PLOS ONE

RESEARCH ARTICLE

# Tracing day-zero and forecasting the COVID-19 outbreak in Lombardy, Italy: A compartmental modelling and numerical optimization approach

**Lucia Russo[1], Cleo Anastassopoulou[2], Athanasios Tsakris[2], Gennaro Nicola Bifulco[3], Emilio Fortunato Campana[4], Gerardo Toraldo[5], Constantinos Siettos[6]***

**1** Consiglio Nazionale delle Ricerche, Istituto delle Scienze e delle Tecnologie per l'Energia e la Mobilità Sostenibile, Napoli, Italy, **2** Department of Microbiology, Medical School, University of Athens, Athens, Greece, **3** Dipartimento di Ingegneria Civile, Edile e Ambientale, Università degli Studi di Napoli Federico II, Napoli, Italy, **4** Consiglio Nazionale delle Ricerche, Dipartimento di Ingegneria, ICT e Tecnologie per l'Energia e i Trasporti, Roma, Italy, **5** Dipartimento di Matematica e Fisica, Università degli Studi della Campania Luigi Vanvitelli, Caserta, Italy, **6** Dipartimento di Matematica e Applicazioni "Renato Caccioppoli", Università degli Studi di Napoli Federico II, Napoli, Italy

* constantinos.siettos@unina.it





**Citation:** Russo L, Anastassopoulou C, Tsakris A, Bifulco GN, Campana EF, Toraldo G, et al. (2020) Tracing day-zero and forecasting the COVID-19 outbreak in Lombardy, Italy: A compartmental modelling and numerical optimization approach. PLoS ONE 15(10): e0240649. https://doi.org/10.1371/journal.pone.0240649

**Editor:** Alberto d'Onofrio, International Prevention Research Institute, FRANCE

**Received:** May 14, 2020

**Accepted:** September 30, 2020

**Published:** October 30, 2020



**Data Availability Statement:** All relevant data are within the manuscript and its Supporting Information files.

**Funding:** We did not receive any specic funding for this study.

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

## Abstract

### Introduction

Italy became the second epicenter of the novel coronavirus disease 2019 (COVID-19) pandemic after China, surpassing by far China's death toll. The disease swept through Lombardy, which remained in lockdown for about two months, starting from the 8th of March. As of that day, the isolation measures taken in Lombardy were extended to the entire country. Here, assuming that effectively there was one case "zero" that introduced the virus to the region, we provide estimates for: (a) the day-zero of the outbreak in Lombardy, Italy; (b) the actual number of asymptomatic infected cases in the total population until March 8; (c) the basic ($R_0$)and the effective reproduction number ($R_e$) based on the estimation of the actual number of infected cases. To demonstrate the efficiency of the model and approach, we also provide a tentative forecast two months ahead of time, i.e. until May 4, the date on which relaxation of the measures commenced, on the basis of the COVID-19 Community Mobility Reports released by Google on March 29.

### Methods

To deal with the uncertainty in the number of the actual asymptomatic infected cases in the total population Volpert et al. (2020), we address a modified compartmental Susceptible/ Exposed/ Infectious Asymptomatic/ Infected Symptomatic/ Recovered/ Dead (SEIIRD) model with two compartments of infectious persons: one modelling the cases in the population that are asymptomatic or experience very mild symptoms and another modelling the infected cases with mild to severe symptoms. The parameters of the model corresponding to the recovery period, the time from the onset of symptoms to death and the time from

exposure to the time that an individual starts to be infectious, have been set as reported from clinical studies on COVID-19. For the estimation of the day-zero of the outbreak in Lombardy, as well as of the "effective" per-day transmission rate for which no clinical data are available, we have used the proposed SEIIRD simulator to fit the numbers of new daily cases from February 21 to the 8th of March. This was accomplished by solving a mixed-integer optimization problem. Based on the computed parameters, we also provide an estimation of the basic reproduction number $R_0$ and the evolution of the effective reproduction number $R_e$. To examine the efficiency of the model and approach, we ran the simulator to "forecast" the epidemic two months ahead of time, i.e. from March 8 to May 4. For this purpose, we considered the reduction in mobility in Lombardy as released on March 29 by Google COVID-19 Community Mobility Reports, and the effects of social distancing and of the very strict measures taken by the government on March 20 and March 21, 2020.

## Results

Based on the proposed methodological procedure, we estimated that the expected day-zero was January 14 (min-max rage: January 5 to January 23, interquartile range: January 11 to January 18). The actual cumulative number of asymptomatic infected cases in the total population in Lombardy on March 8 was of the order of 15 times the confirmed cumulative number of infected cases, while the expected value of the basic reproduction number $R_0$ was found to be 4.53 (min-max range: 4.40- 4.65). On May 4, the date on which relaxation of the measures commenced the effective reproduction number was found to be 0.987 (interquartiles: 0.857, 1.133). The model approximated adequately two months ahead of time the evolution of reported cases of infected until May 4, the day on which the phase I of the relaxation of measures was implemented over all of Italy. Furthermore the model predicted that until May 4, around 20% of the population in Lombardy has recovered (interquartile range: $\sim$10% to $\sim$30%).

## Introduction

The butterfly effect in chaos theory underscores the sensitive dependence on initial conditions, highlighting the importance of even a small change in the initial state of a nonlinear system. The emergence of a novel coronavirus, SARS-CoV-2, that caused a viral pneumonia outbreak in Wuhan, Hubei province, China in early December 2019 has evolved into the COVID-19 acute respiratory disease pandemic due to its alarming levels of spread and severity, with more than 3.5 million cases and 250,000 deaths globally, as of May 7, 2020 ([1, 2]). The seemingly far from the epicenter, old continent became the second-most impacted region after Asia Pacific, mostly as a result of a dramatic divergence of the epidemic trajectory in Italy first, where there have been 214,457 total confirmed infected cases and 29,684 deaths, and then in Spain where there have been 220,325 total confirmed infected cases and 25,857 deaths, as of May 7, 2020 ([1, 2]).

The second largest outbreak outside of mainland China officially started on January 31, 2020, after two Chinese visitors staying at a central hotel in Rome tested positive for SARS-CoV-2; the couple remained in isolation and was declared recovered on February 26 [3]. A 38-year-old man repatriated back to Italy from Wuhan who was admitted to the hospital in

Codogno, Lombardy on February 21 was the first secondary infection case ("patient 1"). "Patient 0" was never identified by tracing the first Italian citizen's movements and contacts. In less than a week, the explosive increase in the number of cases in several bordering regions and in the two autonomous provinces of Trento and Bolzano (the northerner in Italy) placed enormous strain on the decentralized health system. Following a dramatic spike in deaths from COVID-19, Italy transformed into a "red zone", and the movement restrictions were expanded to the entire country on the 8th of March. All public gatherings were cancelled and school and university closures were extended through at least the next month.

In an attempt to assess the dynamics of the outbreak for forecasting purposes, it is important to estimate epidemiological parameters that cannot be computed directly based on clinical data, such as the transmission rate (or as otherwise called "effective contact rate" of the disease and the basic reproduction number, $R_0$. The transmission rate is defined as the product of the probability of transmitting the virus given a contact between a susceptible and an infected individual and the average rate of contacts between susceptibles and infected. $R_0$ is defined as the expected number of exposed cases generated by one infected case in a population where all individuals are susceptible [4].

Since the first confirmed COVID-19 case many mathematical modelling studies have already appeared. The first models mainly focused on the estimation of the basic reproduction number $R_0$ using dynamic mechanistic mathematical models ([5–8]), but also simple exponential growth models (see e.g. [9, 10]). Compartmental epidemiological models like SIR, SIRD, SEIR and SEIRD have been proposed to estimate other important epidemiological parameters, such as the transmission rate and for forecasting purposes (see e.g. [8, 11]). Other studies have used metapopulation models, which include data of human mobility between cities and/or regions to forecast the evolution of the outbreak in other regions/countries far from the original epicenter in China [5, 7, 12, 13], including the modelling of the influence of travel restrictions and other control measures in reducing the spread [14].

Among the perplexing problems that mathematical models face when they are used to estimate epidemiological parameters and to forecast the evolution of the outbreak, two stand out: (a) the uncertainty regarding the day-zero of the outbreak, the knowledge of which is crucial to assess the stage and dynamics of the epidemic, especially during the first growth period, and (b) the uncertainty that characterizes the actual number of the asymptomatic infected cases in the total population (see e.g. [15, 16]).

At this point we should note that what is done until now with dynamical epidemiological models is the investigation of several scenarios including different "days-zero" or just fixing the day-zero and run different levels of asymptomatic cases e.t.c. To cope with the above problems, we herein address a methodological framework that provides estimates for the day-zero of the outbreak and the number of the asymptomatic cases in the total population in a systematic way. Towards this goal, and for our demonstrations, we address a conceptually simple SEIRD model with a total of five compartments, with one of them modelling the asymptomatic infected cases in the population and another one modelling the part of the infected cases that will experience mild to severe symptoms, a significant share of which will be hospitalized, admitted to intensive care units (ICUs) or die from the disease. The proposed approach is applied to Lombardy, the epicenter of the outbreak in Italy. Furthermore, we provide a two-months ahead of time forecast from March 8 (the day of lockdown of all Italy) to May 4 (the first day of the relaxation of the strict isolation measures). The above tasks were accomplished by the numerical solution of a mixed-integer optimization problem using the publicly available data of daily new cases for the period February 21-March 8, and the COVID-19 Community Mobility Reports released by Google on March 29.

# Methodology

## The modelling approach

We address a compartmental SEIIRD model that includes two categories of infected cases, namely the asymptomatic (unknown) cases in the total population and the cases that develop mild to more severe symptoms, a significant share of which are hospitalized, admitted to ICUs and a part of them dies. In agreement with other studies and observations, our modelling hypothesis is that the confirmed cases of infected are only a (small) subset of the actual number of asymptomatic infected cases in the total population [7, 8, 16]. Regarding the confirmed cases of infected as of February 21, a study conducted by the Chinese CDC which was based on a total of 72,314 cases in China, found that about 80.9% of the cases were mild and could recover at home, 13.8% severe and 4.7% critical [17].

On the basis of the above findings, in our modelling approach, it is assumed that the asymptomatic or very mildly symptomatic cases recover from the disease relatively soon and without medical care, while for the other category of infected, on average their recovery lasts longer than the non-confirmed, they may also be hospitalized, admitted to ICUs or die from the disease.

Based on the above, let us consider a well-mixed population of size $N$. The state of the system at time $t$, is described by (see also Fig 1 for a schematic) $S(t)$ representing the number of susceptible persons, $E(t)$ the number of exposed, $I(t)$ the number of asymptomatic infected persons in the total population who experience very mild or no symptoms and recover relatively soon without any other complications, $I_c(t)$ the number of infected cases who may develop mild to more severe symptoms and a significant part of them is hospitalized, admitted to ICUs or dies, $R(t)$ the number of asymptomatic cases in the total population that recover, $R_c(t)$ the number of the recovered cases that come from the compartment of $I_c$ and $D(t)$ the reported number of deaths. For our analysis, and for such a short period, we assume that the total number of the population remains constant. Based on demographic data, the total population of Lombardy is $N = 10m$; its surface area is 23,863.09 $kmq$ and the population density is $\sim$422 (Inhabitants/Kmq).

The rate at which a susceptible ($S$) becomes exposed ($E$) to the virus is proportional to the density of infectious persons $I$. The proportionality constant is the "effective" disease transmission rate, say $\beta = \bar{c}p$, where $\bar{c}$ is the average number of contacts per day and $p$ is the probability of infection upon a contact between a susceptible and an infected. Our main assumption here is that only a fraction, say $\epsilon$ of the actual number of exposed cases $E$ will experience mild to

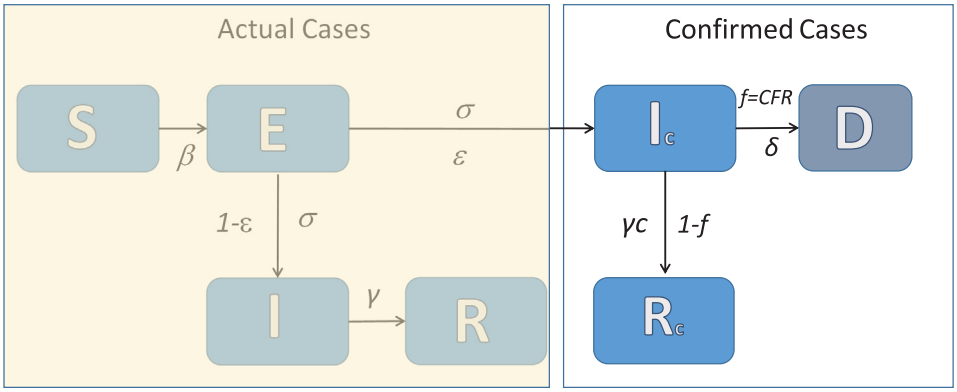


**Fig 1. A schematic of the proposed compartmental SEIIRD model.** The actual number of cases is unknown.

https://doi.org/10.1371/journal.pone.0240649.g001

more severe symptoms denoted as $I_c(t)$ and a significant part of them will be hospitalized, admitted to ICUs or die. Thus, we assume that the infected persons that belong to the compartment $I_c$ go into quarantine at home or they are hospitalized, and, thus, it is assumed that for any practical means they don't transmit further the disease. Here, it should be noted that a wide testing policy may also result in the identification of asymptomatic cases belonging to the compartment $I$ that would then be assigned to compartment $I_c$. However, as a generally reported rule in Italy, tests were conducted only for those who presented for treatment with symptoms like fever and coughing. Thus, people who did not seek medical attention were tested very scarcely [18–20]. Thus, for any practical means the compartment $I_c$ reflects the reported confirmed infected cases.

A fraction of the $I_c$ cases that is given by the fatality ratio $f = D(t)/(I_c(t) + R_c(t) + D(t))$ dies with a mortality rate $\delta$ the inverse of which is the average time from the onset of symptoms to death, while the remaining part $((1 - f))$ of the $I_c$ compartment recovers with a rate $\gamma_c$, the inverse of which corresponds to the average time from the onset of symptoms to full recovery.

We note that while more compartments could conceptually be included, we aimed at keeping a low level of complexity in order to avoid the introduction of more parameters, and thus a model that would suffer from the "curse of dimensionality".

At this point we should note that on March 8, the date of the general lockdown, the number of confirmed infected cases was 3,372, the number of cases in ICUs was 399 and the number of hospitalized persons was 2,616 [21]. That is, until March 8, the number of confirmed cases was approximately equal to the number of hospitalized cases and the cases that were admitted to ICUs. Therefore, until March 8, any difference between the asymptomatic cases as represented by our model by the compartment $I$ and the compartment $I_c$ would approximately reflect a level of under-reporting of the actual asymptomatic cases in the total population. We should also note that in the available data of reported cases [21] there is no distinction between cases that recovered at home and those that recovered at and were dismissed from hospitals. Thus, in the absence of such information, if one were to consider as a separate category the cases that are hospitalized, an extra parameter would have to be introduced (the fraction of recovered cases dismissed from hospitals). On one hand, such a piece of information is not available, and, on the other, such an attempt would add an extra degree of freedom that would need calibration or to be fixed at a certain value; however, due to the small size of the data and the "curse of dimensionality", this would also introduce unnecessary computational burden and further modelling uncertainty.

Thus, our discrete mean field compartmental SEIIRD model reads:

$$S(t) = S(t-1) - \frac{\beta}{N - D(t-1) - R_c(t-1) - I_c(t-1)} S(t-1)I(t-1) \tag{1}$$

$$E(t) = E(t-1) + \frac{\beta}{N - D(t-1) - R_c(t-1) - I_c(t-1)} S(t-1)I(t-1) - \sigma E(t-1) \tag{2}$$

$$I(t) = I(t-1) + (1-\epsilon)\sigma E(t-1) - \gamma I(t-1) \tag{3}$$

$$I_c(t) = I_c(t-1) + \epsilon\sigma E(t-1) - (1-f)\gamma_c I_c(t-1) - f\delta I_c(t-1) \tag{4}$$

$$R(t) = R(t-1) + \gamma I(t-1) \tag{5}$$

$$R_c(t) = R_c(t-1) + (1-f)\gamma_c I_c(t-1) \quad (6)$$

$$D(t) = D(t-1) + f\delta I_c(t-1) \quad (7)$$

The above system is defined in discrete time points $t = 1, 2, \ldots$, with the corresponding initial condition at the very start of the outbreak (day-zero): $S(0) = N - 1$, $I(0) = 1$, $E(0) = 0$ $I_c(0) = 0$, $R(0) = 0$, $R_c(0) = 0$, $D(0) = 0$. The the term $f\delta I_c(t-1)$ in Eq 4 represents the fraction $f$ of the $I_c$ cases that dies with a mortality rate $\gamma$ and the term $(1-f)\gamma_c I_c(t-1)$ represents the complementary part $((1-f))$ of the $I_c$ cases that recovers with a rate $\gamma_c$.

The parameters of the model are:

- $\beta(d^{-1})$ is the "effective" transmission rate of the disease,
- $\sigma(d^{-1})$ is the average per-day "effective" rate at which an exposed person becomes infectious,
- $\gamma(d^{-1})$ is the average per-day "effective" recovery rate within the group of asymptomatic cases in the total population,
- $\gamma_c(d^{-1})$ is the average per-day "effective" recovery rate within the subset of the $I_c$ infected cases that finally recover,
- $\delta(d^{-1})$ is the average per-day "effective" mortality rate within the subset of $I_c$ infected cases that finally die,
- $f$ is the probability that a $I_c$ case will die. Here, this, is given by the "emergent" case fatality ratio, computed as $f = D(t)/(I_c(t) + R_c(t) + D(t))$,
- $\epsilon(d^{-1})$ is the fraction of the actual (all) cases of exposed in the total population that enter to the compartment $I_c$.

Here, we should note the following: as new cases of recovered and dead at each time $t$ appear with a time delay (which is generally unknown but an estimate can be obtained by clinical studies) with respect to the corresponding infected cases, the above per-day rates are not the actual ones; thus, they are denoted as "effective/apparent" rates.

The values of the epidemiological parameters $\sigma$, $\gamma$, $\gamma_c$, $\delta$ that were fixed in the proposed model were chosen based on clinical studies.

In particular, in many studies that use SEIRD models, the parameter $\sigma$ is set equal to the inverse of the mean incubation period (time from exposure to the development of symptoms) of a virus. *However, the incubation period does not generally coincide with the time from exposure to the time that someone starts to be infectious*. Regarding COVID-19, it has been suggested that an exposed person can be infectious well before the development of symptoms [22]. With respect to the incubation period for SARS-CoV-2, a study in China [23] suggests that it may range from 2–14 days, with a median of 5.2 days. Another study in China, using data from 1,099 patients with laboratory-confirmed 2019-nCoV ARD from 552 hospitals in 31 provinces/provincial municipalities suggested that the median incubation period is 4 days (interquartile range: 2 to 7). In our model, as explained above, $1/\sigma$ represents the period from exposure to the onset of the contagious period. Thus, based on the above clinical studies, for our simulations, we have set $1/\sigma = 3$.

Regarding the recovery period, in a study that is based on 55,924 laboratory-confirmed cases, the WHO-China Joint Mission has reported a median time of 2 weeks from onset to clinical recovery for mild cases, and 3-6 weeks for severe or critical cases [24]. Based on the above, and on the fact that within the subset of confirmed cases the mild cases are the 81%

[17], we have set the recovery period for the confirmed cases' compartment to be $\delta_c = 1/21$ in order to balance the recovery period with the corresponding characterization of the cases (mild, severe/critical). The average recovery period of the unreported/non-confirmed part of the infected population, which in our assumptions experiences the disease like the flu or a common cold, is set equal to one week [25], i.e. we have set $\delta = 1/7$. This choice is based also on reports on the serial interval of COVID-19. The serial interval of COVID-19 is defined as the time duration between a primary case-patient (infector) having symptoms and a secondary case-patient having again symptoms. For example, it has been reported that the serial interval for COVID is estimated at 4.4–7.5 days [26]; for the case of Lombardy, the average serial number has been estimated to be 6.6 days [27]. In our model, the $1/\sigma = 3$ period refers to the period from exposure to the onset of the contagiousness. In this period, obviously there are no symptoms. Thus, the serial interval in our model is 7 days (this is the average number of days in which an infectious becomes recovered and no longer transmits the disease). Importantly, there are studies (see e.g. Nushiura et al. [28]) suggesting that a substantial proportion of secondary transmission may occur prior to illness onset. Thus, the 7 days period that we have taken as the average period that an infectious person can transmit the disease before he/she recovers, reflects exactly this period; it refers to the serial interval for the cases that are asymptomatic and for cases with mild symptoms.

Finally, the median time from the onset of symptoms until death for Italy has been reported to be eight days [29], thus in our model we have set $\gamma = 1/8$.

We have set $f = 11\%$ for the optimization. For the forecasting (i.e. for the period March 9 to May 4), the value of $f$ was not fixed but it was computed dynamically each day $t$ through the model simulations as $f = D(t)/(I_c(t) + R_c(t) + D(t))$.

The transmission rate $\beta$, as it cannot be obtained in general by clinical studies, but only by mathematical models, was estimated through the optimization process.

Regarding day-zero in Lombardy, that is also unknown and estimated by the optimization process, what has been officially reported is just the date on which the first infected person was confirmed to be positive for SARS-CoV-2. That day was February 21, 2020, which is the starting date of public data release of confirmed cases.

## Estimation of the day-zero of the outbreak, the scale of data uncertainty and the disease transmission

The day-zero of the outbreak, the per-day "effective" transmission rate $\beta$, and the ratio $\epsilon$ were computed by the numerical solution of a mixed-integer optimization problem with the aid of genetic algorithms to fit the reported data of daily new cases (see the discussion in [30]) from February 21 to March 8, the day of the lockdown of Lombardy.

As already mentioned, on March 8, the number of confirmed infected cases in the population was 3,372, the number of cases in ICUs was 399 and the number of hospitalized persons was 2,616 [21]. That is, until March 8 the number of mild and severe cases that were hospitalized and admitted to ICUs was approximately equal to the number of confirmed infected cases. Thus, for the period of calibration, it is reasonable to assume that for any practical means the number of confirmed cases was approximately the same as for those that experienced mild to more severe symptoms and were admitted for medical care. Thus, until March 8, the parameter $\epsilon$ reflects also the level of under-reporting of the asymptomatic cases in the total population.

Here, for our computations, we have used the genetic algorithm "ga" provided by the Global Optimization Toolbox of Matlab [6] to minimize the following objective function:

$$J(t_0,\beta,\epsilon)=\underset{t_0,\beta,\epsilon}{\operatorname{argmin}}\{\sum_{t=February21}^{March8}(w_1u_t(t_0,\beta,\epsilon|\gamma,\gamma_c,\delta))^2+(w_2g_t(t_0,\beta,\epsilon|\delta,\gamma_c,\delta))^2+(w_3h_t(t_0,\beta,\epsilon|\gamma,\gamma_c,\delta))^2\}, \tag{8}$$

where,

$$\begin{aligned} u_t(t_0,\beta,\epsilon|\delta,\delta_c,\gamma)&=\Delta I_c^{SEIRD}(t)-\Delta I_c(t),\\ g_t(t_0,\beta,\epsilon|\gamma,\gamma_c,\delta)&=\Delta R^{SEIRD}(t)-\Delta R(t),\\ h_t(t_0,\beta,\epsilon|\gamma,\gamma_c,\delta)&=\Delta D^{SEIRD}(t)-\Delta D(t) \end{aligned} \tag{9}$$

$\Delta X^{SEIRD}(t)$, $(X = I, R, D)$ are the new cases resulting from the SEIRD simulator at time $t$.

The weights $w_1$, $w_2$, $w_3$ correspond to scalars serving in the general case as weights to the relevant functions for balancing the different scales between the number of infected, recovered cases and deaths. The convergence tolerance was set to $1.E^{-06}$, the population size (distributed between the lower and upper bounds) was selected as 40 resulting from min(max (10*nvars,40),100) for mixed-integer problems (here nvars = 3), ceil(0.05*PopulationSize) individuals are guaranteed to survive to the next generation, the migration fraction was set to 0.2, while the number of generations that take place between migrations of individuals between subpopulations was set to 20 [6].

At this point we should note that the above optimization problem may in principle have multiple nearby optimal solutions (MNOS). Finding and assessing the information contained from MNOS (known also as niching) is a particular challenging problem [31]. Here, we created a grid of initial guesses within the intervals in which the optimal estimates were sought: for the day-zero ($t_0$) we used a step of 2 days within the interval December 27, 2019 until the 5th of February, 2020 i.e. ±20 days around the 16th of January, for $\beta$ we used a step of 0.05 within the interval (0.3, 0.9) and for $\epsilon$ we used a step of 0.02 within the interval (0.01, 0.29). The numerical optimization procedure was repeated 48 times for each combination of initial guesses. For our computations, we kept the best fitting outcome for each combination of initial guesses.

Next, in order to reveal structured patterns of distributions vs. uniformly random distributions, we fitted the resulting probability distributions of the optimal values using several functions, including the Normal, Log-normal, Weibull, Beta, Gamma, Burr [32], Exponential and Birnbaum-Saunders [33] distributions and kept the one resulting in the maximum Log-likelihood (see in the Supporting Information for more details). For the computed parameters of the corresponding best distributions, we also provide the corresponding 95% confidence intervals. Such fitting can demonstrate that the obtained values are not uniformly distributed.

Finally, we have run simulations based on all obtained values to assess the efficiency of the model and obtained results and the forecasting uncertainty until May 4.

For our computations, we used the parallel computing toolbox of Matlab 2020a [34] utilizing 6 INTEL XEON CPU X5650 cores at 2.66GHz.

## Estimation of the basic and effective reproduction numbers $R_0$, $R_e$ from the SEIIRD model

**Estimation of the basic reproduction number.** Here, we note that we provide an estimation of the basic reproduction number $R_0$ based on the estimation of the total number of (asymptomatic) infected cases in the population. Thus, it is expected that the estimated $R_0$ will be larger than the ones reported using just the confirmed number of cases; the latter may underestimate the actual $R0$.

Initially, when the spread of the epidemic starts, all the population is considered to be susceptible, i.e. $S \approx N$. On the basis of this assumption, we computed the basic reproduction number based on the estimates of the epidemiological parameters computed using the data from the 21st of February to the 8th of March with the aid of the SEIIRD model given by Eqs (1)–(7) as follows.

Note that there are three infected compartments, namely $E$, $I$, $I_c$ and two of them ($E$,$I$) determine the outbreak. Thus, considering the corresponding equations given by Eqs (2), (3) and (4), and that at the very first days of the epidemic $S \approx N$ and $D \approx 0$, the Jacobian of the system as evaluated at the disease-free state reads:

$$\boldsymbol{J} = \frac{\partial(E(t), I(t))}{\partial(E(t-1), I(t-1)} = \begin{bmatrix} 1-\sigma & \beta \\ (1-\epsilon)\sigma & 1-\gamma \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} -\sigma & \beta \\ (1-\epsilon)\sigma & -\gamma \end{bmatrix} \tag{10}$$

The eigenvalues (that is the roots of the characteristic polynomial of the Jacobian matrix) dictate if the disease-free equilibrium is stable or not, that is if an emerging infectious disease can spread in the population. In particular, the disease-free state is stable, meaning that an infectious disease will not result in an outbreak, if and only if all the norms of the eigenvalues of the Jacobian $\boldsymbol{J}$ of the discrete time system are bounded by one. Jury's stability criterion [35] (the analogue of Routh-Hurwitz criterion for discrete-time systems) can be used to determine the stability of the linearized discrete time system by analysis of the coefficients of its characteristic polynomial. The characteristic polynomial of the Jacobian matrix reads:

$$F(z) = a_2 z^2 + a_1 z + a_0, \tag{11}$$

where

$$\begin{aligned} a_2 &= 1 \\ a_1 &= \delta + \sigma - 2 \\ a_0 &= \delta\sigma - \sigma - \beta\sigma - \gamma + \beta\epsilon\sigma + 1. \end{aligned} \tag{12}$$

The necessary conditions for stability read:

$$F(1) > 0, \tag{13}$$

$$(-1)^2 F(-1) > 0. \tag{14}$$

The sufficient conditions for stability are given by the following two inequalities:

$$|a_0| < a_2. \tag{15}$$

The first inequality (13) results in the necessary condition

$$\frac{\beta(1-\epsilon)}{\gamma} < 1. \tag{16}$$

It can be shown that the second necessary condition (14) and the sufficient condition (15) are always satisfied for the range of values of the epidemiological parameters considered here.

Thus, the necessary condition (16) is also a sufficient condition for stability. Hence, the disease-free state is stable, if and only if, condition (16) is satisfied.

Note that in this necessary and sufficient condition (16), the fraction $(1-\epsilon)/\delta$ is the average infection time of the compartment $I$. Thus, the above expression reflects the basic reproduction number $R_0$ which is qualitatively defined by $R_0 = \beta \cdot 1/\textit{infection time}$. Hence, our model results in the following expression for the basic reproduction number:

$$R_0 = \frac{\beta(1-\epsilon)}{\gamma}. \tag{17}$$

Note that for $\epsilon = 0$, the above expression simplifies to $R_0$ for the simple SIR model.

**Estimation of effective reproduction number.** For the estimation of the effective reproduction number $R_e$, representing the average number of secondary infections from an infectious individual when in the population exist already non-susceptible individuals. For the calculation of $R_e$, we now use the next generation matrix approach [36].

The next generation matrix $G$ with elements = $g_{ij}$ is formed by the average number of secondary infections of type i from an infected individual of type j. Formally, it is constructed by:

$$\boldsymbol{G} = \nabla \boldsymbol{F} \cdot \nabla(\boldsymbol{V})^{-1}, \tag{18}$$

where $\boldsymbol{F}$ is the vector containing the transmission rates from the model, and $\mathbf{V}$ is the vector containing the transition rates between the infected compartments. In our model:

$$\boldsymbol{F} = \begin{bmatrix} \beta \frac{S(t-1)}{N-D(t-1)-R_c(t-1)-I_c(t-1)} I(t-1) \\ 0 \end{bmatrix}, \boldsymbol{V} = \begin{bmatrix} \sigma E(t-1) \\ -(1-\epsilon)\sigma E(t-1) + \gamma I(t-1) \end{bmatrix}. \tag{19}$$

The effective reproduction number $R_e$ is the spectral radius, i.e. the dominant eigenvalue of $G$. Thus,

$$R_e(t) = \frac{\beta(1-\epsilon)}{\gamma} \frac{S(t-1)}{N - D(t-1) - R_c(t-1) - I_c(t-1)}. \tag{20}$$

## Forecasting

As discussed, we used the proposed approach to forecast the evolution of the pandemic in Lombardy from March 8 to May 4, i.e. from the first day of lockdown to the first day of the relaxation of the social isolation.

Our estimation regarding the as of March 8 reduction of the "effective" transmission rate was based on the combined effects of prevention efforts and behavioral changes. In particular, our estimation was based on (a) the COVID-19 Community Mobility Reports released by Google on March 29 [37], and, (b) an assessment of the synergistic effects of such control measures as the implementation of preventive containment in workplaces, stringent "social distancing", and the ban on social gatherings, as well as the public awareness campaign prompting people to adopt cautious behaviors to reduce the risk of disease transmission (see

also [38–41]). The effect of the distribution of contacts at home, work, when travelling, and during leisure activities can be also assessed. For example, based on an analysis for the social contacts and mixing patterns relevant to the spread of infectious diseases that was conducted in various countries, it has been found that for Italy, around 20-23% of all physical social contacts during a day are attributed to workplaces, around 17-18% to schools, 2-3% to transportation, 20% to leisure activities, 15-18% to home, 15% to other activities (contacts made at locations other than home, work, school, travel, or leisure) and a 7-8% to contacts made at multiple other locations during the day, not just at a single location [42].

On the basis of the Google COVID-19 Community Mobility Report released on March 29 [37], the average reduction in the mobility in Lombardy during the period February 16-March 8, compared to the period before February 16, was $\sim$15% in retail & recreation activities, $\sim$20% in transit stations, $\sim$12% in workplaces, while it was increased in parks by $\sim$ 11% and was almost the same in groceries and pharmacies. In the period March 9 to March 20, the mobility was reduced by an average of $\sim$73% in retail & recreation activities, by $\sim$75% in transit stations, by $\sim$55% in workplaces, by $\sim$58% in parks and by $\sim$ 32% in groceries and pharmacies. Thus, taking into account the coarse effect of different activities in the physical contact [42], the average reduction was of the order of $\sim$40% in mobility when compared to the period February 21-March 8. In fact, on March 17, based on the release of mobile phone data, the vice-president of Lombardy, announced that the average mobility in the region (for distances more than 500 meters) had been reduced by a $\sim$50% with respect to the period before February 20 [43]. On March 20, the government announced the implementation of even stricter measures that included the closure of all public and private offices, closing all parks, walking only around the residency and not even in pairs, and the prohibition of mobility to second houses [44, 45]. According to the Google COVID-19 Community Mobility Reports [37], from March 20-21 until April 30, activities were reduced by an average of $\sim$87% in retail & recreation activities, by $\sim$84% in transit stations, by $\sim$70% in workplaces, by $\sim$80% in Parks and by $\sim$ 50% in groceries and pharmacies. Thus, taking into account the coarse effect of different activities in the physical contact [42], the average reduction was of the order of $\sim$65% in the mobility when compared to the period of February 21-March 8.

A further reduction may be attributed to behavioral changes [46]. For example, it has been shown that social distancing and cautiousness reduce the disease transmission rate by about 20% [40]. Thus, based on the above, it is reasonable to consider a (1-0.4)(1-0.2) (an average of 40% contribution of the reduction of the mobility and a 20% for the effect of social distancing) reduction in the effective transmission rate for the period March 8-March 19 as compared with the period February 21-March 8. For the period of March 20-21, based on the Google data [37], we considered a reduction of (1-0.65)(1-0.2) (as compared with the period February 21-March 8) reflecting the very strict measures taken then. Based on the above, we attempted a forecasting of the pandemic in Lombardy from March 8 to May 4, the first day of the relaxation of the strict isolation measures.

## Results

As discussed, for our computations we ran 48 times the numerical optimization procedure for each combination of initial guesses based on the daily reported new cases from February 21 to March 8 and for each block of 48 runs, for further analysis we kept the values that yielded the smaller fitting error over all 48 runs. For the period February 21-March 8, which is used for the calibration of the model parameters, the median value of the ratio between the number of new cases of infected and recovered (excluding the zero values) is of the order of 10, while the average value of the ratio of the the number of new infected and deaths is of the order of 20.

**Table 1. Optimal parameter values and interquartiles for day-zero, transmission rate ($\beta$) and fraction ($\epsilon$) of the cases of exposed individuals in the total population that enter to the compartment $I_c$.** The resulting value of $R_0$ along with the minimum and maximum value is also given.

| Parameter | Day-zero | $\beta$ | $\epsilon$ | $R_0$ |
|---|---|---|---|---|
| Mean | January 14 | 0.7 | 0.0707 | 4.53 |
| Interquartiles | January 5—January 23 | 0.665-0.719 | 0.023-0.1 | 4.40-4.65 |
| Distribution Function | Normal | Burr | Birnbaum-Saunders | |
| Parameters (95% CI) | $\mu$ = 1.67 (1.53, 1.80), $\sigma$ = 4.43 (4.33, 4.52) | $\alpha$ = 0.654 (0.653, 0.655), $c$ = 276.061 (248.911, 306.172), $k$ = 0.0537 (0.048, 0.060) | $\mu$ = 0.0492 (0.048, 0.0505), $\alpha$ = 0.934 (0.914, 0.954) | |

https://doi.org/10.1371/journal.pone.0240649.t001

Hence, we have used as weights $w_1 = 1$, $w_2 = 10$, $w_3 = 20$ to balance for the different scales of the number of infected vs. the number of recovered and dead. Other reasonable choices for the values of the weights around these values resulted in similar outcomes.

For all the near-optimal points obtained using the genetic algorithm optimization, the residuals were of the order of $\sim$ 4,750,000. Regarding the values of the optimal parameters, we fitted their cumulative probability distributions using several functions, including the Normal, Log-normal, Weibull, Beta, Gamma, Burr, Exponential and Birnbaum-Saunders functions, and kept the one yielding the maximum Log-likelihood (see in the S1 File).

Table 1 summarizes the mean values of the optimal parameters and their interquartiles, for the day-zero, the transmission rate ($\beta$) and the fraction ($\epsilon$) of the actual cases of exposed in the total population that enter to the compartment $I_c$, and also the information on the values of the parameters of the best fitting distributions.

Note that the optimal values of day-zero were between January 5—January 23 (interquartile range: January 11 to January 18) (see S1 Fig in S1 File), the optimal values of $\beta$ were between 0.636 and 0.86 (interquartile range: 0.665 to 0.719) (see S2 Fig in S1 File), and the optimal values of $\epsilon$ were between 0.01 and 0.249 (interquartile range: 0.023 to 0.1) (see S3 Fig in S1 File). The best fit to the distribution of optimal values of the day-zero was obtained using a Normal CDF with mean 1.67 (i.e. *sim* 2 days before the 16th of January) (95% CI:1.53, 1.80) and variance 4.43 (95% CI: 4.33, 4.52); thus taking the round value at 2 days, the expected day-zero corresponds to January 14 (interquartile range: January 11 to January 18).

The best fit to the distribution of the optimal values of $\beta$, was given by fitting a Burr CDF with $\alpha$ = 0.654 (95% CI: 0.653, 0.655), $c$ = 276.061 (95% CI: 248.911, 306.172), $k$ = 0.0537 (95% CI: 0.048, 0.060) having a mean value of 0.7 (interquartile range: 0.665 to 0.719). Finally, the best fit to the distribution of the optimal values of $\epsilon$ was given by fitting a Birnbaum-Saunders CDF with parameters $\mu$ = 0.0492 (95% CI: 0.048, 0.0505) (scale parameter) and $\alpha$ = 0.934 (95% CI: 0.914, 0.954) (shape parameter), resulting in an expected value 0.07 (interquartile range: 0.023 to 0.1) (see in S1 File).

Thus, based on the derived values of the "effective" per-day disease transmission rate, the basic reproduction number $R_0$ is 4.53 (min-max range: 4.40- 4.65).

Finally, we ran the simulator for all values of the optimal triplets from the corresponding distributions as found by the solution of the optimization problem using the data from February 21 until March 8, then from March 9 to March 20 for validation purposes and from March 29 to May 4 for forecasting purposes.

Figs (2)–(4) depict the simulation results based on the optimal estimates, until the 19th of March. To validate the model with respect to the reported data of confirmed cases from March 9 to March 19, we have considered an (1-0.4)(1-0.2) reduction in the "effective" transmission rate and as initial conditions the values resulting from the simulation on March 8 as described

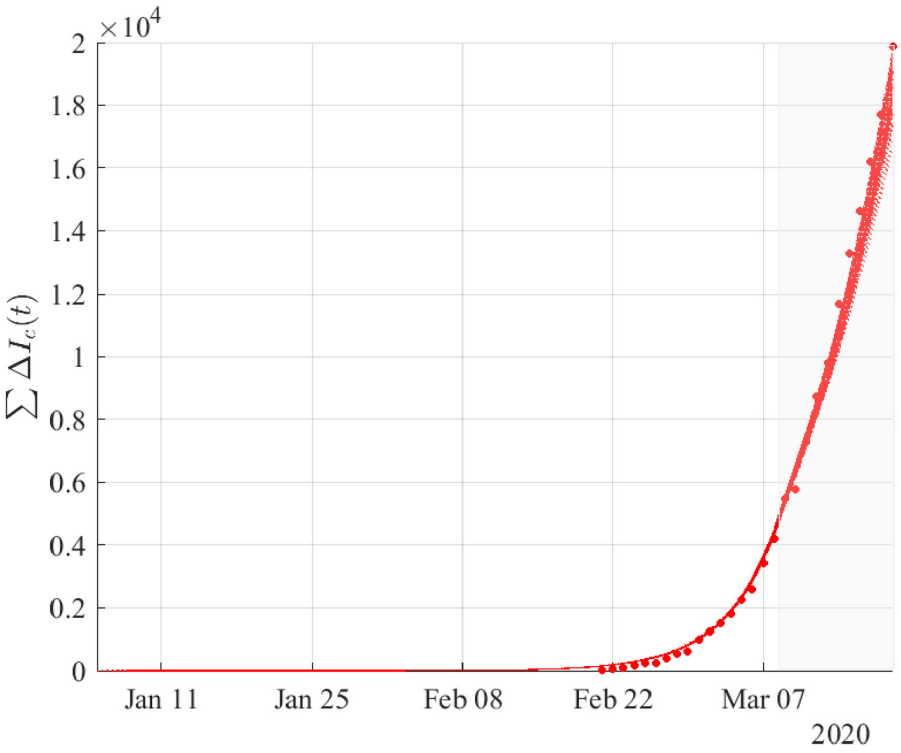


**Fig 2. Cumulative number of infected cases from the compartment $I_c$ resulting from simulations based on the results obtained by fitting the daily new cases of infected from February 21 until the 8th of March.** The validation of the model was performed using the reported data of confirmed cases from March 9 to March 19 (shaded area) by taking (1-0.4)(1-0.2) reduction in the "effective" transmission rate (see in Methodology) to the lockdown of March 8. Dots correspond to the reported data of confirmed cases.

https://doi.org/10.1371/journal.pone.0240649.g002

in the methodology. Thus, the model approximated fairly well the dynamics of the pandemic in the period from February 21 to March 19.

As discussed in the Methodology, we also attempted based on our methodology and modelling approach to forecast the evolution of the outbreak until May 4, the first day of the relaxation of the measures. To do so, as described in the methodology, we have considered a (1-0.65)(1-0.2) reduction in the "effective" transmission rate starting on March 20 (compared to the period February 21-March 8), the day of announcement of even stricter measures in the region of Lombardy (see in Methodology). The result of our forecast is depicted in Fig 5. As shown, the model predicts fairly the evolution of the epidemic two months ahead of March 8 (the model parameters and the day-zero were estimated using the reported data from February 21 to March 8). Note that, regarding the confirmed cases, the mean values of $I_c(t)$ from over all simulations almost coincide with the reported values of the confirmed cases (see top panel of Fig 5).

The reported recovered cases are in the lower part of the model predictions ($R_c(t)$) (see medium panel of Fig 5). Also, the model predicts fairly the total number of deaths until May 4 (see bottom panel of Fig 5. However, at May 4 there is a significant difference between the mean value of deaths as obtained from simulations and the actual number of deaths. This is due to the following reasons. First, the difference that is observed with respect to the reported number of deaths and model forecasts in the period from March 18 to April 15 can be attributed to facts that the model did not take into account, such as the saturation of ICUs in that period, which could potentially lead to a larger number of deaths. Indeed, on March 8, 400

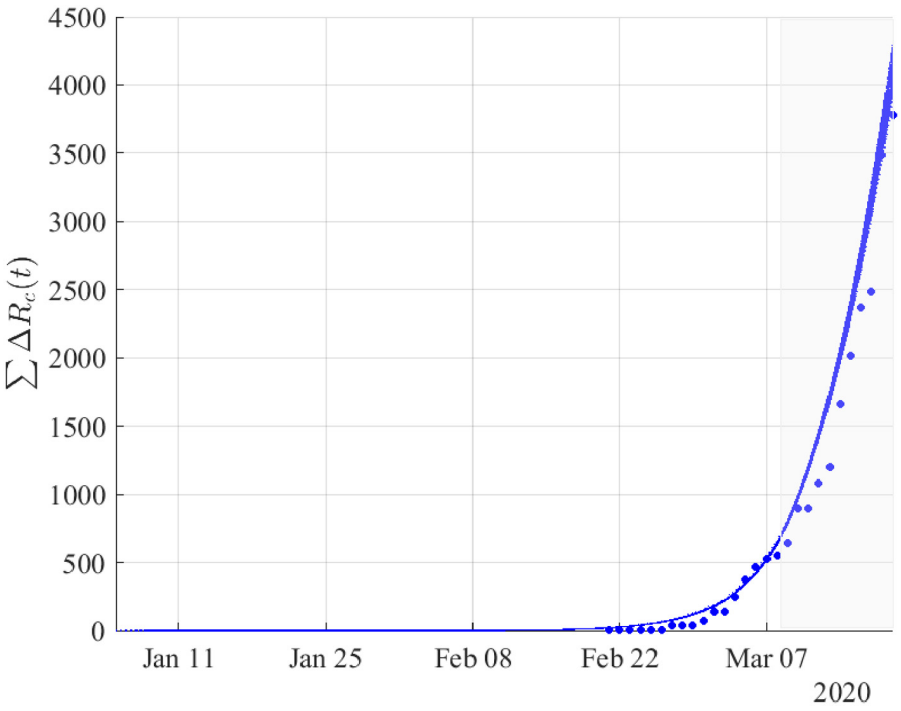


**Fig 3. Cumulative number of recovered cases $R_c(t)$ resulting from simulations based on the results obtained by fitting the daily new cases of recovered from February 21 until the 8th of March.** The validation of the model was performed using the reported data of confirmed cases from March 9 to March 19 (shaded area) by taking (1-0.4)(1-0.2) reduction in the "effective" transmission rate (see in Methodology) to the lockdown of March 8. Dots correspond to the reported data of confirmed cases of recovered.

https://doi.org/10.1371/journal.pone.0240649.g003

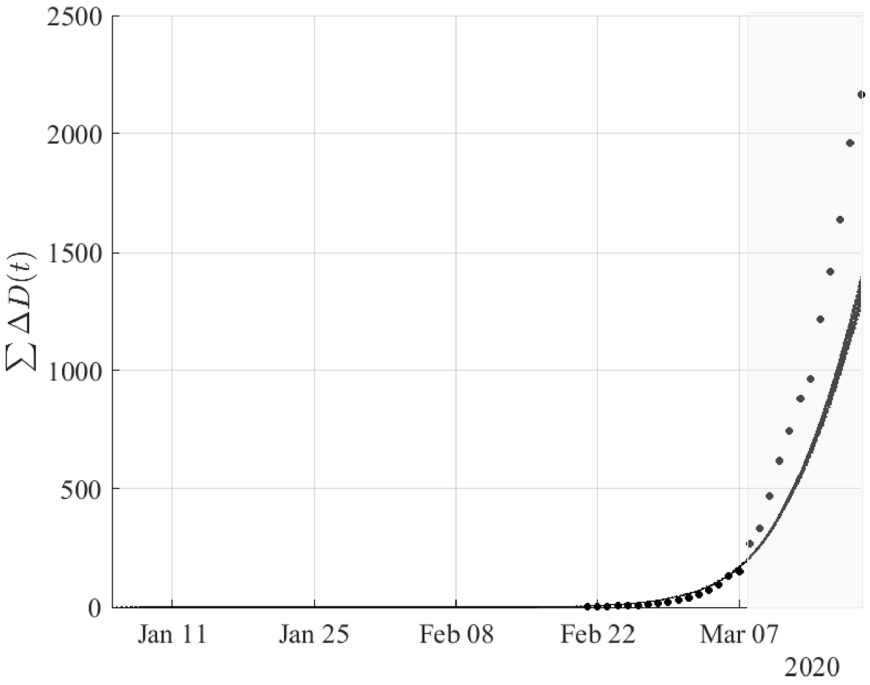


**Fig 4. Cumulative number of deaths resulting from simulations based on the results obtained by fitting the daily new cases of deaths from February 21 until the 8th of March.** The validation of the model was performed using the reported data of confirmed cases from March 9 to March 19 (shaded area) by taking (1-0.4)(1-0.2) reduction in the "effective" transmission rate (see in Methodology) to the lockdown of March 8. Dots correspond to the reported data of confirmed deaths.

https://doi.org/10.1371/journal.pone.0240649.g004

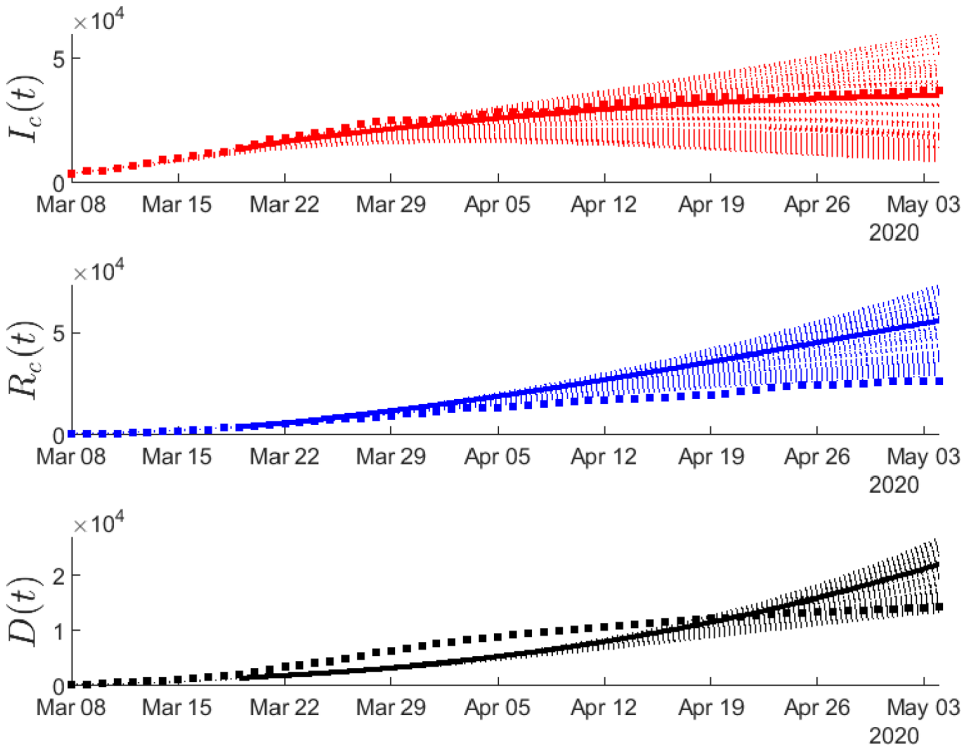


**Fig 5. Estimated number of the infected (aka "confirmed" infected) cases with mild and severe symptoms ($I_c(t)$), recovered ($R_c(t)$) and deaths ($D(t)$) resulting from all simulations from March 8 (the day of the lockdown of all Italy) until May 4 (see in Methodology).** Based on the Google released data (see in Methodology), we considered a (1-0.4)(1-0.2) reduction of the "effective" transmission rate from March 9 until March 19 compared to the period February 21-March 8, and then considering a (1-0.65)(1-0.2) reduction due to the even stricter mobility limitation measures announced by the government on March 20 and March 21 (see in Methodology). Red dots correspond to the reported confirmed infected cases, blue dots correspond to the reported recovered cases and black dots correspond to the reported deaths (up to May 4, 2020). Solid lines correspond to the mean values of $I_c(t)$, $R_c(t)$ and $D(t)$, over all runs (see in Methodology).

https://doi.org/10.1371/journal.pone.0240649.g005

patients were in ICUs, on March 15, 700 patients (an increase of almost 80%), and after 6 days, on March 21, this number increased to 1093 corresponding to an increase of almost 150% with respect to the situation on March 8; the peak was on April 3, with 1381 people in ICUs. As a result, this period was the one with the highest numbers of reported deaths. Namely, the highest number of reported deaths in the region was on March 21 with 546 deaths, followed by ∼540 deaths on March 27, ∼540 deaths on March 28, ∼415 deaths on March 29 and ∼460 deaths on March 30. For comparison, on March 14 the reported number of deaths was 77, while the day after, on March 15, ∼250 new deaths were reported, and on March 18, ∼320 new deaths were reported. However, we note that the model was calibrated with the data reported until March 8. Until March 8, the case fatality ratio was of the order of ∼9-10% and the number of cases admitted to ICUs was relatively small. After March 20 when there were many cases admitted to ICUs, the death toll increased significantly, it almost doubled, thus reaching the ∼17-18% of the confirmed cases. Second, regarding the difference that is observed between the model predictions and actual number of deaths at May 4, this is due to the fact that we did not calibrate the model parameters to take into account the changes in the death and recovery rates. It is expected that when the evolution of the epidemic is abrupt, as it was in Lombardy in the period March 12 until the early of April, due to the saturation of the

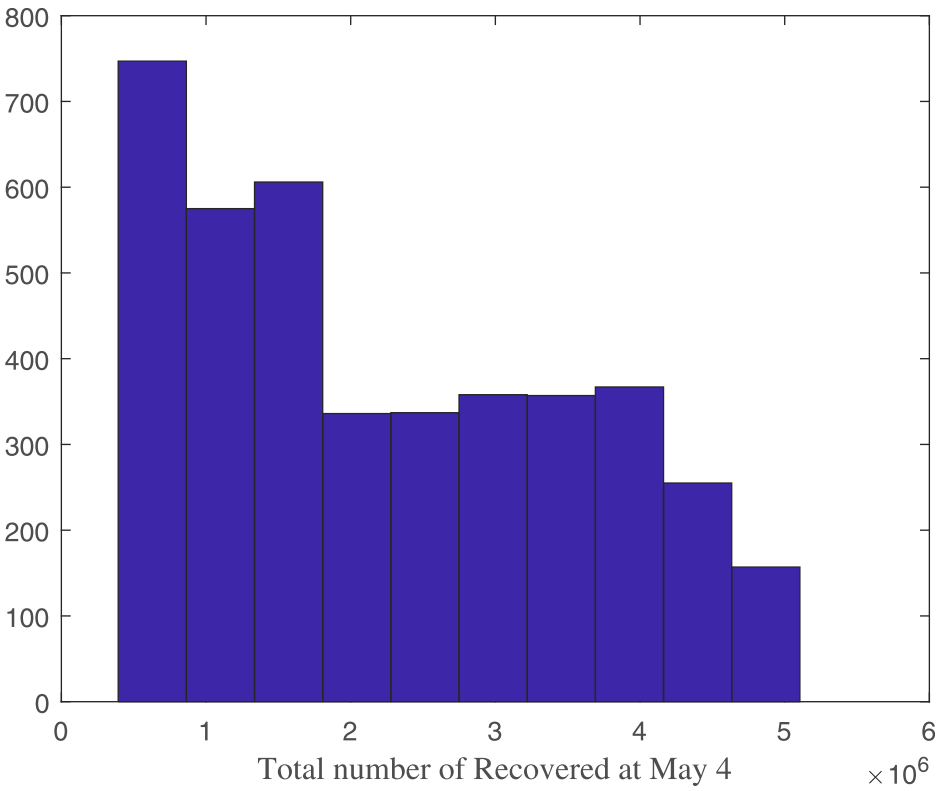


**Fig 6. Histogram of the recovered asymptomatic *R* at May 4 resulting from model simulations.**

https://doi.org/10.1371/journal.pone.0240649.g006

health system and ICUs the mortality rates would be higher and the recovery rates longer. Indeed, during this period, it has been reported, that the median time from the onset of symptoms until death in Lombardy was eight days [29] which a very short period. For example, in other studies it has been reported that the time between symptom onset and death ranged from about 2 weeks to 8 weeks which from two to seven times longer than the one reported in Lombardy [24, 47]. So, this difference explains the difference between simulations and reported number of deaths at May 4. Here, for our demonstrations we have used the information about the epidemiological parameters, that was available at the early phase of the epidemic. Thus, we must underline the very good agreement of the model predictions with the actual number of confirmed infected cases for the entire period March 8-May 4. This number, in contrast to the number of recovered and number of deaths is not shaped/biased by the capacity of ICUs and this is the number on which policy makers should focus in order to decide ahead of time of the necessary resources (e.g. necessary number of beds and ICUs) needed to keep the fatality ratio as low as possible.

The model predicts that until May 4, an average of 20% of the population in Lombardy has already recovered (interquartile range: $\sim$10% to $\sim$30%) (see Fig 6).

Finally, in Fig 7, we report the evolution of the effective reproduction number $R_e$ until May 4.

On May 4 the estimated mean value of $R_e$ was 0.987 (interquartile range: 0.857 to 1.133), thus marking the onset of a critical date for the post-lockdown period.

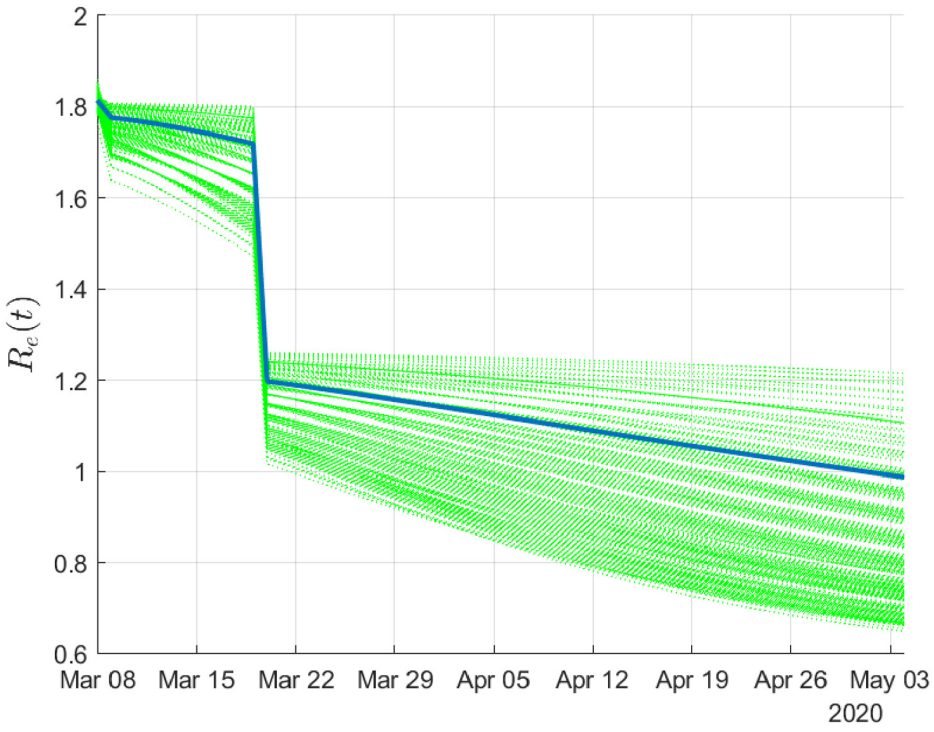


**Fig 7. Evolution of the effective reproduction number.** The solid line depicts the mean value of the distribution.

https://doi.org/10.1371/journal.pone.0240649.g007

## Discussion

The crucial questions about an outbreak is how, when (day-zero), why it started, and if and when it will end. Answers to these important questions would add critical knowledge to our arsenal to combat the pandemic. The tracing of day-zero, in particular, is of outmost importance. It is well known, that minor perturbations in the initial conditions of a complex system, such as the ones of an outbreak, may result in major changes in the observed dynamics. Undoubtedly, a high level of uncertainty for day-zero, as well as the uncertainty in the actual numbers of exposed people in the total population, raise several barriers to our ability to correctly assess the state and dynamics of the outbreak and to forecast its evolution and its end. Such pieces of information would lower the barriers and help public health authorities respond fast and efficiently to the emergency. For example, an over- or under-estimation of day-zero would result in an under- or over-estimation of the the transmission rate $\beta$ and therefore of the basic reproduction number $R_0$ and consequently the effective reproduction number $R_e$. Furthermore, the correct estimation of day-zero is important for the assessment of the number of asymptomatic and actual recovered cases in the total population. This is turn will bias the assessment and ultimately the design of efficient control policies in real time.

This study aimed exactly at shedding more light into this problem. To achieve this goal, we addressed a conceptually simple compartmental SEIIRD model with two infectious compartments in order to bridge the gap between the number of asymptomatic cases in the total population and the cases that will experience mild to more severe symptoms.

What is done until now with mathematical epidemiological models is the investigation of several scenarios, by changing e.g. the (or assuming a fixed) initial day (day-zero), the level of asymptomatic cases etc. Our work, is the first that introduces a methodological framework, to estimate the day-zero as well as the level of asymptomatic cases in the total population in a

systematic way. Following the proposed methodological framework, we found that the day-zero in Lombardy was around the middle of January, a period that precedes by one month the fate of the first confirmed case in the hardest-hit northern Italian region of Lombardy. Interestingly enough, when we submitted a preprint of the work at MEDRXIV on March 20 2020 ([48], another study that was also submitted at the same day at MEDRXIV, that was based on genomic and phylogenetic data analysis, reports the same time period, between the second half of January and early February, 2020, as the time when the novel coronavirus SARS-CoV-2 entered northern Italy [49]. Our analysis further revealed that the actual number of asymptomatic infected cases in the total population in the period until March 8 was around 15 times the number of confirmed infected cases, which until March 8 was also approximately equal to the number of cases that were hospitalized and admitted to ICUs.

Our model and methodological approach assume that there was one effective "zero" infected case that introduced the virus to the region; one could certainly argue that there were more than one cases that introduced the virus to the region on the same day; such scenarios can be investigated in a straightforward manner based on our proposed methodological approach. Furthermore, the proposed approach could be used for the quantification of the uncertainty of the evolving dynamics, taking into account the reported, from clinical studies, distributions of the epidemiological parameters rather than their expected values. A critical point that is connected with the above is that with such a small number of infectious at the initial stage of the simulations, a stochastic model or a hybrid stochastic model could be more realistic in which uncertainty could be modelled in the form of realistic perturbations (see for example [50]). Furthermore, we did not consider the effect of an ongoing sampling strategy in the total population for the estimation of the level of under-reporting (represented in our model with the variable $\epsilon$). As mentioned in the methodology, within the first period of the outbreak the number of confirmed cases was approximately equal to the number of hospitalized cases, i.e. there was no sampling strategy. Furthermore, as also reported for the later period, tests were conducted only for those who seeked medical care and had symptoms like fever and coughing. Thus, people who did not seek for medical attention were tested very scarcely [18–20]. We will consider such type of modelling and analysis in a future work.

Regarding the forecasting in Lombardy from March 8 until May 4 (the first day of relaxation of the measures), we have taken into account the very latest facts on the drop of human mobility, as released by Google [37] until April 30 for the region of Lombardy; these were shaped by the very strict measures announced on March 20-21 that included the closure of all parks, public and private offices and the prohibition of any pedestrian activity, even individually [45]. Our modelling approach approximated fairly well the reported number of infected cases in Lombardy two months ahead of time. The mean value of the evolution of the compartment that in our model reflects the confirmed cases, almost coincides with the reported cases for the entire period from March 8 to May 4. The differences that are observed between the reported number of deaths and simulations as discussed in the results should be attributed to the very short times from the onset of symptoms to death that have been reported in Lombardy at the early phases of the pandemic in the region (8 days instead of 2 to 8 weeks that have been reported in other studies for China that is linked with the saturation of the ICUs [47, 51]. Furthermore another important factor that is missing from the data used is that the number of deaths is largely affected by the criteria of death notification. Indeed as it has been reported, the global coronavirus death toll could be 60% higher than the confirmed [52]. However, this fact did not affect the model predictions with respect to the confirmed cases, but the rate with which the confirmed cases die; the later is also dependent on the rate of the outbreak and the relevant capacity of the ICUs.

To this end, we would like to make a final comment with respect to the basic reproduction number $R_0$, the significance and meaning of which are very often misinterpreted and misused, thereby leading to erroneous conclusions. Here, we found a $R_0 \sim 4.5$, which is higher compared to the values reported by many studies in China, and also in Italy, and in Lombardy in particular. For example, Zhao et al. estimated $R_0$ to range between 2.24 (95% CI: 1.96, 2.55) and 3.58 (95% CI: 2.89, 4.39) in the early phase of the outbreak [9]. Similar estimates were obtained for $R_0$ by Imai et al., 2.6 (95% CI: 1.5, 3.5) [6], Li et al. [53], Wu et al., 2.68 (95% CI: 2.47, 2.86), as well as by Anastassopoulou et al. recently, 3.1 (90% CI: 2.5, 3.7) [8].

Regarding Italy, D'Arienzo and Coniglio [54] used a SIR model to fit the reported data in nine Italian cities and found that $R_0$ ranged from 2.43 to 3.10. In another study, the authors provided an estimate of the basic reproduction number by analyzing the first 5,830 laboratory-confirmed cases. By doing so they estimated the basic reproduction number at 3.1 [27].

First, we would like to stress that $R_0$ is not a biological constant for a disease as it is affected not only by the pathogen, but also by many other factors, such as environmental conditions, demographics, as well as, importantly, by the social behavior of the population (see for example the discussion in [4]). Thus, a value for $R_0$ that is found in one part of the world (e.g. in China), and even in a region of the same country, e.g. in Tuscany, Italy, cannot be generalized as a global biological constant for other parts of the world, or even for other regions of the same country. Obviously, the environmental factors and social behavior of the population in Lombardy are different from the ones, for example, prevailing in Hubei.

Second, most of the studies that provide estimates of the basic reproduction number are based solely on the reported cases, thus the actual number of infected cases in the total population, that may be asymptomatic but transmit the disease, is not considered; this fact may lead to an underestimation of the basic reproduction number. Moreover, in our approach as compared to clinical studies, the computation of $R_0$ comes out as the necessary and sufficient condition of the stability as derived from the proposed model whose parameters are computed based on the available reported data, thus with a delay between the first actual case (see also the discussion in [55].

Our relatively simple conceptually model and approach do not aspire to accurately describe the complexity of the emergent dynamics, which in any case is an overwhelmingly difficult, if not impossible, task in the long run, even with the use of detailed agent-based models. We tried to keep the structure of the model as simple as possible in order to be able to model (in a coarse way) the uncertainty in both the "day-zero" and the number of asymptomatic actual cases in the total population using as few parameters as possible. The results of our analysis have indeed proved that the modelling approach succeeded in providing fair predictions of the evolution of the epidemic two months ahead of time. Such an early assessment would help authorities to evaluate the required measures to control the epidemic, such as the scale of diagnostic tests that have to be performed and the number of ICU beds required. While more complicated models can, in principle, be constructed to take into account more detailed information, such as the number of hospitalized patients and patients in ICUs, for any practical means such an approach would suffer from the "curse of dimensionality" as it would introduce many more parameters that would need calibration based on a relatively small size of data especially at the beginning of an outbreak. An attempt to compute the values of some of these additional parameters, which can only be roughly estimated by clinical studies at the early stages of an emerging novel infectious disease, would introduce additional uncertainty, thereby further complicating matters rather than solving the problem.

To this end, we hope that our conceptually simple, but pragmatic, modelling approach and methodological framework help to provide improved insights into the currently uncontrolled pandemic and to contribute to the mitigation of some of its severe consequences.

## Supporting information

**S1 File.**
(PDF)

## Author Contributions

**Conceptualization:** Constantinos Siettos.

**Data curation:** Cleo Anastassopoulou, Gennaro Nicola Bifulco.

**Formal analysis:** Lucia Russo, Emilio Fortunato Campana, Gerardo Toraldo, Constantinos Siettos.

**Investigation:** Lucia Russo, Cleo Anastassopoulou, Athanasios Tsakris, Gennaro Nicola Bifulco, Emilio Fortunato Campana, Gerardo Toraldo, Constantinos Siettos.

**Methodology:** Lucia Russo, Constantinos Siettos.

**Validation:** Cleo Anastassopoulou, Athanasios Tsakris, Gennaro Nicola Bifulco, Emilio Fortunato Campana, Gerardo Toraldo, Constantinos Siettos.

**Writing – original draft:** Cleo Anastassopoulou, Constantinos Siettos.

**Writing – review & editing:** Lucia Russo, Cleo Anastassopoulou, Athanasios Tsakris, Gennaro Nicola Bifulco, Emilio Fortunato Campana, Gerardo Toraldo, Constantinos Siettos.

## References

1. W.H. Organization, Coronavirus disease 2019 (COVID-19). Situation report 108 (2020). https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200507covid-19-sitrep-108.pdf?sfvrsn=44cc8ed8_2
2. J. H. C. for Health Security, Coronavirus COVID-19 Global Cases by Johns Hopkins CSSE (feb 2020). https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6
3. Carinci F., Covid-19: preparedness, decentralisation, and the hunt for patient zero, BMJ 368. https://www.bmj.com/content/368/bmj.m799
4. Delamater P. L., Street E. J., Leslie T. F., T. Yang Y., Jacobsen K. H., Complexity of the basic reproduction number (r0), Emerging Infectious Diseases 25 (1) (2019) 1–4. https://doi.org/10.3201%2Feid2501.171901
5. J. T. Wu, K. Leung, G. M. Leung, Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in wuhan, china: a modelling study, The Lancet https://doi.org/10.1016%2Fs0140-6736%2820%2930260-9
6. N. Imai, A. Cori, I. Dorigatti, et al., Report 3: Transmissibility of 2019-ncov, Int J Infect Dis https://www.imperial.ac.uk/media/imperial-college/medicine/sph/ide/gida-fellowships/Imperial-2019-nCoV-transmissibility.pdf
7. D. Li, J. Lv, G. Botwin, J. Braun, W. Cao, L. Li, et al., Estimating the scale of covid-19 epidemic in the united states: Simulations based on air traffic directly from wuhan, china, medRxiv https://www.medrxiv.org/content/early/2020/03/08/2020.03.06.20031880
8. Anastassopoulou C., Russo L., Tsakris A., Siettos C., Data-based analysis, modelling and forecasting of the covid-19 outbreak, PLOS ONE 15 (3) (2020) 1–21. https://doi.org/10.1371/journal.pone.0230405
9. S. Zhao, Q. Lin, J. Ran, S. S. Musa, G. Yang, W. Wang, et al., Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in china, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak, Int J Infect Dis https://doi.org/10.1101%2F2020.01.23.916395
10. A. Remuzzi, G. Remuzzi, COVID-19 and italy: what next?, The Lancet https://doi.org/10.1016%2Fs0140-6736%2820%2930627-9
11. W.-K. Ming, J. Huang, C. J. P. Zhang, Breaking down of the healthcare system: Mathematical modelling for controlling the novel coronavirus (2019-nCoV) outbreak in wuhan, china https://doi.org/10.1101%2F2020.01.27.922443

12. H.-Y. Yuan, M. P. Hossain, M. M. Tsegaye, X. Zhu, P. Jia, T.-H. Wen, et al., Estimating the risk on outbreak spreading of 2019-ncov in china using transportation data https://www.medrxiv.org/content/early/2020/02/04/2020.02.01.20019984

13. P. M. De Salazar, R. Niehus, A. Taylor, C. O. Buckee, M. Lipsitch, Using predicted imports of 2019-ncov cases to determine locations that may not be identifying all imported cases https://www.medrxiv.org/content/early/2020/02/11/2020.02.04.20020495

14. M. Chinazzi, J. T. Davis, M. Ajelli, C. Gioannini, M. Litvinova, S. Merler, et al., The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak, Science (2020) eaba9757 https://doi.org/10.1126%2Fscience.aba9757

15. Volpert V., Banerjee M., d'Onofrio A., Lipniacki T., Petrovskii S., Tran V.C., Coronavirus–scientific insights and societal aspects, Mathematical Modelling of Natural Phenomena 15 (2020) E2. https://doi.org/10.1051%2Fmmnp%2F2020010

16. CNBC, Current US coronavirus cases are "just the tip of the iceberg,– former USAID director says. (2020). https://www.cnbc.com/2020/03/05/us-coronavirus-cases-just-the-tip-of-the-iceberg-ex-usaid-director.html

17. T. N. C. P. E. R. E. Team, The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (covid-19) in china, 2020, China CDC Weekly 2 (2020) 113.http://weekly.chinacdc.cn//article/id/e53946e2-c6c4-41e9-9a9b-fea8db1a8f51

18. T. Guardian, In one Italian town, we showed mass testing could eradicate the coronavirus (2020). https://www.theguardian.com/commentisfree/2020/mar/20/eradicated-coronavirus-mass-testing-covid-19-italy-vo

19. I. messaggero, Coronavirus, stop ai test facili: tampone solo a chi ha i sintomi (2020). https://www.ilmessaggero.it/italia/coronavirus_controlli_tampone_sintomi_ultime_notizie_27_febbraio-5077073.html

20. I. Post, Perch& x00E9; non stiamo facendo pi& x00F9; tamponi? (2020). https://www.ilpost.it/2020/04/04/coronavirus-tamponi-test-problemi-numero/

21. M. della Salute, Covid-19—Situazione in Italia (2020). http://www.salute.gov.it/portale/nuovocoronavirus/dettaglioContenutiNuovoCoronavirus.jsp?area=nuovoCoronavirus&id=5351&lingua=italiano&menu=vuoto

22. C. for Disease Control, Prevention, How COVID-19 Spreads (2020). https://www.cdc.gov/coronavirus/2019-ncov/about/transmission.html

23. Q. Li, X. Guan, P. Wu, e. a. Wang, Early transmission dynamics in wuhan, china, of novel coronavirus infected pneumonia, New England Journal of Medicine 0 (0) (0) null. https://doi.org/10.1056/NEJMoa2001316

24. W. H. Organization, Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19) (2020). https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf

25. t. F. Fernando Duarte, BBC. [link]. https://www.cdc.gov/flu/symptoms/symptoms.htm

26. Anderson R. M., Heesterbeek H., Klinkenberg D., Hollingsworth T. D., How will country-based mitigation measures influence the course of the COVID-19 epidemic?, The Lancet 395 (10228) (2020) 931–934. https://doi.org/10.1016%2Fs0140-6736%2820%2930567-5

27. D. Cereda, M. Tirani, F. Rovida, V. Demicheli, M. Ajelli, P. Poletti et al., The early phase of the covid-19 outbreak in lombardy, italy (2020). arXiv:2003.09320.

28. Nishiura H., Linton N. M., Akhmetzhanov A. R., Serial interval of novel coronavirus (COVID-19) infections, International Journal of Infectious Diseases 93 (2020) 284–286. https://doi.org/10.1016%2Fj.ijid.2020.02.060

29. I. Istituto Superiore di Sanit Characteristics of COVID-19 patients dying in Italy Report based on available data on March 20th, 2020 (2020). https://www.epicentro.iss.it/coronavirus/bollettino/Report-COVID-2019_20_marzo_eng.pdf

30. King A. A., de Cellès M. D., Magpantay F. M. G., Rohani F. M. G., Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to ebola, Proceedings of the Royal Society B: Biological Sciences 282 (1806) (2015) 20150347. https://doi.org/10.1098%2Frspb.2015.0347 PMID: 25833863

31. M. Preuss, Niching methods and multimodal optimization performance, in: Natural Computing Series, Springer International Publishing, 2015, pp. 115–137. https://doi.org/10.1007%2F978-3-319-07407-8_5

32. Burr I. W., Cumulative frequency functions, The Annals of Mathematical Statistics 13 (2) (1942) 215–232. https://doi.org/10.1214%2Faoms%2F1177731607

33. Birnbaum Z., Saunders S., A new family of life distributions, Journal of Applied Probability 6 (02) (1969) 319–327. https://doi.org/10.1017%2Fs0021900200032848

34. The Mathworks, Inc., Natick, Massachusetts, MATLAB R2018b (2018).

35. E. I. Jury, L. Stark, V. V. Krishnan, Inners and stability of dynamic systems, IEEE Transactions on Systems, Man, and Cybernetics SMC-6 (10) (1976) 724–725. https://doi.org/10.1109%2Ftsmc.1976.4309436

36. Diekmann O., Heesterbeek J. A. P., Roberts M. G., The construction of next-generation matrices for compartmental epidemic models, Journal of The Royal Society Interface 7 (47) (2009) 873–885. https://doi.org/10.1098%2Frsif.2009.0386

37. Google, Community Mobility Reports (2020). https://www.google.com/covid19/mobility/

38. W. H. Organization, Advancing the right to health: the vital role of law, World Health Organization, 2016.

39. ping Su C., de Perio M. A., Cummings K. J., McCague A.-B., Luckhaupt S. E., Sweeney M. H., Case investigations of infectious diseases occurring in workplaces, united states, 2006–2015, Emerging Infectious Diseases 25 (3) (2019) 397–405. https://doi.org/10.3201%2Feid2503.180708

40. Caley P., Philp D. J., McCracken K., Quantifying social distancing arising from pandemic influenza, Journal of The Royal Society Interface 5 (23) (2007) 631–639. https://doi.org/10.1098%2Frsif.2007.1197

41. Fong M. W., Gao H., Wong J. Y., Xiao J., Shiu E. Y., Ryu S., et al., Nonpharmaceutical measures for pandemic influenza in nonhealthcare settings—social distancing measures, Emerging Infectious Diseases 26 (5). https://doi.org/10.3201%2Feid2605.190995

42. Mossong J., Hens N., Jit M., Beutels P., Auranen K., Mikolajczyk R., et al., Social contacts and mixing patterns relevant to the spread of infectious diseases, PLoS Medicine 5 (3) (2008) e74. https://doi.org/10.1371%2Fjournal.pmed.0050074

43. S. Tg4, Coronavirus Lombardia, Sala: "Il 40% ancora si sposta. Si vede dalle celle telefoniche (2020). https://tg24.sky.it/cronaca/2020/03/17/coronavirus-spostamenti-lombardia.html

44. R. News, Coronavirus, nuove restrizioni. A Milano militari per le strade. Posti di blocco a Roma. (2020). http://www.rainews.it

45. R. News, Coronavirus, ordinanza Lombardia con nuove limitazioni (2020). https://www.rainews.it

46. Manfredi P., D'Onofrio A. (Eds.), Modeling the Interplay Between Human Behavior and the Spread of Infectious Diseases, Springer New York, 2013. https://doi.org/10.1007%2F978-1-4614-5474-8

47. D. Baud, X. Qi, K. Nielsen-Saines, D. Musso, L. Pomar, G. Favre, Real estimates of mortality following COVID-19 infection, The Lancet Infectious Diseases https://doi.org/10.1016%2Fs1473-3099%2820%2930195-x

48. L. Russo, C. Anastassopoulou, A. Tsakris, G. N. Bifulco, E. F. Campana, G. Toraldo, et al., Tracing day-zero and forecasting the covid-19 outbreak in lombardy, italy: A compartmental modelling and numerical optimization approach., medRxiv arXiv: https://www.medrxiv.org/content/early/2020/05/13/2020.03.17.20037689.full.pdf

49. G. Zehender, A. Lai, A. Bergna, L. Meroni, A. Riva, C. Balotta, et al., Genomic characterisation and phylogenetic analysis of sars-cov-2 in italy, Journal of Medical Virology n/a (n/a). arXiv: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jmv.25794

50. d'Onofrio A. (Ed.), Bounded Noises in Physics, Biology, and Engineering, Springer New York, 2013. https://doi.org/10.1007%2F978-1-4614-7385-5

51. W. H. Organization, WHO Statement Regarding Cluster of Pneumonia Cases in Wuhan, China (2020). https://www.who.int/china/news/detail/09-01-2020-who-statement-regarding-cluster-of-pneumonia-cases-in-wuhan-china

52. F. Times, Global coronavirus death toll could be 60% higher than reported | Free to read (2020). https://www.ft.com/content/6bd88b7d-3386-4543-b2e9-0d5c6fac846c

53. Q. Li, X. Guan, P. Wu, et al., Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus Infected Pneumonia (2020). https://doi.org/10.1088%2F0951-7715%2F16%2F2%2F308

54. M. D'Arienzo, A. Coniglio, Assessment of the SARS-CoV-2 basic reproduction number, r0, based on the early phase of COVID-19 outbreak in italy, Biosafety and Health. https://doi.org/10.1016/j.bsheal.2020.03.004

55. Cori A., Ferguson N. M., Fraser C., Cauchemez S., A new framework and software to estimate time-varying reproduction numbers during epidemics, American Journal of Epidemiology 178 (9) (2013) 1505–1512. https://doi.org/10.1093%2Faje%2Fkwt133