RESEARCH ARTICLE

A data-driven model to describe and forecast the dynamics of COVID-19 transmission

Henrique Mohallem Paiva[1]*, Rubens Junqueira Magalhães Afonso[2,3], Igor Luppi de Oliveira[1], Gabriele Fernandes Garcia[1]

1 Institute of Science and Technology (ICT), Federal University of São Paulo (UNIFESP), São José dos Campos, SP, Brazil, **2** Institute of Flight System Dynamics, Department of Aerospace and Geodesy, Technical University of Munich (TUM), Garching bei München, Bavaria, Germany, **3** Department of Electronics Engineering, Aeronautics Institute of Technology (ITA), São José dos Campos, SP, Brazil

* hmpaiva@unifesp.br

Citation: Paiva HM, Afonso RJM, de Oliveira IL, Garcia GF (2020) A data-driven model to describe and forecast the dynamics of COVID-19 transmission. PLoS ONE 15(7): e0236386. https://doi.org/10.1371/journal.pone.0236386

Editor: Martin Chtolongo Simuunza, University of Zambia, ZAMBIA

Received: April 13, 2020

Accepted: July 6, 2020

Published: July 31, 2020

Data Availability Statement: All relevant data either are within the manuscript or may be downloaded from the ECDC website: www.ecdc.europa.eu/en.

Funding: Dr Rubens Junqueira Magalhães Afonso received a postdoctoral fellowship (#88881.145490/2017-01) from a joint program between the Brazilian governmental agency CAPES and the Alexander Von Humboldt foundation, from the Federal Ministry for Education and Research of Germany. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de

Abstract

This paper proposes a dynamic model to describe and forecast the dynamics of the coronavirus disease COVID-19 transmission. The model is based on an approach previously used to describe the Middle East Respiratory Syndrome (MERS) epidemic. This methodology is used to describe the COVID-19 dynamics in six countries where the pandemic is widely spread, namely China, Italy, Spain, France, Germany, and the USA. For this purpose, data from the European Centre for Disease Prevention and Control (ECDC) are adopted. It is shown how the model can be used to forecast new infection cases and new deceased and how the uncertainties associated to this prediction can be quantified. This approach has the advantage of being relatively simple, grouping in few mathematical parameters the many conditions which affect the spreading of the disease. On the other hand, it requires previous data from the disease transmission in the country, being better suited for regions where the epidemic is not at a very early stage. With the estimated parameters at hand, one can use the model to predict the evolution of the disease, which in turn enables authorities to plan their actions. Moreover, one key advantage is the straightforward interpretation of these parameters and their influence over the evolution of the disease, which enables altering some of them, so that one can evaluate the effect of public policy, such as social distancing. The results presented for the selected countries confirm the accuracy to perform predictions.

1 Introduction

The geographic spread of a novel coronavirus (SARS-CoV-2) in Wuhan, China, in December 2019, characterized the emergence of a severe acute respiratory syndrome, afterward named COVID-19 [1–3]. Studies show that SARS-CoV-2 has a rapidly human-to-human and asymptomatic transmission, mainly by respiratory droplets, which makes it more contagious than Severe Acute Respiratory Syndrome (SARS) and Middle East Respiratory Syndrome (MERS), other well-known coronaviruses diseases, despite the contamination similarity. The new

Nível Superior - Brasil (CAPES) - Finance Code 001. The other authors received no specific funding for this work. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

identified virus is closely related to SARS-CoV (79%) and MERS-CoV (50%) [1],[4]. Concerning the clinical aspects of COVID-19, the most common symptom reported was fever, followed by cough. A severe onset of the disease can lead to death due to alveolar damage and respiratory failure [1],[3],[5],[6].

Under these circumstances, COVID-19 has been spread through over 100 countries and all of the continents in a couple of months after its first confirmed case and it was declared, by the World Health Organization (WHO), a public health emergency of international matter in January 2020 [1],[7]. As of April 11th, 2020, about 1.6 million cases and 100.000 deaths were confirmed globally, showing an expressing breakthrough in Europe (839257 confirmed and 70565 deaths) and the region of the Americas (536664 confirmed and 19294 deaths) in comparison to the outbreak area (118549 confirmed and 4017 deaths). The most affected countries are the United States of America, Italy, Spain, China, Germany and France [8].

In the past two decades, there have been two coronavirus epidemics that also led to global health consternation, SARS in 2003 and MERS around 2012. It is noteworthy that COVID-19 has already killed more people than both of those diseases combined [9],[10]. Other epidemics have also ravaged the world and been considered an international emergency: H1N1 (2009), poliomyelitis (2014), Ebola (2014, 2019) and Zika (2016) [1].

Meanwhile, scientists have always had an essential role to play in the study of the dynamics of infectious diseases, particularly in mathematical modeling. The study of epidemiological mathematical models leads to deep understanding of the dynamics of epidemics, being an important tool to assess the potential effects of preventive and controlled measures, especially when their characteristics are still unclear [11],[12].

Recent works about COVID-19 models present different approaches to describe the proportions of its transmission and future numbers. The most common ones are the Susceptible-Infectious-Recovered/Death (SIRD) model [1],[13] and the Susceptible-Exposed-Infectious-Recovered (SEIR) model [14–16]; both derived from the Susceptible-Infectious-Recovered (SIR) pioneer model described by Kermack and McKendrick in 1927 [14–17]. Other works describe COVID-19 utilizing an exponential family model [18]; a second derivative model [19]; and Susceptible, Un-quarantined infected, Quarantined infected, Confirmed infected (SUQC) model [20].

These models are relevant to understanding the factors that alter the evolution of the disease. Furthermore, such models can be used to forecast evolution of the pandemic. In planning policies, a forecast of the number of infected and deceased individuals is of paramount importance. However, the models mentioned in the paragraph above do not account other classes, such as hospitalized individuals. The health infrastructure was placed under extreme pressure by the COVID-19 pandemic. Therefore, being able to forecast the number of hospitalized individual rises as an important task. Measures to mitigate the amount of simultaneously hospitalized individuals can be designed and evaluated to manage the facilities so that the deceased due to lack of proper treatment are minimized.

In the present paper, a SEIR model that includes the deceased and hospitalized is proposed to describe the dynamics of COVID-19. The availability of up-to-date data regarding the number of infected and deceased each day enables the estimation of the parameters so as to match the output of the model with the data from each country. Provided that, for each country two phases are assumed: (i) when no measures are taken such as reduction of social interaction and (ii) when public policies act to reduce the spread of the disease. Only the parameters linked to the transmissibility rates are allowed to change between these two phases. Moreover, the basic reproduction number is deduced for the proposed model and calculated from the parameters in the two phases.

One key feature of the proposed model is the straightforward connection between the parameters and their influence in the evolution of the disease. Therefore, the physical meaning of the parameters is clear, as opposed to black-box approaches and others that do not include the classes of individuals such as infected, recovered and so on, but rather focus on fitting curves. Therefore, it enables adjusting the parameters separately, each of which reflect real-world policy/behavior changes. For instance, transmissibility by infected individuals is a model parameter, which is affected by policy such as social distancing. This relationship allows one to perform simulations of different scenarios to predict the evolution of the disease under varied degrees of social distancing. This, in turn, is helpful to evaluate which policy is more promising. Moreover, the inclusion of the hospitalized class is another feature not present in the aforementioned models.

The remaining sections of this paper are divided as follows. The proposed model is presented in Section 2.1, whereas the basic transmission number is calculated as function of the model parameters in Section 2.2 and the parameter estimation problem is formally defined in Section 2.3. The results are presented in Section 3 and discussed in Section 4. Concluding remarks are given in Section 5.

2 Materials and methods

2.1 Model

Our proposed model is of the SEIR type and is inspired by the one used in [21] to successfully model the evolution of the Middle Eastern Respiratory Syndrome (MERS) coronavirus dynamic in the outbreak in South Korea in 2015. On the other hand, we introduce one extra class not present in [21], namely the class of deceased people, which play a relevant role in the evolution of the COVID-19. Therefore, the proposed model divides the population of interest in seven classes, as shown in Table 1. It is clear that each variable in Table 1 cannot assume negative values, as each represents the amount of individuals in a class.

The total population is represented by the symbol N. The introduction of the number of deceased is important in the case of COVID-19 due to two main effects: (i) in formulating policies, it is paramount to be able to predict the amount of deceased persons; (ii) these individuals are removed from the infected population and thus do not contribute to generate new infections. Albeit relevant in absolute terms and in proportion to the number of infected individuals, the deceased are proportionally low as compared to the total population, which justifies the adoption of Assumption 1 in our model.

Assumption 1 *The population is deemed constant, i.e., N is not altered throughout the simulation of the model.*

Table 1. Classes of the proposed model.

Symbol	Meaning
$S \geq 0$	Susceptible.
$E \geq 0$	Exposed.
$I \geq 0$	Infectious Symptomatic.
$A \geq 0$	Asymptomatic.
$H \geq 0$	Hospitalized.
$R \geq 0$	Recovered.
$D \geq 0$	Deceased.

https://doi.org/10.1371/journal.pone.0236386.t001

Assumption 1 is reasonable, as the time interval for the simulation is short in terms of demographic changes and the amount of deceased people by the disease itself is not large enough to significantly alter the population of a country, as corroborated by the data.

Focus is placed on the relevant parameters to obtain a model that represents the main characteristics of the dynamics with minimal additional complexity. Following the development in [21], Assumption 2 deems zoonotic transmission not relevant for modeling purposes.

Assumption 2 *Zoonotic transmission is not considered within the proposed model.*

Indeed, although it is suspected that the origin of the first cases is zoonotic, the evolution of the transmission has, since the first few hundred cases, in China been among humans only. In the other countries studied in the present work there has been no report of zoonotic transmissions.

Another relevant difference between our model and the one in [21] is the inclusion of an infection rate dependent on the asymptomatic individuals. It has been noticeable that many asymptomatic individuals transmit the virus SARS-CoV-2, therefore we introduced a term in the generation of new infections that reflects this fact. However, the infection rate of these individuals is not considered the same as that of infected symptomatic nor that of hospitalized ones, as we introduced a different coefficient for this rate in the generation of new infections.

Assumption 3 *Deceased individuals came either from class H (hospitalized) or I (infected), but not from class A (asymptomatic).*

Assumption 3 is made because our model parameters are estimated based on real data and the data available do not include the number of asymptomatic individuals that perish from the disease. The reported deceased come from the infected (I) or hospitalized (H) classes.

The resulting model under the aforementioned assumptions is given in (1), where we aimed at keeping the notation as close as possible to [21] to simplify comparison. The equations are presented in their general form; nevertheless, in order to represent Assumption 3, the value of δ_A is assumed equal to zero.

$$\begin{aligned}
\dot{S} = &\quad -S\frac{\beta(I+\ell_a A+\ell H)}{N},\\
\dot{E} = &\quad S\frac{\beta(I+\ell_a A+\ell H)}{N}-\kappa E,\\
\dot{I} = &\quad \kappa\rho E-(\gamma_a+\gamma_I+\delta_I)I,\\
\dot{A} = &\quad \kappa(1-\rho)E-\mu A,\\
\dot{H} = &\quad \gamma_a I-(\gamma_r+\delta_H)H,\\
\dot{R} = &\quad \gamma_I I+\gamma_r H+\mu(1-\delta_A)A,\\
\dot{D} = &\quad \delta_H H+\delta_I I+\mu\delta_A A.
\end{aligned} \tag{1}$$

The meaning of the parameters in (1) is given in Table 2. A block diagram representation depicting the relationship between the variables in the model (1) is shown in Fig 1.

2.2 Basic reproduction number

The basic reproduction number $\mathcal{R}_0$ is the ratio of new infections from one single infected individual [22] in a totally susceptible population. It has theoretical value to understand how "infectious" a disease is, as larger $\mathcal{R}_0$ indicates more spreading.

Introducing the state variable

$$\mathbf{x}=[x_1 \quad x_2 \quad x_3 \quad x_4 \quad x_5 \quad x_6 \quad x_7]^\top=[S \quad E \quad I \quad A \quad H \quad R \quad D]^\top, \tag{2}$$

Table 2. Parameters of the proposed model.

Symbol	Meaning
$\beta \geq 0$	human-to-human transmission rate per unit time (day)
$\ell \geq 0$	relative transmissibility of hospitalized patients
$\ell_a \geq 0$	relative transmissibility of asymptomatic infected
$\kappa \geq 0$	rate at which an individual leaves the exposed class by becoming infectious (symptomatic or asymptomatic)
$\rho \geq 0$	proportion of progression from exposed class E to symptomatic infected class I
$\gamma_a \geq 0$	rate at which symptomatic individuals are hospitalized
$\gamma_I \geq 0$	recovery rate without being hospitalized
$\gamma_r \geq 0$	recovery rate of hospitalized patients
$\mu \geq 0$	rate of asymptomatic infectious that no longer transmit, becoming either recovered or deceased
$\delta_A \geq 0$	proportion of progression from asymptomatic class A to deceased class D
$\delta_H \geq 0$	death rate of hospitalized patients
$\delta_I \geq 0$	death rate of infected patients

https://doi.org/10.1371/journal.pone.0236386.t002

one may rewrite the system of ordinary differential equations (ODEs) (1) as

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}). \tag{3}$$

This nonlinear dynamic may have equilibrium points, i.e., points $\mathbf{x}_{eq}$ such that $\mathbf{f}(\mathbf{x}_{eq}) = \mathbf{0}$. Particularly interesting is the so-called disease-free equilibrium point in Definition 1.

Definition 1 *A disease-free equilibrium point (DFE)* $\bar{\mathbf{x}}$ *is an equilibrium point of the dynamic* (3) *such that* $\bar{x}_1 = S = N$ *and* $\bar{x}_i = 0, i > 1$.

It is easy to verify that (1) has a DFE.

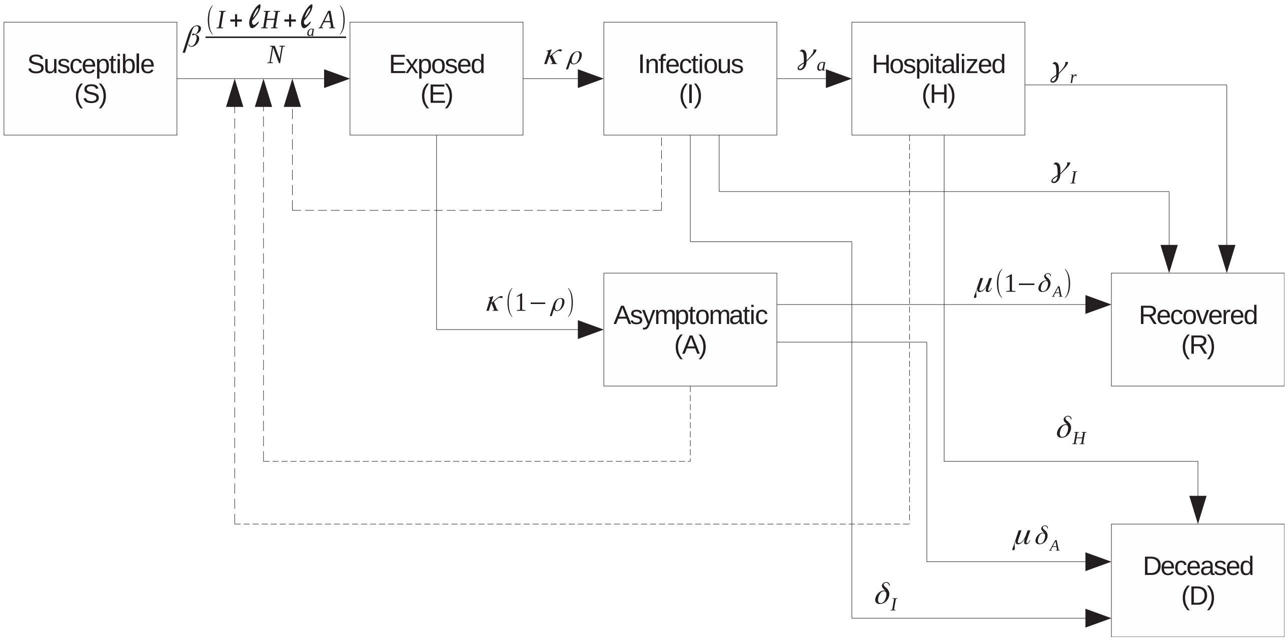

Fig 1. Block diagram representation of the model (1).

https://doi.org/10.1371/journal.pone.0236386.g001

Linearization of the model around the DFE is used in [23] to study the relationship between $\mathcal{R}_0$ and stability. It has been demonstrated that if $\mathcal{R}_0 > 1$, then the disease-free equilibrium point is unstable, i.e., an infected individual is enough to remove the system from the neighborhood of the DFE and infection of the population is possible. On the other hand, $\mathcal{R}_0 < 1$ entails local asymptotic stability.

The authors argue in [23] that the definition of $\mathcal{R}_0$ in terms of the model is not purely dependent on the model equations only. In fact, one must arbitrarily select which are the "infected" classes (state components of **x** in our notation). For our analysis we consider the infected states to be $x_2 = E$, $x_3 = I$, $x_4 = A$, and $x_5 = H$. The equations of interest for the analysis are the ODEs related to these state variables, namely

$$\dot{\mathbf{x}}_I = \begin{bmatrix} \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \\ \dot{x}_5 \end{bmatrix} = \begin{bmatrix} x_1 \dfrac{\beta(x_3 + \ell_a x_4 + \ell x_5)}{N} - \kappa x_2 \\ \kappa \rho x_2 - (\gamma_a + \gamma_I + \delta_I) x_3 \\ \kappa(1-\rho) x_2 - \mu x_4 \\ \gamma_a x_3 - (\gamma_r + \delta_H) x_5 \end{bmatrix} = \mathbf{f}_I(\mathbf{x}), \tag{4}$$

where $\mathbf{x}_I = [x_2 \quad x_3 \quad x_4 \quad x_5]^\top$ are the infected states. One may now separate the dynamics $\mathbf{f}_I(\mathbf{x})$ in two terms $\mathbf{f}_I(\mathbf{x}) = \mathcal{F}_I(\mathbf{x}) - \mathcal{V}_I(\mathbf{x})$, where $\mathcal{F}_I$ are the (positive) new infections and $\mathcal{V}_I$ are the transitions between classes. From (4) one has

$$\mathcal{F}_I(\mathbf{x}) = \begin{bmatrix} x_1 \dfrac{\beta(x_3 + \ell_a x_4 + \ell x_5)}{N} \\ 0 \\ 0 \\ 0 \end{bmatrix} \text{ and } \mathcal{V}_I(\mathbf{x}) = \begin{bmatrix} \kappa x_2 \\ -\kappa \rho x_2 + (\gamma_a + \gamma_I + \delta_I) x_3 \\ -\kappa(1-\rho) x_2 + \mu x_4 \\ -\gamma_a x_3 + (\gamma_r + \delta_H) x_5 \end{bmatrix}. \tag{5}$$

Calculating the Jacobian of $\mathcal{F}_I$ and $\mathcal{V}_I$ with respect to $\mathbf{x}_I$ at $\bar{\mathbf{x}}$ yields

$$\frac{\partial \mathcal{F}_I}{\partial \mathbf{x}_I}(\bar{\mathbf{x}}) = \begin{bmatrix} 0 & \beta & \beta \ell_a & \beta \ell \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{6}$$

and

$$\frac{\partial \mathcal{V}_I}{\partial \mathbf{x}_I}(\bar{\mathbf{x}}) = \begin{bmatrix} \kappa & 0 & 0 & 0 \\ -\kappa\rho & \gamma_a + \gamma_I + \delta_I & 0 & 0 \\ -\kappa(1-\rho) & 0 & \mu & 0 \\ 0 & -\gamma_a & 0 & \gamma_r + \delta_H \end{bmatrix}. \tag{7}$$

Assuming for the time being that the parameters are positive (an assumption that will be verified in the results in Section 3), $\frac{\partial \mathcal{V}_I}{\partial \mathbf{x}_I}(\bar{\mathbf{x}})$ is non-singular, thus we may calculate the inverse as

$$\left[\frac{\partial \mathcal{V}_I}{\partial \mathbf{x}_I}(\bar{\mathbf{x}})\right]^{-1} = \begin{bmatrix} \frac{1}{\kappa} & 0 & 0 & 0 \\ \frac{\rho}{\gamma_a+\gamma_I+\delta_I} & \frac{1}{\gamma_a+\gamma_I+\delta_I} & 0 & 0 \\ \frac{1-\rho}{\mu} & 0 & \frac{1}{\mu} & 0 \\ \frac{\gamma_a \rho}{(\gamma_r+\delta_H)(\gamma_a+\gamma_I+\delta_I)} & \frac{\gamma_a}{(\gamma_r+\delta_H)(\gamma_a+\gamma_I+\delta_I)} & 0 & \frac{1}{\gamma_r+\delta_H} \end{bmatrix}. \quad (8)$$

The so-called next generation matrix is defined in [23] as $\frac{\partial \mathcal{F}_I}{\partial \mathbf{x}_I}(\bar{\mathbf{x}})\left[\frac{\partial \mathcal{V}_I}{\partial \mathbf{x}_I}(\bar{\mathbf{x}})\right]^{-1}$ and $\mathcal{R}_0$ is then determined as the spectral radius of the next generation matrix. From (6) and (8) one obtains

$$\frac{\partial \mathcal{F}_I}{\partial \mathbf{x}_I}\left[\frac{\partial \mathcal{V}_I}{\partial \mathbf{x}_I}(\bar{\mathbf{x}})\right]^{-1} = \begin{bmatrix} \beta\left(\frac{\rho}{\gamma_a+\gamma_I+\delta_I} + \ell_a \frac{1-\rho}{\mu} + \frac{\ell \gamma_a \rho}{(\gamma_r+\delta_H)(\gamma_a+\gamma_I+\delta_I)}\right) & \beta\left(\frac{1}{\gamma_a+\gamma_I+\delta_I} + \frac{\ell \gamma_a}{(\gamma_r+\delta_H)(\gamma_a+\gamma_I+\delta_I)}\right) & \frac{\beta \ell_a}{\mu} & \frac{\beta \ell}{\gamma_r+\delta_H} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}. (9)$$

The next generation matrix in (9) has only one positive eigenvalue and the remaining three are null. Therefore, the spectral radius of $\frac{\partial \mathcal{F}_I}{\partial \mathbf{x}_I}\left[\frac{\partial \mathcal{V}_I}{\partial \mathbf{x}_I}(\bar{\mathbf{x}})\right]^{-1}$ is

$$\mathcal{R}_0 = \beta\left(\frac{\rho}{\gamma_a+\gamma_I+\delta_I} + \ell_a \frac{1-\rho}{\mu} + \frac{\ell \gamma_a \rho}{(\gamma_r+\delta_H)(\gamma_a+\gamma_I+\delta_I)}\right). \quad (10)$$

2.3 Parameter estimation

The parameters are estimated in order to match the total number of cases C and deceased D obtained from the model with the available data for each country, herein termed C_{real} and D_{real}, respectively.

The cumulative number of infected individuals C at time t is obtained as follows:

$$C(t) = \int_0^t \kappa \rho E(\tau)\, d\tau, \quad (11)$$

where the integrand is the positive part of $\dot{I}$ in (1), which represents the number of new infected symptomatic individuals that enter class I per time unit.

As in the case of the MERS spread in South Korea reported in [21], some parameters of the model are allowed to assume two different values at two different periods, namely β, ℓ_a and ℓ, which are related to the rate of contagion of the population. These parameters can be affected by changes in policies by the authorities and reflect the control of the spread of the virus.

Therefore, one has $\beta = \beta(t)$, $\ell_a = \ell_a(t)$, and $\ell = \ell(t)$ defined as:

$$\beta(t) = \begin{cases} \beta_1, & 0 \leq t < T \\ \beta_2, & t \geq T \end{cases}$$
$$\ell_a(t) = \begin{cases} \ell_{a,1}, & 0 \leq t < T \\ \ell_{a,2}, & t \geq T \end{cases} \tag{12}$$
$$\ell(t) = \begin{cases} \ell_1, & 0 \leq t < T \\ \ell_2, & t \geq T \end{cases}$$

where T the is phase change time given in days and is also estimated. The data C_{real} and D_{real} are given as sequences $C_{real}(i)$ and $D_{real}(i)$ for $i = 0, 1, 2, \ldots, \bar{T}$, where 0 represents the day the number of infected people in the country reached 500 and $\bar{T}$ is the number of days of data used in the parameter estimation, which may vary for each country. All the remaining parameters are assumed constant.

The estimation is carried out via a constrained optimization problem enunciated as

Problem 1

$$\mathbf{p}* = \arg\min \sum_{i=0\bar{T}} i \left\{ \left[\frac{C(i) - C_{real}(i)}{C_{real}(\bar{T})} \right]^2 + \left[\frac{D(i) - D_{real}(i)}{D_{real}(\bar{T})} \right]^2 \right\} \tag{13}$$

subject to

$$\mathbf{p} \quad = [\, T \quad \beta_1 \quad \beta_2 \quad \ell_1 \quad \ell_2 \quad \ell_{a,1} \quad \ell_{a,2} \quad \kappa \quad \rho \quad \gamma_a \quad \gamma_I \quad \gamma_r \quad \mu \quad \delta_H \quad \delta_I \quad I(0) \,]^\top, \tag{14}$$

$$\mathbf{p} \geq \mathbf{0}. \tag{15}$$

In (15) the inequality in considered element-wise. The result $\mathbf{p}^*$ is the vector of optimal parameter values, i.e., values that minimize (13) while satisfying the constraints (14) and (15). This optimization problem is solved using a Sequential Quadratic Program algorithm [24], adopting as starting point the parameters used in [21] to describe the MERS epidemic.

2.4 Data source

The COVID-19 data used in this paper, with the number of infected and deceased people in each country, were downloaded from the website of the European Centre for Disease Prevention and Control (ECDC) [25] on June 19th, 2020.

3 Results

The resulting parameters from the estimation are shown in Table 3, where the numbers in subscript "1" and "2" indicate whether the parameter refers to the first or second period, respectively. The basic reproduction number was also calculated for both periods and is depicted separately in Table 3, since it is not a model parameter.

The data of number of cumulative infected individuals and deceased per day as well as the model output for these variables are shown in Figs 2, 3, 4, 5, 6 and 7, for China, Italy, Spain, France, Germany, and the USA, respectively.

Table 3. Results of the parameter estimation for the selected countries.

Parameter	Country					
	China	Italy	Spain	France	Germany	USA
T (days)	18	30	23	26	24	51
β_1	0.334	0.189	0.382	0.298	0.135	0.303
β_2	0.140	0.081	0.160	0.129	0.055	0.130
ℓ_1	0.673	8.000	7.690	8.000	4.800	0.851
ℓ_2	0.135	8.000	6.490	8.000	1.130	0.851
$\ell_{a.1}$	8.000	0.649	3.900	8.000	4.900	4.090
$\ell_{a.2}$	8.000	0.649	3.900	8.000	4.900	0.820
κ	0.440	0.284	0.362	0.309	0.578	1.330
ρ	0.053	0.270	0.102	0.033	0.021	1.010
γ_a	0.503	0.224	0.116	0.300	0.542	0.055
γ_I	0.263	0.040	0.063	0.020	0.050	0.296
γ_r	0.141	0.240	0.281	0.131	0.036	0.018
μ	1.640	0.146	1.030	1.530	0.302	0.828
δ_H	0.008	0.023	0.019	0.029	0.003	0.00029
δ_I	0.003	0.023	0.016	0.018	0.002	0.023
$\mathcal{R}_{0.1}$	1.62	2.00	2.09	1.98	2.47	2.90
$\mathcal{R}_{0.2}$	0.66	0.86	0.84	0.85	0.91	1.26

https://doi.org/10.1371/journal.pone.0236386.t003

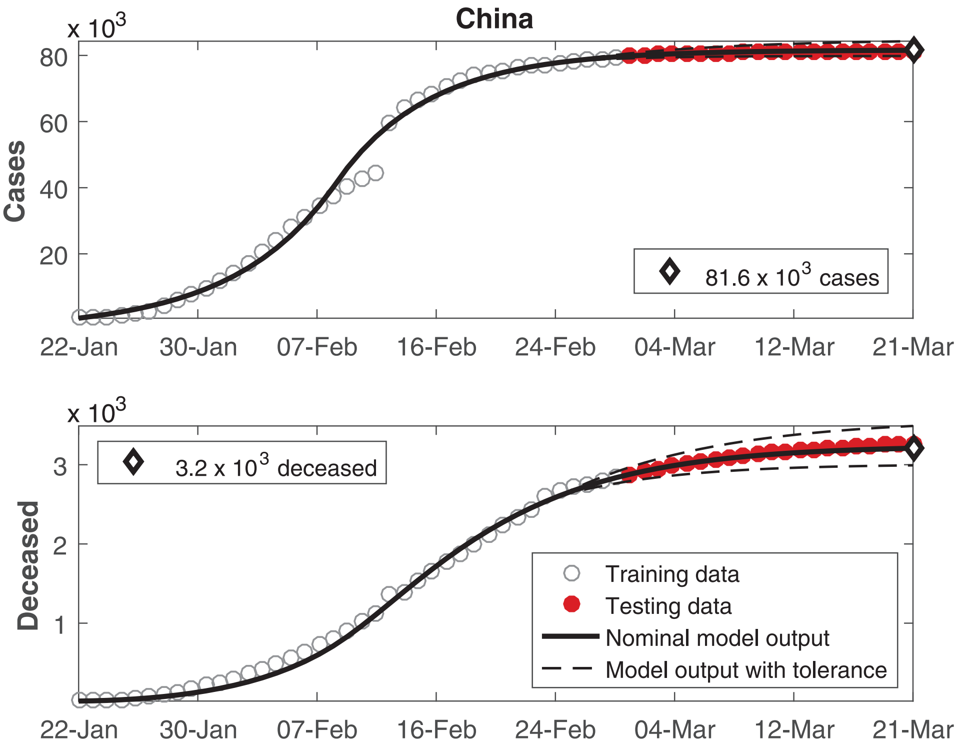

Fig 2. Cumulative number of infected individuals *C* and number of deceased individuals *D* for China.

https://doi.org/10.1371/journal.pone.0236386.g002

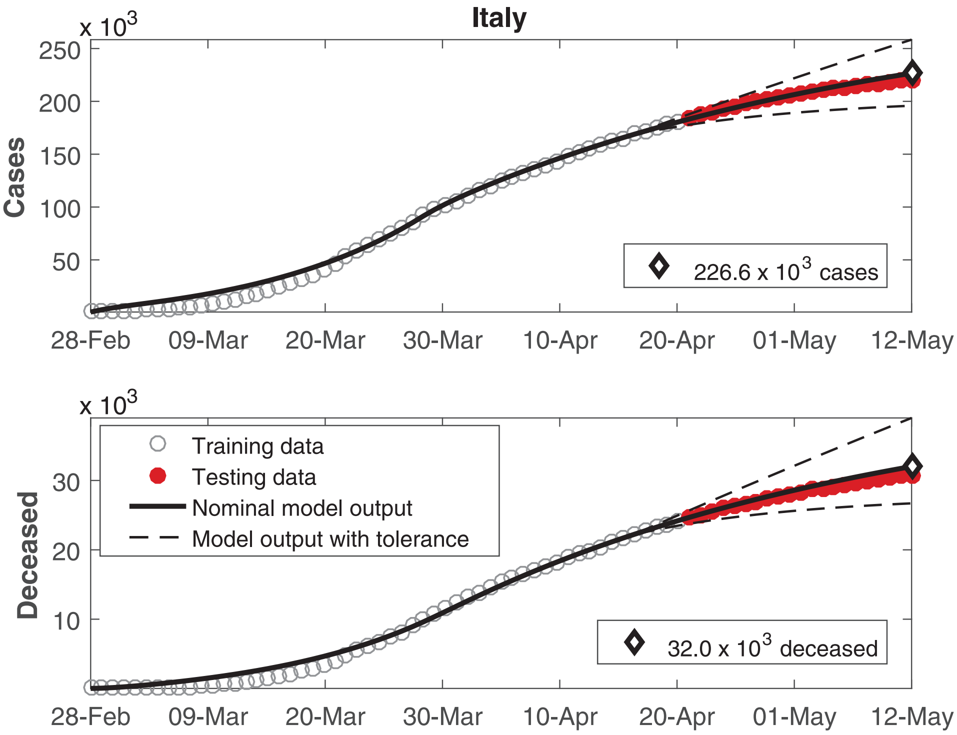

Fig 3. Cumulative number of infected individuals *C* and number of deceased individuals *D* for Italy.

https://doi.org/10.1371/journal.pone.0236386.g003

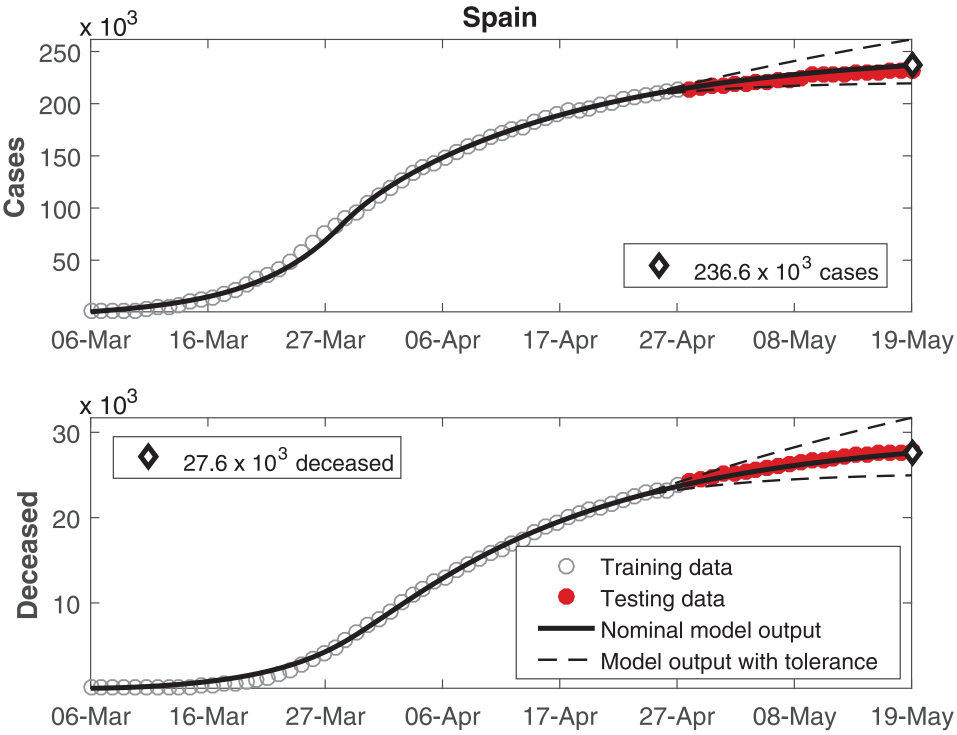

Fig 4. Cumulative number of infected individuals *C* and number of deceased individuals *D* for Spain.

https://doi.org/10.1371/journal.pone.0236386.g004

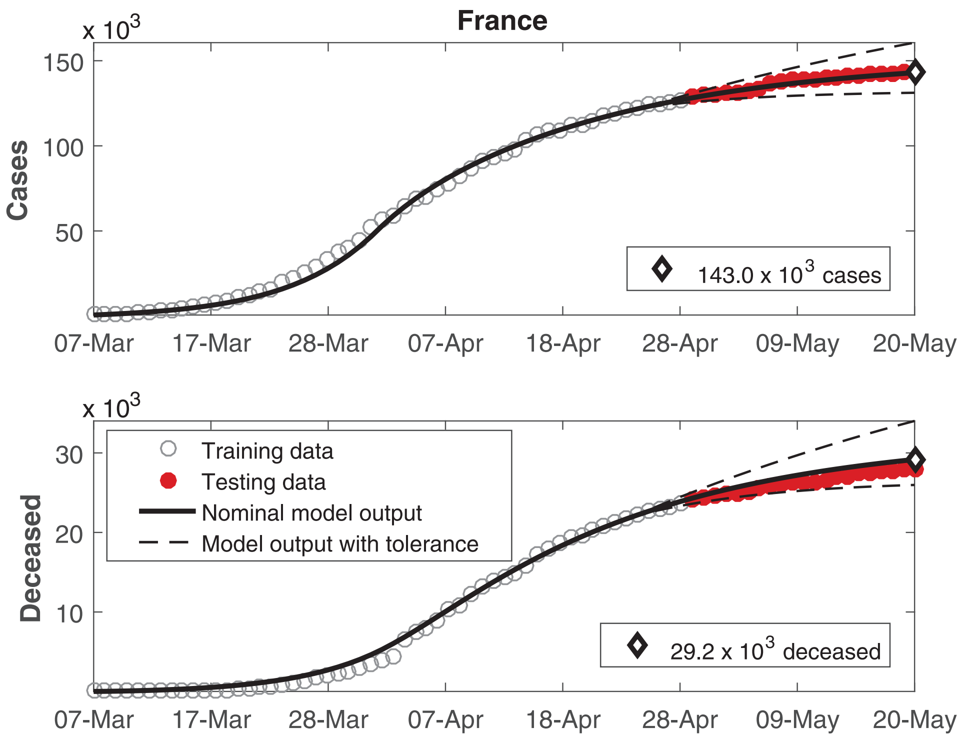

Fig 5. Cumulative number of infected individuals *C* and number of deceased individuals *D* for France.

https://doi.org/10.1371/journal.pone.0236386.g005

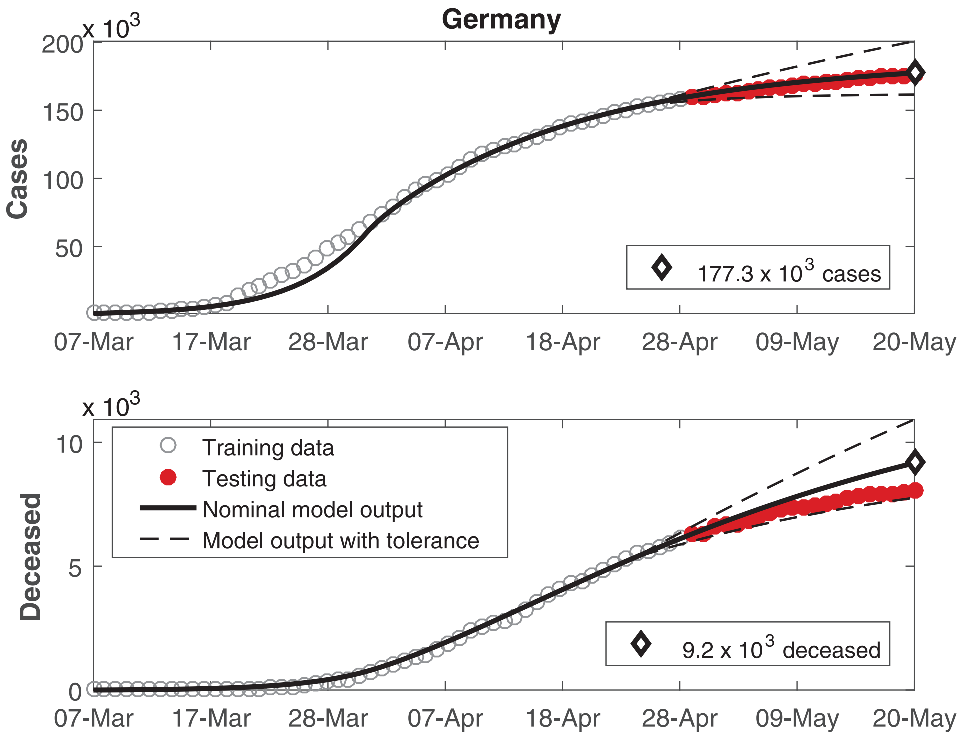

Fig 6. Cumulative number of infected individuals *C* and number of deceased individuals *D* for Germany.

https://doi.org/10.1371/journal.pone.0236386.g006

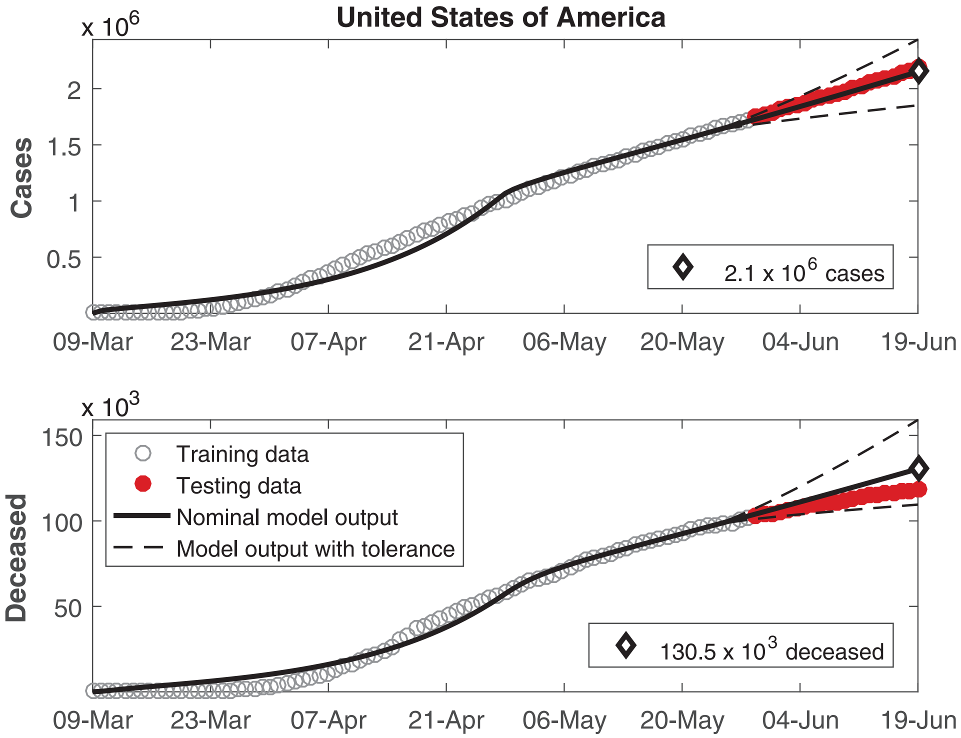

Fig 7. Cumulative number of infected individuals *C* and number of deceased individuals *D* for the United States of America.

https://doi.org/10.1371/journal.pone.0236386.g007

In these figures, data from the last 21 days are used as testing data, in order to allow a comparison between the model predictions and the real outcome of the disease. In these cases, the black continuous lines represent the predictions with the nominal parameters from Table 3 and the dashed lines are variations when selected parameters are changed by ±30%.

4 Discussion

In Figs 2 to 7, it can be seen that there is a good match between the model output and the observed data. Furthermore, the model presents accurate forecasts of the disease progression. There are some variations between the nominal prediction and the observed data, but within the established tolerances. The uncertainty in the prediction increases as more time has passed since the date when the model was calibrated; these uncertainties are reduced if the model parameters are updated periodically.

Although an overall good match was obtained for China (Fig 2), a mismatch can be seen on the four days before 12-feb. This difference is ascribed to the fact that the country changed the notification methodology on 12-feb [26], reporting more than 15000 cases on a single day, which caused a discontinuity in the cumulative cases on this date. The model cannot reproduce such discontinuity, for it was not designed to consider changes in the nature of observed data.

In general, the parameters reported in Table 3 are consistent (in terms of magnitude order) with the ones described previously in [21] for the MERS outbreak in South Korea. It is important to remark that the values of the parameters are expected to be different among the various countries, as the contagion, hospitalization and death rates are affected by social habits, demographic conditions, governmental action, quality and coverage of the health systems, etc.

The value of $\mathcal{R}_0$ is of particular interest. It is expected that $\mathcal{R}_0 > 1$ in the first phase during the spread and $\mathcal{R}_0 < 1$ after the authorities take measures to control the spread. A reduction in ℓ_a combined with a reduction in β shows where the focus to reduce contagion is most effective: reducing social contact at all cost so that both symptomatic but also importantly asymptomatic individuals stop transmitting the virus.

The estimated parameters are specially trustworthy for countries that already reached the asymptotic behavior of the curves. For countries where the epidemic is still ongoing, it is expected that the parameter estimates are more sensitive to new data.

According to our estimations, five countries have reached a second phase with $\mathcal{R}_0 < 1$, meaning that the epidemic is under control and converging to an equilibrium. On the other hand, the USA have also reached a second stage with a lower $\mathcal{R}_0$, but still higher than 1, indicating that the number of infections is decelerating but is not under control yet.

A δ_I value higher than δ_H was to be expected, indicating a lower rate of mortality in hospitalized patients. However, that was observed only in the USA. In the other countries in this study, the value of δ_H is higher or very similar to the value of δ_I. In these cases, this behavior might be justified by the most critical cases being treated in the hospitals, therefore being associated to a higher mortality rate. It is interesting to remark that the predictions of the number of deceased in Germany in Fig 6 and the USA in Fig 7 presented a larger mismatch both as compared to the other countries as well as compared to the number of cases within Germany and the USA themselves. Just as social distancing, more careful hygiene and other such measures can reduce the transmissibility, it might be expected that the development of protocols for treatments using the knowledge developed by medical experts might reduce the death ratios beyond model predictions. In this case, the parameters related to the deceases might demand another phase for estimation.

The relatively high values of ℓ and ℓ_a indicate that the contagion rate is higher with hospitalized patients and asymptomatic individuals. The higher transmissibility in the hospitals was to be expected, whereas the higher contagion rate of asymptomatic individuals may be ascribed to them not taking the same precautions as symptomatic people to avoid spreading the disease. The value of β should not be analyzed alone, for it provides a better insight when combined with the values of ℓ and ℓ_a.

In all countries, it can be seen that β indeed decreased in the second phase. However, for some cases ℓ_a remained constant. One conjecture for such results lies in the lack of data for estimating the parameters: the used database from ECDC does not include asymptomatic individuals. In some countries where testing on a larger scale took place, one could try to use such data to more accurately estimate the parameters associated with the asymptomatic class. However, this data is bound to be less trustworthy that those for infected symptomatic individuals and deceased, for which the data are collected daily and one can say that the reports are very close to the reality. On the other hand, the asymptomatic will remain an estimate as long as a representative part of the population is not tested. Moreover, this would have to be repeated at regular intervals. Such repeated large scale testing is impractical in reality, therefore the lack of data covering asymptomatic infected individuals with the same regularity and within tight confidence intervals as those for symptomatic and deceased.

Reducing the values of the parameters β, ℓ_a, and ℓ are the focus of the authorities in controlling the spread, therefore allowing them to change throughout the estimation is a means to address changes promoted by governmental action. This is a point where the proposed model and estimation are useful for prediction. The model enables not only prediction with the current estimates of the parameters, but also simulations with changes in these parameters. This permits decision makers to assess the impact of changes in terms of the rate of social contact avoidance that is necessary to enforce in a country as means to control the spread. Lately, many ways to measure the social contact reduction have been employed, such as using data from movements of cell phones.

5 Conclusion

This paper proposed a SEIR model to describe the COVID-19 epidemics. Its parameters were estimated with a numerical optimization algorithm, based on data from the European Centre for Disease Prevention and Control. An analysis was presented for the six countries where the pandemic is widely spread: China, Italy, Spain, France, Germany, and the USA. A good match between theoretical and observed data was achieved and a forecast was presented.

Since the model is data-driven, one drawback is that it cannot be used for countries where the epidemic is at a very early stage. Nevertheless, in the lack of local data, the pandemic behavior can still be estimated using parameters from another country. However, in this case, the uncertainty in the estimations would be higher.

It should be pointed out that forecasting the future is always imprecise, and that predictions are better for the near future. The outcome of the epidemic might be significantly altered by changes in governmental policy, such as enforcing or releasing measures to reduce social contact, or by other factors such as overload of the health care systems. Therefore, the model parameters should be updated as soon as new data become available. Although this paper presented a three-week ahead forecast, the authors recommend to update such parameters at least on a weekly basis.

Reducing the basic reproduction number $\mathcal{R}_0$ below one is known to represent that the spread is under control. Our developed formula for $\mathcal{R}_0$ shows a strong influence on the human-to-human transmission rate β, followed by the relative transmissibility of hospitalized patients ℓ and of the asymptomatic infected ℓ_a. By evaluating the estimated values authorities can grasp which actions are more likely to yield more meaningful results. For example, considering that for a country the value of β has stalled, whereas ℓ_a is still relatively high, measures to reduce the social contact of asymptomatic individuals appear as promising alternatives, instead of simply isolating the symptomatic individuals.

The forecasts are useful for planning purposes. For instance, when authorities must decide whether to invest in augmenting the capacity of hospitals in the coming weeks, it is of paramount importance to have a precise forecast of the number of hospitalized individuals. Besides enabling that, our proposed model involves parameters that have straightforward meaning related to the spread of the disease. Therefore, not only a nominal case can be considered, but the parameters may be varied within reasonable bounds under the scrutiny of specialists to yield worst and best case predictions, enhancing the awareness level of authorities in the process of decision making.

Future work includes defining a fixed target date and value for the maximal number of cumulative infected individuals and then optimize the parameters β, ℓ_a, and ℓ so as to determine whether authorities should aim at more or less stringent social contact control measures, for instance. Another opportunity for future enhancement would be to allow for variations in δ_H to model the overload of the hospitals. This, however, requires available data of the number of hospitalized individuals and available infrastructure.

Acknowledgments

The authors acknowledge the European Centre for Disease Prevention and Control (ECDC) for making the COVID-19 data publicly available and for allowing its use for research purposes.

Author Contributions

Conceptualization: Henrique Mohallem Paiva, Rubens Junqueira Magalhães Afonso.

Data curation: Henrique Mohallem Paiva, Rubens Junqueira Magalhães Afonso, Igor Luppi de Oliveira, Gabriele Fernandes Garcia.

Formal analysis: Henrique Mohallem Paiva, Rubens Junqueira Magalhães Afonso, Igor Luppi de Oliveira, Gabriele Fernandes Garcia.

Investigation: Igor Luppi de Oliveira, Gabriele Fernandes Garcia.

Methodology: Henrique Mohallem Paiva, Rubens Junqueira Magalhães Afonso.

Project administration: Henrique Mohallem Paiva.

Supervision: Henrique Mohallem Paiva.

Validation: Henrique Mohallem Paiva, Rubens Junqueira Magalhães Afonso.

Visualization: Henrique Mohallem Paiva, Rubens Junqueira Magalhães Afonso.

Writing – original draft: Henrique Mohallem Paiva, Rubens Junqueira Magalhães Afonso, Igor Luppi de Oliveira, Gabriele Fernandes Garcia.

Writing – review & editing: Henrique Mohallem Paiva, Rubens Junqueira Magalhães Afonso, Igor Luppi de Oliveira, Gabriele Fernandes Garcia.

References

1. Lai CC, Shih TP, Ko WC, Tang HJ, Hsueh PR. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges. International Journal of Antimicrobial Agents. 2020; 55(3): 105924. https://doi.org/10.1016/j.ijantimicag.2020.105924 PMID: 32081636
2. Linton NM, Kobayashi T, Yang Y, Hayashi K, Akhmetzhanov AR, Jung S-M, et al. Incubation Period and Other Epidemiological Characteristics of 2019 Novel Coronavirus Infections with Right Truncation: A Statistical Analysis of Publicly Available Case Data. Journal of Clinical Medicine. 2020; 9(2):538. https://doi.org/10.3390/jcm9020538
3. Shim E, Tariq A, Choi W, Lee Y, Chowell G. Transmission potential and severity of COVID-19 in South Korea. International Journal of Infectious Diseases. 2020. Available from: https://doi.org/10.1016/j.ijid.2020.03.031
4. Anastassopoulou C, Russo L, Tsakris A, Siettos C. Data-based analysis, modelling and forecasting of the COVID-19 outbreak. PLoS ONE. 2020 15(3): e0230405. https://doi.org/10.1371/journal.pone.0230405 PMID: 32231374
5. Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, et al. Clinical Characteristics of 138 Hospitalized Patients with 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China. JAMA—Journal of the American Medical Association. 2020, 323(11): 1061–1069. https://doi.org/10.1001/jama.2020.1585
6. Xu Z, Shi L, Wang Y, Zhang J, Huang L, Zhang C, et al Pathological findings of COVID-19 associated with acute respiratory distress syndrome. The Lancet Respiratory Medicine, 2020; 2600(20): 19–21.
7. Guan WJ, Ni ZY, Hu Y, Liang WH, Ou CQ, He JX, Du B. Clinical characteristics of coronavirus disease 2019 in China. New England Journal of Medicine. 2020; 382(18):1708–1720. https://doi.org/10.1056/NEJMoa2002032 PMID: 32109013
8. World Health Organization. Coronavirus disease 2019 (COVID-19): situation report, 82. 2020. Available from: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200411-sitrep-82-covid-19.pdf?sfvrsn=74a5d15_2 [Accessed 11 April 2020].
9. Guarner J. Three Emerging Coronaviruses in Two Decades: The Story of SARS, MERS, and Now COVID-19. American Journal of Clinical Pathology. 2020; 153(4): 420–421. https://doi.org/10.1093/ajcp/aqaa029
10. Mahase E. Coronavirus covid-19 has killed more people than SARS and MERS combined, despite lower case fatality rate. BMJ (Clinical Research Ed.). 2020 Feb; 368:m641.

11. Chatterjee K, Chatterjee K, Kumar A, Shankar S. Healthcare impact of COVID-19 epidemic in India: A stochastic mathematical model. Medical Journal Armed Forces India. 2020. Available from: https://doi.org/10.1016/j.mjafi.2020.03.022.

12. Khan T, Ullah Z, Ali N, Zaman G. Modeling and control of the hepatitis B virus spreading using an epidemic model. Chaos, Solitons and Fractals. 2019; 124: 1–9. https://doi.org/10.1016/j.chaos.2019.04.033

13. Fanelli D, Piazza F. Analysis and forecast of COVID-19 spreading in China, Italy and France. Chaos, Solitons and Fractals. 2020, 134, 109761. https://doi.org/10.1016/j.chaos.2020.109761

14. Fang Y, Nie Y, Penny M. Transmission dynamics of the COVID-19 outbreak and effectiveness of government interventions: A data-driven analysis. Journal of Medical Virology. 2020; 92(6):645–659. https://doi.org/10.1002/jmv.25750 PMID: 32141624

15. Lin Q, Zhao S, Gao D, Lou Y, Yang S, Musa SS, et al. A conceptual model for the coronavirus disease 2019 (COVID-19) outbreak in Wuhan, China with individual reaction and governmental action. International Journal of Infectious Diseases. 2020, 93: 211–216. https://doi.org/10.1016/j.ijid.2020.02.058 PMID: 32145465

16. Peng L, Yang W, Zhang D, Zhuge C, Hong L. Epidemic analysis of COVID-19 in China by dynamical modeling. arXiv. 2020; 1–18. Available from: http://arxiv.org/abs/2002.06563

17. Jardon-Kojakhmetov H, Kuehn C, Pugliese A, Sensi M. A geometric analysis of the SIR, SIRS and SIRWS epidemiological models. arXiv. 2020; 1–29. Available from: http://arxiv.org/abs/2002.00354

18. Petropoulos F, Makridakis S. Forecasting the novel coronavirus COVID-19. PLoS One, 2020; 15(3): e0231236. https://doi.org/10.1371/journal.pone.0231236 PMID: 32231392

19. Chen X, Yu B. First two months of the 2019 Coronavirus Disease (COVID-19) epidemic in China: real-time surveillance and evaluation with a second derivative model. Global Health Research and Policy. 2020, 5(1): 1–9.

20. Zhao S, Chen H Modeling the epidemic dynamics and control of COVID-19 outbreak in China. Quantitative Biology, 2020; 8(1): 11–19. https://doi.org/10.1007/s40484-020-0199-0

21. Kim Y, Lee S, Chu C, Choe S, Hong S, Shin Y. The characteristics of Middle Eastern respiratory syndrome coronavirus transmission dynamics in South Korea. Osong public health and research perspectives. 2016 Feb; 7(1):49–55. https://doi.org/10.1016/j.phrp.2016.01.001 PMID: 26981343

22. Diekmann O, Heesterbeek JAP, Metz JA. On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations. Journal of mathematical biology. 1990 Jun; 28(4):365–382. https://doi.org/10.1007/BF00178324 PMID: 2117040

23. Van den Driessche P, Watmough J. Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. Mathematical biosciences. 2002 Nov-Dec; 180(1-2):29–48. https://doi.org/10.1016/S0025-5564(02)00108-6 PMID: 12387915

24. Gill PE, Wong E. Sequential quadratic programming methods. Society for Industrial and Applied Mathematics. 2012. In: Mixed integer nonlinear programming (pp. 147–224). Springer, New York, NY.

25. European Centre for Disease Prevention and Control (ECDC). Available from: https://www.ecdc.europa.eu/en [Accessed 19 June 2020].

26. Tan W, Ellyatt H China confirms 15,152 new coronavirus cases, 254 additional deaths. Consumer News and Business Channel (CNBC). 2020. Available from: https://www.cnbc.com/2020/02/13/coronavirus-latest-updates-china-hubei.html [Accessed 11 April 2020].