A Novel Heuristic Global Algorithm to Predict the COVID-19 Pandemic Trend

Panagiotis G. Asteris[1]*, Maria Douvika[1], Christina Karamani[1], Athanasia Skentou[1], Tryfon Daras[2], Liborio Cavaleri[3], Danial Jahed Armaghani[4], Katerina Chlichlia[5] and Theoklis E. Zaoutis[6]

[1] Computational Mechanics Laboratory, School of Pedagogical and Technological Education, 14121 Heraklion, Athens, Greece. Email: panagiotisasteris@gmail.com

[2] School of Environmental Engineering, Technical University of Crete, Chaniá, Greece

[3] Department of Civil, Environmental, Aerospace and Materials Engineering (DICAM), University of Palermo, Palermo, Italy

[4] Department of Civil Engineering, Faculty of Engineering, University of Malaya, Kuala Lumpur, Malaysia

[5] Laboratory of Molecular Immunology, Department of Molecular Biology and Genetics (MBG), Democritus University of Thrace, Alexandroupolis, Greece

[6] Division of Infectious Diseases, The Children's Hospital of Philadelphia, Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA

[*] Corresponding author: Prof. Panagiotis G. Asteris, email: panagiotisasteris@gmail.com

Summary

Background Mathematical models are useful tools to predict the course of an epidemic. The present manuscript proposes a heuristic global algorithm for predicting the COVID-19 pandemic trend.

Methods The proposed method utilizes a Gaussian-function-based algorithm for estimating how the temporal evolution of the pandemic develops by predicting daily COVID-19 deaths, for up to

10 days, from the day the prediction is made. This dataset, the number of daily deaths in each country or region, encapsulates information about (a) the quality of the health system of each country or region, (b) the age profile of the country's/region's population, and (c) environmental and other conditions.

Findings The validity of the proposed heuristic global algorithm has been tested in the case of China (at different temporal stages of the pandemic), a country where the disease trend seems to have run its course. It has been applied to ten countries/states/cities, namely California, Germany, China, Greece, Iran, Lombardia/Italy, New York, Sweden, United Kingdom and USA, for each one of which predictions have been obtained. The method has also been applied to the United States as a whole, as well as to the states of New York and California, in order to investigate how the pandemic is developing in different parts of the same country.

Interpretation Based on the predicted findings, the proposed algorithm seems to offer a robust and reliable method for revealing the SARS-CoV-2 temporal dynamics and disease trend, as such, can be a useful tool for the relevant authorities.

Introduction

In January 2020, the novel aggressive coronavirus SARS-CoV-2 was identified as the causative agent of an outbreak of viral pneumonia in Wuhan/China, the coronavirus disease 2019 (COVID-19). The outbreak of COVID-19 has already spread to more than 200 countries and has been officially declared a global pandemic [1]. The number of confirmed cases increases drastically every day as well as the number of deaths.

In response to COVID-19, governments have implemented several regulations that constrain personal freedoms (physical distancing) and restrict their economies, placing approximately 3 billion people under lockdown. To follow the transmission dynamics, there is a big demand in early diagnosis with a race in developing and approving tests for early and accurate molecular diagnosis of the infection. Thus, there is an urgent need in early prediction in order to reduce the risk for transmission worldwide in all locations. The task governments and national authorities now are facing is to apply recommendations and decisions for rapid strengthening of outbreak surveillance and control efforts. It is very important to be able to estimate and predict the virus spread and be able to decode viral characteristics of the infection and spread pattern in many countries worldwide, taking into consideration all relevant aspects, in order to take robust decisions for political interventions and control measures. Mathematical models are used to forecast the course of the epidemic. In the light of the above, with reference to the mortality data in each country, the aim of the current study was to develop a novel robust and reliable global algorithm for estimating and predicting the COVID-19 pandemic outburst for up to 10 days after the prediction date in different locations worldwide.

Methods

Assumptions and Data sources

During the study of the development of the COVID-19 pandemic, the daily total number of confirmed deaths due to COVID-19 for each location have been recorded and utilized further. The selection of daily deaths was based on the authors' assumption that mortality rates provide more accurate and reliable data as opposed to recordings of the number of daily infected

individuals. The credibility of the latter is restricted due to the fact that the actual real situation needs systematic and thorough study of many data based on statistical rules, which render the study time-demanding and high-priced. In addition, the daily mortality rate contains additional information on many crucial parameters that influence the pandemic transmission trend and spread. Among the parameters that are included in the death recordings are the following:

- The climate conditions of each country, state or region
- The quality of the health-care system
- The experience/level of the medical staff / health care workers
- The age distribution of the population (demographic structure)

In addition to the above, the main assumption during the design of this algorithm was the observation that the mortality rate, in particular the death numbers in the respective populations, follow a normal distribution. Even though daily recording might not be the case for optimal normal distribution, it is important to notice that the selection of death recordings every 2 days, or 3 days etc. leads - almost always – to an optimal normal distribution. Following this assumption, the simulation of the pandemic spread was investigated for a variety of different scenarios of mortality rate recordings and the best setting giving the best results and predictions was selected.

Based on the above, the final overall database was based on the combination of three individual databases. Data for the sum of the countries were obtained from the database Worldometer [2], for USA were derived from the COVID Tracking project [3], while for the Italian cities from GitHub, Inc. [4].

Proposed Heuristic Algorithm

Analyzing the official data from China (daily Coronavirus-19 incidents and deaths), the country where the pandemic began and which now appears to have largely overcome it, one can easily see that these data can be closely approximated/expressed using a suitable Gaussian curve (or equivalently a proper normal distribution density function). In addition, by studying the evolution of the pandemic and the course of the events/restrictions in this country, and taking into account that almost all European and other world countries have taken similarly strict restrictive measures, we assume that if something does not change dramatically, and not taking in account possible population or climate/environmental differences, the development of Coronavirus-19 pandemic will be similar in (most) European and world countries i.e data concerning the epidemic will be expressed using a proper normal distribution.

A Gaussian function is a function of the form: $y = f(x) = Ae^{-\frac{(x-\mu)^2}{2\sigma^2}}$ The graph of the function is a symmetrical bell-shaped curve centered at the position $x = \mu$, A is the height of the peak and the variance σ^2 controls its width. On both sides of the peak, the tails of the curve quickly fall of and approach the x-axis (asymptote). Our algorithm is trying to determine in each case (country/state) the optimal normal curve (for daily deaths), by calculating the parameters A, μ, σ^2, i.e by fitting to the given data the "best" possible normal curve. The optimality is given w.r.t well known statistical indices.

More precisely, the main steps of the algorithm are (through a triple loop):

1. (for A / first inner loop) We start from a given value of A, and with step 1, we continue up to a certain value (desired accuracy) depending on the maximum value of our available data (deaths)

2. (for σ^2 / second inner loop) $A = \frac{1}{\sqrt{2\pi}\sigma}$ so peaking A, σ^2 can be calculated. Next, the algorithm uses a probability value p, starting from p=0.85, and with step 0.01, continues up to 0.99999 (it is known that $P = px100\%$ of the data under a normal distribution curve lie inside the interval $[\mu + z_{q/2}s, \mu - z_{q/2}s]$,where $q = 1 - p$ and $z_{q/2} = \Phi^{-1}(q/2)$, Φ the cumulative distribution function of the standard normal distribution ($N0,1$) and Φ^{-1} its inverse function. This interval is used to fit the actual data (using a proper transformation).

3. (for μ / third inner loop) We start from a value of μ =20 and we continue, with step 1 (day), up to a value of μ =60 (we observed for example that in the case of China, the phenomenon lasted for about 60 days with an average (peak day of deaths) in about the 30th day).

4. As a result, of the algorithm application, a great deal of (proper) normal distributions are being created, by calculating in each time the theoretical/ experimental values of the corresponding normal distribution. Finally, these values are being compared with the empirical values (actual deaths data) and the “best” possible curve is being selected using a number of indices (smallest possible differences between theoretical and empirical data).

Performance Assessment

The reliability and accuracy of the developed best fit Gaussian curves for each one prediction were evaluated using Pearson's correlation coefficient R and the root mean square error (RMSE). RMSE presents information on the short-term efficiency which is a benchmark of the difference of predicated values in relation to the experimental values. The lower the RMSE, the more accurate is the evaluation. The Pearson's correlation coefficient R measures the variance that is interpreted by the model, which is the reduction of variance when using the model. R values ranges from 0 to 1 while the model has healthy predictive ability when it is near to 1 and is not analyzing whatever when it is near to 0. These performance metrics are a good measure of the overall predictive accuracy.

Methodology

The present section outlines the methodology used to investigate the spread of COVID-19 in a country or parts of it such as a state, city or region. In particular, the methodology is being presented here step by step, as it was conducted and applied in the case of the investigation of the spread of the epidemic in China. In fact, given that the epidemic in China proceeded of the epidemic in other countries, provides the possibility of applying the proposed algorithm both at the beginning of the phenomenon, in its next phase which is usually characterized by a strong dynamic and finally at its peak where the dynamic of the phenomenon begins to fade as it is usual in dynamic phenomena.

The main characteristics/steps of the proposed methodology are:

- In each step of the study of the phenomenon, the optimal normal distribution is calculated using the proposed algorithm and based on (up to the moment of application) available data.

- It was deemed necessary that, the first assessment be made 14 days after the first death record. The period of two weeks is considered necessary to characterize, in a reliable way, the beginning of the phenomenon (initial conditions in the light of the dynamic phenomena in the field of engineering)

- In each time step, following the 14 day period from the first death record, the optimal data simulation curve is calculated with the use of the proposed algorithm. Figure 1 shows the optimal curve that best simulates the data (number of deaths) that precede the time of the prediction (12 2-day periods or 24 days).

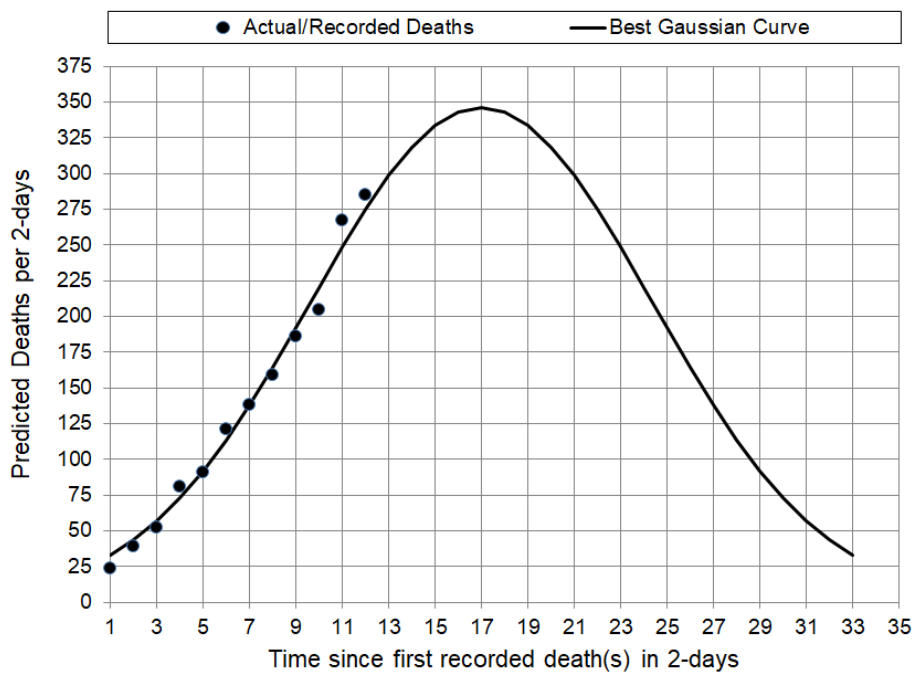

Figure 1. Prediction of the best Fit curve of the actual deaths at the time of prediction

- The same procedure is being applied on a daily basis and based on the values of the maximum number of deaths and on the time when this maximum is attained we plot the curve of Figure 2 This figure is very useful as it illustrates how the phenomenon evolves by providing us with an estimate of when the phenomenon is expected to peak, even with the number of deaths. This information is especially useful to the authorities because it helps them prepare accordingly to deal with it.

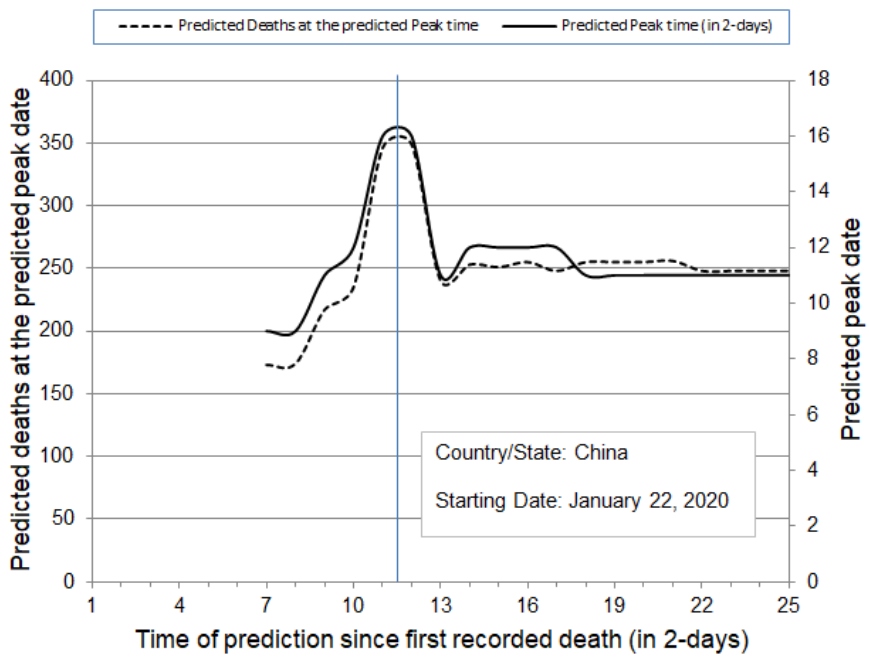

Figure 2. Prediction of number of deaths *(dashed line), in 2-day time intervals, and peak-date of deaths (solid line) for the country of China*

- By predicting the optimal curve of Figure 1, we are also provided with information about the dynamics of the phenomenon. In particular, knowing the parameters of the distribution (sigma, mi and fitting probability) its area is calculated which shows the total predicted number of deaths. Based on the percentage change in the number of deaths as a

function of time, the change in the dynamics of the phenomenon is defined (Figure 3). In this figure it is intensely demonstrated that the COVID-19 phenomenon is a predominantly dynamic phenomenon with clear dynamic characteristics which oscillates strongly during its transition to the peak and then dissipates.

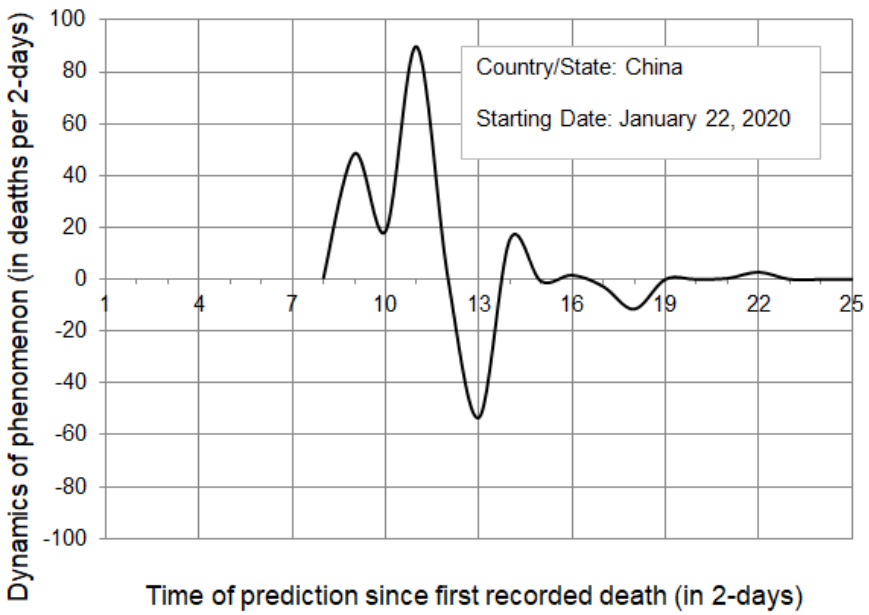

Figure 3. Prediction of the dynamics of COVID-19 phenomenon for the country of China

- In addition to the above useful estimates and the revelation of the dynamic characteristics of the phenomenon using the proposed heuristic algorithm, it is possible to reliably predict the expected number of deaths for the next 10 days Figure 4. Simultaneously, with the estimated expected number of deaths we get an estimate for its higher and lower limits. Based on a comprehensive study in all ten countries and the results that were

procured and, will be presented below, these limits were confirmed for all countries and cities as well as the difference between the predicted and actual deaths.

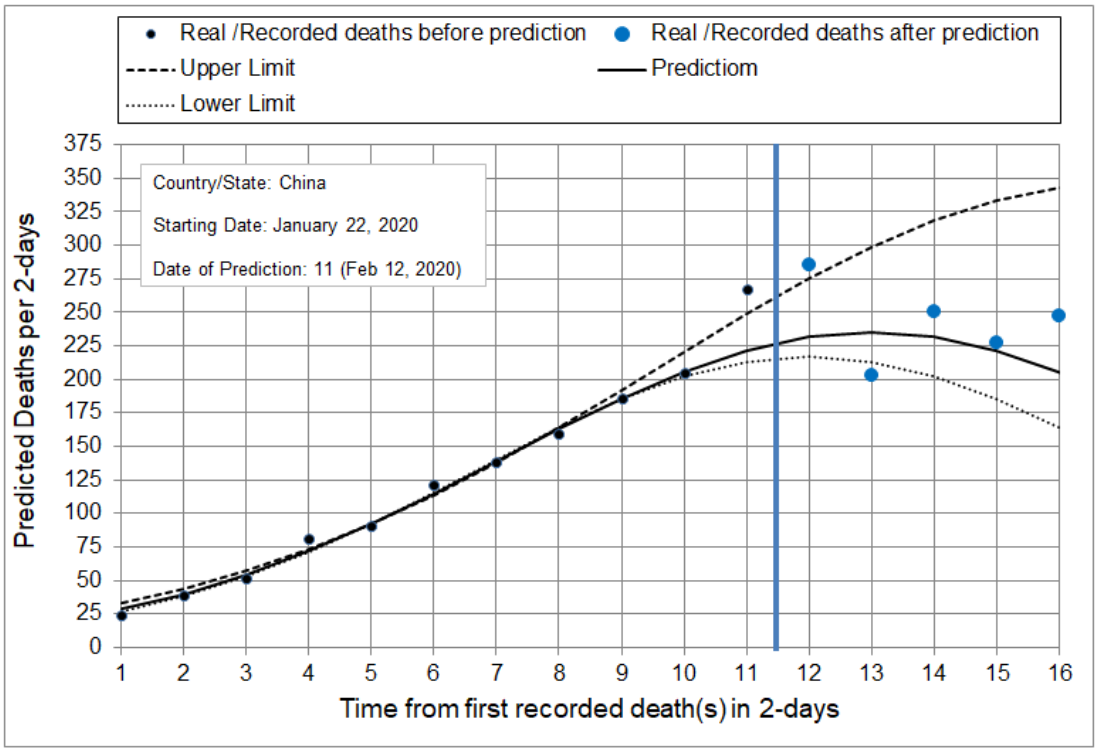

Figure 4. Prediction of number of deaths, in 2-day intervals, for the next ten days starting February 12, 2020, for the country of China. Black dots represent actual data until the day in which the algorithm made the prediction. Blue dots represent actual data after the day in which the algorithm made the prediction.

Results and Discussion

In the light of the above, a new computer software has been developed in Computational Mechanics Laboratory, School of Pedagogical and Technological Education, Athens, Greece under the supervision of Prof. Asteris. Utilizing this software through implementation of the heuretic algorithm, the behavioral development of the epidemic was investigated in ten different geographical locations. In particular, COVID-19 development trend was examined in California,

Germany, Greece, Iran, New York, Sweden, United Kingdom and United States. The investigation was implemented in two stages. In the first stage during the previous time periods, the data as well as the daily mortality rates and results were known. This stage was selected to evaluate the phenomen, but more importantly to document the proposed heuretic global algorithm and the respective methodology. In the second stage, predictions were made for which the results are unknown. In detail, for the total of the above-mentioned locations the number of daily deaths for the next 10 consecutive days, from the start of 13th of April 2020. The results are presented in detail for every situation in tables and in figures in supplementary materials.

The main results of our study are the following:

- The proposed algorithm was confirmed absolutely for the total predictions implemented for the first phase, where data and results of daily mortality rates were known.
- The proposed methodology provides an upper and a lower estimation limit, which was confirmed for the total cases of the first stage predictions.
- For the predictions of the secondary stage and more specifically from the 13th until the 22nd of April the confirmation or disproof is expected based on the respective predictions provided by the Institute for Health Metrics and Evaluation (IHME). These were performed on the 12th of April, the same day of the current study, The upper and lower estimation limit is in our predictions much smaller than the respective limits provided by IHME as evident in
- Table ***1***.
- Noteworthy, in confirmation of the above-mentioned statements, using the diagrams where all the development of the phenomenon is presented, additional information is

provided for the efficacy and efficiency of the governmental restrictive measures (lockdown) for further handling/dealing with them. A typical example is/indicates the case of China, where the dynamics impact diminishes 23 days after the implementation of restrictive measures.

Nr	Country	Present work		IHME	
		Minimum	Maximum	Minimum	Maximum
1	California	42	101	124	185
2	Germany	50	119	461	756
3	Greece	0	1	25	28
4	Iran	20	23	NA	NA
5	Lombardia	7	17	3	37
6	New York	176	234	376	1970
7	Sweden	193	388	377	531
8	UK	446	717	1668	4008
9	USA	1058	2652	4017	6620

Table 1. Minimum and maximum range of predicted daily deaths, compared with corresponding predictions by Institute for Health Metrics and Evaluation (IHME)

Conclusions

The aim of the current study was to develop a novel algorithm for estimating and predicting the COVID-19 pandemic outburst for up to 10 days after the prediction date.

Although so far there are available reports on country-specific models, taking into consideration the specific features of each country or region, this novel global algorithm is of particular importance as it applies in all 10 different locations examined with exceptional characteristics and relevance. Noteworthy, since there are significant differences in many aspects among countries as well as unique local transmission dynamics, there are difficulties in designing an algorithm that is able to predict with high confidence the outcome of the outbreak for the next

period of time in all examined locations. Interestingly, this global prediction tool applies to all countries tested although every country has different internal political characteristics in response to coronavirus crisis (big variation in how well governments are responding), and as such the impact of SARS-CoV-2 infection pattern is not evenly distributed.

This information, by providing a tool for estimating and predicting the development of the pandemic, is of high relevance and paramount importance for governments and local authorities for taking key decisions as whether to extend quarantine or relax social distancing control measures. While disease is growing exponentially, the health-care system faces several burdens. Based on the proposed predictions, governments can be prepared, plan and act immediately to ensure adequate health-care and reduced mortality risk due to COVID-19.

The proposed algorithm is expected to make a substantial contribution to engineering problems, where it is frequent that the parameters of a multitude of engineering problems follow a normal distribution. Authors also believe that since data/parameters referring to other "related families of viruses" of COVID-19 appear to have a normal distribution, the proposed algorithm will be universally applicable. The authors have begun to investigate in this direction and very soon results will be presented on a companion paper.

References

1. WHO Rolling updates on coronavirus disease (COVID-19): WHO characterizes COVID-19 as a pandemic. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen

   Date accessed: March 30, 2020
2. https://www.worldometers.info/about/
3. https://covidtracking.com/about-data
4. https://github.com/pcm-dpc/COVID-19/tree/master/schede riepilogative/regioni?fbclid=IwAR1XzjDqpNWgF4o94ysxYfv1k-7HTv1zERT33EAGq9NJ4gupbTAygTQEXsk

Tables

Table 2 Predicted daily deaths, in the time interval April13 to April 22, 2020, obtained from our model compared with corresponding predictions by Institute for Health Metrics and Evaluation (IHME)

Country	Model	Prediction	April									
			13	14	15	16	17	18	19	20	21	22
California	Proposed Algorithm	Upper Limit	77	77	87	87	98	98	108	108	118	118
		Prediction	60	60	62	62	63	63	62	62	60	60
		Lower Limit	35	35	31	31	27	27	22	22	17	17
	IHME	Upper Limit	167	166	177	185	186	192	192	179	154	137
		Prediction	45	46	48	50	51	51	52	51	50	48
		Lower Limit	2	1	2	3	5	7	8	10	11	13
	Actual (reported)	Prediction										
Germany	Proposed Algorithm	Upper Limit	173	173	141	141	109	109	79	79	54	54
		Prediction	113	113	80	80	52	52	32	32	18	18
		Lower Limit	54	54	33	33	18	18	9	9	4	4
	IHME	Upper Limit	829	766	735	703	667	634	593	547	524	494
		Prediction	272	229	224	217	210	201	192	182	173	164
		Lower Limit	73	39	39	39	39	39	38	37	36	33
	Actual (reported)	Prediction										
Greece	Proposed Algorithm	Upper Limit	3	2	2	2	2	1	1	1	1	
		Prediction	3	2	2	2	2	1	1	1	1	
		Lower Limit	2	1	1	1	1	1	1	0	0	
	IHME	Upper Limit	27	28	29	29	30	30	30	30	30	30
		Prediction	8	8	9	9	10	10	11	11	11	11
		Lower Limit	2	2	2	2	2	3	3	3	3	3
	Actual (reported)	Prediction										
Iran	Proposed Algorithm	Upper Limit	98	86	86	74	74	63	63	52	52	

		Prediction	98	86	86	74	74	63	63	52	52	
		Lower Limit	76	63	63	51	51	41	41	32	32	
	IHME	Upper Limit										
		Prediction										
		Lower Limit										
	Actual (reported)	Prediction										
Lombardia	Proposed Algorithm	Upper Limit	97	97	68	68	46	46	30	30	19	19
		Prediction	97	97	68	68	46	46	30	30	19	19
		Lower Limit	80	80	54	54	34	34	21	21	12	12
	IHME	Upper Limit	107	67	54	43	34	26	20	15	11	8
		Prediction	107	67	54	43	34	26	20	15	11	8
		Lower Limit	70	37	30	24	19	15	12	9	7	5
	Actual (reported)	Prediction										
New York	Proposed Algorithm	Upper Limit	775	704	704	600	600	479	479	359	359	
		Prediction	770	699	699	596	596	476	476	357	357	
		Lower Limit	599	488	488	366	366	254	254	162	162	
	IHME	Upper Limit	2163	1952	1742	1518	1292	1079	886	716	575	453
		Prediction	743	651	595	536	474	412	353	297	246	200
		Lower Limit	193	142	142	140	135	130	118	104	91	77
	Actual (reported)	Prediction										
Sweden	Proposed Algorithm	Upper Limit	197	197	239	239	285	285	335	335	388	388
		Prediction	52	52	40	40	28	28	18	18	11	11
		Lower Limit	4	4	2	2	1	1	0	0	0	0
	IHME	Upper Limit	420	422	425	428	432	446	498	520	554	613
		Prediction	150	142	150	159	169	179	191	204	218	234

		Lower Limit	15	40	46	50	55	61	67	72	76	82
	Actual (reported)	Prediction										
		Upper Limit	1180	1180	1227	1227	1227	1180	1180	1180	1051	1051
	Proposed Algorithm	Prediction	967	967	924	924	924	808	808	808	646	646
		Lower Limit	734	734	627	627	627	482	482	482	334	334
UK		Upper Limit	4255	4115	3953	3706	3431	3118	2797	2484	2167	1868
	IHME	Prediction	1156	1079	1067	1040	1002	952	895	830	760	687
		Lower Limit	247	177	191	201	207	209	210	213	208	200
	Actual (reported)	Prediction										
		Upper Limit	2965	2965	2965	3434	3434	3434	3813	3813	3813	4061
	Proposed Algorithm	Prediction	2755	2755	2755	3097	3097	3097	3321	3321	3321	3400
		Lower Limit	1907	1907	1907	1844	1844	1844	1667	1667	1667	1409
USA		Upper Limit	7084	6771	6564	6278	5960	5631	5312	4989	4684	4386
	IHME	Prediction	2150	1953	1928	1888	1838	1779	1715	1648	1582	1518
		Lower Limit	464	292	307	326	340	350	357	362	365	369
	Actual (reported)	Prediction										

IHME: Institute for Health Metrics and Evaluation

Figures

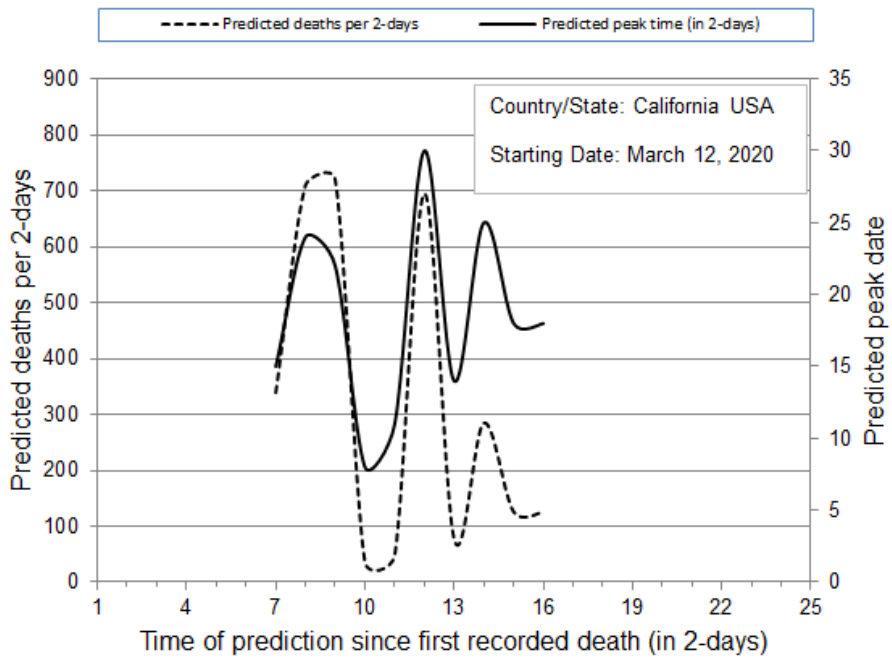

Figure 5. Prediction of number of deaths (dashed line), in 2-day time intervals, and peak-date of deaths (solid line) for the state of California, USA

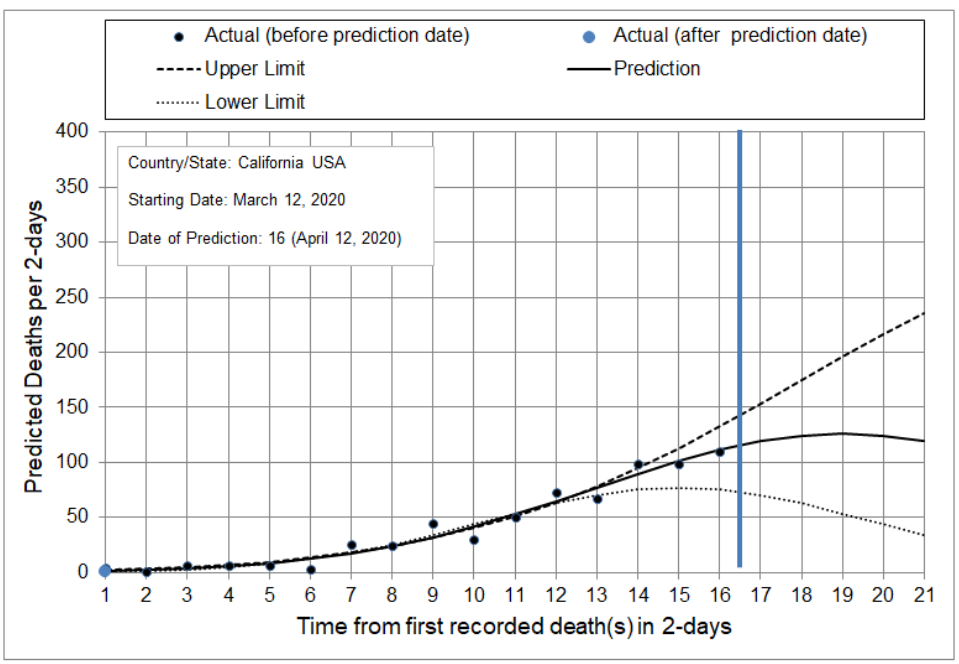

Figure 6. Prediction of number of deaths, in 2-day intervals, for the next ten days starting April 13, 2020, for the state of California, USA. Black dots represent actual data until the day in which the algorithm made the prediction. Blue dots represent actual data after the day in which the algorithm made the prediction.

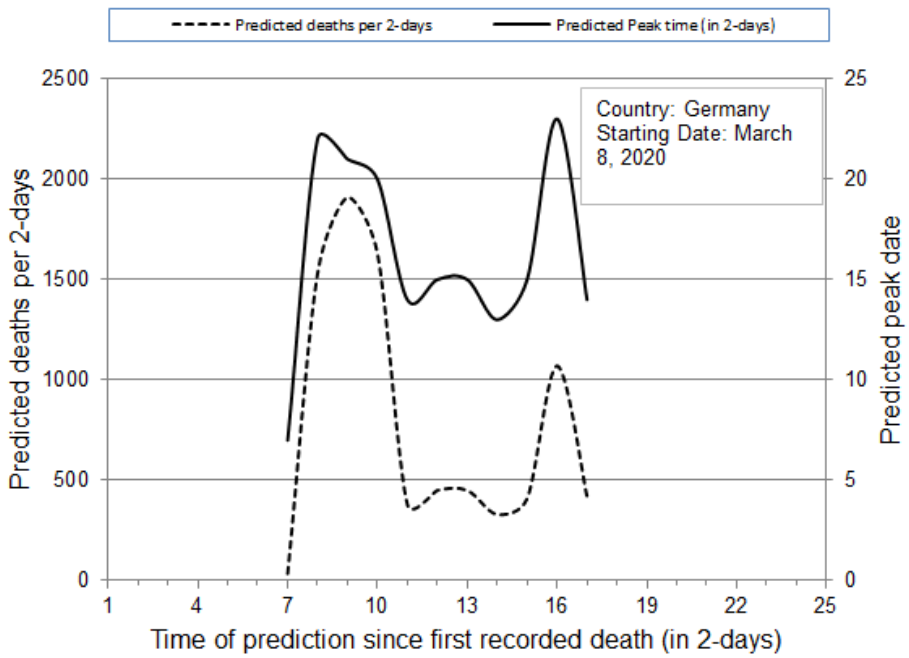

Figure 7. Prediction of number of deaths (dashed line), in 2-day time intervals, and peak-date of deaths (solid line) for the country of Germany

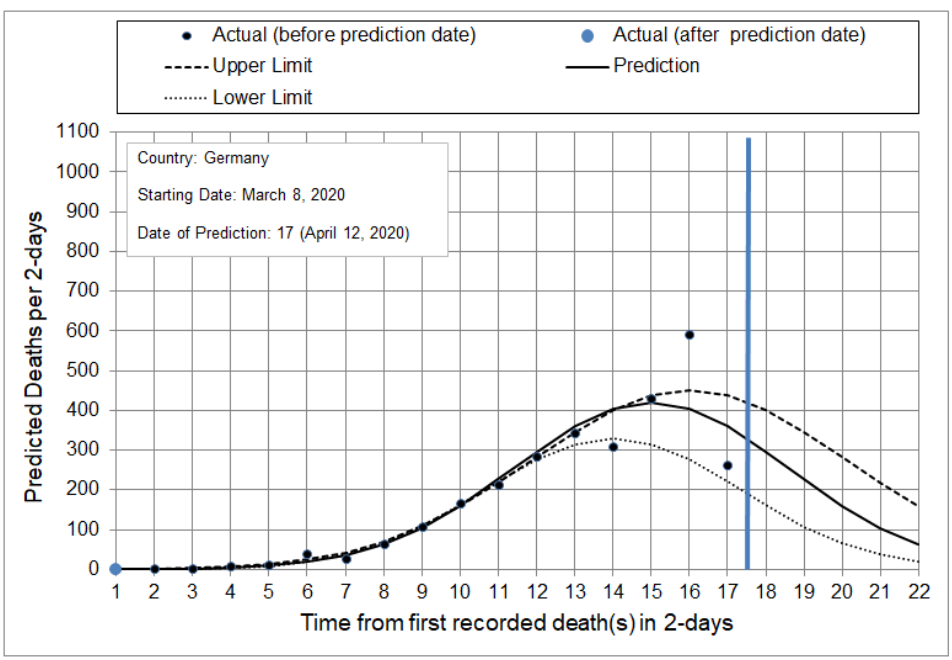

Figure 8. Prediction of number of deaths, in 2-day intervals, for the next ten days starting April 13, 2020, for the country of Germany. Black dots represent actual data until the day in which the algorithm made the prediction. Blue dots represent actual data after the day in which the algorithm made the prediction.

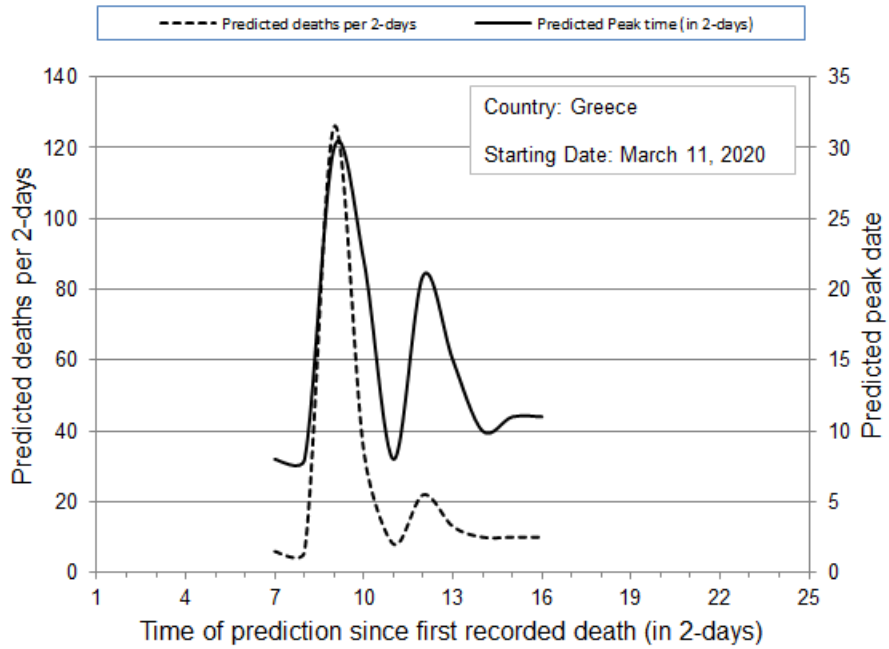

Figure 9. Prediction of number of deaths (dashed line), in 2-day time intervals, and peak-date of deaths (solid line) for the country of Greece

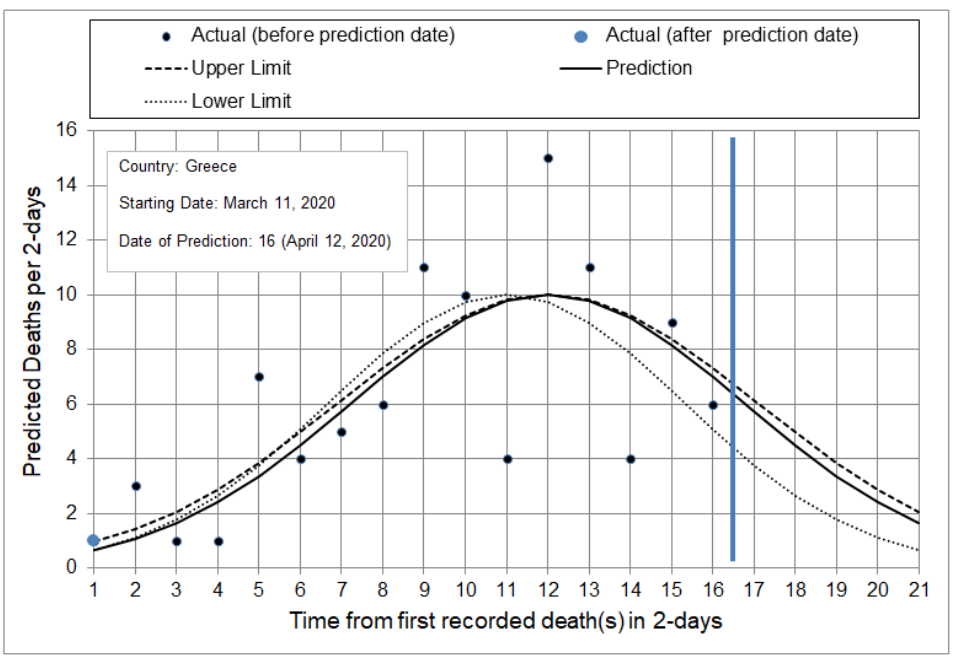

Figure 10. Prediction of number of deaths, in 2-day intervals, for the next ten days starting April 13, 2020, for the country of Greece. Black dots represent actual data until the day in which the algorithm made the prediction. Blue dots represent actual data after the day in which the algorithm made the prediction.

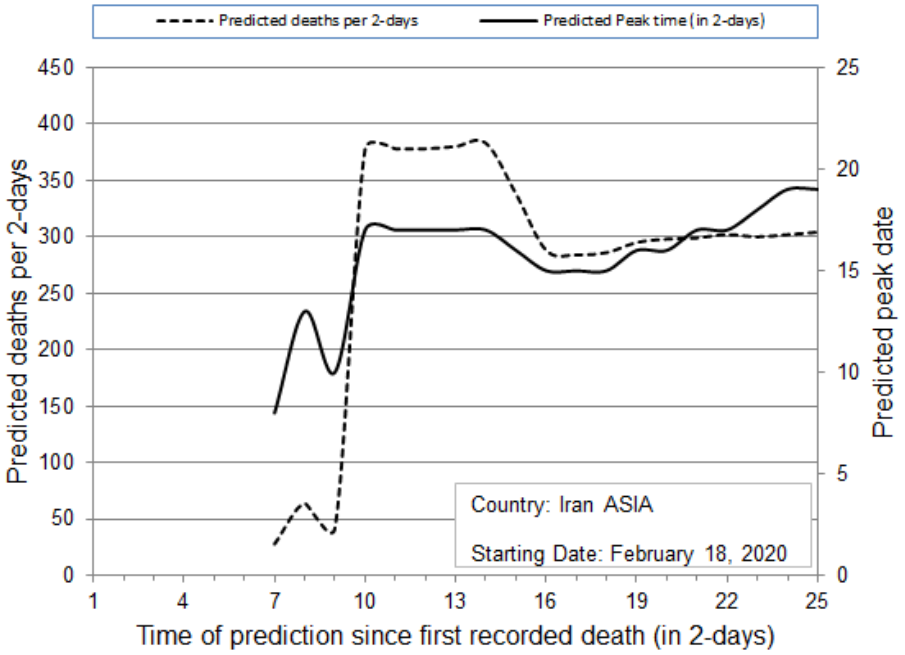

Figure 11. Prediction of number of deaths (dashed line), in 2-day time intervals, and peak-date of deaths (solid line) for the country of Iran

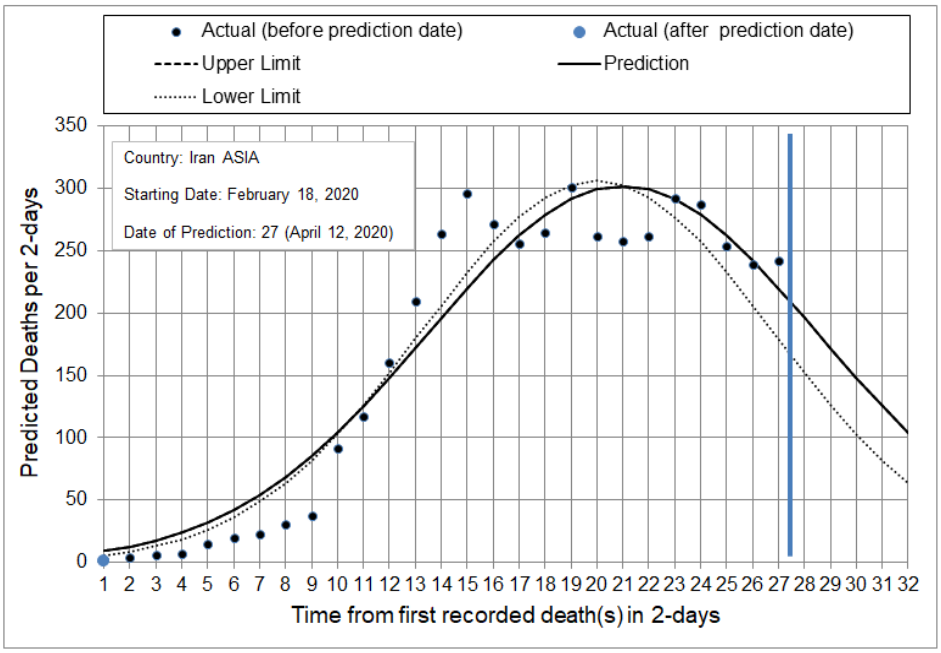

Figure 12. Prediction of number of deaths, in 2-day intervals, for the next ten days starting April 13, 2020, for the country of Iran. Black dots represent actual data until the day in which the algorithm made the prediction. Blue dots represent actual data after the day in which the algorithm made the prediction.

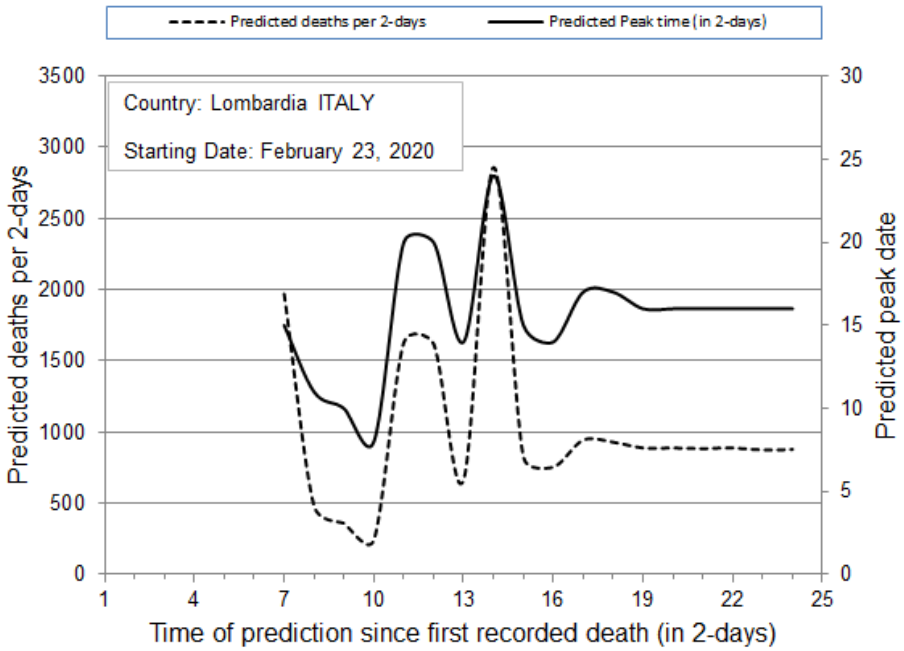

Figure 13. Prediction of number of deaths (dashed line), in 2-day time intervals, and peak-date of deaths (solid line) for the district of Lombardia, Italy

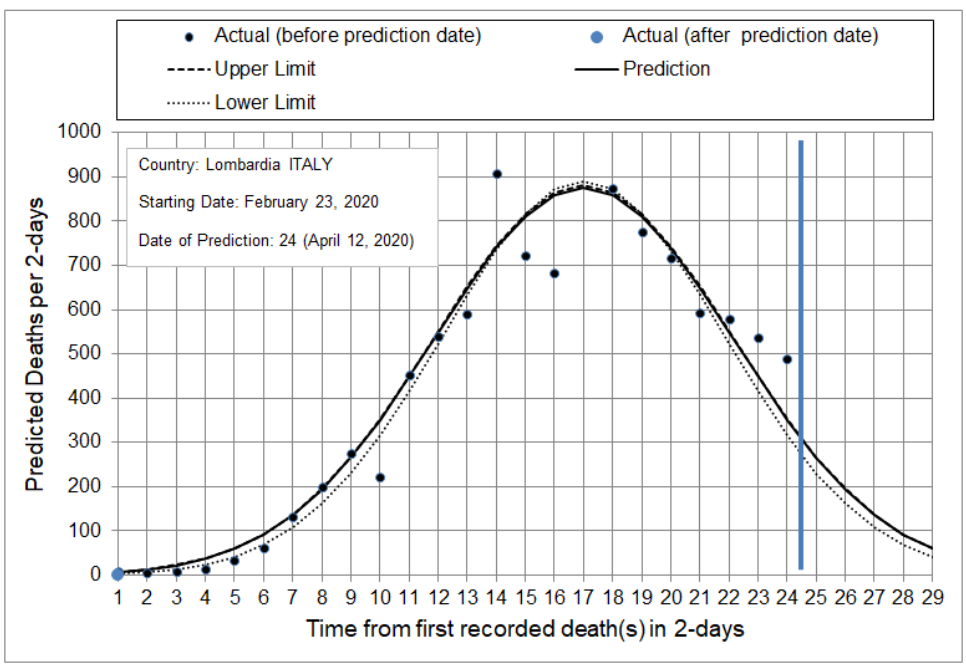

Figure 14. Prediction of number of deaths, in 2-day intervals, for the next ten days starting April 13, 2020, for the district of Lombardia, Italy. Black dots represent actual data until the day in which the algorithm made the prediction. Blue dots represent actual data after the day in which the algorithm made the prediction.

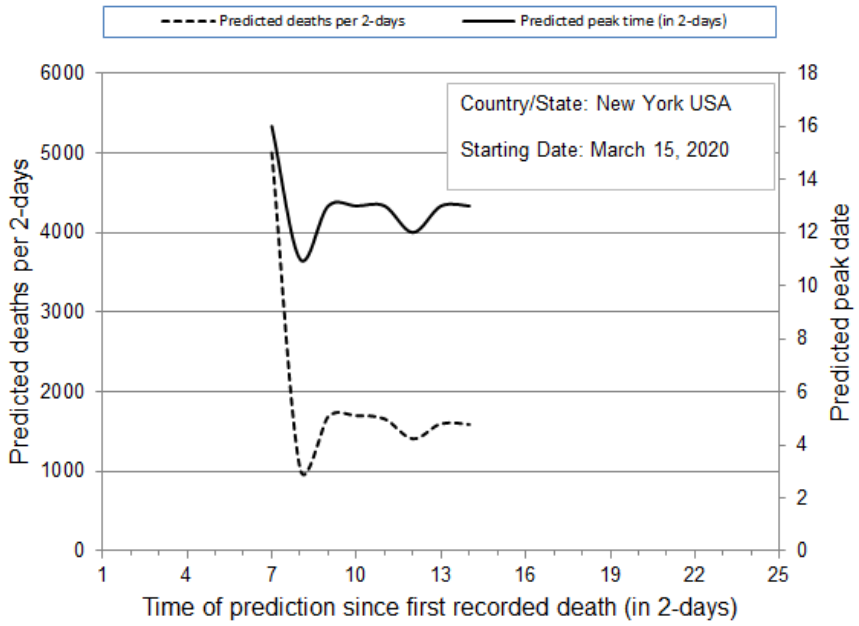

Figure 15. Prediction of number of deaths (dashed line), in 2-day time intervals, and peak-date of deaths (solid line) for the state of New York, USA

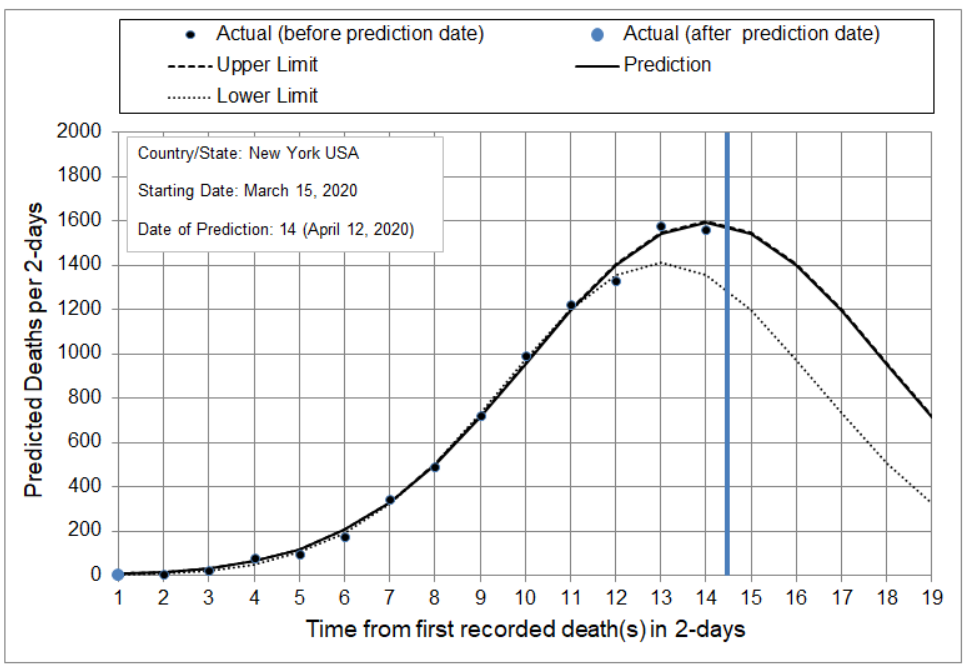

Figure 16. Prediction of number of deaths, in 2-day intervals, for the next ten days starting April 13, 2020, for the state of New York, USA. Black dots represent actual data until the day in which the algorithm made the prediction. Blue dots represent actual data after the day in which the algorithm made the prediction.

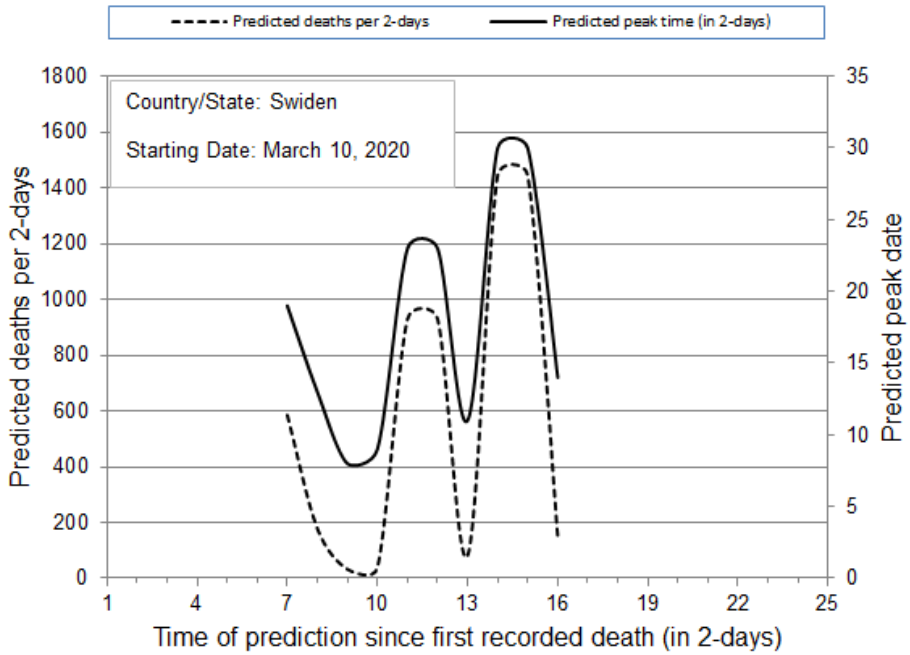

Figure 17. Prediction of number of deaths (dashed line), in 2-day time intervals, and peak-date of deaths (solid line) for the country of Sweden

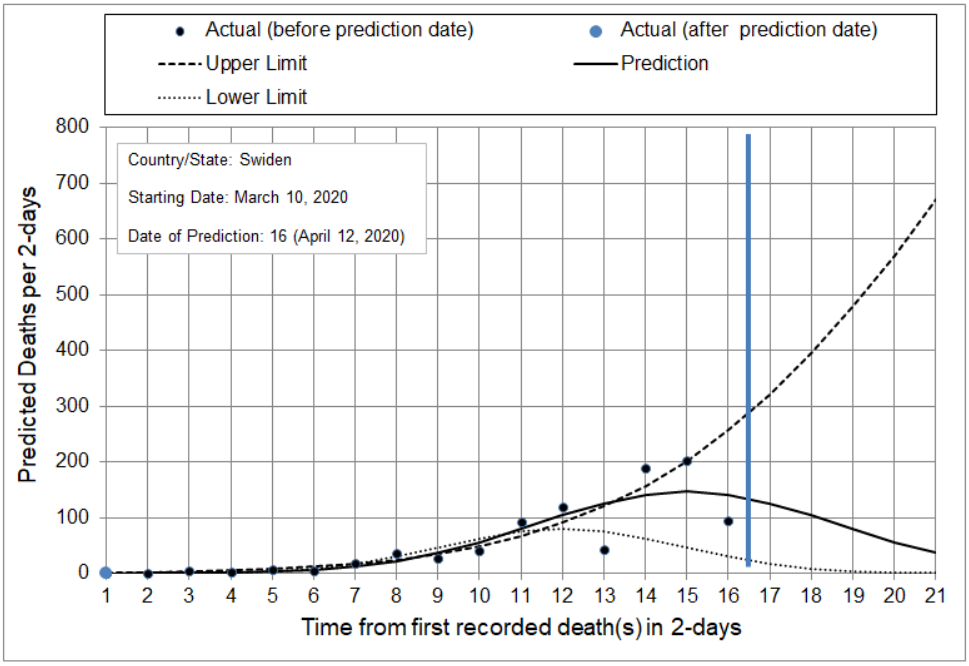

Figure 18. Prediction of number of deaths, in 2-day intervals, for the next ten days starting April 13, 2020, for the country of Sweden. Black dots represent actual data until the day in which the algorithm made the prediction. Blue dots represent actual data after the day in which the algorithm made the prediction.

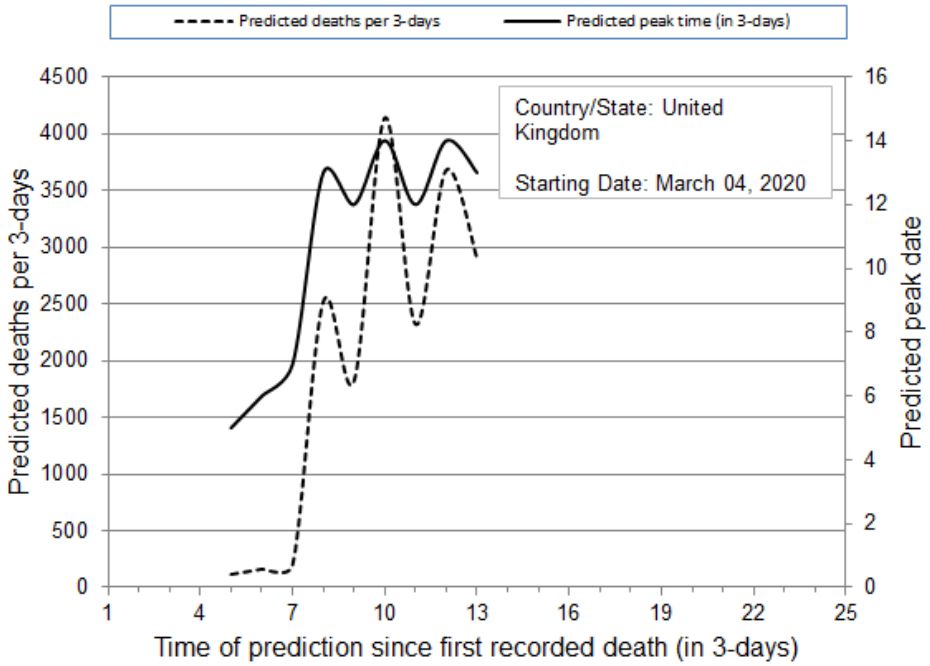

Figure 19. Prediction of number of deaths (dashed line), in 3-day time intervals, and peak-date of deaths (solid line) for the country of United Kingdom

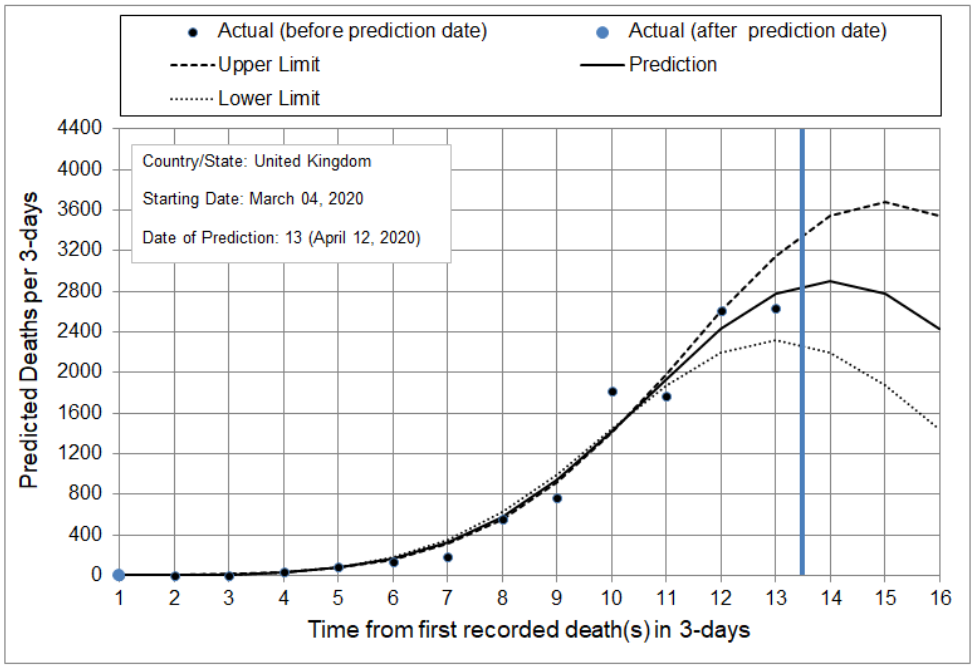

Figure 20. Prediction of number of deaths, in 3-day intervals, for the next ten days starting April 13, 2020, for the country of United Kingdom. Black dots represent actual data until the day in which the algorithm made the prediction. Blue dots represent actual data after the day in which the algorithm made the prediction.

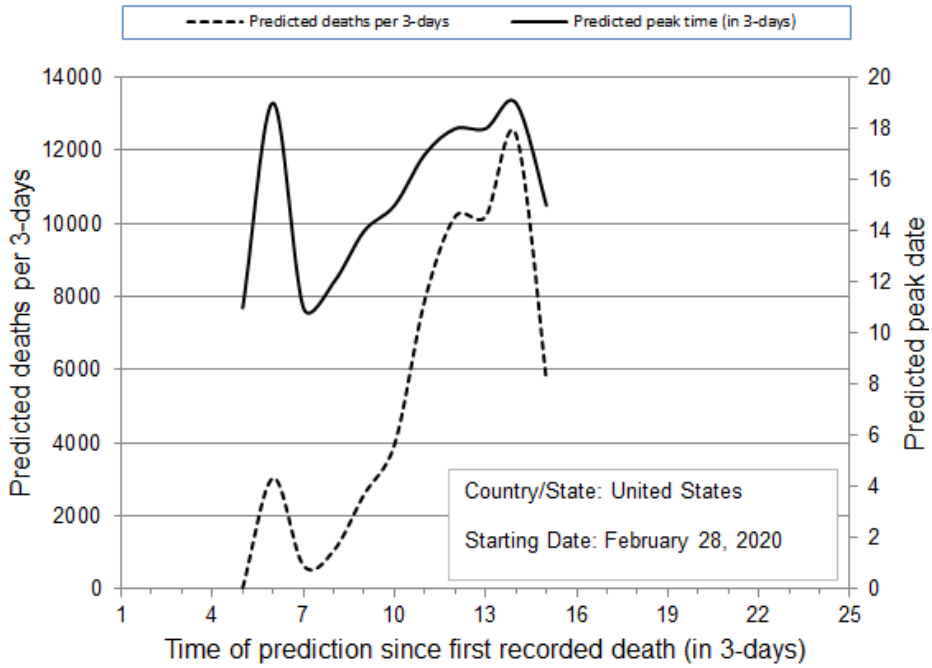

Figure 21. Prediction of number of deaths (dashed line), in 3-day time intervals, and peak-date of deaths (solid line) for the country of USA

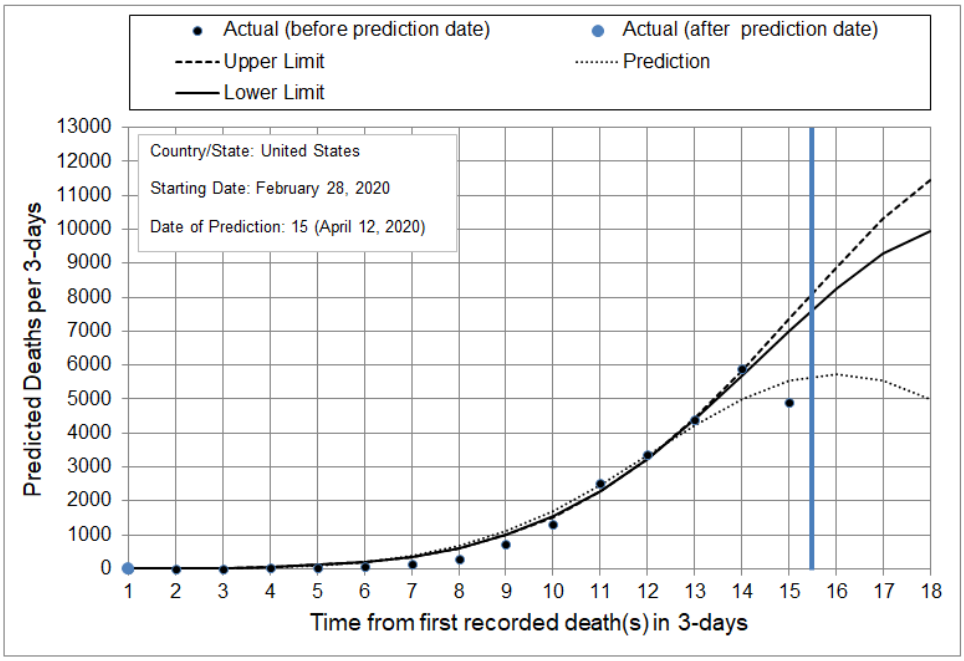

Figure 22. Prediction of number of deaths, in 3-day intervals, for the next ten days starting April 13, 2020, for the country of USA. Black dots represent actual data until the day in which the algorithm made the prediction. Blue dots represent actual data after the day in which the algorithm made the prediction.