Proceedings of the 2020 Winter Simulation Conference
K.-H. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Thiesing, eds.

USING SIMPLE DYNAMIC ANALYTIC FRAMEWORK TO CHARACTERIZE AND FORECAST EPIDEMICS

Amna Tariq
Kimberlyn Roosa
Gerardo Chowell

Department of Population Health Sciences
School of Public Health
Georgia State University
140 Decatur St. Suite 400
Atlanta, GA 30303, USA

ABSTRACT

Mathematical modeling provides a powerful analytic framework to investigate the transmission and control of infectious diseases. However, the reliability of the results stemming from modeling studies heavily depend on the validity of assumptions underlying the models as well as the quality of data that is employed to calibrate them. When substantial uncertainty about the epidemiology of newly emerging diseases (e.g. the generation interval, asymptomatic transmission) hampers the application of mechanistic models that incorporate modes of transmission and parameters characterizing the natural history of the disease, phenomenological growth models provide a starting point to make inferences about key transmission parameters, such as the reproduction number, and forecast the trajectory of the epidemic in order to inform public health policies. We describe in detail the methodology and application of three phenomenological growth models, the generalized-growth model, generalized logistic growth model and the Richards model in context of the COVID-19 epidemic in Pakistan.

1 INTRODUCTION

Emerging novel pathogens with life threatening transmission potential in humans have motivated the development and implementation of sophisticated computational approaches and mathematical models to estimate the transmission parameters, assess the impact of interventions, and generate forecasts (Colizza et al. 2006; Balcan et al. 2009; Merler et al. 2015; Chinazzi et al. 2020). The output from these models can be useful to design intervention strategies according to a local context, allocate resources and inform public health policies (Chretien et al. 2015). With the current outbreak of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2 virus), the causative agent of coronavirus disease 2019 (COVID-19), having reached pandemic proportions, the investigation of the global situation may provide the key to understanding the transmissibility potential of the disease in different regions across the globe (Balcan et al. 2009). Multiple epidemic models, ranging from classical compartmental models based on differential equations to agent-based computational models, have been used to simulate, calibrate and generate epidemic forecasts of SARS-CoV-2 in order to understand its transmission dynamics (Pell et al. 2018). These models provide a framework to understand the underlying disease transmission mechanisms at different spatial and temporal scales and vary in complexity based on the number of parameters and equations characterizing the dynamic states of the systems (Chowell 2017). While highly detailed computational models have become increasingly popular to model detailed disease processes, simple dynamic models that capture a variety of empirical growth trajectories with only a few parameters (e.g.,

978-1-7281-9499-8/20/$31.00 ©2020 IEEE

growth rate, scaling of growth, total outbreak size) provide a powerful approach to characterize transmission dynamics and generate short-term forecasts (Chowell et al. 2016).

Incomplete and inaccurate information during the early transmission stages of an infectious disease pathogen hinders the availability of detailed and reliable epidemiological data and disease-specific epidemiological parameters, and in turn, the application of mechanistic models of disease transmission and control. Moreover, reporting of infections is often influenced by the contribution of asymptomatic infections, local testing capacity, reporting delays, efficiency of surveillance systems, and the level of burden on health care facilities (Balcan et al. 2009). Therefore, phenomenological growth models can serve as a starting point to capture the empirical patterns of epidemics and provide estimates of early transmission potential to gain an understanding of the evolution of the outbreak and generate short-term forecasts of the epidemic trajectory (Chowell et al. 2016). It is worth noting that outbreaks of rapid dissemination often only spread during a few generations of disease transmission, so epidemic assessment using phenomenological forecasting models is crucial during the early phase of the outbreak to estimate the potential disease burden and approximate the scale of interventions required to contain the epidemic (Pell et al. 2018).

In this paper, we employ the established dynamic phenomenological growth models which have provided a good description of multiple outbreak trajectories for a number of infectious diseases including SARS, pandemic influenza, Ebola and the current COVID-19 pandemic (Wang et al. 2012; Chowell et al. 2019; Yan and Chowell 2019; Roosa et al. 2020a). Here we apply these models in near real time to the ongoing COVID-19 epidemic in Pakistan, a country located in South Asia neighboring China, the epicenter of the COVID-19 pandemic. The first cases of COVID-19 in Pakistan were reported in late February 2020, followed by a sudden upsurge in case counts in mid-March that led to a countrywide lockdown along with multiple other restrictions including a ban on group gatherings and meetings. These social distancing measures proved successful at containing the rapid rise in case incidence, and by May 22, 2020, countrywide lockdowns had to be lifted (Ali et al. 2020) to provide some oxygen to a suffocating economy. Since then, the country has re-imposed strategic lockdowns in regions reporting the highest number of cases (India 2020). We apply the generalized growth model (GGM), generalized logistic growth model (GLM) and the Richards growth model to illustrate the methodology and generate short term forecasts in near real time to understand the epidemic trend of the ongoing COVID-19 outbreak in Pakistan. We also assess the early transmission potential of SARS-CoV-2 based on the early epidemic trajectory, which can guide the intensity of interventions and inform public health policies.

2 METHODS

2.1 COVID-19 Incidence Data

We use daily time series of polymerase chain reaction (PCR) confirmed COVID-19 case data as of July 22, 2020 from the publicly available COVID-19 dashboard, set up by the National Institute of Health, Pakistan (Pakistan 2020). We retrieved the daily number of COVID-19 countrywide cases reported on the dashboard from the published epidemic curve. We also estimate the daily COVID-19 positivity rate i.e. the total number of positive test results from the total number of tests conducted.

2.2 Models

We use three phenomenological growth models including a two parameter GGM, a three parameter GLM, and Richards growth model to forecast the epidemic trajectory of COVID-19 in Pakistan. The general form of a phenomenological growth model is given by:

$$\frac{dx_i}{dt} = f_i(x_1, \dots\dots, x_n; \Theta), i = 1, \dots\dots, n,$$

where $\frac{dx_i}{dt}$ denotes the rate of change of the system state x_i, and Θ=$(\theta_1, \ldots\ldots, \theta_m)$ is the set of model parameters that characterizes the state of systems in the model (Lara-Díaz et al. 2019).

2.3 Generalized Growth Model

The generalized growth model (GGM) is a simple model that characterizes the early ascending phase of the epidemic. Previous studies have highlighted the occurrence of early sub-exponential growth patterns in various infectious disease outbreaks. This model allows for the relaxation of exponential growth by modulating a "scaling of growth parameter", p, which allows the model to capture a range of epidemic growth profiles (Viboud et al. 2016). The GGM is given by the following differential equation:

$$\frac{dC(t)}{dt} = C'(t) = rC(t)^p.$$

In this equation $C'(t)$ describes the incidence curve over time t, $C(t)$ describes the cumulative number of cases at time t, p∈[0,1] is a "deceleration or scaling of growth" parameter and r is the growth rate. This model represents constant incidence over time if p=0 and exponential growth for cumulative cases if p =1. If p is in the range 0< p <1, then the model indicates sub-exponential or polynomial growth dynamics (Chowell et al. 2016; Viboud et al. 2016; Chowell 2017).

2.4 Generalized Logistic Growth Model

The generalized logistic growth model (GLM) is an extension of the simple logistic growth model that allows for capturing a range of epidemic growth profiles, including sub-exponential and exponential growth dynamics. The GLM characterizes epidemic growth through the intrinsic growth rate r, a dimensionless "deceleration of growth" parameter p, and the final epidemic size, k_0. The deceleration parameter modulates the epidemic growth patterns including sub-exponential growth (0< p <1), constant incidence (p =0) and exponential growth dynamics (p =1). The GLM is given by the following differential equation:

$$\frac{dC(t)}{dt} = rC^p(t)(1 - \frac{C(t)}{k_0}),$$

where $\frac{dC(t)}{dt}$ describes the incidence over time t, and the cumulative number of cases at time t is given by $C(t)$ (Chowell et al. 2016; Pell et al. 2018; Shanafelt et al. 2018).

2.5 Richards Growth Model

The Richards growth model is also an extension of the simple logistic growth model and relies on 3 parameters. It extends the simple logistic growth model by incorporating a scaling parameter, a, that measures the deviation from the symmetric simple logistic growth curve (Richards 1959; Wang et al. 2012; Chowell 2017). The Richards model is given by the differential equation:

$$\frac{dC(t)}{dt} = rC(t)\left[1 - \left(\frac{C(t)}{k_0}\right)^a\right],$$

where $C(t)$ represents the cumulative case count at time t , r is the growth rate, a is a scaling parameter and k_0 is the final epidemic size.

2.6 Parameter Estimation

To illustrate the fitting and 20-day ahead forecasts using the models described above, we calibrate the GGM to the daily incidence curve by dates of reporting in Pakistan using time series data that is available from

March 10–May 11, 2020. We calibrate the GLM and Richards growth model to the daily incidence curve by dates of reporting in Pakistan using time series data that is available from March 10–June 3, 2020.

The uncertainty in the parameter estimates for dynamical systems can arise as a result of noise in the data or the underlying assumptions for the model employed to infer the parameter estimates. We will now focus on how to account for the uncertainty in the parameter estimates arising due to the noise in the data by assuming a particular error structure i.e. the negative binomial distribution.

Model parameters are estimated by a non-linear least square fitting of the model solution to the incidence data by the date of reporting. This is achieved by searching for the set of model parameters $\widehat{\Theta} = (\widehat{\Theta}_1, \widehat{\Theta}_{2,....}\, \Theta_m)$ that minimizes the sum of squared differences between the observed data $y_{ti} = y_{t1}, y_{t2}, \ldots . y_{tn}$ and the corresponding mean incidence curve given by $f\left(t_i, \widehat{\Theta}\right)$, where $\widehat{\Theta} = (r, p)$ corresponds to the set of parameters of the GGM, $\widehat{\Theta} = (r, p, k_0)$ corresponds to the set of model parameters of the GLM, and $\widehat{\Theta} = (r, a, k_0)$ corresponds to the set of model parameters for the Richards growth model. The objective function for the best fit solution of the model, $f(t_i, \widehat{\Theta})$ is given by:

$$\widehat{\Theta}=\arg\min \sum_{i=1}^{n} (f(t_i, \Theta) - y_{t_i})^2,$$

where t_i is the time stamp at which the time series data are observed and n is the total number of data points available for inference. This way, $f\left(t_i, \widehat{\Theta}\right)$ gives the best fit to the time series data y_{t_i}. We estimate the best fit solution for the GLM and Richards growth model by initializing the parameter estimates for the nonlinear least squares method over a range of feasible parameters derived from a uniform distribution using Latin hypercube sampling. Whereas for the GGM we provide a reasonable guess of the initial parameter estimates for the nonlinear least squares method. We fix the initial condition according to the first data point. We also observe the temporal variation of the residuals (i.e. the difference between the best fit model solution and the time series data) to assess the quality of model fit (Chowell 2017).

Next, we utilize a parametric bootstrapping approach, assuming a negative binomial error structure in the data, to derive uncertainty of the parameter estimates as previously described (Efron and Tibshirani 1993; Chowell et al. 2006; Chowell 2017). We assume the variance to be ~24.8 times the mean for GGM and ~44.5 times the mean for GLM and Richards model based on our examination of the variability in the data. Our calibration results represent M = 200 resampled data sets, which we refit each model to in order to obtain M new parameter estimates. Model fits are used to obtain 95% confidence intervals for each parameter.

Each of the M model fits is extended through a 20 day forecasting period and then used to generate m=30 data curves with negative binomial error structure; thus, these 6000 (M×m) curves are used to generate 95% prediction intervals for the 20-day ahead forecasts (Chowell 2017).

The MATLAB code required to fit and forecast the epidemic trajectories using the GGM, GLM and Richards growth model is available upon request from the authors.

2.7 Performance Metrics

In order to demonstrate the performance metrics, we evaluate the performance of our two models, Richards growth model and the GLM by calibrating both models from March 10–June 3, 2020 and forecasting 20-day ahead to assess their capability to describe short term incidence patterns. For the calibration performance, we compare the model fit to the reported case data through the calibration period, and for forecasting performance, we compare forecasts with the incidence data reported 20 days ahead of the last date of the calibration period.

In order to compare the quality of the fits using different models as well as the performance of short term forecasts, we analyze four performance metrics namely the mean absolute error (MAE), the mean squared error (MSE), the coverage of the 95% prediction intervals, and the mean interval score (MIS) (Gneiting and Raftery 2007).

The mean absolute error and mean squared error quantify the mean deviations of the model to the observed data and are given by the following equations:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |f\left(t_i, \widehat{\Theta}\right) - y_{t_i}|,$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(f\left(t_i, \widehat{\Theta}\right) - y_{t_i})^2.$$

In these equations y_{t_i} is the time series of the incidence curve describing the epidemic wave, t_i is the time stamp, $\widehat{\Theta}$ is the set of best-fit model parameters, and n equals the number of data points in the calibration period. For the forecasting period, n = 20 for 20 days ahead forecast.

To assess the model uncertainty and performance of prediction intervals, we assess the prediction interval coverage and the mean interval score. Prediction interval (PI) coverage is defined as the fraction of reported data points that fall within 95% prediction interval calculated as:

$$PI\ coverage = \frac{1}{n}\sum_{t=1}^{n} \mathbf{1}\{Y_t > L_t \ \cap \ Y_t < U_t\}.$$

In this equation, L_t and U_t are the lower and upper bounds of the 95% prediction intervals respectively, n is the length of the time period, Y_t are the data, and 1 is an indicator variable that equals 1 if Y_t is within the specified interval and 0 otherwise.

We also assess the mean interval score (MIS), which considers the coverage and width of 95% prediction interval, given by the following equation:

$$MIS = \frac{1}{h}\sum_{i=1}^{n}\left(U_{t_i} - L_{t_i}\right) + \frac{2}{0.05}(L_{t_i} - y_{t_i})I\{y_{t_i} < L_{t_i}\} + \frac{2}{0.05}\left(U_{t_i} - y_{t_i}\right)I\{y_{t_i} > U_{t_i}\}.$$

In this equation L_{t_i} and U_{t_i} are the lower and upper bounds of the 95% prediction interval, $y_{t_i}$are the data and I is an indicator function that equals 1 if y_{t_i} is in the specified interval and 0 otherwise (Gneiting and Raftery 2007). Therefore, if the PI coverage is 1, the MIS is the average width of the interval across each time point. For two models with an equivalent PI coverage, a lower MIS indicates narrower intervals.

2.8 Reproduction Number, R_t, from Case Incidence using GGM

The effective reproduction number, R_t, is defined as the average number of secondary cases generated by a primary case at time t during the outbreak. This is an important measure that can influence the intensity of interventions required to contain an epidemic (Anderson and May 1991; Nishiura et al. 2010; Chowell et al. 2015). Estimates of effective R_t indicate if the disease transmission continues (R_t>1) or if the active disease transmission declines (R_t<1). Therefore, in order to contain an outbreak, we need to maintain R_t<1. We estimate the reproduction number by calibrating the GGM to the early growth phase of the epidemic (63 days) (Viboud et al. 2016). We model the generation interval of SARS-CoV-2 assuming a gamma distribution with (i) a mean of 5.2 days and a standard deviation of 1.72 days based on refs (Ganyani et al. 2020) and (ii) a mean of 4.41 days and a standard deviation of 3.17 days based on refs. (Nishiura et al. 2020; You et al. 2020). We estimate the growth rate parameter, r, and the deceleration of growth parameter, p, as described above. In order to estimate the reproduction number, we simulate the progression of incidence cases I_i at calendar time t_i from the calibrated GGM and apply the discretized probability distribution of the generation interval denoted by ρ_i to the renewal equation (Nishiura and Chowell 2009; Paine et al. 2010; Nishiura and Chowell 2014).

$$R_{\mathrm{t}_i} = \frac{I_i}{\sum_{j=0}^{i}(I_{i-j}\rho_j)}$$

The numerator represents the total new cases I_i, and the denominator represents the total number of cases that contribute to generating the new cases I_i at time $t = \mathrm{i}$. This way, R_t, represents the average number of secondary cases generated by a single case at time t. Next, we derive the uncertainty bounds around the curve of R_t directly from the uncertainty associated with the parameter estimates (*r, p*) as described above. We estimate R_t for 200 simulated curves assuming a negative binomial error structure where variance is assumed to be ~24.8 times the mean (Chowell 2017). The MATLAB code required to estimate the reproduction number using GGM is available upon request from the authors.

3 RESULTS

The COVID-19 epidemic trajectory displays broadly an unimodal pattern with a majority of the cases concentrated between May 29-June 30, 2020. A total of 269,173 cases have been reported as of July 22, 2020 (Figure 1). The average number of new cases reported in Pakistan was estimated at ~92 cases per day in March 2020, followed by an increase to an average of ~4700 cases per day in June 2020, and then a decline to an average of ~2532 cases per day in July 2020. Subsequently, the COVID-19 positivity rate in Pakistan has fluctuated between ~0.4-25.8% over the course of five months (March-July, 2020). The monthly average COVID-19 positivity rate was ~8.2% in March 2020, ~10.5% in April 2020, ~14.9% in May 2020, ~18.9% in June 2020 and ~11.2% in July 2020 (Figure 2).

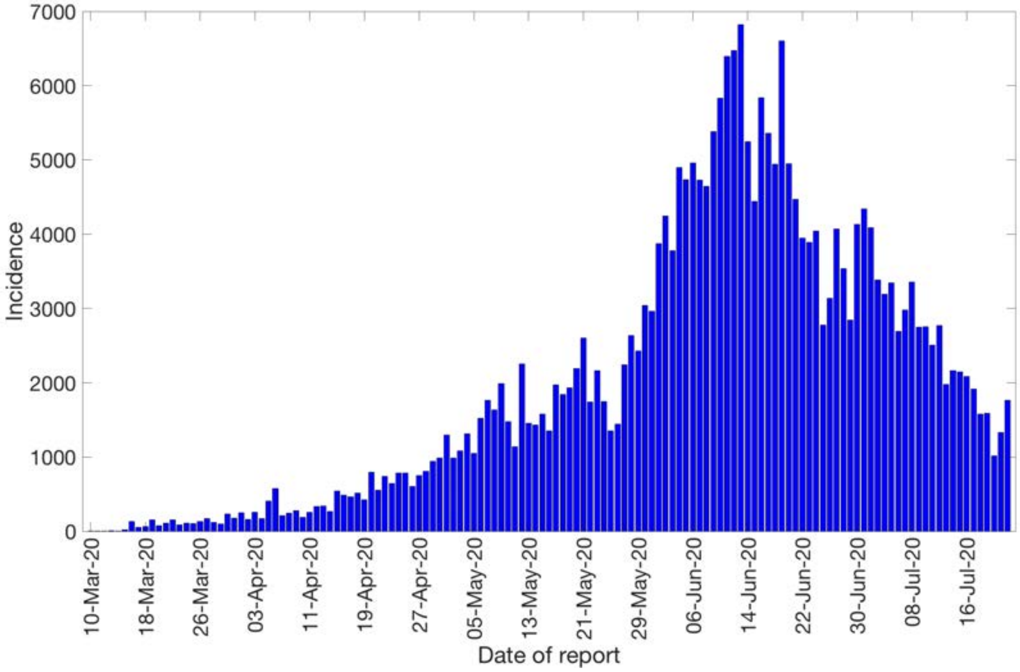

Figure 1: The daily curve of new COVID-19 cases reported in Pakistan as of July 22, 2020.

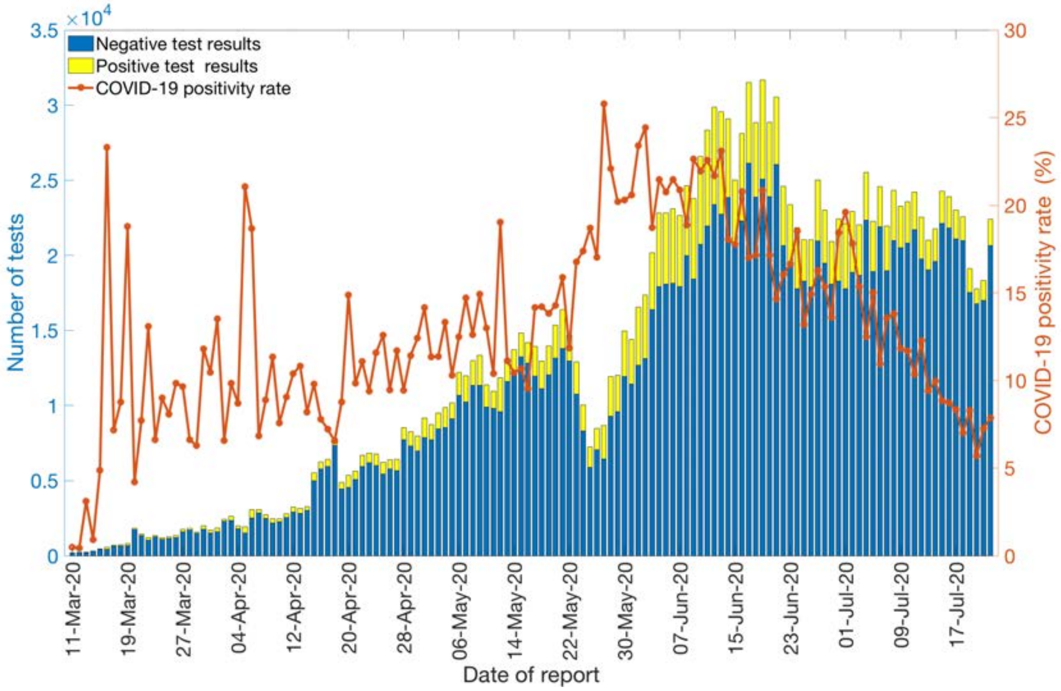

Figure 2: Testing and positivity rates of COVID-19 epidemic in Pakistan as of July 22, 2020. The blue bars indicate the negative test results, the yellow bars indicate the positive test results and the orange solid line indicates the COVID-19 positivity rate.

3.1 GGM Model Fit and 20-day ahead Forecast

The GGM fits well to the early growth phase of the epidemic, and the residuals display a random scatter, indicating there is not a systematic deviation of the model from the data, which could suggest the model is not appropriate for the data. The best model fit yielded the growth rate, *r,* estimate at 1.0 (95% CI: 0.81, 1.2) and the scaling of growth parameter estimate, *p*, at 0.72 (95% CI: 0.69, 0.75), indicating early polynomial growth dynamics of COVID-19. The scaling of growth parameter, *p*, is also well-identified with a narrow CI (Table 1). The 20-day ahead average forecast generated from the GGM calibrated from March 10-May 11, 2020 projects a reasonable ascending trajectory of the epidemic and shows that Pakistan could accumulate ~48486 (95% PI: 38019, 61373) additional cases in the next 20 days (between May 12-May 31, 2020) (Figure 3).

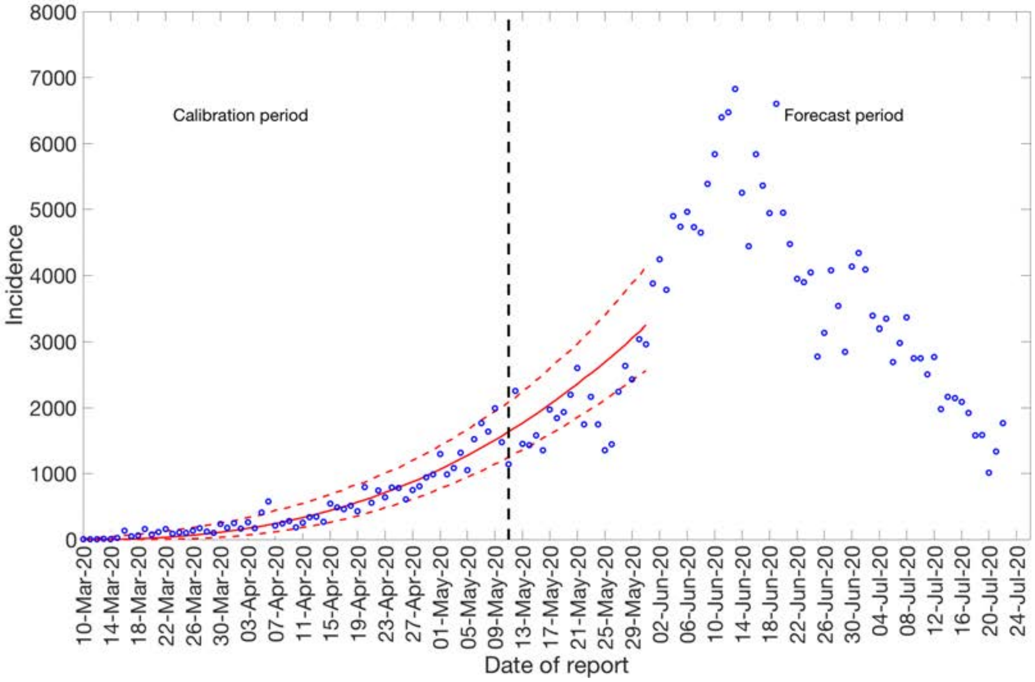

Figure 3: 20-days ahead forecast of the COVID-19 epidemic in Pakistan by calibrating the GGM model from March 10-May 11, 2020. Blue circles correspond to the data points, the red solid line indicates the best model fit and the red dashed lines represent the 95% prediction interval. The vertical black dashed line represents the time of the start of the forecast period.

3.2 GLM Model Fit and 20-day ahead Forecast

The GLM calibrated from March 10–June 3, 2020 provides a reasonably good fit to the data. The best model fit yielded a growth rate, *r,* estimate at 0.92 (95% CI: 0.7, 1.1), the scaling of growth parameter estimate, p , at 0.73 (95% CI: 0.70, 0.77) and the final epidemic size, k_0, estimate at 6.1 e+07 (95% CI: 3.4 e+05, 1 e+08). The parameters p and r are well-identified with a much wider CI for the final epidemic size (Table 1). The 20-day ahead average forecast generated from the GLM model calibrated from March 10-June 3, 2020 predicts that Pakistan could accumulate a total of ~82910 (95% PI: 58997, 105299) additional cases in the next 20 days (between June 4- June 23, 2020) (Figure 4), with the actual reported case count lying between the 95% PI observing in retrospect. While the model predicts continued epidemic growth, the lower bound of the 95% PI also includes a downturn or slowing down of the growth.

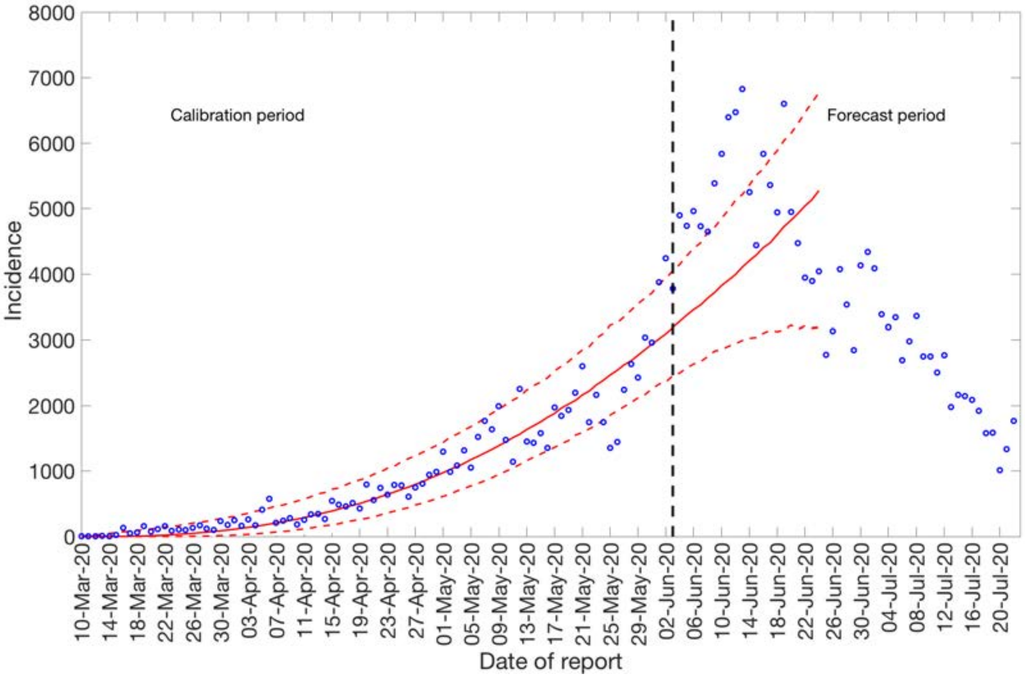

Figure 4: 20-days ahead forecast of the COVID-19 epidemic in Pakistan by calibrating the GLM model from March 10-June 3, 2020. Blue circles correspond to the data points, the red solid line indicates the best model fit and the red dashed lines represent the 95% prediction interval. The vertical black dashed line represents the time of the start of the forecast period.

3.3 Richards Model Fit and 20-day ahead Forecast

The Richards model fit to the data calibrated from March 10–June 3, 2020 indicates that Richards model is unable to explain the early dynamics of the COVID-19 epidemic, as it systematically under-predicts the first half of the epidemic curve. The best model fit yielded a growth rate, *r*, estimate at 0.56 (95% CI: 0.47, 0.64), the final epidemic size, k_0, estimate at 2.7 e+05 (95% CI: 2.3 e+05, 3.1 e+05) and the scaling parameter, *a*, estimate at 0.055 (95% CI: 0.048, 0.068). All the parameters are well identified with narrow CI's (Table 1). Richards model calibrated from March 10-June 3, 2020 provides an under-predicted 20-day ahead average forecast, as the model predicted a downturn in incidence cases while daily cases were still increasing. The model predicted that Pakistan could accumulate a total of ~58452 (95% PI: 44400, 77222) additional cases in the next 20 days (between June 4-June 23, 2020) (Figure 5).

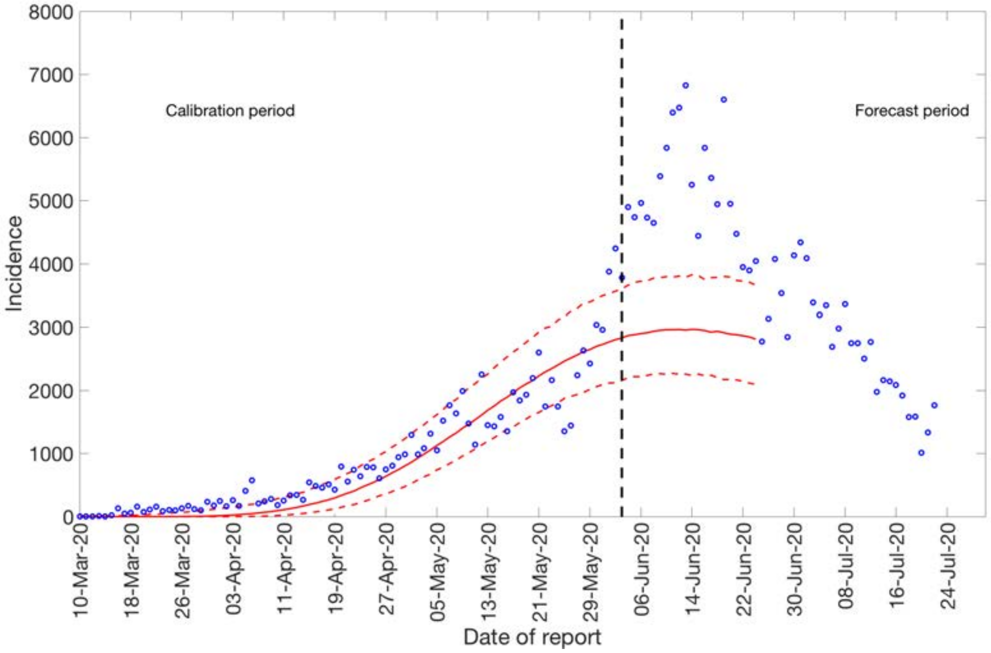

Figure 5: 20-days ahead forecast of the COVID-19 epidemic in Pakistan by calibrating the Richards model from March 10-June 3, 2020. Blue circles correspond to the data points, the red solid line indicates the best model fit and the red dashed lines represent the 95% prediction interval. The vertical black dashed line represents the time of the start of the forecast period.

Table 1: GGM, GLM and Richards model parameter estimates. Mean parameter estimates are presented with the 95% confidence intervals obtained from the 200 bootstrap realizations.

Model	*r* (95% CI)	*p* (95% CI)	$\boldsymbol{k_0}$ (95% CI)	*a* (95% CI)
GGM	1.0 (95% CI: 0.81, 1.2)	0.72 (95% CI: 0.69, 0.75)	-	-
GLM	0.92 (95% CI: 0.7, 1.1)	0.73 (95% CI: 0.70, 0.77)	6.1 e+07 (95% CI: 3.4 e+05, 1 e+08)	-
Richards	0.56 (95% CI: 0.47, 0.64)	-	2.7 e+05 (95% CI: 2.3 e+05, 3.1 e+05)	0.055 (95% CI: 0.048, 0.068)

3.4 Performance Metrics

We compare the calibration (March 10-June 3, 2020) and 20-day ahead short term forecasting performance of the two models, GLM (Figure 4) and the Richards model (Figure 5) in Table 2. The GLM yielded the best fit to the daily incidence curve of COVID-19 in Pakistan based on each of the four performance metrics. The RMSE was estimated to be 4.3 times smaller, MAE was estimated to be 1.4 times smaller and the MIS was estimated to be 1.7 times smaller for the GLM compared to the Richards growth model. The 95% PI coverage for the GLM model was 1.6 times the Richards model.

For the short-term forecasting, again the GLM provided better overall performance compared to the Richards model. The RMSE was estimated to be 1.5 times smaller, the MAE was estimated to be 1.6 times smaller and the MIS was estimated to be 2.5 times smaller for the GLM compared to the Richards model. The 95% PI coverage for the GLM was 40 percent better than the Richards model (Table 2).

Table 2: Calibration and forecasting performance of the GLM and Richards growth model, calibrating data from March 10- June 3, 2020 and forecasting 20 days ahead.

Model	RMSE	MAE	PI coverage	MIS
		Calibration period		
GLM	22.0	183.1	87.2%	1618.0
Richards	94.9	253.8	53.5%	2790.8
		Forecasting period		
GLM	1432.6	1241.3	40%	27760
Richards	2186.2	2016.5	0%	70262

3.5 Reproduction Number

We estimate the reproduction number of COVID-19 for the first 63 epidemic days in Pakistan obtained using the two estimates of generation interval as described in the methods. There were minor differences in results using the two estimates of generation interval (as mentioned in the methods), indicating that estimation results are not sensitive to small differences in this parameter (Table 3). The incidence curve displays sub-exponential growth dynamics with a scaling of growth parameter, *p*, estimated at ~0.72 (95% CI: 0.69, 0.75) and the growth rate, *r*, estimated at ~1.0 (95% CI: 0.82, 1.2). The reproduction number for the early transmission phase was estimated to be *R*~1.2 (95% CI: 1.2, 1.3) (Table 3) indicating sustained SARS-CoV-2 transmission in the region.

Table 3: Mean estimates and the corresponding 95% confidence intervals for the basic reproduction number, growth rate and the scaling of growth parameter during the early growth phase as of May 11, 2020.

Parameters	Generation interval: mean of 5.2 days and a standard deviation of 1.72 days based on refs (Ganyani et al. 2020)	Generation interval: mean of 4.41 days and a standard deviation of 3.17 days based on refs. (Nishiura et al. 2020; You et al. 2020)
Reproduction number, *R*	1.2 (95% CI: 1.2, 1.3)	1.2 (95% CI: 1.2 , 1.2)
Growth rate, *r*	1.0 (95% CI: 0.82, 1.2)	1.0 (95% CI: 0.82, 1.2)
Scaling of growth parameter, *p*	0.72 (95% CI: 0.69,0.75)	0.72 (95% CI: 0.69, 0.74)

4 DISCUSSION

In this manuscript we have described and illustrated the application of three relatively simple dynamical growth models to the COVD-19 epidemic trajectory in Pakistan. Estimates of the reproduction number for the early ascending phase of the epidemic with an *R* estimated at ~1.2 implies sustained COVID-19 transmission in the region. This estimate is slightly lower compared to other reproduction numbers estimated in different geographic areas including Brazil, Peru, China, Korea, South Africa and Iran that lie in the range of 1.5-7.1 (Felix and Fontenele 2020; Hwang et al. 2020; Masjedi et al. 2020; Mbuvha and Marwala 2020; Mizumoto et al. 2020; Munayco et al. 2020; Muniz-Rodriguez et al. 2020; Read et al. 2020; Shim et al. 2020; Wu et al. 2020). The initial scaling of growth parameter in Pakistan indicates a sub-exponential growth trend (p~0.7), and the poor fit of the Richards model to the early growth phase confirms the need for sub-exponential growth, as the assumption of exponential growth resulted in over-estimation for the first half of the epidemic curve (Figure 5). In the context of a highly susceptible population, it is likely that the spatial heterogeneity in the risk of viral infection could have contributed to the polynomial growth pattern. This estimate of the scaling of growth parameter is consistent with sub-exponential growth dynamics of COVID-19 that have been observed in Singapore ($p \sim 0.7$), Korea ($p \sim 0.76$) and other Chinese provinces excluding Hubei ($p \sim 0.67$) (Roosa et al. 2020b; Shim et al. 2020; Tariq et al. 2020).

The GGM and the GLM provide a good fit to the reported time series data, whereas, the Richards model fails to capture the early dynamics of the epidemic. However, the Richards model is able to capture the later part of the epidemic (May 5-June 3, 2020), though the short term forecast underestimates the observed incidence curve. A retrospective examination of the 20-day ahead forecasts generated from the GGM and GLM showed reasonable estimations of case incidence, with the actual reported case incidence covered by the 95% prediction interval. The actual case count between May 12-May 31, 2020 was reported at ~40379. This lied within the 95% PI of the 20-day ahead GGM forecast (95% PI: 38019, 61373) . Similarly the GLM forecasted an average of ~82910 (95% PI: 58997, 105299) cases, whereas the actual case count was reported ~104573 between June 4- June 23, 2020, lying within the 95% PI. However, the Richards model seems to under predict the number of COVID-19 cases. Richards model forecasted an average of only ~58452 (95% PI: 44400, 77222) cases from June 4-June 23, 2020. Therefore, it is reasonable to assume that the suitable dynamic phenomenological growth models provide reasonable estimations of short term forecasts in near real time.

We have used daily series of reported COVID-19 case incidence data from Pakistan captured by the NIH surveillance system to calibrate the models and forecast the epidemic trajectory. As expected, the surveillance system only captures a fraction of the total number of SARS-CoV-2 infections as a substantial number of infections remain asymptomatic and can only be detected via broad testing and tracing strategies (Mizumoto et al. 2020; Nishiura et al. 2020). Our study implies that, in the absence of reliable information about the transmission mechanisms of an emerging infection and the effects of control intervention, simple phenomenological models can provide an early assessment of the potential scope of outbreaks in near real-time and serve as useful tools to generate short term forecasts of epidemic growth in real time (Chowell et al. 2016; Shanafelt et al. 2018). However, availability of timely case reporting is required so that projecting the epidemic in the future becomes worthwhile. Our study also shows promising results for forecasting the temporal evolution of COVID-19 epidemic in Pakistan using all the three models. We can further extend this work to a provincial level analysis expanding the forecasts to multiple geographic areas. Although, phenomenological models cannot replace mechanistic models that are amenable to incorporate different routes of disease transmission and realistic distributions for epidemiological parameters when the appropriate epidemiological data is available. These mechanistic models could be useful for long and short term forecasts as well as assess the impact of intervention strategies against COVID-19. However, our simple approach grounded on an empirical analytic framework that only requires estimating a few parameters has performed well in other settings to study the transmission dynamics of Ebola, Zika, and foot and mouth disease (Chowell et al. 2016; Shanafelt et al. 2018). These models have proved useful as the first response mathematical modeling toolkit to address the transmission dynamics of outbreaks (Pell et al. 2018).

Our study is not exempt from limitations. We use data by the date of reporting for fitting the models to the data, however, it is more accurate to use the data by the dates of symptom onset or after adjusting for reporting delays. Moreover, cases were not stratified as imported and local cases, therefore, we estimate reproduction number assuming all infections contribute in the same way to the transmission dynamics. Another limitation is the variable testing rates and testing strategies of COVID-19 that can add uncertainty to the results. Some countries like Korea have tested extensively for coronavirus whereas other countries including Italy, Spain and Pakistan prioritized testing to severe cases or limited populations.

In summary the mathematical and statistical methodology presented in this tutorial provides flexible and powerful tools to characterize and predict the trajectory of epidemics with quantified uncertainty.

ACKNOWLEDGMENTS

We acknowledge Sushma Dahal for her valuable input. This research was partially funded by the National Science Foundation under grant NSF #2026860.

REFERENCES

Ali, M., M. Imran, and A. Khan. 2020. "Analysis and Prediction of the COVID-19 outbreak in Pakistan." https://www.medrxiv.org/content/10.1101/2020.06.21.20136341v1, accessed 29th July.

Anderson, R. M., and R. M. May. 1991. *Infectious Diseases of Humans*.1st ed. Oxford: Oxford Univeristy Press.

Balcan, D., H. Hu, B. Goncalves, P. Bajardi, C. Poletto, J. J. Ramasco, D. Paolotti, N. Perra, M. Tizzoni, W. Van den Broeck, V. Colizza, and A. Vespignani. 2009. "Seasonal Transmission Potential and Activity Peaks of the New Influenza A(H1N1): A Monte Carlo Likelihood Analysis Based on Human Mobility". *BioMed Central Medicine* 7(1): 45.

Chinazzi, M., J. T. Davis, M. Ajelli, C. Gioannini, M. Litvinova, S. Merler, A. Pastore y Piontti, K. Mu, L. Rossi, K. Sun, C. Viboud, X. Xiong, H. Yu, M. E. Halloran, I. M. Longini, and A. Vespignani. 2020. "The Effect of Travel Restrictions on the Spread of the 2019 Novel Coronavirus (COVID-19) Outbreak". *Science* 368(6489): 395-400.

Chowell, G. 2017. "Fitting Dynamic Models to Epidemic Outbreaks with Quantified Uncertainty: A Primer for Parameter Uncertainty, Identifiability, and Forecasts". *Infectious Disease Modelling* 2(3): 379-398.

Chowell, G., C. E. Ammon, N. W. Hengartner, and J. M. Hyman. 2006. "Transmission Dynamics of the Great Influenza Pandemic of 1918 in Geneva, Switzerland: Assessing the Effects of Hypothetical Interventions". *Journal of Theoretical Biology* 241(2): 193-204.

Chowell, G., F. Abdirizak, S. Lee, J. Lee, E. Jung, H. Nishiura, and C. Viboud. 2015. "Transmission Characteristics of MERS and SARS in the Healthcare Setting: A Comparative Study". *BioMed Central Medicine* 13(1): 210.

Chowell, G., D. Hincapie-Palacio, J. Ospina, B. Pell, A. Tariq, S. Dahal, S. Moghadas, A. Smirnova, L. Simonsen, and C. Viboud. 2016. "Using Phenomenological Models to Characterize Transmissibility and Forecast Patterns and Final Burden of Zika Epidemics". *Public Library of Science Currents* 8: ecurrents.outbreaks.f14b2217c2902f2453d9320a2243a2235b9583.

Chowell, G., L. Sattenspiel, S. Bansal, and C. Viboud. 2016. "Mathematical Models to Characterize Early Epidemic Growth: A Review". *Physics of Life Reviews* 18: 66-97.

Chowell, G., A. Tariq, and J. M. Hyman. 2019. "A Novel Sub-Epidemic Modeling Framework for Short-term Forecasting Epidemic Waves". *BioMed Central Medicine* 17(1): 164.

Chretien, J-P., S. Riley, and D. B. George. 2015. "Mathematical Modeling of the West Africa Ebola Epidemic". *ELife* 4: e09186.

Colizza, V., A. Barrat, M. Barthélemy, and A. Vespignani. 2006. "The Role of the Airline Transportation Network in the Prediction and Predictability of Global Epidemics". *Proceedings of the National Academy of Sciences* 103(7): 2015-2020.

Efron, B., and R. J. Tibshirani. 1993. *An Introduction to the Bootstrap*. 1st ed. New York: Chapman & Hall.

Felix, F. H. C., and J. B. Fontenele. 2020. "Instantaneous R Calculation for COVID-19 Epidemic in Brazil". https://www.medrxiv.org/content/10.1101/2020.04.23.20077172v1, accessed 12th July.

Ganyani, T., C. Kremer, D. Chen, A. Torneri, C. Faes, J. Wallinga, and N. Hens. 2020. "Estimating the Generation Interval for Coronavirus Disease (COVID-19) Based on Symptom Onset Data, March 2020". *Eurosurveillance* 25(17): 2000257.

Gneiting, T., and A. E. Raftery. 2007. "Strictly Proper Scoring Rules, Prediction, and Estimation". *Journal of the American Statistical Association* 102(477): 359-378.

Hwang, J., H. Park, S-H. Kim, J. Jung, and N. Kim. 2020. "Basic and Effective Reproduction Numbers of COVID-19 Cases in South Korea Excluding Sincheonji Cases". https://www.medrxiv.org/content/10.1101/2020.03.19.20039347v2, accessed 30th July.

Times of India. 2020. Smart Lockdown Policy Working, Says Pakistan PM Imran as Covid-19 Near 275,000. https://timesofindia.indiatimes.com/world/pakistan/smart-lockdown-policy-working-says-pakistan-pm-imran-as-covid-19-near-275000/articleshow/77201024.cms, accessed 27th July.

Lara-Díaz, L., R. Bürger, and G. Chowell. 2019. "Comparative Analysis of Phenomenological Growth Models Applied to Epidemic Outbreaks". *Mathematical Biosciences and Engineering* 16: 4250-4273.

Masjedi, H., J. F. Rabajante, F. Bahranizadd, and M. H. Zare. 2020. "Nowcasting and Forecasting the Spread of COVID-19 in Iran." https://www.medrxiv.org/content/10.1101/2020.04.22.20076281v1, accessed 20th July.

Mbuvha R., and T. Marwala. 2020. "Bayesian inference of COVID-19 spreading rates in South Africa". *Public Library of Science One* 15(8): e0237126.

Merler, S., M. Ajelli, L. Fumanelli, M. F. C. Gomes, A. P. Y. Piontti, L. Rossi, D. L. Chao, I. M. Longini, M. E. Halloran, and A. Vespignani. 2015. "Spatiotemporal Spread of the 2014 Outbreak of Ebola Virus Disease in Liberia and the Effectiveness of Non-pharmaceutical Interventions: A Computational Modelling Analysis". *The Lancet Infectious Diseases* 15(2): 204-211.

Mizumoto, K., K. Kagaya, and G. Chowell. 2020. "Early Epidemiological Assessment of the Transmission Potential and Virulence of Coronavirus Disease 2019 (COVID-19) in Wuhan City: China, January-February, 2020". *BioMed Central Medicine* 18(1): 217

Mizumoto, K., K. Kagaya, A. Zarebski, and G. Chowell. 2020. "Estimating the Asymptomatic Proportion of Coronavirus Disease 2019 (COVID-19) Cases on Board the Diamond Princess Cruise Ship, Yokohama, Japan, 2020". *Euro surveillance* 25(10): 2000180.

Munayco, C. V., A. Tariq, R. Rothenberg, G. G. Soto-Cabezas, M. F. Reyes, A. Valle, L. Rojas-Mezarina, C. Cabezas, M. Loayza, G. Chowell, D. C. Garro, K. M. Vasquez, E. S. Castro, I. S. Ordinola, J. M. Mimbela, K. M. Cornejo, F. C. Quijano, L. La Torre Rosillo, L. O. Ibarguen, M. V. Dominguez, R. V. Gonzalez Seminario, M. C. Silva, M. S. Dreyfus, M. L. Pineda, M. Durand, N. Janampa, J. Chuquihuaccha, S. M. Lizarbe, D. E. Cusi, I. M. Pilco, A. Jaramillo, K. Vargas, O. Cabanillas, J. Arrasco, M. Vargas, and W. Ramos. 2020. "Early Transmission Dynamics and Control of COVID-19 in a Southern Hemisphere Setting: Lima-Peru, February 29th-March 30th, 2020". *Infectious Disease Modelling* 5: 338-345.

Muniz-Rodriguez, K., I. C-H. Fung, S. Ferdosi, S. Ofori, Y. Lee, A. Tariq, and G. Chowell. 2020. "Severe Acute Respiratory Syndrome Coronavirus 2 Transmission Potential, Iran, 2020". *Emerging Infectious Diseases Journal* 26(8): 1915.

Nishiura, H., and G. Chowell. 2009. "The Effective Reproduction Number as a Prelude to Statistical Estimation of Time-dependent Epidemic Trends". *Mathematical and Statistical Estimation Approaches in Epidemiology*, edited by G.Chowell, J.M.Hyman, L.M.A.Bettencourt, and C.Castillo-Chavez, 103-121. Dordrecht: Springer.

Nishiura, H., and G. Chowell. 2014. "Early Transmission Dynamics of Ebola Virus Disease (EVD), West Africa, March to August 2014". *Euro surveillance* 19(36) :pii=20894.

Nishiura, H., G. Chowell, H. Heesterbeek, and J. Wallinga. 2010. "The Ideal Reporting Interval for an Epidemic to Objectively Interpret the Epidemiological Time Course". *Journal of the Royal Society, Interface* 7(43): 297-307.

Nishiura, H., T. Kobayashi, T. Miyama, A. Suzuki, S. M. Jung, K. Hayashi, R. Kinoshita, Y. Yang, B. Yuan, A. R. Akhmetzhanov, and N. M. Linton. 2020. "Estimation of the Asymptomatic Ratio of Novel Coronavirus infections (COVID-19)". *International Journal of Infectious Diseases* 94: 154-155.

Nishiura, H., N. M. Linton, and A. R. Akhmetzhanov. 2020. "Serial Interval of Novel Coronavirus (COVID-19) Infections". *International Journal of Infectious Diseases* 93: 284-286.

Paine, S., G. Mercer, P. Kelly, D. Bandaranayake, M. Baker, Q. Huang, G. Mackereth, A. Bissielo, K. Glass, and V. Hope. 2010. "Transmissibility of 2009 Pandemic Influenza A(H1N1) in New Zealand: Effective Reproduction Number and Influence of Age, Ethnicity and Importations". *Euro surveillance* 15(24) :19591.

Government of Pakistan. 2020. COVID-19 dashboard. http://covid.gov.pk/stats/pakistan, accessed 27th June.

Pell, B., Y. Kuang, C. Viboud, and G. Chowell. 2018. "Using Phenomenological Models for Forecasting the 2015 Ebola Challenge". *Epidemics* 22: 62-70.

Read, J. M., J. R. Bridgen, D. A. Cummings, A. Ho, and C. P. Jewell. 2020. "Novel Coronavirus 2019-nCoV: Early Estimation of Epidemiological Parameters and Epidemic Predictions". https://www.medrxiv.org/content/10.1101/2020.01.23.20018549v2, accessed 27th July.

Richards, F. J. 1959. "A Flexible Growth Function for Empirical Use". *Journal of Experimental Botany* 10(2): 290-301.

Roosa, K., Y. Lee, R. Luo, A. Kirpich, R. Rothenberg, J. M. Hyman, P. Yan, and G. Chowell. 2020a. "Short-term Forecasts of the COVID-19 Epidemic in Guangdong and Zhejiang, China: February 13-23, 2020". *Journal of Clinical Medicine* 9(2): 596.

Roosa, K., Y. Lee, R. Luo, A. Kirpich, R. Rothenberg, J. M. Hyman, P. Yan, and G. Chowell. 2020b. "Real-time Forecasts of the COVID-19 Epidemic in China from February 5th to February 24th, 2020". *Infectious Disease Modelling* 5: 256-263.

Shanafelt, D. W., G. Jones, M. Lima, C. Perrings, and G. Chowell. 2018. "Forecasting the 2001 Foot-and-Mouth Disease Epidemic in the UK". *Public Health Official Journal of EcoHealth Alliance* 15(2): 338-347.

Shim, E., A. Tariq, W. Choi, Y. Lee, and G. Chowell. 2020. "Transmission Potential and Severity of COVID-19 in South Korea". *International Journal of Infectious Diseases* 93: 339-344.

Tariq, A., Y. Lee, K. Roosa, S. Blumberg, P. Yan, S. Ma, and G. Chowell. 2020. "Real-time Monitoring the Transmission Potential of COVID-19 in Singapore, March 2020". *BioMed Central Medicine* 18(1): 166.

Viboud, C., L. Simonsen, and G. Chowell. 2016. "A Generalized-Growth Model to Characterize the Early Ascending Phase of Infectious Disease Outbreaks". *Epidemics* 15: 27-37.

Wang, X. S., J. Wu, and Y. Yang. 2012. "Richards model revisited: Validation by and Application to Infection Dynamics". *Journal of Theoretical Biology* 313: 12-19.

Wu, J. T., K. Leung, and G. M. Leung. 2020. "Nowcasting and Forecasting the Potential Domestic and International Spread of the 2019-nCoV Outbreak Originating in Wuhan, China: A Modelling Study". *The Lancet* 395(10225): 689-697.

Yan, P., and G. Chowell. 2019. *Quantitative Methods for Investigating Infectious Disease Outbreaks*. 1st ed. Berlin, Germany: Springer International Publishing.

You, C., Y. Deng, W. Hu, J. Sun, Q. Lin, F. Zhou, C. H. Pang, Y. Zhang, Z. Chen, and X-H. Zhou. 2020. "Estimation of the time-varying reproduction number of COVID-19 outbreak in China". *International Journal of Hygiene and Environmental Health* 228:113555.

AUTHOR BIOGRAPHIES

AMNA TARIQ is a current doctoral student at Georgia State University School of Public Health, Atlanta, Georgia. She is a Second Century Initiative (2CI) Fellow under the big data cluster. Her research focuses on infectious disease dynamics and mathematical epidemiology. Particularly, her research interests include mathematical modeling of infectious diseases and exploring quantitative methods for analyzing disease outbreaks to better understand and predict the ongoing and upcoming infectious disease epidemics such as Ebola, Dengue and COVID-19 pandemic. Her email address is atariq1@student.gsu.edu.

KIMBERLYN ROOSA is a recent graduate from Georgia State University School of Public Health, Atlanta, Georgia, where she was a Second Century Initiative Fellow focusing on research in mathematical epidemiology. Particularly, her research has included computational analyses, such as parameter identifiability analyses and uncertainty quantification, as well as real-time application of mathematical models to characterize and predict infectious disease outbreak trends, including Ebola in the Democratic Republic of the Congo and the ongoing COVID-19 pandemic. Her email address is kroosa1@student.gsu.edu.

GERARDO CHOWELL is a professor of mathematical epidemiology and chair of the Department of Population Health Sciences at Georgia State University School of Public Health, Atlanta, Georgia. He also holds an external affiliation as a Senior Research Fellow at the Division of International Epidemiology and Population Studies at the Fogarty International Center, National Institutes of Health. As a mathematical epidemiologist, he develops and applies mathematical and statistical methods to investigate the transmission dynamics and control of infectiuos diseases. His email address is gchowell@gsu.edu.