Transactions of the Indian National Academy of Engineering (2020) 5:109–115
https://doi.org/10.1007/s41403-020-00112-y

ORIGINAL ARTICLE

Temporal Dynamics of COVID-19 Outbreak and Future Projections: A Data-Driven Approach

Rajesh Ranjan[1]

Received: 14 April 2020 / Revised: 18 May 2020 / Accepted: 22 May 2020 / Published online: 6 June 2020

Abstract

Long-term predictions for an ongoing epidemic are typically performed using epidemiological models that predict the timing of the peak in infections followed by its decay using non-linear fits from the available data. The curves predicted by these methods typically follow a Gaussian distribution with a decay rate of infections similar to the climbing rate before the peak. However, as seen from the recent COVID-19 data from the US and European countries, the decay in the number of infections is much slower than their increase before the peak. Therefore, the estimates of the final epidemic size from these models are often underpredicted. In this work, we propose two data-driven models to improve the forecasts of the epidemic during its decay. These two models use Gaussian and piecewise-linear fits of the infection rate respectively during the deceleration phase, if available, to project the future course of the pandemic. For countries, which are not yet in the decline phase, these models use the peak predicted by epidemiological models but correct the infection rate to incorporate a realistic slow decline based on the trends from the recent data. Finally, a comparative study of predictions using both epidemiological and data-driven models is presented for a few most affected countries.

Keywords COVID-19 · Coronavirus · India · Epidemiology

Introduction

In recent days, Coronavirus disease 2019 (COVID-19) has emerged as an unprecedented challenge before the world. This disease is caused by a novel coronavirus SARS-CoV-2, for which there is no specific medication or vaccine approved by medical authorities. This disease is transmitted by inhalation or contact with infected droplets or fomites, and the incubation period may range from 2 to 14 days (Wu and McGoogan 2020). This disease can be fatal to the elderly patients (about 27% for 60+ age groups), and those with underlying co-morbid conditions (Yang et al. 2020). As of May 15, 2020, there have been about 4.6 million confirmed cases of COVID-19 and about 300,000 reported deaths globally.

A realistic estimate of intensity and temporal distribution of this epidemic can be beneficial to design key strategies to regulate the quarantine as well as to prepare for social and economic consequences due to lockdown. However, as seen from the recent literature (Roda et al. 2020), the predictions by epidemiological models for an ongoing spread are often unreliable as they do not accurately capture the dynamics of COVID-19 in the absence of established parameters. In this work, we propose data-driven models for COVID-19 decay purely based on characteristics of COVID-19 spread, and thus include the effects of lockdown and other key factors. For subsequent discussions, the default year is 2020 and all the statistics are based on data till May 15, 2020, unless otherwise specified.

First, we examine the dynamics of COVID-19, before and after the lockdown as shown in Fig. 1a. The abscissa indicates the days shifted by the date when the lockdown was imposed or other intervention measures were taken (see list in Ranjan 2020b). Thus, the four phases indicate: (1) Early slow epidemic growth ($t < t_{-1}$), (2) initial exponential growth ($t_{-1} < t < t_0$) typical of an epidemic, (3) continuing exponential growth during lockdown based on the incubation period of SARS-CoV-2 ($t_0 < t < t_1 \equiv t_0 + 14$) and (4) expected deceleration phase ($t > t_1$). In Fig. 1a, both China and South Korea (SK) show a very rapid arrest of the COVID-19 growth post interventions ($t > t_1$), while other

✉ Rajesh Ranjan
ranjan.25@osu.edu

[1] Department of Mechanical & Aerospace Engineering, The Ohio State University, Columbus, OH 43210, USA

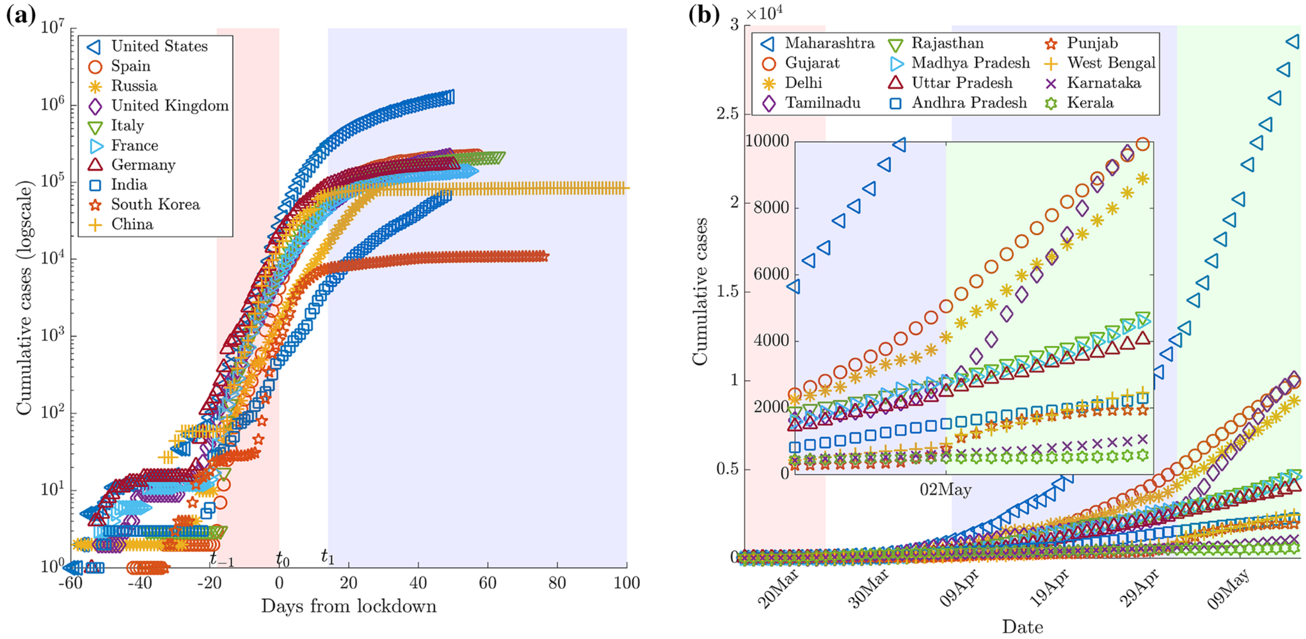

Fig. 1 Dynamics of COVID-19 before and after interventions for key countries in the world(**a**), and most affected states in India (**b**)

countries display just a slowdown evident by the change in slope. Further, the growth curves for India and Russia are much more rapid in this phase compared to those for other countries .

The differences in COVID-19 spread among geographical regions after the lockdown can be better visualized on a linear scale as shown in Fig. 2a. Most of the countries considered in the figure took social distancing measures before the end of March, so it is expected that the effects of interventions should become visible latest by mid-April. Both the US and the UK exhibit linear growth in this period, while other European countries show initial linear growth followed by a slow flattening (Ansumali and Prakash 2020). The curves for India and Russia are closer to exponential

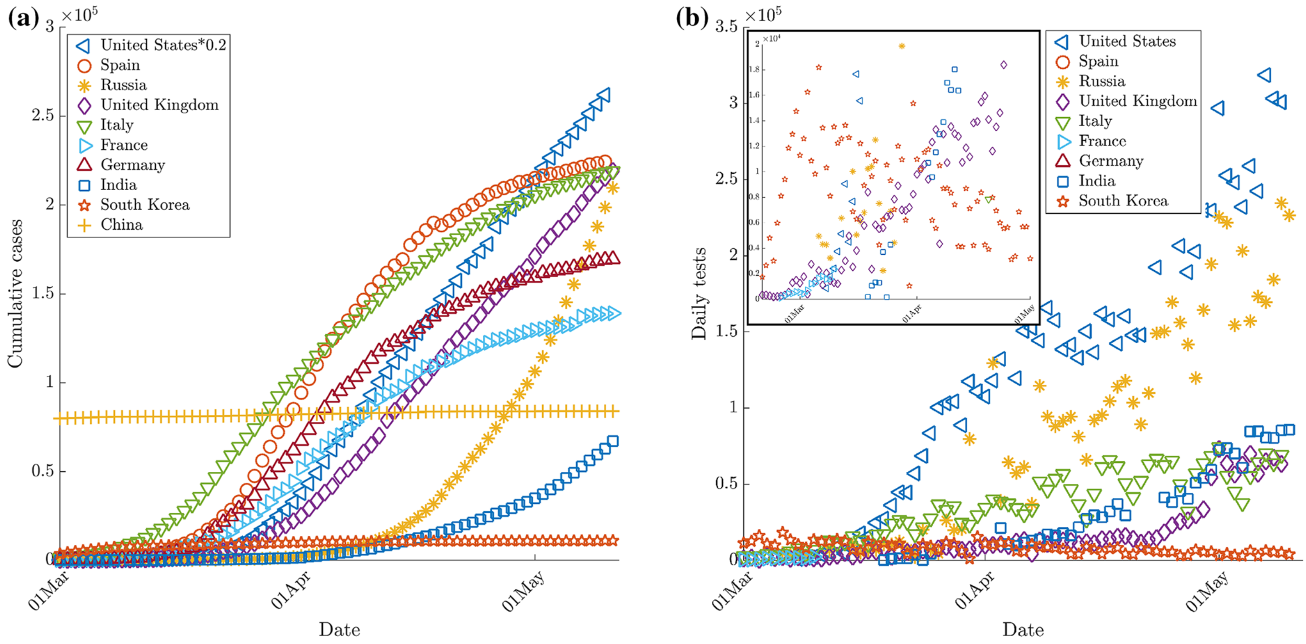

Fig. 2 (**a**) Growth curves, (**b**) daily number of tests for different countries

growth. A relatively low number of infections for India despite exponential growth reflects the early measures of India in international flight suspension and lockdown.

Very different trends of the curve after the lockdown indicate a disparity in compliance levels of social distancing measures. For example, in the US, each state follows its norm of intervention, and social distancing measures are imposed on different dates. An implication of this is that when the initially most impacted states like New York and New Jersey started showing signs of flattening in late April, other states like Illinois, Massachusetts and California displayed surges in the number of cases, thereby keeping the overall growth in the US on a linear course.

The case of India is also compelling as it first displayed a strong impact of lockdown (close to 70% compliance, as suggested in Ranjan 2020a) despite a few local outbreaks. This led to a linear growth for some time, but an escalation in early May cases put India on a near exponential course. To further examine this, we plot COVID-19 distribution in key affected states in India in Fig. 1b. The time-series data is divided into four periods, with the first three being before, during and after lockdown similar to that in Fig. 1a and the last one from May 2 (green shade) when a surge in the number of cases in many states put India onto the exponential course. Fig. 1b, we note a varying distribution of COVID-19 among Indian states much like in the US, just four states - Maharashtra, Gujarat, Tamilnadu, and New Delhi contribute to about 70% of the total cases.

Among these, the most affected states, Maharashtra and Gujarat are on the course of exponential growth. Delhi, Tamilnadu and West Bengal show an initial arrest of the growth (blue shade), followed by later and more recent local outbreaks as marked by a discontinuity in the slope (see green shade in inset). Several other states, including Uttar Pradesh, Kerala and Karnataka display good control over the epidemic, while other states are in the linear regime throughout from t_1 (beginning of the blue shade).

Since the predictive models depend significantly on data, an important aspect to consider is that the number of reported infections does not truly reflect the actual outbreak of COVID-19. The data on infection rate are often limited by the countries' testing capability, which in turn is related to the availability of testing kits, size of healthcare professionals per population, and infrastructure. Further, the asymptomatic population is often excluded in testing strategies adopted by most countries. To elucidate this, we show the daily number of tests for key countries in Fig. 2b. We generally note that the increase in the number of tests with time is very closely related to the infections shown in Fig. 2a, as expected. A small number of reported cases for India in March could be due to inadequate testing at that time. Therefore the predictions by models using data from that period had considerable uncertainty (Ranjan 2020b; Singh and Adhikari 2020).

We briefly discuss the implication of rigorous testing in COVID-19 control by carefully examining the South Korean data, shown in the inset in Fig. 2b. Unlike most countries with the number of tests increasing slowly during the initial phase of the outbreak, the reverse is seen for South Korea. The response of SK to the outbreak was quick and they ran the most comprehensive and well-organized testing program in the world from February (Fig. 2b) when the outbreak was still not severe. This, combined with large-scale efforts to isolate infected people and trace and quarantine their contacts, lead to successful control of the outbreak. For comparison, SK, the US, and India respectively have 14500, 33000 and 1550 tests per million inhabitants on May 15.

A final but most important factor affecting the outbreak and the predictions is the epidemiology of COVID-19 in different geographical regions. The values of epidemiological

Table 1 Characteristics of COVID-19 spread and projected epidemic size until middle of August, 2020

				Projected size ($\times10^{-5}$)			
	Characteristic ratios (%)			Epidemiological		Data-driven	
Country	PR	CRR	CFR	SIR	SEIQRDP	MGDM	PLDM
USA[2]	13.4	22.1	6.0	20.2	15.88	18.68	25.28
Russia[1]	4.1	23.2	0.9	4.11	5.56	6.74	5.83
UK	10.1	–	14.4	2.60	2.68	3.24	4.11
Spain[3]	9.0	68.9	10.0	2.26	2.19	2.66	2.64
Italy	7.8	53.7	14.1	2.23	2.18	2.64	2.48
France	13.0	33.7	15.3	1.77	1.79	2.18	1.96
Germany	5.6	86.8	4.5	1.73	1.65	2.01	1.86
India[1,2]	4.0	35.6	3.2	2.22	6.61	7.13	6.19

[1] High uncertainty in projections as COVID-19 curve displays near exponential growth

[2] Uncertainty due to varying intervention measures among states

[3] Based only on the PCR data

parameters such as transmission rate, recovery rate, and basic reproduction number (Liu et al. 2020) depend on many social and environmental factors and are dissimilar in different regions. For an ongoing outbreak, the epidemiology is not fully established, but available data can provide meaningful insights. We report three characteristic ratios: positivity ratio (PR), case recovery ratio (CRR) and case fatality ratio (CFR) to roughly correlate with the epidemiological parameters: the rates of infection β, recovery γ, and mortality μ respectively (Hethcote 2020). PR is the total number of infections for a given number of tests. CRR and CFR are respectively, the number of recovered and deceased cases as a fraction of total infections. These values are reported in percentages in Table 1. There is a high disparity in these values among different regions. India and Russia have the lowest PR $\simeq$ 4%, compared to the very high values in the US ($\simeq$ 14%), UK ($\simeq$ 10%), and France ($\simeq$ 13%). A very low PR for India could be due to the factors like warmer climate as well as humidity (O'Reilly et al. 2020), a large proportion of the young in the total population, and possible immunity due to BCG vaccinations (Curtis et al. 2020) and malarial infections (Goswami et al. 2020). CRRs for Germany ($\simeq$ 87%) and Spain ($\simeq$ 69%) are highest, but it is expected that the value of CRR in countries currently in the acceleration phase will improve with time. The case fatality ratio is very high for France($\simeq$ 15%), UK($\simeq$ 14%) and Italy($\simeq$ 14%) compared to the world average of 2–3%. A high ratio may be due to a higher percentage of the elderly population in these countries.

It is clear from the above discussion that the epidemiologies of COVID-19, as well as the impact of social distancing for different countries, are very dissimilar. Further, there is an inhomogenous COVID-19 spread within a country as seen for the US and India. All of these factors make the modeling of this epidemic during its progress very challenging. Typically, epidemiological models such as a logistic or a compartmental model are preferred for modeling the later stages. However, these models are highly dependent on initial conditions and underlying unknown epidemiological parameters, incorrect estimation of which can give completely different results. A further concern with these models is the prediction of the decay rate of infections, which is generally high compared to the recent trends (Ranjan 2020a). Therefore, in this work, we propose two data-driven models for the predictions that incorporate the slow decay of the epidemic post-lockdown and provide more realistic estimates. Projections for key affected countries are presented using these data-driven as well as epidemiological models.

COVID-19 data used in this study are taken from various sources. Modeling is based on the time-series data from Johns Hopkins University Coronavirus Data Stream, which combines World Health Organization (WHO) and Centers for Disease Control and Prevention (CDC) case data. Data on tests are taken from 'Our World in Data' source that compiles data from the European Centre for Disease Prevention and Control (ECDC). Time-series data for Indian states are taken from github.com/covid19india.

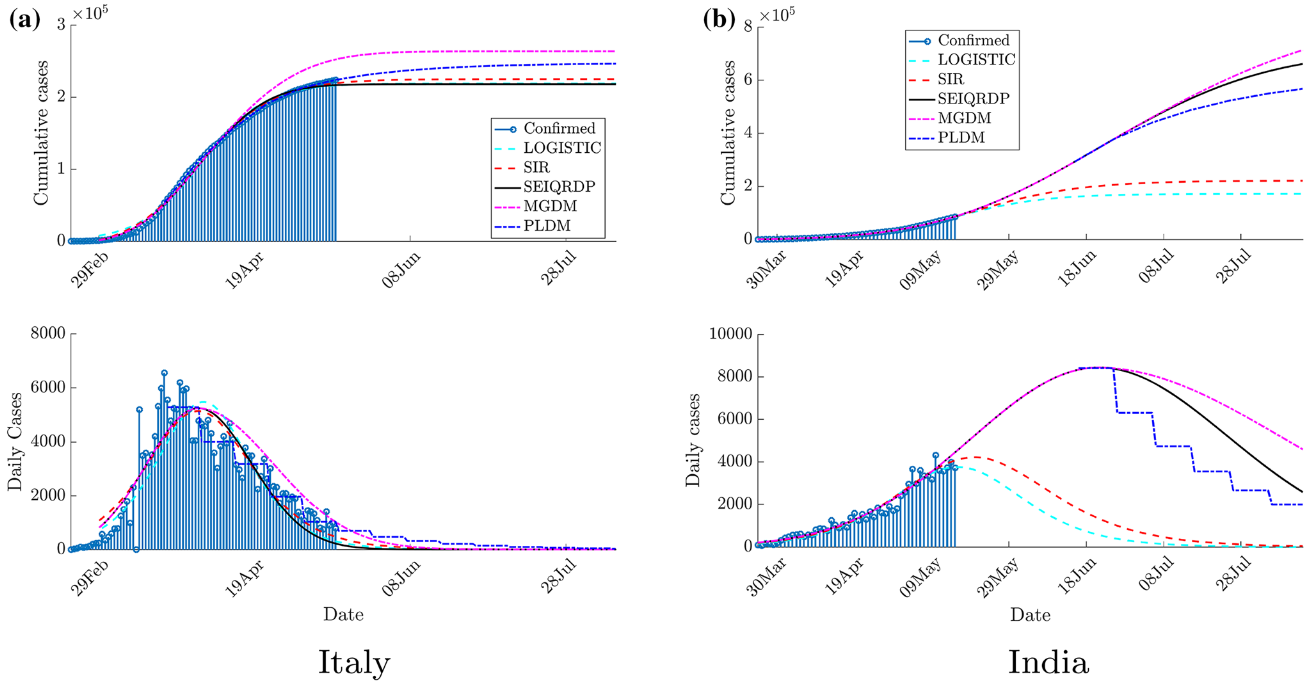

Fig. 3 Projections using epidemiological and proposed data-driven models

Data-Driven Models

Typically for an ongoing epidemic, epidemiological models estimate the underlying parameters based on fit from available data and then use simple ordinary differential equations to predict the day of the peak and the decay rate. To illustrate the limitations of these models, we show predictions for Italy using a logistic model (Ma 2020), as well as two compartmental epidemiological models- SIR (Hethcote 2020) and Generalized SEIR (SEIQRDP) (Peng et al. 2020) in Fig. 3a. Open-source MATLAB codes developed by Batista (2020) and Cheynet (2020) are used for SIR and SEOQRDP models respectively.

As the epidemic has passed its peak in Italy, a key parameter for estimation is the decay rate. A close examination of the daily cases in Fig. 3a shows that all the three models predict a faster decay rate as the curve is nearly symmetric around the peak. This distribution leads to an under-prediction of size as well as the duration of the epidemic. Although not shown for every geographical region considered in this paper, this is true for most of the predictions.

Modified Gaussian Decay Model (MGDM)

We shall describe the first data-driven model to improve the predictions. As shown in Fig. 3a, the infection rate (daily cases) predicted by epidemiological models follow a nearly normal distribution with $\frac{\mathrm{d}I}{\mathrm{d}t} = a\exp\left(-\left(\frac{t-\mu}{\sigma}\right)^2\right)$, where μ represents the day with the peak number of cases, σ is spread around this day from the beginning of the epidemic to the end, and a is a constant that determines the number of cases. Because of the nearly symmetric distribution of the curve, the decline rate is typically predicted as the negative of the climb rate.

To make the predictions closer to actual values in the deceleration phase, we introduce a new parameter η, which changes the variance σ of this distribution after the peak to make the decline rate more realistic. Hence, the new distribution in this Modified Gaussian Decay Model (MGDM) is given by $\frac{\mathrm{d}I}{\mathrm{d}t} = \zeta a\exp\left(-\left(\frac{t-\mu}{\eta\sigma}\right)^2\right)$, where the pre-multiplication factor ζ ensures that the number of infections on the peak day remains unaltered.

The dash-dot magenta curve in Fig. 2a shows the distribution of MGDM. The infection rate, in this case, is closer to the actual values during decay and generally provides the upper limit of estimated total cases, as seen from the difference in the cumulative cases. Parameters for fitted normal distribution with $R^2 = 0.9999$, RMSE=4.09 and 95% confidence bounds are: $a = 5273(5272, 5275), \mu = -5.572(-5.723, -5.421), \sigma = 248.1(248, 248.2)$. The modifications due to data gives $\eta = 1.4, \zeta = 0.996$. The final epidemic size by this model as well as SIR and SEIQRDP models are given in Table 1. Note that, a Gaussian fit of the infection rate can be directly used in MGDM for regions with sufficient data in the deceleration phase, and a prediction from epidemiological model is not necessary.

Piecewise-Linear Decay Model (PLDM)

We shall now discuss the second data-driven model. As discussed earlier, recent trends indicate that the lockdown arrests the initial exponential growth but a linear regime persists after that, and then a prolonged decay follows. We propose that this decay can be modeled better with several linear segments than an exponential or a Gaussian curve that gives a fast decline with a relatively small tail. Piecewise-Linear Decay Model (PLDM) incorporates these dynamics. The cumulative data in the deceleration phase is collected and then divided into equal segments. An optimal piecewise linear fit in a least-squares sense is then obtained. The slopes of these linear fits are m_i, where $i = 1, ..., N$, N being the number of segments. The ratios of slopes are then computed, $\alpha_i = \frac{m_i}{m_{i-1}}$, with $\alpha_i < 1$, and a slope factor $\bar{\alpha}$ is calculated by taking an average of the last three ratios. This factor is then used to predict the slope of the next future segment such as $m_{N+1} = \bar{\alpha} m_N$ and so on. Figure 3a includes the predictions from this model for Italy. Data during the decay phase, between Mar 21 and May 15, have been divided into five equal segments of eleven days. The modeling gives $\bar{\alpha} = 0.6734$, which is used to predict the slopes of future linear segments of the same sizes (see bottom panel in the figure). As evident from Fig. 3a, while the prediction of the final estimate size for both MGDM and PLDM is similar, PLDM predicts the cumulative curve more closely and has a more gradual decay.

Predictions

Table 1 shows the projections for key countries using both the epidemiological (SIR and SEIQRDP) as well as data-driven models until the middle of August. Though not shown for individual cases, it is ensured in every case that the logistic fits are statistically significant with $R^2 > 0.98$ and p-value < 0.0001. Parameters of data-driven models (η, ζ in MGDM, and $\bar{\alpha}, m_N$ in PLDM) are directly obtained from the data in the deceleration phase when available. For countries, where the infection rate is still growing, predictions from the SEIQRDP model are used as a baseline, and parameters of data-driven models are taken from the fit used in Italy. As expected, there is higher uncertainty for these countries.

Table 2 Prediction of epidemic size for key states in India ($\times 10^{-4}$)

State	Maharashtra	Gujarat	New Delhi	Tamilnadu	Madhya Pradesh
Range	6.6–17.4	1.27–2.23	1.28–3.32	1.06–2.34	0.59–0.92
State	Rajasthan	Punjab	Uttar Pradesh	West Bengal	Karnataka
Range	0.55–0.84	0.22–1.77	0.47–0.77	0.33–1.70	0.09–1.10

All European countries in Table 1 except UK, where the outbreak is already in the decline phase, show a good convergence of epidemic sizes i.e., predictions from the epidemiological models are not very different as shown in the case of Italy (Fig. 3a). Likewise, estimates from both the data-driven models are very close and are higher than those from epidemiological models as expected. Both MGDM and PLDM suggest the equilibrium to be expected towards the end of July. Predictions for these regions using the data-driven models are fairly reliable provided there is no new outbreak.

For the US, there is an uncertainty due to fluctuations in the recent data, which in turn is due to different epidemiology and a differential impact of stay-at-home order among different states (Ranjan 2020a). For the UK, the epidemic is in the linear growth stage with mild signs of decay recently. Therefore, there is high unreliability in the prediction of the peak. Nevertheless, forecasts by different models in these cases can provide an estimate of the expected range.

For India and Russia, the growths are still close to exponential, and therefore there is a significant disparity in predictions by different epidemiological models. Figure 3b illustrates the uncertainty in such cases by showing the projections for India by different models. While the Logistic and SIR models predict the peak close to each other, SEIQRDP model shows continuing growth till the middle of June before the decline begins. This difference leads to a significantly higher epidemic size with SEIQRDP (0.66 million) than those with logistic (0.17 million) and SIR (0.22 millions) models. A critical difference between the SIR and SEIQRDP models implemented here is that in SIR, the population N considered is just the number of susceptible persons before the outbreak, while the entire population of the region is taken as the population size in SEIQRDP. As results from the SEIQRDP model are used as a baseline for data-driven models, estimates from the latter are also in the higher range.

These models are then used for statewise projections in India. Table 2 gives the lower and upper range value of the estimated epidemic size as calculated by all the models.

As expected, the highest contribution comes from four key states: Maharashtra, Gujarat, Delhi and Tamilnadu, who are on the exponential growth (Fig. 1b). Also, the projections have the highest uncertainty in these regions among the states listed in the table. If these states can control the epidemic and new outbreaks do not appear in other states, it is expected that the optimistic scenario for India shown by the SIR model in Fig. 3b can be realized.

The final epidemic size of the entire world is difficult to estimate without getting individual estimates of all the countries. This is because the global trend of total infections is still on an accelerating stage with new countries (Brazil, Peru, Canada) reporting surge in the number of cases.

Concluding Remarks

Epidemiological models such as logistic and compartmental models are generally used to predict the total size and duration of COVID-19. However, these models generally do not account for the precise change in dynamics due to different interventions, or a new outbreak, and therefore estimate unrealistic epidemic size. We show that the COVID-19 curves for different countries after the lockdown are very dissimilar with four primary distributions: linear, exponential, and slow and fast flattening. Further, within a country, the characteristics of spread among states may be different. Therefore, to account for differences in dynamics, a locally data-driven approach for modeling may be more suitable. Two data-driven models for the decay of COVID-19 based on recent trends- one based on skewed Gaussian distribution and the other by using a piecewise linear fit—are proposed. These models generally provide a more realistic estimate of the epidemic size than epidemiological models for regions in the deceleration phase, with the piecewise linear model predicting a more gradual decay. For countries (like India and Russia) still in the growth stage, these data-driven models use predictions from epidemiological models as a baseline and impose corrections using parameters obtained from an available data with a realistic decline rate. The uncertainty in predictions for such cases is higher. The paper also highlights that the reported data on infections is not an accurate representation of actual outbreak, and is limited by the testing capacity. Therefore, estimations given by these models could still be optimistic and should be used with caution. A periodic evaluation of characteristics of COVID-19 spread, and thus a revision of projections is necessary.

Acknowledgements The author would like to thank Prof. Datta Gaitonde for his support and encouragement during these unprecedented times. The critical inputs from Prof. Roddam Narasimha in improving the manuscript are gratefully acknowledged. Contributions from Dr. Sudheendra N R Rao (Scientific Advisor, Organization for Rare

Diseases India) and Dr. Deepti Chugh (The Ohio State University) are duly acknowledged.

References

Ansumali S, Prakash MK (2020) A very flat peak: Exponential growth phase of covid-19 is mostly followed by a prolonged linear growth phase, not an immediate saturation. medRxiv, https://doi.org/10.1101/2020.04.07.20055772. https://www.medrxiv.org/content/early/2020/04/11/2020.04.07.20055772

Batista M (2020) fitviruscovid19, matlab central file exchange. retrieved march 31,2020. https://www.mathworks.com/matlabcentral/fileexchange/74658-fitviruscovid19

Cheynet E (2020) Generalized seir epidemic model (fitting and computation).retrieved april 8, 2020. https://www.github.com/ECheynet/SEIR

Curtis N, Sparrow A, Ghebreyesus TA, Netea MG (2020) Considering bcg vaccination to reduce the impact of covid-19. The Lancet 395(10236):1545–1546

Goswami RP, Mittal DK, Goswami RP (2020) Interaction between malarial transmission and bcg vaccination with covid-19 incidence in the world map: a changing landscape human immune system? medRxiv. https://doi.org/10.1101/2020.04.03.20052563. https://www.medrxiv.org/content/early/2020/04/08/2020.04.03.20052563

Hethcote HW (2020) The mathematics of infectious diseases. SIAM Rev 42(4):599–653

Liu Y, Gayle AA, Wilder-Smith A, Rocklöv J (2020) The reproductive number of covid-19 is higher compared to sars coronavirus. J Travel Med 27(2):1–4

Ma J (2020) Estimating epidemic exponential growth rate and basic reproduction number. Infect Dis Modelling 5:129–141. https://doi.org/10.1016/j.idm.2019.12.009

O'Reilly KM, Auzenbergs M, Jafari Y, Liu Y, Flasche S, Lowe R (2020) Effective transmission across the globe: the role of climate in covid-19 mitigation strategies. Lancet Planet Health 4(5):e172. https://doi.org/10.1016/S2542-5196(20)30106-6

Peng L, Yang W, Zhang D, Zhuge C, Hong L (2020) Epidemic analysis of covid-19 in china by dynamical modeling. arXiv preprint arXiv:2002.06563

Ranjan R (2020) Estimating the final epidemic size forcovid-19 outbreak using improved epidemiological models. medRxiv, https://doi.org/10.1101/2020.04.12.20061002. https://www.medrxiv.org/content/early/2020/04/20/2020.04.12.20061002

Ranjan R (2020) Predictions for covid-19 outbreak in india using epidemiologicalmodels.medRxiv, (2020). https://doi.org/10.1101/2020.04.02.20051466. https://www.medrxiv.org/content/early/2020/04/06/2020.04.02.20051466

Roda WC, Varughese MB, Han D, Li MY (2020) Why is it difficult to accurately predict the covid-19 epidemic? Infect Dis Modelling 5:271–281

Singh R, Adhikari R (2020) Age-structured impact of social distancing on the COVID-19 epidemic in India. arXiv preprint arXiv:2003.12055

Wu Z, McGoogan JM (2020) Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the chinese center for disease control and prevention. Jama 323(13):1239–1242

Yang J, Zheng Y, Gou X, Pu K, Chen Z, Guo Q, Ji R, Wang H, Wang Y, Zhou Y (2020) Prevalence of comorbidities in the novel wuhan coronavirus (covid-19) infection: a systematic review and meta-analysis. Int J Infect Dis 94:91–95