**METHODS**

We used three forecasting models (i.e., simple exponential smoothing (SES), auto-regressive integrated moving average, and automatic time-series forecasting models), to identify the global trend of rCFR for COVID-19. Second, we used the Mann-Kendall (M-K) trend analysis to identify existence of any trend and the direction of the trend (increasing or decreasing). Finally, we developed generalized linear mixed model (GLMM) of explanatory variables to identify whether the variables have any relationship between the country’s rCFR of COVID-19. All these three different approaches helped us to make a plausible conclusion on the global trend of COVID-19 CFR and factors affecting the CFR of COVID-19 in different phase of pandemic. All analyses were carried out using the statistical software R version 3.5.2.2.

**COVID-19 data**

The necessary COVID-19 related data, including daily new cases, daily new deaths, total deaths, and total deaths per million, vaccination, and total cases from the WHO daily COVID-19 situation reports of 210 countries were collected from January 01, 2020 to August 31, 2021. The ARIMA, SES and Prophet models were fitted for the full dataset.1

**Reported case-fatality rate (rCFR)**

We estimated cumulative rCFR COVID-19 as the number of deaths per 100 COVID-19 confirmed cases. As the number of cases and deaths both are a fraction of total cases or deaths, we considered the term as reported CFR or simply as rCFR.2,3

**Time series model to predict the trend**

We performed three time-series model including SES, ARIMA and Prophet models to identify the global trend of rCFR for COVID-19. We selected all these time series models as the outcome variable (cumulative rCFR) are dependent on the previous records and all these three models can take this into account. Using the time series models with the reported COVID-19 data, we forecasted trends for the prospective 10-days and visualizing in the figure. SES was used as a benchmark to compare the performance of the ARIMA and Prophet models. We also used M-K trend analysis to identify the daily or weekly cumulative trend (increasing or decreasing) of COVID-19 rCFR.

***Simple Exponential Smoothing:***

Simple exponential smoothing is one of the familiar methods for forecasting procedures.14 The SES is a short-term forecasting model that assumes data fluctuates around a relatively stable mean.15 For infectious diseases in general, this method has been shown to be reasonably accurate and reliable.16–18 It takes into account the more recent observations and exponentially reduces the weights of older observations.19 The SES model for this study had been carried out using R package ‘fpp2’.20

***Auto-Regressive Integrated Moving Average (ARIMA):***

We performed an ARIMA model to forecast the trend of global weekly cumulative rCFR. The ARIMA model is an exploratory, data-oriented method that allows the user to fit an appropriate model adapted from the structure of the data itself. 21 This model assumes that the time series values are linearly related and intends to extract local patterns by eliminating high-frequency noise from the data.22

The benefit of ARIMA models is the ability to adjust to dynamically oriented system which evolve over time by updating the model to forecast the system's future state based on recent events.23 The ARIMA model for this study had been carried out using R package ‘forecast’.24

***Automatic Forecasting time-series model (Prophet):***

We also performed a decomposable automatic forecasting time-series model called ‘Prophet’ using R package “prophet” to predict the 10-days fatality rate and compared with rCFR.25 The Prophet model ignores the temporal dependence of the data. Moreover, the irregular observations are allowed in the data set and the model fits very quickly.26 It is also robust for missing data and generally manages outliers well.27 There are three main features of the model, i.e., trend, seasonality, holidays. It can be represented as,

*Y (t) = g(t) + s(t) + h(t) + ∈t*

where the model parameters g(t), s(t), h(t), ∈t is piecewise linear curve for modelling non-periodic changes in time series, periodic changes, the effects of holidays with irregular schedules considered in the model by some parameters, respectively. The error term accounts for any unexpected changes that the model does not account for.27

***Mann-Kendall (M-K) trend:***

We used weekly cumulative rCFR data and performed the M-K trend test to identify the trend of COVID-19 rCFR for both pre-peak and post-peak period.28

The M-K method is a non-parametric test that provides an indicator of whether there is a monotonous trend and whether there is a positive or negative trend.28 . The M-K test statistic is robust when dealing with non-normally distributed data, censored data, and time series with missing values because it is calculated by ranks and sequences of time series rather than the original values.29

In addition, the Sen’s slope test was applied to determine the changes in COVID-19 rCFR in both periods.30 M-K and Sen’s slope trend analysis had been carried out using R package ‘trend’.31

**Empirical evaluation**

The ARIMA and Prophet models are empirically assessed by comparing their results to benchmarks in predicting the rCFR.. This benchmark permitted us to assess the performance gains made by their counterparts.32 The SES also allows the most appropriate non-seasonal model for each series, allowing for any kind of error or trend component. Then, we analyse and compare the performance of the studied time series models with some of the commonly used measures to evaluate the prediction significance including coefficient of determination (R2), root mean square error (RMSE), and mean absolute error (MAE).

**Outcome and predictor variables**

We used rCFR as the outcome variable, we also collected and used several predictors data from the World Bank and other UN sources such as population density,5 percentage of people above 65 years of age,6 Gross Domestic Product (GDP),7 worldwide governance indicators (WGI)8 and Global Health Security Index (GHSI),9 the prevalence of obesity 10 in our analyses. We also included country-specific prevalence of diabetes and cardiovascular disease to explain the variation of COVID-19 rCFR. The GHSI index scored between 0 and 100 to indicate the country’s capacity for early detection and reporting for epidemics.9 The WGI scored between -2.5 and 2.5, where -2.5 indicates the weakest and 2.5 indicates the strongest governance performance.8 The median age of the diagnosed people (daily) is an important variable which we could not include in the model as these data are not publicly available for most countries of the world. The OxCGRT systematically collects information from 148 countries on several policy responses that governments have taken, scores stringency of such measures, and aggregates them into a common Stringency Index (SI) in a daily basis [10, 11]. In the Our World in Data4, the SI was calculated by using nine response indicators including school closures, workplace closures, and travel bans, rescaled to a value from 0 to 100 (100 = strictest). The value of the SI on any given day is the average value of these nine indicators. Thus, the index reports a number that reflects the overall stringency of the government’s response. Higher index indicates higher overall response level. A detailed description of the calculation of SI can be found in a file (https://www.bsg.ox.ac.uk/sites/default/files/Calculation%20and%20presentation%20of%20the%20Stringency%20Index.pdf).

**Statistical analysis**

We observed that the rCFR of COVID-19 has changed over time (**Fig. 1**). Using time-series model alone would not allow us to identify the reason behind the increasing and decreasing trend of COVID-19 rCFR. We explored whether the relationship between the rCFR of COVID-19 and country-level explanatory variables vary over time through generalized linear mixed models. As the trend of rCFR in both periods is different, we ran generalized linear mixed models to investigate the association between possible explanatory variables and tried to investigate which variables affect the most in both periods separately.

**Generalized Linear Mixed Models**

Generalized Linear Models have been largely used to model the behavior of single subjects. As the name suggests, GLMs are a generalization of Linear Models (LMs) to response variables that are not normally distributed11. The GLMM is an extension of the GLM that allows the analysis of clustered categorical data, as in the case of repeated responses from different subjects12. Some of the advantages of the GLMM analysis are: (a) the GLMM takes the whole ensemble of the responses as input data, (b) it separately estimates the variability of fixed and random effects, and (c) it allows an easier assessment of the goodness of fit13. The fixed component usually estimates the effect of interest, such as the experimental effect, whereas the random component estimates the heterogeneity between clusters (i.e., between subjects). In this way, we estimate a single model across all subjects, but we allow each subject to have a different variability and a different sample size14. The main advantage of GLMM is that it separates the levels of the models to account for the group effect nesting the lower level observations. In this study, locations are treated as the second level which group sequential observations within the same area, and independent variables are treated as repeated observations at the lower level. While the location data are assumed to be time-invariant, the independent data are assumed to be universal over the whole study areas at a certain time point. A reduction in the value of random effects represents more variation in the dependent variable and is explained by the selected variables (fixed effects). This model has a log link as stated in the top line of the summary. The model describes beta distribution family that has a logit link. We conducted GLMMs using the R software.

**RESULTS**

More than 233.77 million cumulative confirmed cases and 4.78 million deaths had been documented globally and the global rCFR of COVID-19 is reported as 2.05% as of September 30th, 2021. The weekly global cumulative rCFR of COVID-19 reached a peak at 7.23% during the 17th Epidemiological week (April 22-28, 2020). The top five countries with COVID-19 rCFR are Yemen (18.74%), Peru (9.23%), Mexico (7.74%), Sudan (7.51%), and Syria (7.24%) (**Fig. 2 and S1**). The weekly mean cumulative rCFR was 3.61% (95% CI: 2.87-4.36).

In the SES model, we found a constant trend between observed and predictive global rCFR of COVID-19 with a R2, RMSE and MAE being 99.62%, 0.10 and 0.05, respectively (**Table 1 and Fig. 3**). In the ARIMA and Prophet Model, we found a strong declining trend between observed and predictive global rCFR of COVID-19 with a R2, RMSE and MAE value of 99.94% and 99.66%, 0.04 and 0.09, and 0.01 and 0.04, respectively (**Table 1**). In terms of accuracy, ARIMA model performed better over Prophet and SES model (with better R2, RMSE and MAE value). The coefficient of determination of the ARIMA model was the larger and errors are lower than Prophet and benchmark SES model. According to the forecast in both models, the ratio of COVID-19 rCFR is expected to decrease considerably in the coming 10 days. The forecasting of global cumulative rCFR of COVID-19 for each model are shown in **Fig. 3.**

In M-K trend analysis, we found a negative trend of cumulative rCFR (p <0.001 and tau = -0.82). In Sen’s slop test, the slope was -0.04 (95% CI: -0.05 to -0.03) (**Table 1**).

**Figure 3.** Top: Observed and predicted daily worldwide daily rCFR using Simple Exponential Smoothing (SES) model. Middle: Observed and predicted daily worldwide daily cumulative rCFR using Auto-Regressive Integrated Moving Average (ARIMA) model. Bottom: Observed and predicted daily worldwide daily cumulative rCFR using Automatic Forecasting time-series model (Prophet). The black dots indicate observed data, blue line indicate the predictive CFR and shaded area indicate 95% confidence interval of predicted CFR.

|  |
| --- |
| C:\Users\PHD_Mamoni\AppData\Local\Microsoft\Windows\INetCache\Content.Word\DRplotDailySES.TIFF |
| C:\Users\PHD_Mamoni\AppData\Local\Microsoft\Windows\INetCache\Content.Word\DRplotDailyProph.tiff |
| E:\ResearchProject\Jamal Sir\Vaccination\Update2\DRplotDailyARIMA.tiff |

**Table 1.** The summary of Simple Exponential Smoothing (SES), Auto-Regressive Integrated Moving Average (ARIMA), Automatic forecasting time-series model (Prophet), Mann-Kendall (M-K) trend and Sen’s slope analysis. The SES, ARIMA and Prophet models used daily cumulative case-fatality rate (CFR) data whereas the M-K trend analysis and Sen’s slop used weekly cumulative CFR data. The Kendall’s Tau value permits a comparison of the strength of correlation between two data series (here, week of the year 2020 and rCFR) 28.

|  |  |  |  |
| --- | --- | --- | --- |
| **Method & Period** | **R2** | **RMSE** | **MAE** |
| ***Simple Exponential Smoothing*** | | | |
| Overall | 99.62% | 0.10 | 0.05 |
|  |  |  |  |
| ***Auto-Regressive Integrated Moving Average*** | | | |
| Overall ARIMA (5,2,1) | 99.94% | 0.04 | 0.01 |
|  |  |  |  |
| ***Automatic Forecasting time-series model*** | | | |
| Overall | 99.66% | 0.09 | 0.04 |
| ***Mann-Kendell trend analysis*** | | | |
|  | **tau** | **P** | |
|  | -0.82 | <0.001 | |
| ***Sen’s slop test*** | | | |
|  | **Sen’s Slope** | **95% CI** | |
|  | -0.04 | -0.05 to -0.03 | |

*RMSE: Root Mean Square Error; MAE: Mean Absolute Error*

The results of all GLMMs are summarized in Table. The estimated effect of each variable is presented in relative risk (RR) and its significance is shown by its p-value. The 95% confidence intervals are also provided.

The table includes the various covariates and the random intercept in the model. The intraclass correlation coefficient (ICC) of .648 was calculated by dividing the variance of the random effect by the total variance. Thus, the spatial unit effects account for approximately 64.8% of the total variance of weekly rCFR, which suggests Moderate reliability on location effects on weekly rCFR. It is also to be noted that with the introduction of a random intercept, vaccination, population density, GDP, weeks and stringency index has significantly negative effects on weekly rCFR and the percentage of people aged 65 and above, GHSI, WGI, and obesity have significantly positive effects.

In the generalized linear mixed model, the percentage of people aged 65 years or above the age of the population of the country were significantly positively (RR: 1.18, 95% CI: 1.01–1.39) associated with COVID-19 rCFR (Table). The COVID-19 vaccination (0.34 [0.16–0.71]), GDP (0.75 [0.65–0.88]), and stringency index (0.86 [0.85–0.87]) was negatively significantly associated with the COVID-19 rCFR.

Table: Factors associated with rCFR of COVID-19 using generalized linear mixed model

|  |  |  |  |
| --- | --- | --- | --- |
| **Variables** | **RR** | **95% CI** | **P-value** |
| Vaccination | 0.335 | 0.159- 0.705 | 0.004 \*\* |
| Weeks | 0.885 | 0.808- 0.970 | 0.009 \*\* |
| The percentage of people aged 65 and above | 1.181 | 1.003- 1.390 | 0.046 \* |
| Population density | 0.888 | 0.792- 0.996 | 0.042 \* |
| GDP | 0.753 | 0.647- 0.877 | 0.002 \*\*\* |
| GHSI | 1.121 | 0.952- 1.321 | 0.171 |
| WGI | 1.059 | 0.914- 1.228 | 0.443 |
| Obesity (%) | 1.129 | 0.997- 1.277 | 0.055 |
| Stringency index | 0.861 | 0.849- 0.873 | <0.001 \*\*\* |
| Vaccination: Weeks | 1.252 | 1.057- 1.482 | 0.009 \*\* |
|  |  |  |  |
| **Groups Name** | **Variance** | **Std.Dev.** |  |
| Location (Intercept) | 0.3645 | 0.6038 |  |
| Weeks (Intercept) | 0.0405 | 0.2012 |  |
|  |  |  |  |
| **AIC** | -161199.0 | **Conditional R2** | 0.717 |
| **BIC** | -161097.5 | **Marginal R2** | 0.197 |
| **RMSE** | 0.0002244 | **ICC** | 0.648 |

*Note. RR = relative risk; CI = confidence interval.*

*\*p < 0.1. \*\*p < .05. \*\*\*p < .01.*

Reference:

1. Max Roser, Hannah Ritchie, E. O.-O. and J. H. Coronavirus Pandemic (COVID-19) - Statistics and Research - Our World in Data. https://ourworldindata.org/coronavirus (2020).

2. Mohammadpour, S., Torshizi Esfahani, A., Halaji, M., Lak, M. & Ranjbar, R. An updated review of the association of host genetic factors with susceptibility and resistance to COVID-19. *Journal of Cellular Physiology* (2020) doi:10.1002/jcp.29868.

3. Hasan, M. N. *et al.* The Global Case-Fatality Rate of COVID-19 Has Been Declining Since May 2020. *Am. J. Trop. Med. Hyg.* (2021) doi:10.4269/ajtmh.20-1496.

4. Data, O. W. in. COVID-19: Stringency Index, Oct 2, 2021. https://ourworldindata.org/grapher/covid-stringency-index (2021).

5. World Data Bank. Population density (people per sq. km of land area). *World Development Indecators* https://data.worldbank.org/indicator/EN.POP.DNST (2018).

6. The World Bank. Population ages 65 and above, total. *World Bank Data* https://data.worldbank.org/indicator/SP.POP.65UP.TO.ZS (2018).

7. The World Bank. GDP (current US$) | Data. *The World Bank* https://data.worldbank.org/indicator/NY.GDP.MKTP.CD (2018).

8. WGI. WGI 2019 Interactive. *WGI 2019* 1 http://info.worldbank.org/governance/wgi/ (2019).

9. Nuclear Threat Initiative (NTI) & Johns Hopkins Center for Health Security (JHU). GHS Index. https://www.ghsindex.org/ (2020).

10. WHO. *Global health services data, obesity and overweight*. (2020).

11. Moscatelli, A., Mezzetti, M. & Lacquaniti, F. Modeling psychophysical data at the population-level: The generalized linear mixed model. *J. Vis.* **12**, 26–26 (2012).

12. Samur, A. A., Coskunfirat, N. & Saka, O. Comparison of predictor approaches for longitudinal binary outcomes: application to anesthesiology data. *PeerJ* **2**, (2014).

13. Harrison, X. A. *et al.* A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ* **6**, (2018).

14. Bell, A., Fairbrother, M. & Jones, K. Fixed and random effects models: making an informed choice. *Qual. Quant. 2018 532* **53**, 1051–1074 (2018).