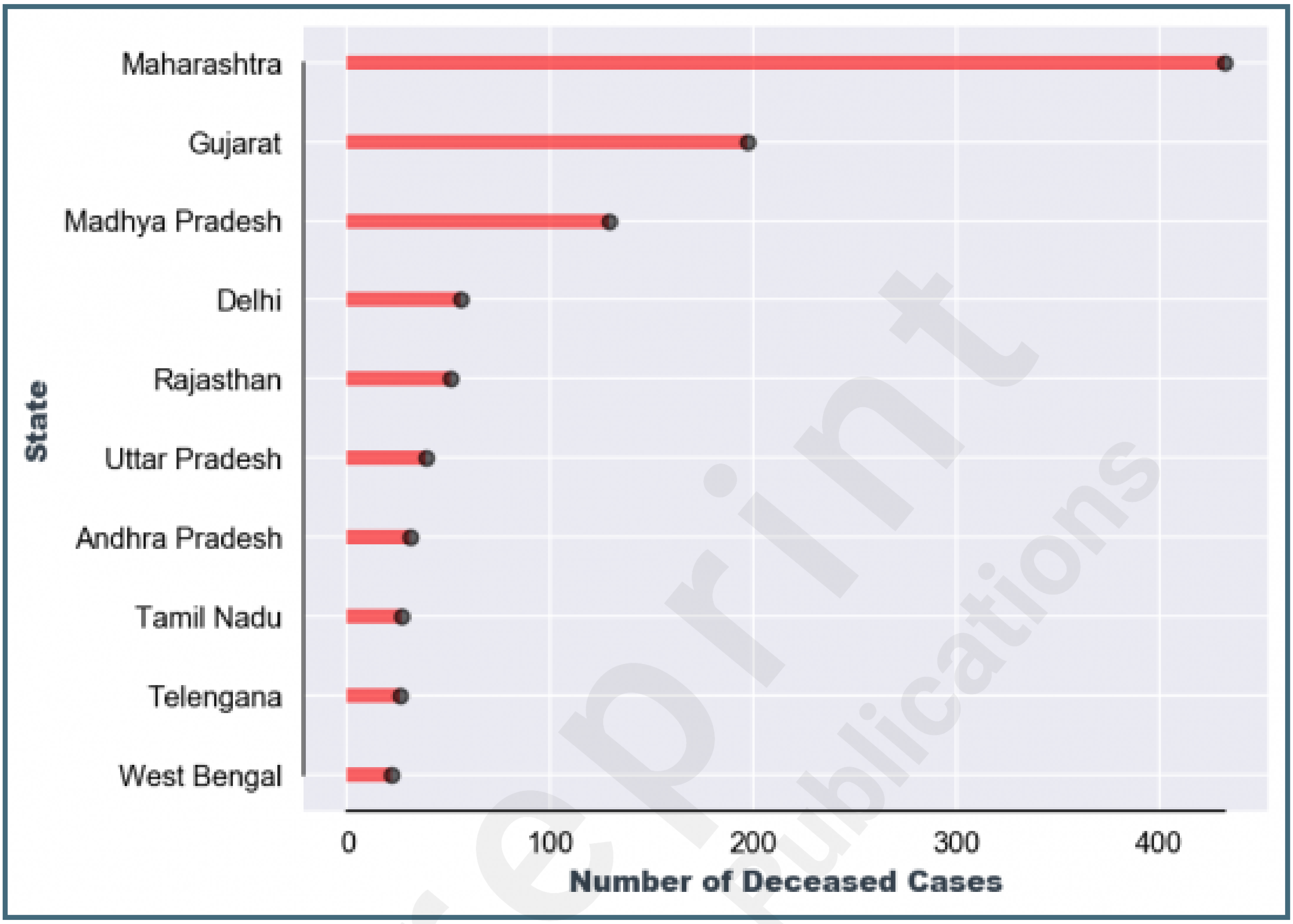
Count of affected Indian states and Union Territories.



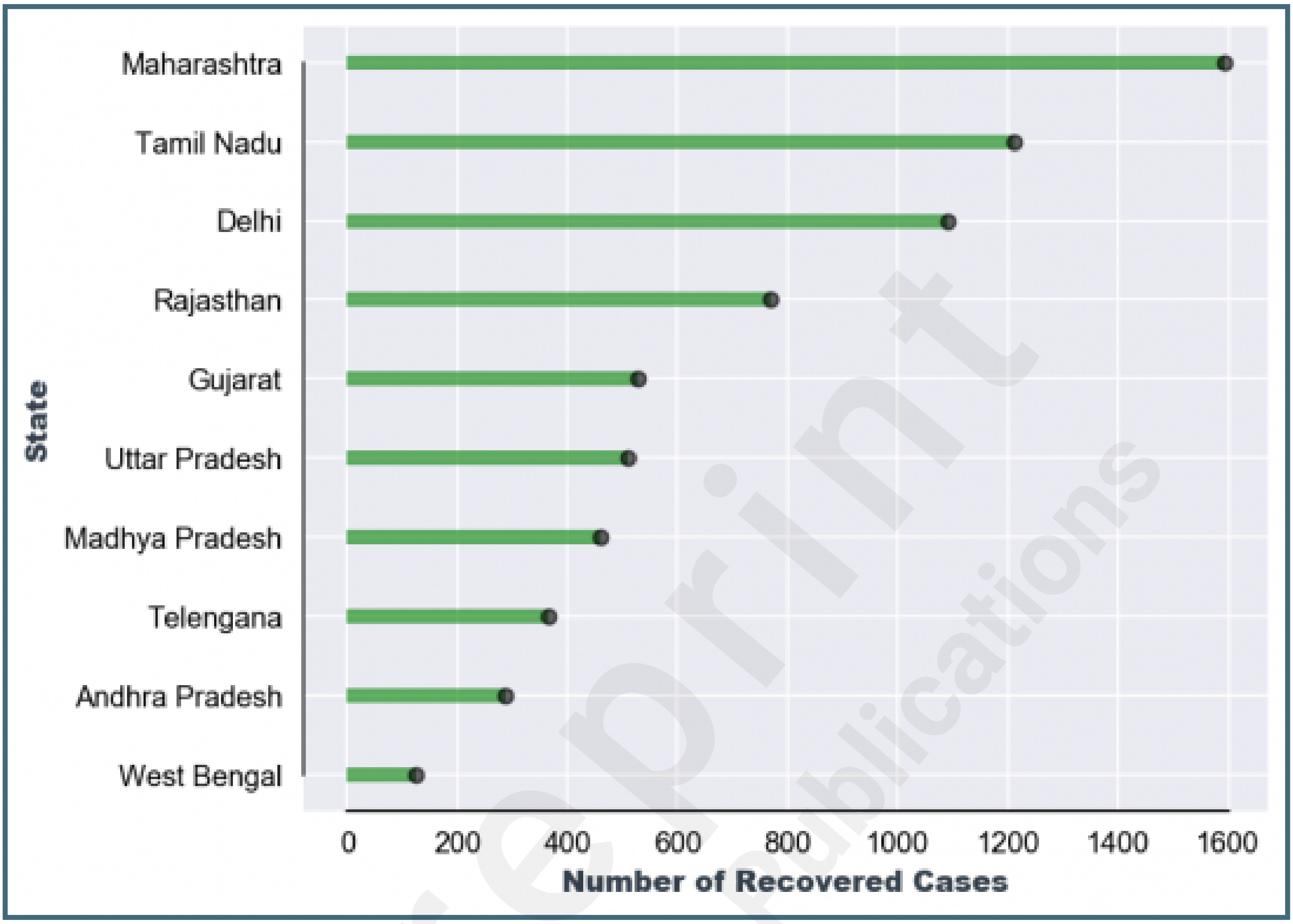
Top ten Indian States affected by COVID-19 (Confirmed Cases).



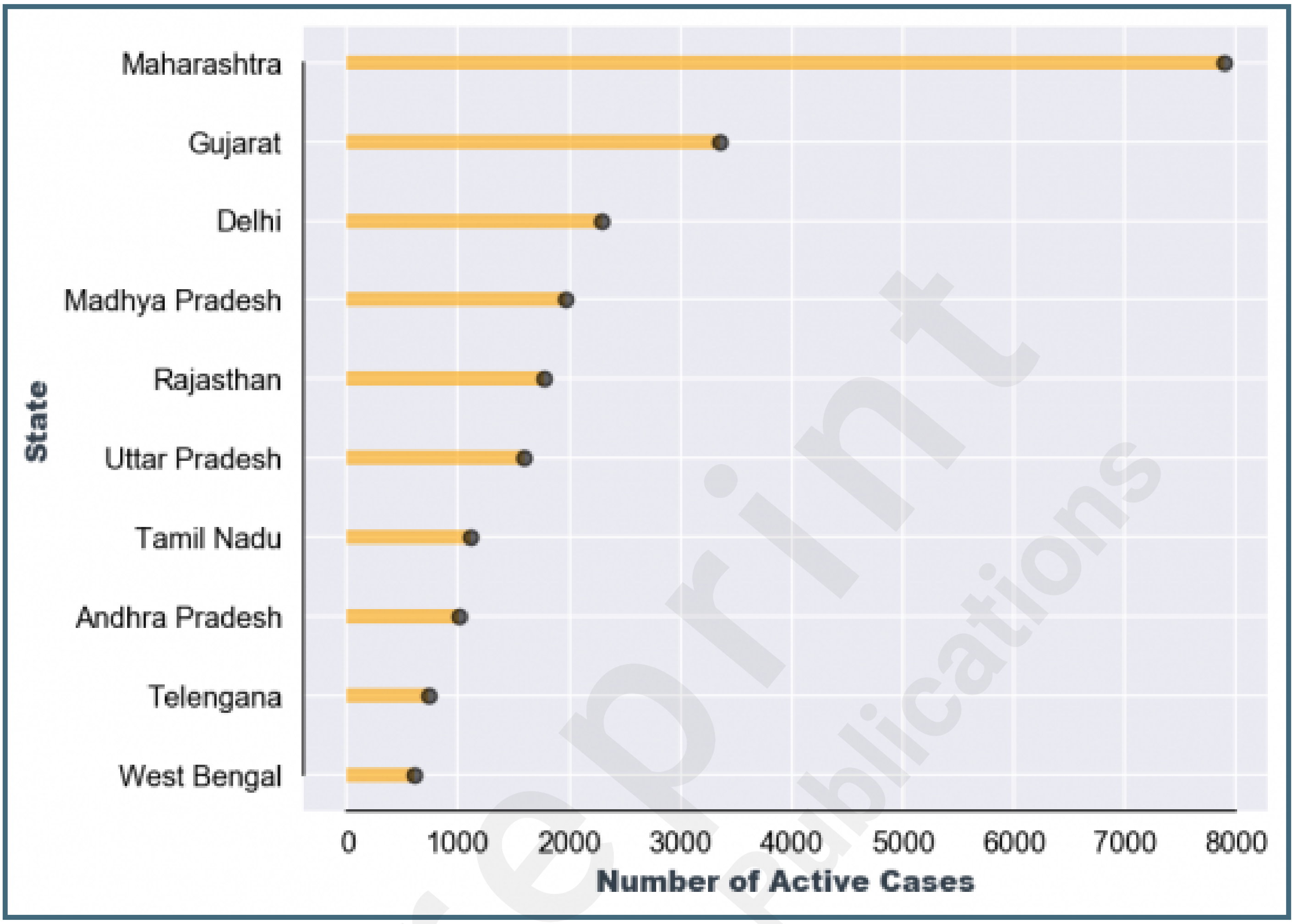
Top ten Indian States affected by COVID-19 (Recovered Cases).



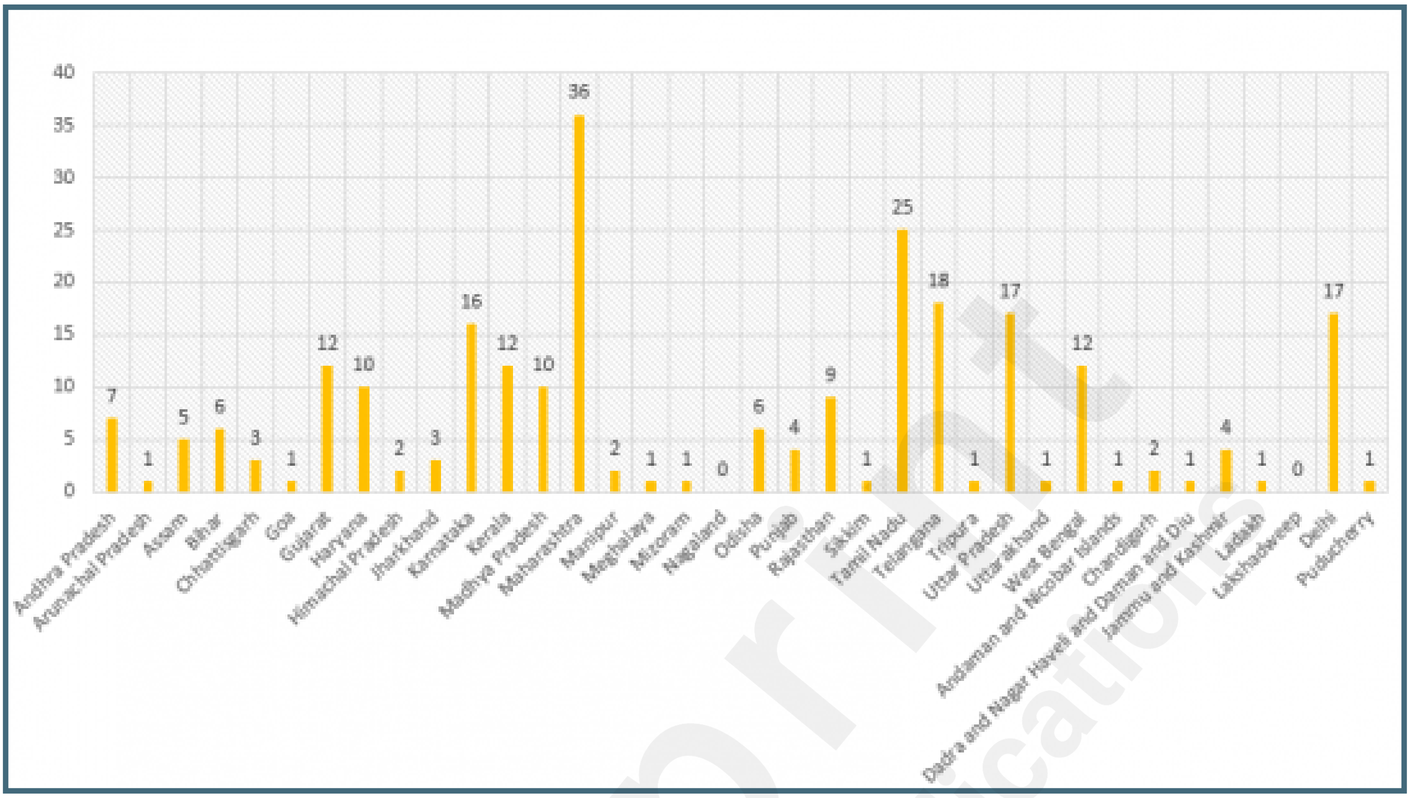
Top ten Indian States affected by COVID-19 (DeceasedCases).



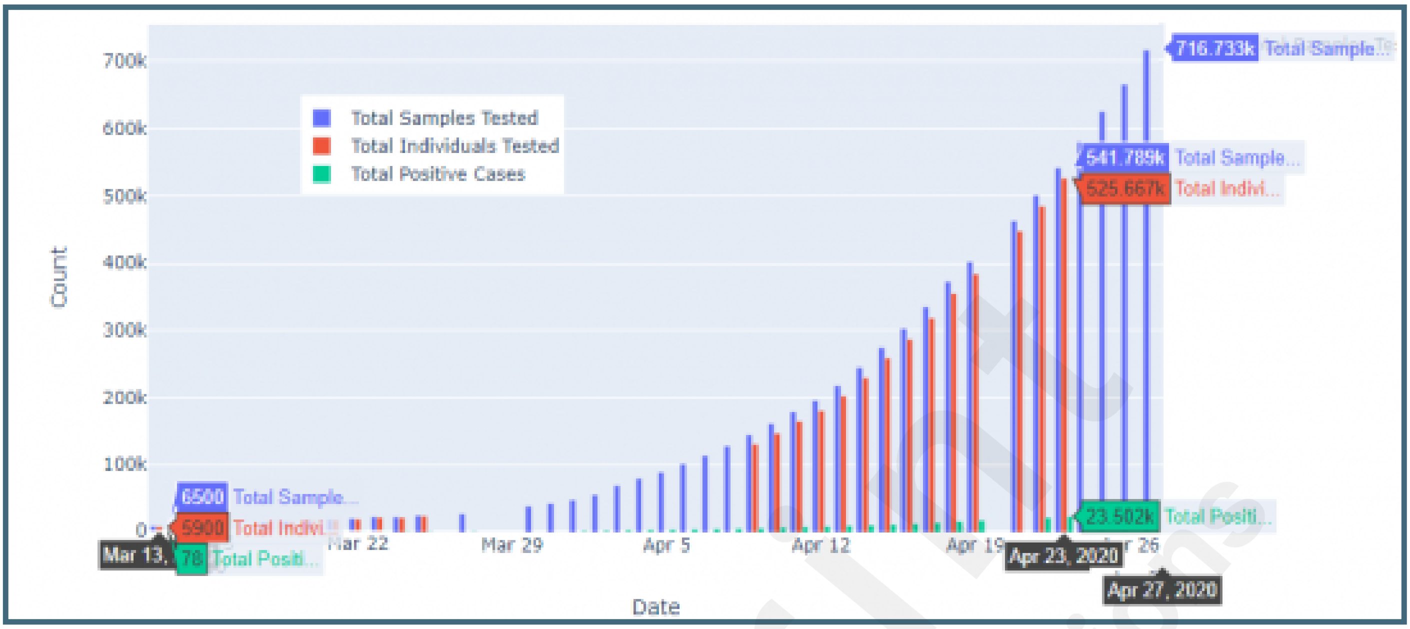
Top ten Indian States affected by COVID-19 (Active Cases).



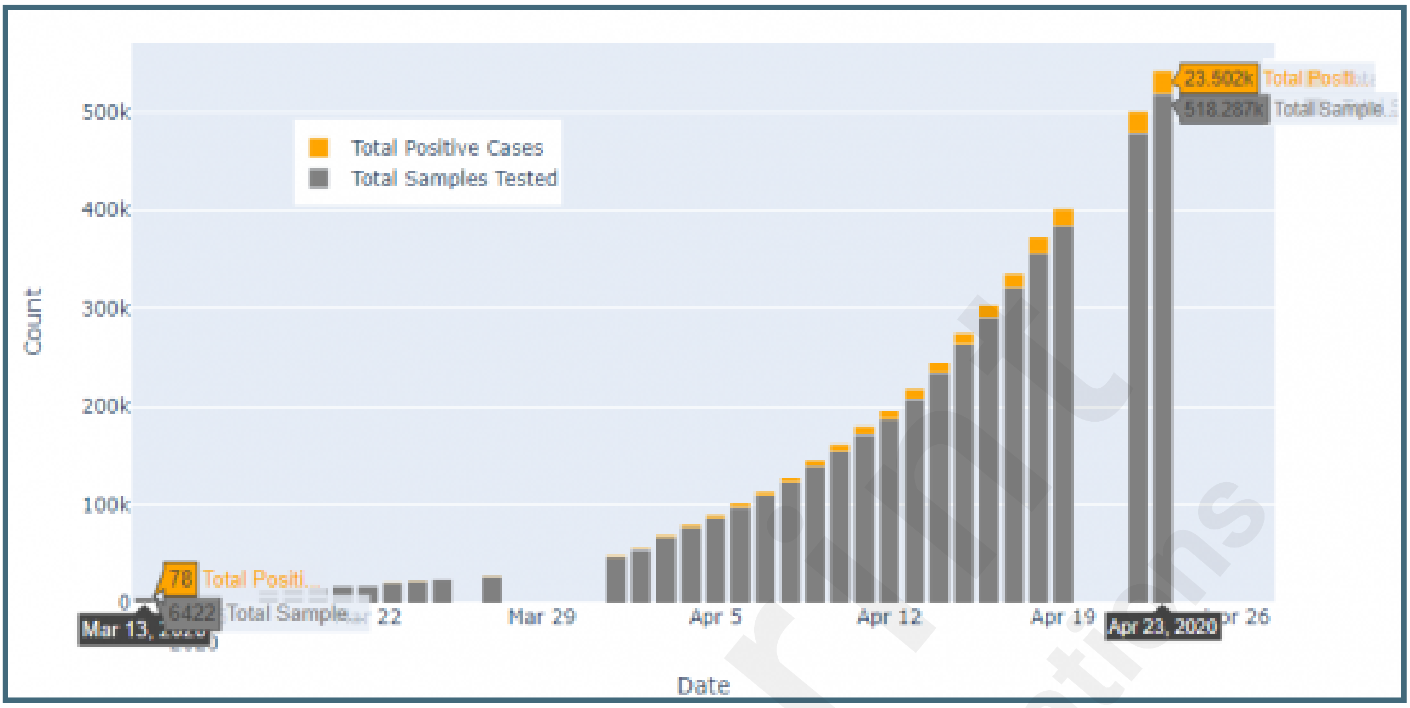
Count of COVID-19 Testing Laboratories within India.



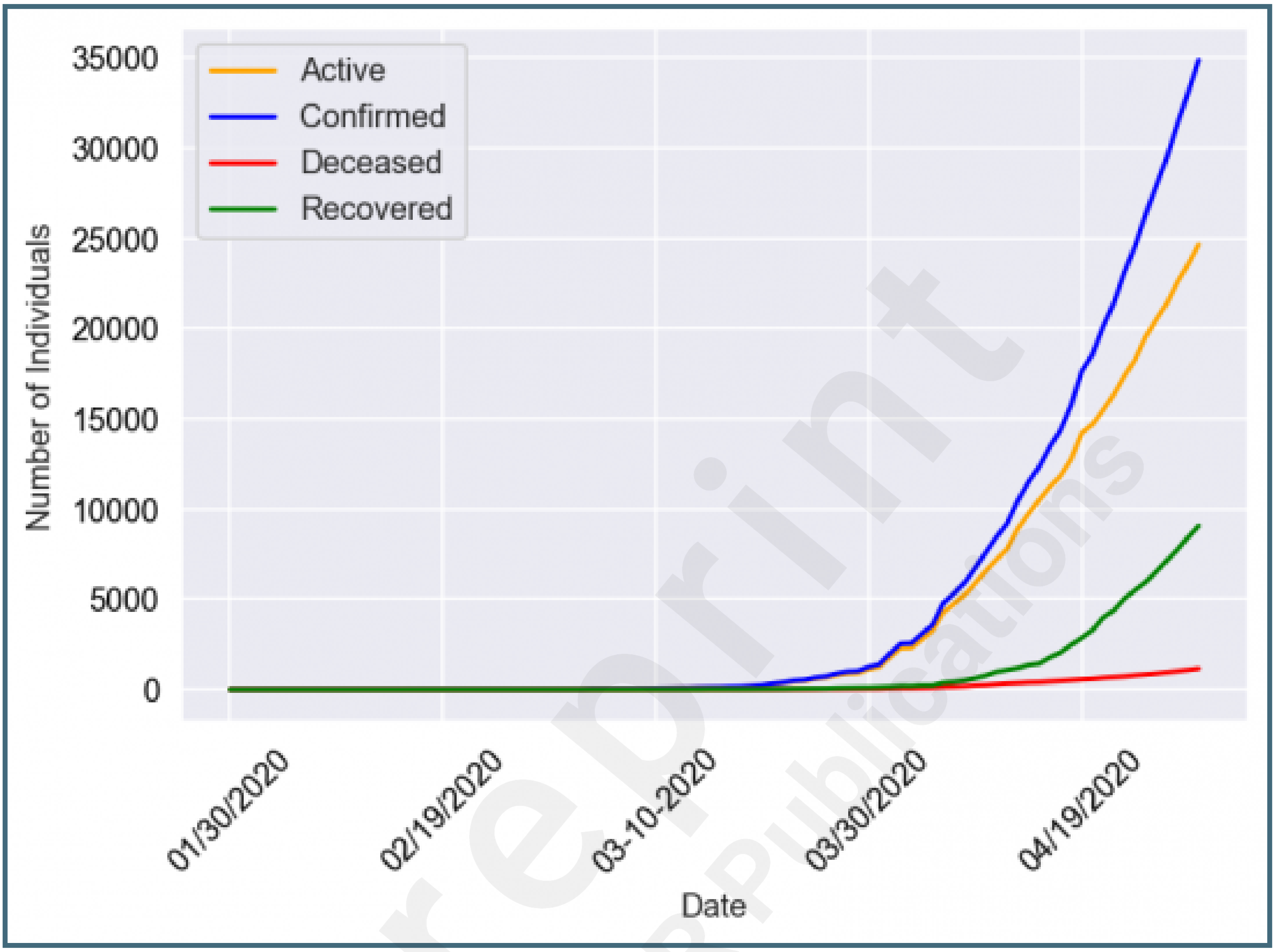
Total number of samples tests conducted for COVID-19.



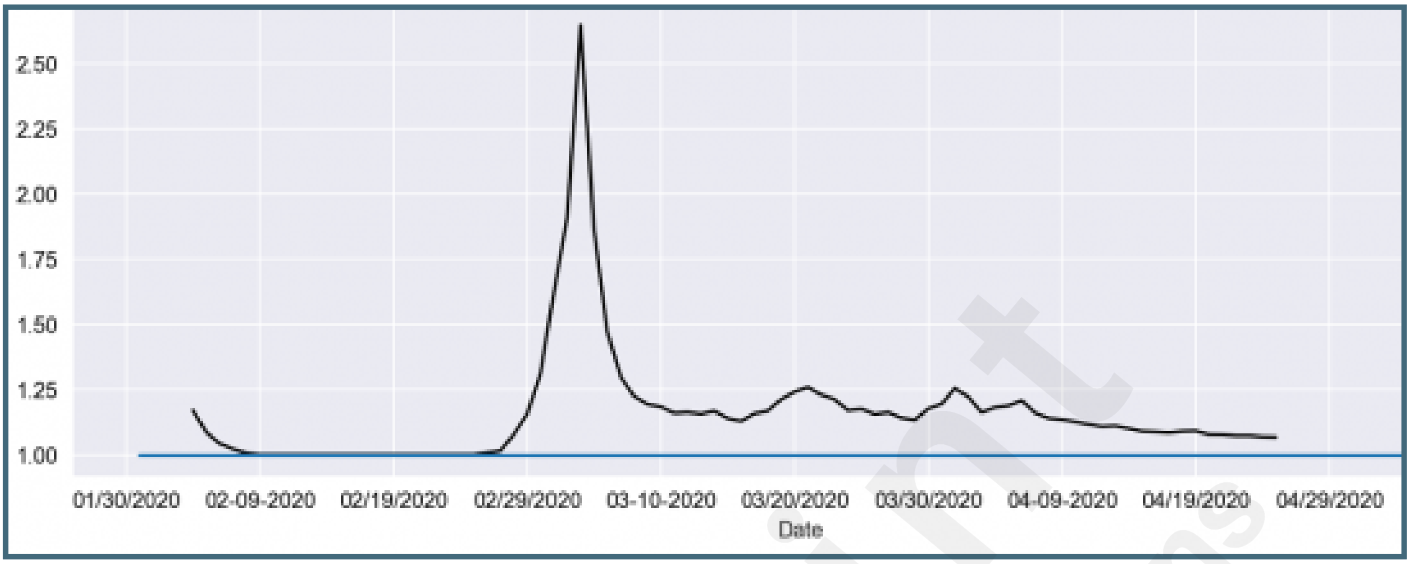
Total positive cases out of total samples tested.



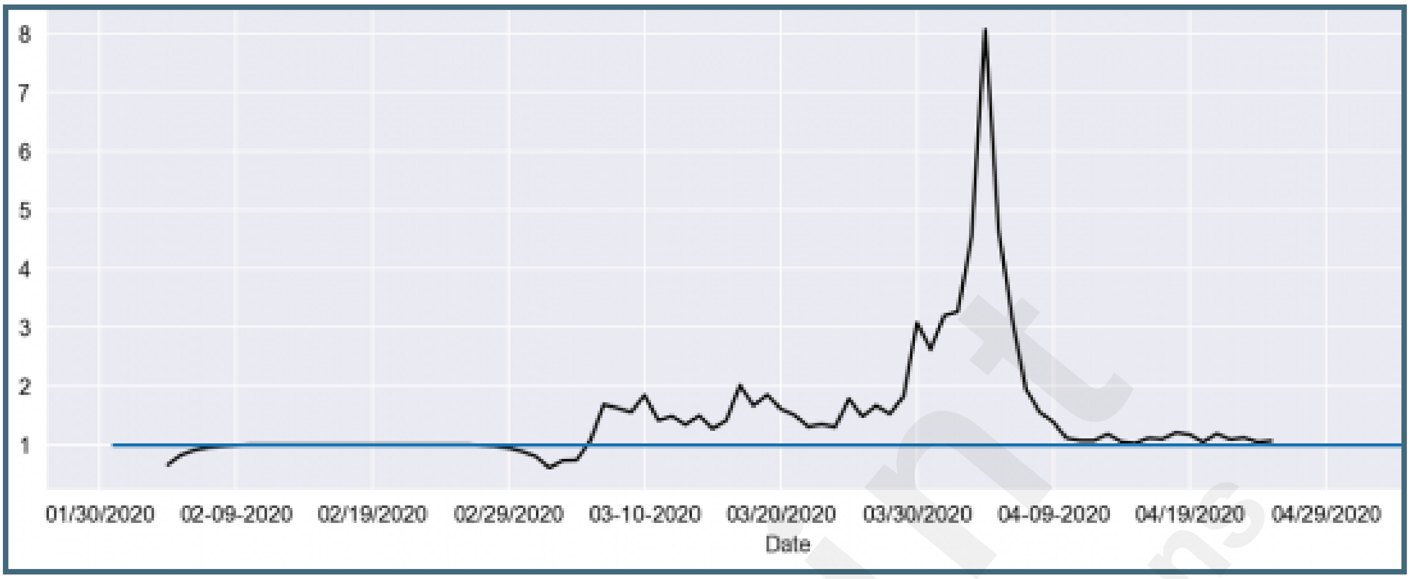
92-days Plot.


35000
30000
25000
20000
15000
10000
5000
0
Number of Individuals
Active
Confirmed
Deceased
Recovered
01/30/2020
02/19/2020
03-10-2020
03/30/2020
04/19/2020
Date

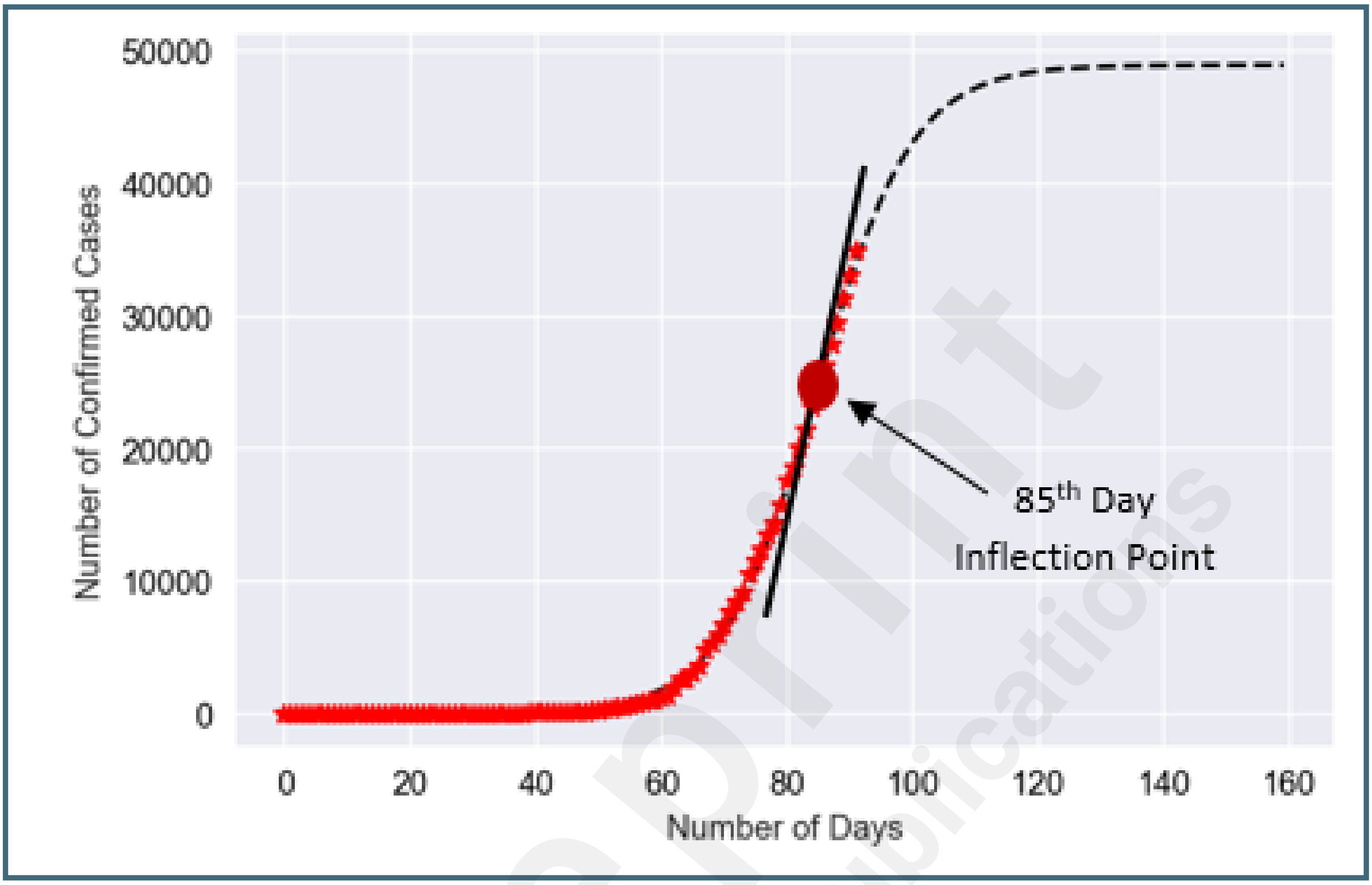
Growth Ratio.



Growth Factor.



Logistic Growth Curve.



Untitled.
URL: https://asset.jmir.pub/assets/27d05810d717291c8cb013d51895db0f.docx

Untitled.
URL: https://asset.jmir.pub/assets/adfa8df6624a4ea9050938e992675bcc.docx