

Comparative Study of Coronavirus Pandemic in the United States and India Through Predictive Modeling

Ruth R Robert

Dallas VA Medical Centre, USA

Suresh Kumar Sharma

Panjab University, Chandigarh, India

Raul Rivera

Dallas VA Medical Centre, USA

Gwen Robinson

Dallas VA Medical Centre, USA

ABSTRACT

Objective: This study aims to develop predictive models of the coronavirus pandemic for the US and India and also to do a comparison of various parameters of COVID-19 for the two countries. The USA and India have faced serious challenges during the COVID-19 pandemic and many lives have been lost. Both countries differ significantly in terms of population density, healthcare infrastructure, and other socioeconomic factors. The population density is much higher in India than in the USA. The rationale for comparing the USA and India using the generalized additive model (GAM) is to see how disparities in the two countries can be accounted for by socio-economic parameters and healthcare infrastructure. Generalized Additive Model (GAM) is a crucial tool for predicting mortality due to various COVID-19 risk factors at any given time point. GAM is flexible and robust because it can capture both linear and nonlinear effects by adjusting the degrees of freedom. It also allows us to include covariates that may initially smoothen, and subsequently, the effect of the main variable can be added to predict mortality. The predictive model is useful in providing valuable insights to enhance resource allocation, risk management, planning, and resource allocation, ultimately resulting in better pandemic management and improved health outcomes. **Method:** The data is extracted from 1 April 2020 to 30 April 2022 (25 months) for a total of 790 days. It consists of: daily confirmed cases, active cases, total confirmed or cumulative cases, recovered cases, mortality per day, cumulative mortality, and true active cases. The association between daily confirmed cases and mortality was established using the generalized additive model (GAM) with natural spline smoothers at (12,2,2) degrees of freedom. **Results:** During this period, there were 1017381 deaths in the US and 523807 in India, with mortality rates of 1339 and 689 per day for the US and India, respectively. It has been observed that the lag effects of daily confirmed cases (DCC) on mortality are significant. The relative risk shows a significant increase in mortality for every 1000 daily confirmed cases in the US and India. **Conclusion:** Using the applications of the GAM model, it has been observed that the lag effects

of daily confirmed cases (DCC) on mortality were significant. Monthly and yearly comparisons were also significant.

Keywords: Generalized additive model, smoothers, active cases, mortality, degrees of freedom, time series analysis.

1.1 INTRODUCTION

Coronavirus (COVID-19) is an infectious disease caused by the SARS-CoV-2 virus. The COVID-19 global pandemic has generated an abundance of research since its outbreak. In early December 2019, an outbreak of coronavirus (COVID-19) occurred in Wuhan City, Hubei Province, China. On January 30, 2020, the World Health Organization (WHO) declared the outbreak a public health emergency of international concern. On February 14, 2020, WHO confirmed 49,053 laboratory cases of coronavirus and 1,381 deaths globally. In the early reports, WHO warned that people infected with the virus may experience mild to moderate respiratory illness and may recover without requiring special treatment. However, some may become seriously ill and require medical attention. Older people and those with underlying medical conditions like cardiovascular disease, diabetes, chronic respiratory disease, or cancer are more likely to develop serious illnesses. The virus can spread from an infected person's mouth or nose in small liquid particles when they cough, sneeze, speak, sing, or breathe. These particles range from larger respiratory droplets to smaller aerosols (Nyberg et al., 2021; Bager et al., 2021; Patone et al., 2021; Novel Coronavirus-MOHFW; GOI 2020; Chin et al., 2020). Every continent in the world has been affected by this highly contagious disease in over 200 countries worldwide. It is also important to note the incubation period of the coronavirus. The 'incubation period' means the time between catching the virus and the time when symptoms of the disease begin to emerge. This is a critical period because when people do not know they have the disease, they may not be as vigilant about being careful not to spread it. Most estimates of the incubation period for COVID-19 range from 1 to 14 days, though the virus most commonly surfaces with symptoms around day five or day six. From April 2020 to April 2022, the US is the worst-affected country, with 1017381 deaths and an average of 1339 deaths per day, followed by Brazil. India is the third-most affected nation, with 523807 deaths and an average of 689 deaths per day, followed by Russia, Mexico, and the United Kingdom. In the US, the five most affected states are California, Texas, Florida, New York, and Illinois. The most affected states in India are Maharashtra, Kerala, Karnataka, Tamil Nadu, and Andhra Pradesh. In the entire world, the most affected regions are the Americas, European Region, South-East Asia Region, West Pacific Region, and Mediterranean Region. In the US and India, acute effects also need to be studied so that a comprehensive policy can be evolved for the prevention of the health effects of coronavirus. Therefore, a time-series study was carried out to determine the association of daily confirmed cases with mortality in the presence of other important indicators.

In Section 1.2, material and methods have been listed. Statistical tools used for analysis have been mentioned in Section 2. Section 3 focuses on results and interpretation in depth. It includes descriptive statistics, trends for various variables, box plots, and a comparison of US India for daily confirmed cases (DCC), total confirmed cases (TCC), active cases (AC), total recoveries (TR), mortality, and cumulative (total) mortality. Monthly and yearly comparisons were also made in this section. The development of the GAM model for the US and India is described in Section 4. This includes smoothers, partial autocorrelation functions

(PACF), residual plots, risk ratios (RR), and predicted plots for the US and India. The conclusion and discussion follow in Section 5. In Section 6, the limitations of the study have been discussed.

1.2 MATERIALS AND METHODS

In this study, the secondary data was extracted from the global-dataset website <https://www.kaggle.com/datasets/josephassaker/covid19-global-dataset> (available for all countries) for the period from April 2020 to April 2022, that is, for 25 months (760 days). Some information was also extracted from other websites (given in references). The date-wise data is concerned with the following variables of interest:

- TCC= Total Confirmed Cases
- DCC=Daily Confirmed Cases
- AC =Active Cases
- TR =Total Recoveries
- TAC =True Active Cases (TCC-TR)
- Mortality = daily mortality (Number of deaths per day) Total mortality = cumulative deaths.

The association between mortality and daily confirmed cases has been established by using Generalized Additive Model (GAM) by smoothing the effects of covariates, that is day effect, active cases, and true active cases (De Boor 1978). Partial Autocorrelation Function (PACF) plots and Residual Plots are useful for studying the lag effects and patterns in time series data (Eilers 1996, Hastie and Tibshirani (1986, 1990), Green and Silverman 1994, Gu 2002).

Partial Autocorrelation Function (PACF) Plots

A characteristic feature of time series is that the observations are ordered through time; a consequence of this is autocorrelation, that is, an underlying pattern between observations from one-time period to the next within a time series. We can define autocorrelation at lag k as the correlation between observations k periods apart. This can be estimated for lags $k = 0, 1, 2 \dots$ as

$$r_k = \frac{\sum_{t=k+1}^n (x_t - \bar{x})(x_{t-k} - \bar{x})}{\sum_{t=1}^n (x_t - \bar{x})^2},$$

where x_t denotes the observation at time t and $\bar{x}$ denotes the sample mean of the series $\{x_t\}$. The partial autocorrelation at lag k is defined as the partial correlation between observations k periods apart. That is, it is the correlation between x_t and x_{t-k} after taking out the effect of the intervening observations $x_{t-1}, \dots, x_{t-k+1}$. It can be estimated using a fast-computational algorithm. The collection of autocorrelations at lags $k = 1, 2, \dots$ is known as the autocorrelation function (ACF). Similarly, the collection of partial autocorrelations is known as the partial autocorrelation function (PACF).

Residual Plot

A residual plot is a graph that shows the residuals on the vertical axis and the independent variable on the horizontal axis. If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a non-

linear model is more appropriate. The difference between the observed value of the dependent variable (y) and the predicted value ($\hat{y}$) is called the residual (e). A plot of residuals versus predicted responses is essentially used to spot possible heteroscedasticity in time series.

2 STATISTICAL ANALYSIS

The data were examined for quality and consistency by cross-checking different websites. Descriptive statistics, i.e., mean, standard deviation, minimum, maximum, and range, were computed. Trend analyses, box plots, and bar diagrams were also drawn for comparison between the US and India. A generalized additive model (GAM) was fitted in R statistical software with the Quasi-Poisson link function for mortality as the dependent variable and daily confirmed cases as the main independent variable with a penalized or natural spline with (12,2,2) degrees of freedom (df). The smoothers were added for day effects, active cases, and true active cases. It has been observed that the GCV score for Penalized and Natural Splines does not differ significantly, and therefore it has been decided to use only the Natural Spline model with (12,2,2) df. The base model includes smoothers at the initial stage:

$$\text{Log (E (mortality))} = s (\text{day effect, df}) + s (\text{active cases, df}) + s (\text{true active cases, df}).$$

The optimal degrees of freedom (df) were determined using the approach suggested by Aggarwal et al. (2014). Once the base model had been established, the main variable, daily confirmed cases (DCC), was added to the model.

The adopted model was:

$$\text{Log (E (mortality))} = \text{daily confirmed cases} + s (\text{day effect}) + s (\text{true active cases}) + s (\text{active cases})$$

- Generalized Additive Model (GAM) with natural spline in R
- Quasi-Poisson function with mortality as a dependent variable
- Smoothers for day effect, true active cases, and active cases
- Exposure at single-day lags of 0 to 5 days.

The interpretation of the model is based on a generalized cross-validation (GCV) score, Rsquare (adjusted) value, deviance explained, and scale estimate. The higher value of R^2 (adjusted) and a smaller value of the GCV score confirm the validity of the model.

3 RESULTS ALONG WITH INTERPRETATION

3.1 Descriptive Statistics

During the 760-day period from April 1, 2020, to April 30, 2022, there were 1017381 deaths in the US and 523807 in India. In this period, the mortality rate on average was 1339 per day in the US and 689 per day in India. The summary statistic, which includes the mean, standard deviation, minimum, maximum, and range for the US and India, is shown in Table 1 for all the variables.

Table 1: Descriptive Statistics for US and India (n=760)

Variables	Mean	Standard Deviation	Minimum	Maximum	Range
Total Confirmed cases	32565977.93	25040002.86	238519	83135714	82897195
Daily Confirmed cases	109111.92	137944.29	8530	909610	901080
Active cases	3480687.47	3426953.88	210415	17935434	17725019
Total Recoveries	28562152.48	23266564.20	21476	80689561	80668085
Total mortality	523137.98	299818.26	6628	1022713	1016085
Daily mortality	1338.66	879.52	151	4352	4201
True Active cases	4003825.45	3564428.28	217043	18852703	18635660
Descriptive Statistics for India (n=760)					
Total Confirmed cases	20110167.21	15227430.90	1998	43079188	43077190
Daily Confirmed cases	56681.30	82343.16	489	414433	413944
Active cases	554856.91	737521.49	1791	3741179	3739388
Total Recoveries	19288102.95	15025983.75	148	42536253	42536105
Total mortality	267207.34	191106.33	59	523843	523784
Daily mortality	689.22	946.81	0	5093	5093
Actual Active cases	822064.26	762150.68	1850	3997144	3995294

The monthly averages for all the variables have been computed, and the following inferences have been drawn.

Description for Monthly Average (US):

The total confirmed cases (TCC) and total recoveries (TR) are close to each other and have shown an increasing trend for almost all months, which means the recovery rate was also fast. However, there were spikes in mortality, particularly on December 20 and January 21, when around 2806 and 3244 deaths per day were recorded. The active cases and true active cases were very high on January 22 and February 22, and that led to high mortality in these months (Figure 1).

Description for Monthly Average (India):

A similar pattern for India was observed for TCC and TR. However, there was a spike in mortality, particularly on September 20, when around 1131 deaths per day were recorded. With the decline of the first wave by September 20, mortality began to fall. In May 2021 (during the onset of the second wave), mortality began to rise, and on April 21 and May 21, an astonishing number, i.e., almost 1689 and 4398 deaths, respectively, per day, were reported. On June 21, the number of deaths also went up to 1897 deaths per day. The active cases and true active cases were very high on April 21 and May 21, and it resulted in high mortality during this period (Figure 2).

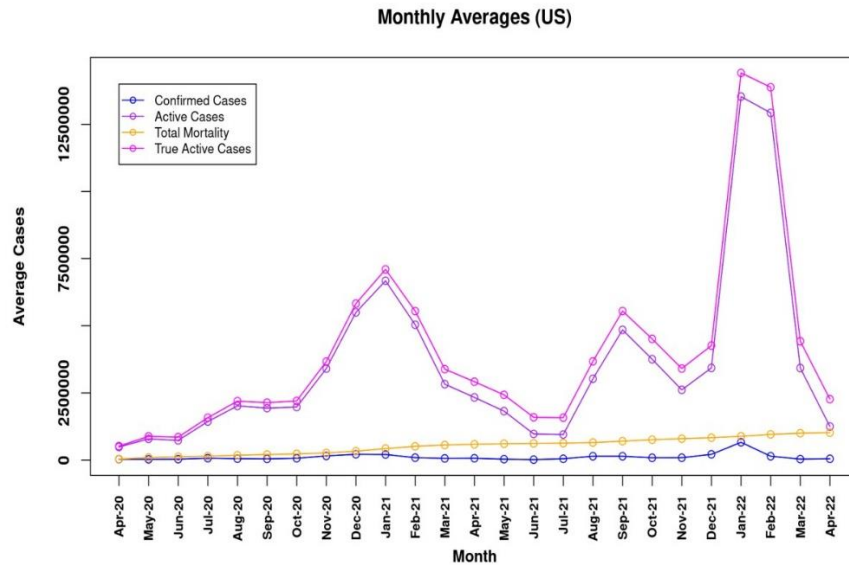


Figure 1: Trend of Monthly Averages for DCC, AC, Total Mortality and TAC (US)

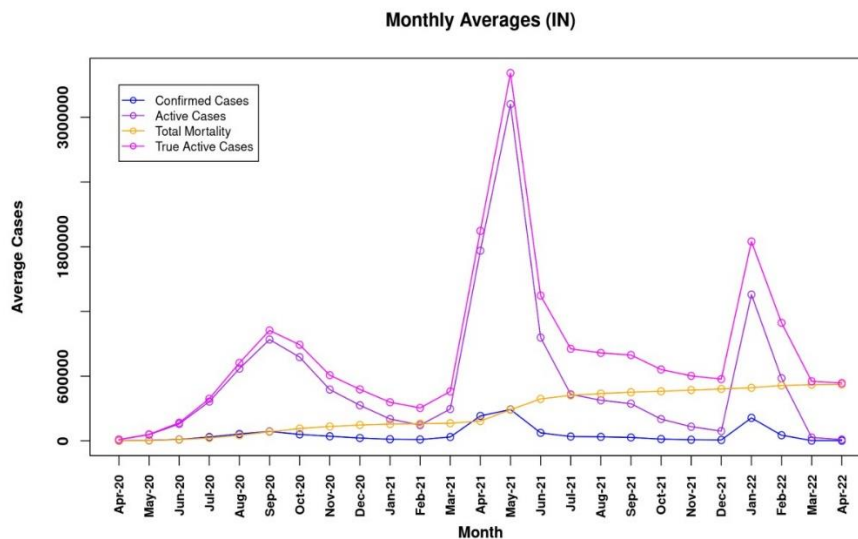


Figure 2: Trend of Monthly Averages for DCC, AC, Total Mortality and TAC (India)

3.2 Month-wise Comparison of the US and India

The comparative study of US and India for each month shows that there were significant differences for all months for TCC ($p < .0001$), DCC ($p < .0001$, except for months Oct-20 and July-21, $p > .05$), AC ($p < .0001$, except June-21, $p > .05$), TR ($p < .0001$, except October-20, November-20, $p > .05$), Total Mortality ($p < .0001$), Daily Mortality ($p < .0001$, except October-20, $p > .05$) and TAC ($p < .0001$).

Year-wise Comparison of the US and India for each Variable:

Year-wise comparisons (2020 vs. 2021; 2020 vs. 2022; 2021 vs. 2022) for TCC, DCC, AC, TR, Total Mortality, and Daily Mortality show that there was a significant difference between the US and India ($p < .0001$).

Total Confirmed Cases (TCC):

The comparison of the US and India was also made separately for the years 2020, 2021, and 2022. The analysis of variance (ANOVA) results show that all three years differ significantly for TCC (US: $F = 4384$, $p = .0001$, India: $F = 1313.56$, $p = .0001$). Post-hoc multiple comparisons were also significant for the US and India ($p = .0001$).

Daily Confirmed Cases (DCC):

All three years for DCC (US: $F = 58.41$, $p = .0001$, India: $F = 12.17$, $p = .0001$) differ significantly for the US and India, respectively; however, multiple comparisons for India show that the years 2021 and 2022 do not differ significantly ($p = .893$).

Active Cases (AC):

For AC, all three years (US: $F = 169.26$, $p = .0001$, India: $F = 10.0$, $p = .0001$) differ significantly for the US and India; however, multiple comparisons for India show that the years 2020 and 2022 do not differ significantly ($p = .337$).

Total Recoveries (TR):

For TR, the comparison of the US and India was also made separately for the years 2020, 2021, and 2022. The analysis of variance (ANOVA) results show that all three years differ significantly for TR (US: $F = 3093$, $p = .0001$, India: $F = 1268.84$, $p = .0001$). Post-hoc multiple comparisons were also significant for the US and India ($p = .0001$).

Total Mortality:

The analysis of variance (ANOVA) results show that all three years differ significantly (US: $F = 3249.40$, $p = .0001$, India: $F = 1077.84$, $p = .0001$). Post-hoc multiple comparisons were also significant for the US and India ($p = .0001$).

Daily Mortality:

For Daily Mortality ANOVA results for the US were non-significant ($F = .796$, $p = .452$). However, ANOVA results for India were significant ($F = 24.12$, $p = .0001$). Post-hoc multiple comparisons were also significant for India ($p = .0001$).

True Active Cases (TAC):

For this variable, all three years for TAC (US: $F = 216.80$, $p = .0001$, India: $F = 47.46$, $p = .0001$) differ significantly for the US and India; however, multiple comparisons for India show that the years 2021 and 2022 do not differ significantly ($p = .946$). The five most affected states in the US are California, Texas, Florida, New York, and Illinois, and in India, the most affected states are Maharashtra, Kerala, Karnataka, Tamil Nadu, and Andhra Pradesh. In terms of total cases and deaths, the most affected states in the US are California and Texas, and in India, it is Maharashtra and Kerala. However, the death ratio was higher for Florida in the US and Maharashtra in India.

3.3 Variation Plots for the US and India for Daily Mortality**Description of US Mortality:**

It can be seen from the box plots (Figure 3) that variations were high during the outbreak of the pandemic between April 20 and May 20. Again, peaks were observed during the months from December 20 to February 21, and then January 22 to February 22. This indicates that

there are variations in terms of deaths in these months. Most of the box plots are negatively skewed, indicating that more deaths occurred during the first half of the month. There is uniformity of deaths on May 20, November 20, February 21, March 21, November 21, and February 22, as the median line is approximately in the centre.

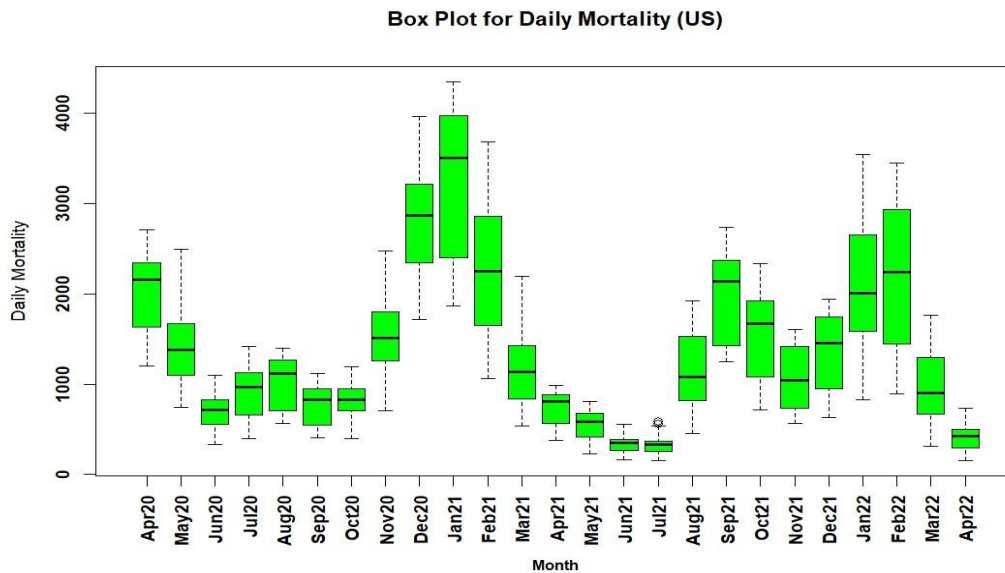


Figure 3: Month-wise Mortality for US

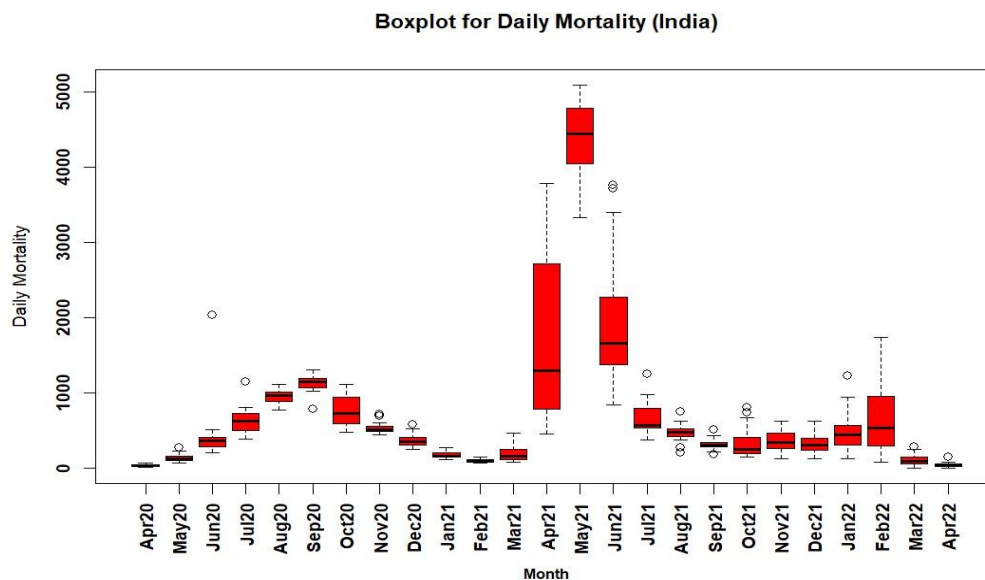


Figure 4: Month-wise Mortality for India

Description of India Mortality:

There are some outliers marked as (o) in May 20, June 20, and July 20 in the year 2020 (Figure 4). This may be due to reporting errors or because of the backlog of cases that have been reported for these days. Cases started to rise on April 21 and reached their maximum on May 21; after that, there was a downfall on June 21. On February 22, there was also a sudden

rise, and it went down on March 22 and April 22. Outliers were also observed in some other months. On October 20, April 21, June 21, July 21, and February 22, more deaths occurred in the second half of the month, as the median line is towards the first quartile. Since the median is not in the centre for any month, therefore, no month has a uniform number of deaths.

4 DEVELOPMENT OF GAM MODEL

The additive model generalizes the linear model by modeling the expected value of Y as

$$E[Y | X_1, \dots, X_k] = S_0 + S_1(X_1) + \dots + S_k(X_k) = S_0 + \sum_{i=1}^k S_i(X_i).$$

This model is a non-parametric model where the usual regression coefficients $\beta_0, \beta_1, \dots, \beta_k$ in the linear regression model are replaced by smoothers $S_1, \dots, S_k$ (where S_i is an unspecified non-parametric function). A Smoother is an arbitrary function that appears as a smooth curve through a scatter plot of points and is estimated in a nonparametric fashion. Sometimes, an additive model may contain a combination of parametric linear functions and non-parametric smooth functions.

4.1 GAM Model

If we combine additive models and a generalized linear model, we have the notion of Generalized Additive Model (GAM) written as

$$g(E[Y | X_1, \dots, X_k]) = S_0 + \sum_{i=1}^k S_i(X_i).$$

Generalized Additive Model consists of a random component, an additive component and a link function relating these two components. The response Y, the random component is assumed to possess a density in the exponential family. We define

$$\eta = S_0 + \sum_{i=1}^k S_i(X_i).$$

Note that $S_1, \dots, S_k$ are smooth functions and they are additive in nature. Finally, the relationship between mean μ of the response variable and η is defined through a link function $g(\mu) = \eta$ where $\mu = E[Y | X_1, \dots, X_k]$.

4.2 Generalized Cross Validation

Cross Validation (CV) works by leaving points (x_i, y_i) and estimating the smoother at x_i based on remaining $(n-1)$ points. The cross-validation sum of squares is

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^n \{y_i - \hat{f}_{\lambda}^{-i}(x_i)\}^2$$

where $\hat{f}_{\lambda}^{-i}(x_i)$ indicates the fit at x_i computed by leaving out the i^{th} data point. There is a simple way to define $\hat{f}_{\lambda}^{-i}(x_i)$, given only smoother matrix $\mathbf{S}_{\lambda}$. The Generalized Cross Validation (GCV) can be computed from the expression

$$GCV(\lambda) = \frac{1}{n} \sum_{i=1}^n \left\{ \frac{y_i - \hat{f}_{\lambda}(x_i)}{1 - \text{tr}(\mathbf{S}_{\lambda})/n} \right\}^2,$$

where $\hat{f}_{\lambda}(x_i) = \sum_{j=1}^n S_{ij}(\lambda) y_j$ and $\mathbf{S} = \{S_{ij}\}$ is an $n \times n$ matrix called smoother matrix.

where $\hat{f}_{\lambda}^{-i}(x_i) = \sum_{j=1, j \neq i}^n S_{ij}(\lambda) y_j$ and $\mathbf{S} = \{S_{ij}\}$ is an $n \times n$ matrix called smoother matrix. Fitting

of the BASE model is carried out with Natural Splines (NS) at (12,2,2) df by considering smoothers for day effect, true active cases (TAC), and active cases (AC). Note that TAC and AC may have an indirect effect on mortality and, therefore, we need to control them by smoothing. After that, the main variable, daily confirmed cases (DCC), is added to the model. We have worked out the PACF and residual plots for lag0, along with the predicted plot of mortality in R for the US and India. The relative risk (RR) along with 95% confidence intervals was computed from lag0 to lag5 separately for the US and India.

4.3 Smoothing Plots

In this section, smoothing plots for daily confirmed cases (DCC) and total confirmed cases (TCC) have been plotted simultaneously for the US and India for comparison purposes. For these variables, the US curve (except for a few days) is higher than that of India. The smoothers generally act as covariates, and after controlling their effect, the actual relationship between the dependent variable (mortality) and the independent variable (daily confirmed cases) can be determined. They also help us identify outliers present in the data. For smoothing plots, the optimal degrees of freedom (df) have been worked out by considering different knots (break points). The smoothers for DCC (df = 60) and TCC (df = 30) have been drawn with optimal degrees of freedom for all the 790 observations for the US and India. Smooth plots are presented in Figure 5 for the US and India. In DCC, peaks were obtained for US between 600 to 700 days while it was obtained for 350 to 450 days for India. The high peak for the US was also observed between 200 and 300 days. The smooth curve for total (cumulative) confirmed cases (TCC) in the US is much higher than in India, which means total confirmed cases in the US were much higher in the US than in India.

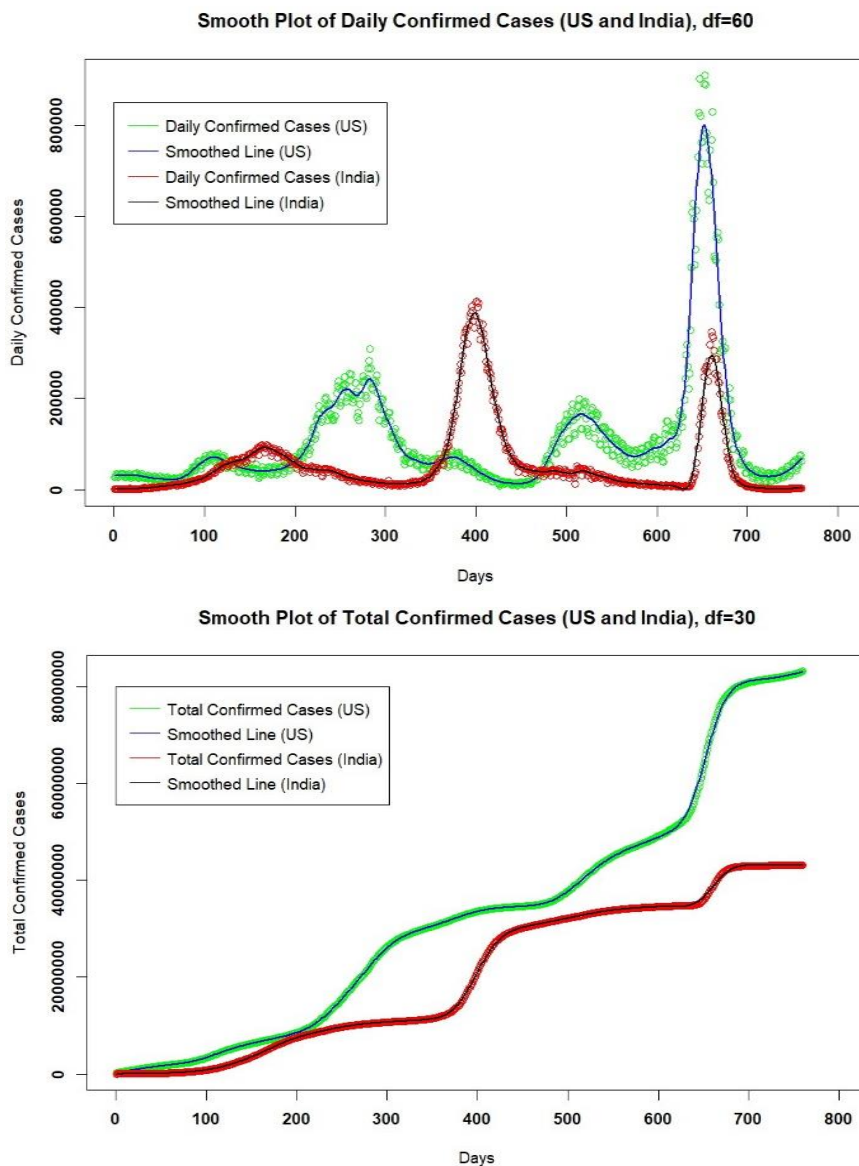


Figure 5: Smooth Plots for Daily Confirmed Cases and Total Confirmed Cases for US and India

4.4 Predictive Model for US

Model: NS (12,2,2) Family: Quasi-Poisson; Link function: log

Formula: Mortality $\sim$ DCC + s (Day, k = 12 + 1, fx = T, bs = "cr") + s (TAC, k = 2 + 1, fx = T, bs = "cr") + s (AC, k = 2 + 1, fx = T, bs = "cr")

Table 2: Significance of Main Variable and Smoothers (US)

Significance of Main Variable				
	Estimate	Std. Error	t-value	Pr(> t)
(Intercept)	6.911	.02043	80.156	< 0.00001*
DCC	0.0000009877	0.0000001496	6.601	<0.0001*
Significance of Smooth Terms (US)				
	Edf	Ref. df	F	p-value
s(Day)	12	12	76.18	< 0.00001*

s(TAC)	2	2	250.56	< 0.00001*
s(AC)	2	2	248.73	< 0.00001*
R-sq.(adj) = 0.806 Deviance explained = 82.7%				
GCV = 97.933 Scale est. = 92.974, n = 760				

It is evident from Table 2 that the effect of daily confirmed cases (DCC) on mortality is highly significant ($p < 0.0001$). The smoothers for day effect ($p < 0.00001$), TAC ($p < 0.0001$), and AC ($p < 0.00001$) are also significant. The deviance explained by the model is 82.7%, with a GCV score of 97.933. The relative risk, along with 95% confidence intervals, is given in Table 3. The relative risk has been computed for the increase in every 1000 daily confirmed cases. For an increase in the number of daily confirmed cases, the expected number of deaths amounts to 1.00099 (approximately) for the US at lag 0.

Table 3: Relative risk along Lag Effects with 95% confidence intervals (US)

Lag	GCV	β -coefficient	Stand. Error	p-value	RR	Lower	Upper
0	97.93	0.0000009877	0.0000001496	<0.00001*	1.000988188	1.000694725	1.00128174
1	102.95	0.0000009877	0.0000001496	<0.00001*	1.00098819	1.000694725	1.001281737
2	98.44	0.0000001343	0.0000001565	0.391	1.00013431	0.999827575	1.000441137
3	95.457	0.0000009114	0.0000001637	<0.00001*	0.99908902	0.998768507	0.999409626
4	100.96	0.0000012450	0.0000001727	<0.00001*	0.99875577	0.998417761	0.999093903
5	100.41	0.0000004983	0.0000001815	0.00619*	0.99950182	0.999146325	0.99985745

The unprecedented challenges caused by the COVID-19 pandemic demand timely action. However, due to the complex nature of different variants from time to time, a lag is likely to exist for several days, generally 0 to 14 days (quarantine period). Lag0 stands for same-day mortality, lag1 stands for the stands for the impact of DCC on mortality the next day, and so on. To get the estimate RR, we carried out the lag analysis from lag0 to lag5, and the results are presented in Table 3 with the GCV score, relative risk (RR), and 95% confidence intervals. It is evident from Table 3 that the lag effect is non-significant for lag2 ($p = .391$), that is, the second day, but it is significant on all other days until lag5. The PACF plot at lag0 (Figure 6) shows that mortality is high on the 5th, 6th, and 7th days for the US.

4.5 Residuals, PACF, and Predicted Plot for Mortality (US): NS (12, 2, 2) Model

There is a significant effect of daily confirmed cases (DCC) on mortality after smoothing the effects of the day (day-effect), true active cases (TCC), and active cases (AC). The points in the residual plots are symmetrically distributed over and above the baseline zero, leading to a justification of the model. The predicted plot of mortality at lag0 can be used to determine episodes of mortality during a particular day (Figure 6). A prediction for episodes of mortality seems to be reasonably well made, as can be seen from the predicted plot. The effect of daily confirmed cases on mortality can not only be seen on a particular day, but it may last for months. The PACF plots (Figure 6) show that mortality is high (i.e., significant) during the first day at lag 0, and after that, it comes out to be significant on the 5th, 6th, and 7th days, which means the effect of daily confirmed cases on mortality has a severe effect for the first seven days. Moreover, the predicted plot closely fits the actual data (Figure 7). Hence, the NS (12, 2, 2) model works well with US data.

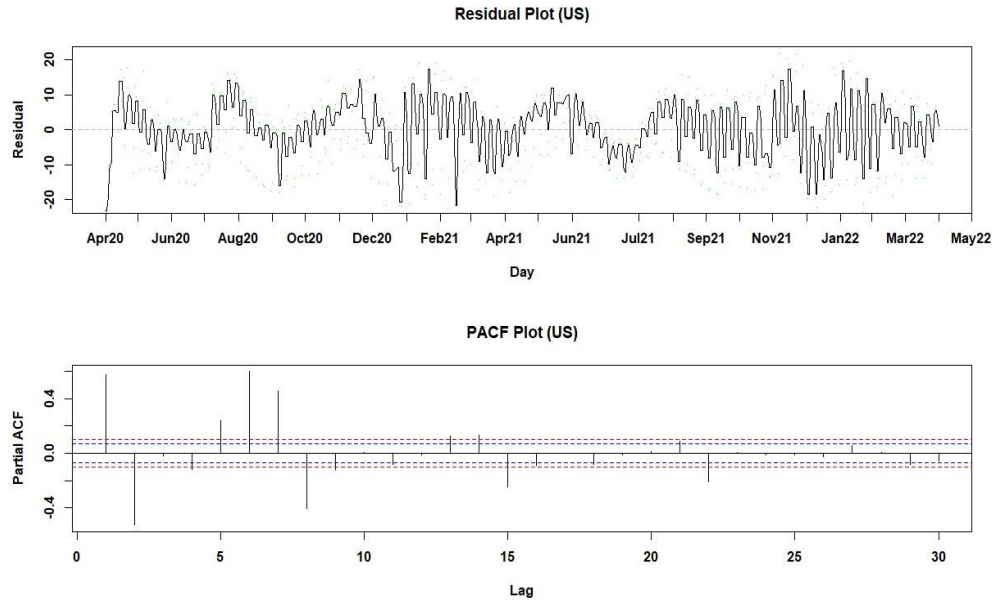


Figure 6: Residual and PACF Plots for US

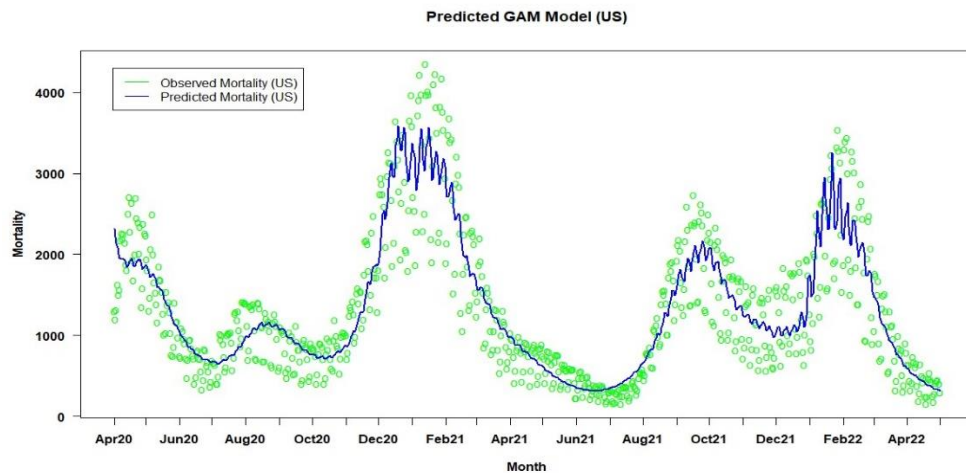


Figure 7: Predicted Generalized Additive Model for US

4.6 Predictive Model for India

Model: NS (12,2,2) Family: Quasi-Poisson; Link function: log

Formula: Mortality ~ DCC + s (Day, k = 12 + 1, fx = T, bs = "cr") + s (TAC, k = 2 + 1, fx = T, bs = "cr") + s (AC, k = 2 + 1, fx = T, bs = "cr")

Table 4: Significance of Main Variable and Smoothers (India)

Significance of Main Variable				
	Estimate	Std. Error	t-value	Pr(> t)
(Intercept)	6.045	0.03357	180.36	< 0.00001*
DCC	0.000002143	0.0000005189	4.13	< 0.00001*
Significance of Smooth Terms (India)				
	Edf	Ref. df	F	p-value

s(Day)	12	12	97.94	< 0.00001*
s(TAC)	2	2	52.51	< 0.00001*
s(AC)	2	2	38.91	< 0.00001*
R-sq.(adj) = 0.972 Deviance explained = 96.1%				
GCV = 33.819 Scale est. = 39.02, n = 760				

From Table 4, it follows that daily confirmed cases (DCC) are highly associated with mortality ($p < 0.00001$). The smoothers for the day effect ($p < 0.00001$), TAC ($p < 0.0005$), and AC ($p < 0.00001$) are significant. The deviance explained (DE) by the model is 96.1%, with a GCV score of 33.819. The relative risk along with 95% confidence intervals and the GCV score for lag0 to lag5 are presented in Table 5.

Table 5: Relative risk along Lag Effects with 95% confidence intervals (India)

Lag	GCV	β -coefficient	Stand. Error	p-value	RR	Lower	Upper
0	33.81	0.0000021430	0.0000005189	<.00001*	0.997859295	0.9968449	0.9988746
1	32.73	0.0000035930	0.0000005770	<0.00001*	0.996413447	0.995287	0.997541
2	33.61	0.0000029780	0.0000006350	<0.00001*	0.99702643	0.995786	0.998268
3	34.59	0.0000013660	0.0000006724	0.0425*	0.998634933	0.99732	0.999952
4	34.85	0.000000693	0.0000006838	0.919	0.999930662	0.998591	1.001272
5	34.28	0.0000022220	0.0000006679	0.000921*	1.00222447	1.000913	1.003537

4.7 Residuals, PACF and Predicted Plots for Mortality (India): NS (12,2,2) Model

For India also, there is a significant effect of daily confirmed cases (DCC) on mortality after smoothing the effects of day, TAC, and AC. The predicted plot of mortality at lag0 can be used to determine episodes of mortality during a particular day. The PACF plots (Figure 8) show that mortality is high (i.e., significant) during the first day at lag0, and after that, it again comes out to be significant on the 6th and 7th days, which means the effect of daily confirmed cases on mortality has a severe effect on these days. Figure 9 shows that the predicted plot closely fits the actual data. Hence, the NS (12, 2, 2) model also fits well with Indian data.

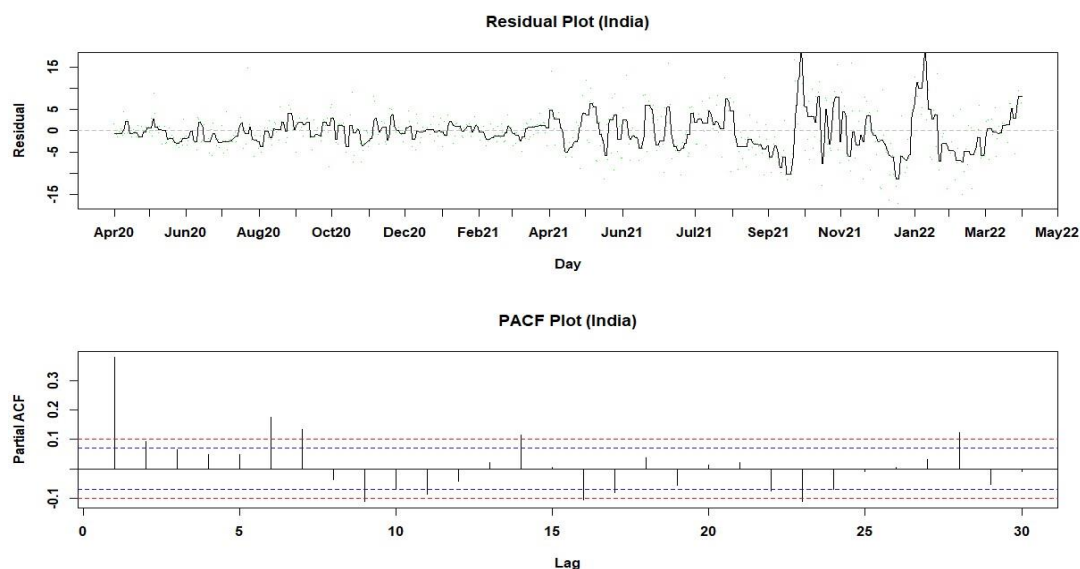


Figure 8: Residual and PACF Plots for India

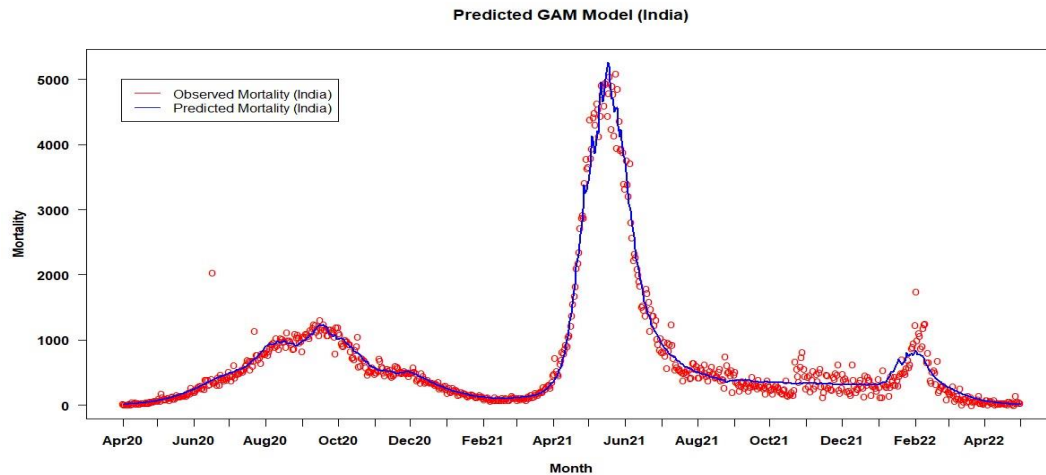


Figure 9: Predicted Generalized Additive Model for India

Future Predictions for Mortality:

The validity of the model has been assessed by taking 15 random observations for which the values of the covariates, viz., day, TAC, and AC, were available. Actual mortality was also available. The predicted and actual mortality are shown in Table 6, and there is not much difference between actual and predicted mortality.

Table 6: Model Validation and Future Predictions

Predicted Mortality by GAM and Actual Mortality						
Day	TCC	AC	DCC	Predicted Mortality	Actual Mortality	(Observed-Predicted) ² /Predicted
1	579691	445763	38532	493.93	483	0.241866
2	763536	705384	70068	944.78	941	0.015124
3	580594	442867	44276	500.88	501	0.000029
4	331951	174735	12569	133.88	131	0.061954
5	12311	11824	922	37.08	38	0.022826
6	404543	252195	19046	254.37	251	0.044647
7	1831761	1461216	114537	2848.89	2807	0.615949
8	699611	651413	67066	957.18	971	0.199537
9	1059981	962287	82767	1072.52	1057	0.224584
10	579691	445763	38532	493.93	483	0.241866
11	1536847	1156892	93896	2249.65	2275	0.285654
12	21367	20484	1607	58.01	56	0.069645
13	1831761	1461216	114537	2848.89	2807	0.615949
14	933166	860857	90600	1043.90	1067	0.51117
15	1092962	1003070	87382	1196.66	1155	1.450333
Total						4.601133

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Predicted})^2}{\text{Predicted}} = 4.601133$$

The value of Chi-square statistics at 14 degrees of freedom is non-significant ($p = .99$). Thus, the model satisfies the goodness-of-fit test. For the above table, we also computed G^2 Statistics = 1.9056 with a p-value of 0.999 to verify the null hypothesis that there is no significant difference between observed and predicted values.

5 CONCLUSION AND DISCUSSION

The COVID-19 pandemic has currently expanded its roots to more than 200 countries in the world. Morbidity and mortality are not only threats to humans, but their impact on the economy is indirectly affecting us. The present study has been carried out to see the trends of coronavirus in the US and India. The data collected from various websites (mentioned in the references) has been subjected to statistical analysis. Different statistical tools, including descriptive statistics, graphs (showing trends), box-plots for variation, and the GAM model for predictions, were used. The following conclusions follow from this study:

During the 760 days between April 1, 2020, and April 30, 2022, there were 1017381 deaths in the US and 523807 in India. In this period, the mortality rate on average was 1339 per day in the US and 689 per day in India. There were spikes in mortality, particularly in December-20 and January-21, when around 2806 and 3244 deaths per day were recorded for the US, but for India, it was in April-21 and May-21, where almost 1689 and 4398 deaths per day were reported, and in June-21 also, the number of deaths went up to 1897. The five most affected states in the US are California, Texas, Florida, New York, and Illinois, and in India, the most affected states are Maharashtra, Kerala, Karnataka, Tamil Nadu, and Andhra Pradesh. The death ratio was higher for Florida State in the US and in India for the State of Maharashtra. The comparative study of the US and India for each month shows that there was a significant difference for all months for TCC ($p < .0001$), DCC ($p < .0001$, except for months Oct-20 and July-21, $p > .05$), AC ($p < .0001$, except June-21, $p > .05$), TR ($p < .0001$, except October-20, November-20, $p > .05$), Total Mortality ($p < .0001$), Daily Mortality ($p < .0001$, except October-20, $p > .05$), and TAC ($p < .0001$). Year-wise comparisons (2020 vs. 2021; 2020 vs. 2022; 2021 vs. 2022) for TCC, DCC, AC, TR, Total Mortality, and Daily Mortality show that there was a significant difference between the US and India ($p < .0001$). TAC also differs for the US in all three years; however, for India, it does not differ significantly for the years 2021 and 2022. The variation in mortality was high at the outbreak of the pandemic on April 20 and May 20.

Again, peaks were observed during the months from December 20 to February 21, and then January 22 to February 22 for the US. In India, there are some outliers on May 20, June 20, and July 20 for the year 2020. This may be due to reporting errors or because of the backlog of cases that have been reported these days. COVID-19 has affected day-to-day life, and its effects on health, business, the economy, and social factors are visible. The mortality was high among people who were suffering from co-morbidities, particularly chronic diseases like asthma, cardiovascular, liver, kidney, diabetes, etc. This information is based on recent studies related to COVID-19. A generalized additive model (GAM) was performed to see the effect of daily confirmed cases on mortality. Fitting of the BASE model is carried out with Natural Splines (NS) at (12,2,2) df by considering the day effect, active cases, and true active cases as covariates. In time series data, the problem of autocorrelation is likely to occur; therefore, smoothing was done for covariates with appropriate degrees of freedom. At various lags, most of the smoothers were found to be significant. It has been observed that the lag effects of daily confirmed cases (DCC) on mortality were significant. Mortality was high on the 5th, 6th, and 7th

days for the US and the 6th and 7th days for India. This is evident from partially autocorrelation function (PACF) plots. Analysis was performed for lag0 to lag5, and the risk ratio (RR) along with confidence intervals were computed. For the US, the lag effects of the second day (lag1) and India's fifth day (lag4) were non-significant.

6 LIMITATIONS OF THE STUDY

The data for sex, age, and co-morbidities was not available, and therefore separate predictions for males and females were not possible. Sensitivity analysis for different age groups (≤ 20 years, 20–50 years, and >50 years) could not be performed. Due to the non-availability of specific co-morbidities, age-sex and disease-specific analyses could not be performed.

Author Contributions

All the authors have made equal contributions.

Conflict of Interest

The authors declare no competing financial and non-financial interests.

Acknowledgment

The authors are thankful to VA Medical Centre, Dallas, Texas, USA, and Panjab University, Chandigarh, India, for their support in carrying out this joint research project.

Funding

The project has been approved by the IRB at the VA Medical Center in Dallas, Texas, but no funding has been provided to carry out this research work.

Disclaimers

The opinions expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the US Government.

Data Availability Statement

The data for the USA and India is available with the authors. It is also accessible through the website.

<https://www.kaggle.com/datasets/josephassaker/covid19-global-dataset>

References

1. Nyberg, T., Twohig, K. A., Harris, R. J., Seaman, S. R., Flannagan, J., Allen, H. & Presanis, A. M. (2021). Risk of hospital admission for patients with SARS-CoV-2 variant B. 1.1. 7: cohort analysis. *bmj*, 373.
2. Bager, P., Wohlfahrt, J., Fonager, J., Rasmussen, M., Albertsen, M., Michaelsen, T. Y., & Krause, T. G. (2021). Risk of hospitalization associated with infection with SARSCoV-2 lineage B. 1.1. 7 in Denmark: an observational cohort study. *The Lancet Infectious Diseases*, 21(11), 1507-1517.
3. Ministry of Health and Family Welfare (GOI). (2020). Novel Coronavirus-MOHFW. Retrieved from <http://www.mohfw.gov.in/>
4. Chin, A. W., Chu, J. T., Perera, M. R., Hui, K. P., Yen, H. L., Chan, M. C., & Poon, L. L. (2020). Stability of SARS-CoV-2 in different environmental conditions. *The Lancet Microbe*, 1(1), e10.

5. Aggarwal, R., Sharma, S. K., & Jain, K. (2014) Optimal Choice of Splines and Knots in TPSPLINE and TRANSREG Procedures. *IOSR Journal of Mathematics*, 10(1), Ver -II, 42-52.
6. De Boor, C., & Splines, A. P. G. T. (1978). Springer-Verlag. New York, 7.
7. Eilers, P. H., & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical science*, 11(2), 89-121.
8. Green, P. J., & Silverman, B. W. (1993). Nonparametric regression and generalized linear models: a roughness penalty approach. Crc Press.
9. Gu, C., & Gu, C. (2013). Smoothing spline ANOVA models (Vol. 297). New York: Springer.
10. Hastie, T., & Tibshirani, R. (1987). Generalized additive models: some applications. *Journal of the American Statistical Association*, 82(398), 371-386.
11. Hastie, T.J., & Tibshirani, R.J. (1990). Generalized Additive Models. Monographs on Statistics and Applied Probability (No. 43). New York, NY: Chapman and Hall.
12. Patone M., Xue, W. M., Lahiru H., et al. (2021). Risks of myocarditis, pericarditis, and cardiac arrhythmias associated with COVID-19 vaccination or SARS-CoV-2 infection.

Websites

<https://www.kaggle.com/datasets/josephassaker/covid19-global-dataset>

Links for USA

https://github.com/CSSEGISandData/COVID19/tree/master/csse_covid_19_data/csse_covid_19_daily_reports_us

<https://covidtracking.com/data/download>

<https://app.powerbigov.us/view?r=eyJrIjoiMjA2ZThiOWUtM2FhNS00MGY5LWFMmYj>

[UtNmQwNTQ3Nzg5N2I2IiwidCI6ImU0YTM0MGU2LWI4OWUtNGU2OC04ZWZhL](https://app.powerbigov.us/view?r=eyJrIjoiMjA2ZThiOWUtM2FhNS00MGY5LWFMmYj)

[TE1NDRkMjcwMzk4MCJ9](https://app.powerbigov.us/view?r=eyJrIjoiMjA2ZThiOWUtM2FhNS00MGY5LWFMmYj)

Links for India <https://ourworldindata.org/>___<https://www.thetoptens.com/countries-with-covid-19-coronavirus-cases/>_<https://prsindia.org/covid-19/cases>