BMJ Open

Novel model prediction time-to-event analysis: data validation and estimation of 200 million cases in the global COVID-19 epidemic

Ali Rezania,[1] Elaheh Ghorbani,[2] Davood Hassanian-Moghaddam,[1] Farnaz Faeghi,[2] Hossein Hassanian-Moghaddam [3,4]

To cite: Rezania A, Ghorbani E, Hassanian-Moghaddam D, *et al*. Novel model prediction time-to-event analysis: data validation and estimation of 200 million cases in the global COVID-19 epidemic. *BMJ Open* 2023;**13**:e065487. doi:10.1136/bmjopen-2022-065487

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2022-065487).

Received 08 June 2022
Accepted 14 December 2022

For numbered affiliations see end of article.

Correspondence to
Dr Hossein Hassanian-Moghaddam;
hassanian@sbmu.ac.ir

ABSTRACT

Objectives Assessment of recuperation and death times of a population inflicted by an epidemic has only been feasible through studying a sample of individuals via time-to-event analysis, which requires identified participants. Therefore, we aimed to introduce an original model to estimate the average recovery/death times of infected population of contagious diseases without the need to undertake survival analysis and just through the data of unidentified infected, recovered and dead cases.

Design Cross-sectional study.

Setting An internet source that asserted from official sources of each government. The model includes two techniques—curve fitting and optimisation problems. First, in the curve fitting process, the data of the three classes are simultaneously fitted to functions with defined constraints to derive the average times. In the optimisation problems, data are directly fed to the technique to achieve the average times. Further, the model is applied to the available data of COVID-19 of 200 million people throughout the globe.

Results The average times obtained by the two techniques indicated conformity with one another showing p values of 0.69, 0.51, 0.48 and 0.13 with one, two, three and four surges in our timespan, respectively. Two types of irregularity are detectable in the data, significant difference between the infected population and the sum of the recovered and deceased population (discrepancy) and abrupt increase in the cumulative distributions (step). Two indices, discrepancy index (DI) and error of fit index (EI), are developed to quantify these irregularities and correlate them with the conformity of the time averages obtained by the two techniques. The correlations between DI and EI and the quantified conformity of the results were −0.74 and −0.93, respectively.

Conclusion The results of statistical analyses point out that the proposed model is suitable to estimate the average times between recovery and death.

STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ Data were drawn from the Kaggle website comprised of cumulative infected, recovered and deceased population affected by COVID-19.
⇒ We used the curve fitting technique to propose a multimodal delayed outcome-based compartmental technique.
⇒ The process of curve fitting functions as a denoising agent to facilitate achieving a better compartmental model.
⇒ The study does not have individual-level data, nor an individual assessment of digital access, use or competency.
⇒ The study is limited by being cross-sectional rather than interventional or prospective.

INTRODUCTION

Virus epidemics have been known as one of the major health issues leading to a high mortality rate in human communities throughout history. The Spanish influenza emerged in 1918, caused about 50 millions deaths for just over 2 years.[1] Also, since the early reports of HIV infection in 1980, more than 36 million of deaths have been reported around the world due to virus infection until the end of 2020.[2] Tragically, not only these older global epidemics but also the local spread of SARS, MERS and Ebola viruses in recent years causing disruptions in functions of societies, implies that the health authorities are not ready yet for such crises.[3]

The COVID-19 is a new strain that has not been previously identified in human. Most people infected with the COVID-19 virus, the cause of the recent pandemic across the globe would experience mild-to-moderate respiratory illness and recover without requiring special treatment. Older people, and those with underlying medical problems such as cardiovascular diseases, diabetes, chronic respiratory diseases and cancer, are more likely to develop serious illnesses.[4] Since December 2019, a growing number of cases for the COVID-19, a worldwide health disaster of this time, has been discovered initially in China.

A variety of models and simulations have developed as important decision tools that can be useful to control human and animal diseases.[5] However, since each disease exhibits its own particular biological characteristics, the models need to be adapted to each specific case in order to be able to tackle real situations.[6]

Among the epidemiological models, the most common model used to study the spread of COVID-19 are the compartmental models. In such models, there is a susceptible population, which is assumed to be equal to the population of whichever region is being examined minus the number of people that have previously had the disease. Some of the susceptible individuals get infected in each period, where the rate of infection is a function of the number of infected individuals as well as other factors that shift the rate of transmission. Finally, infected individuals, according to the selected model, either move to the removed compartment or again to the susceptible compartment.**7 8**

Some prevalent forms of compartmental models are SIR, SI, SIS, SIRS, SEIR, SEIRS, MSIR, MSEIR and MSEIRS, among others. In all these acronyms, S, I, R, E and M stand for susceptible, infected, removed, exposed (individuals already exposed to the disease but not infected yet) and maternal (those with maternal immunity), respectively.[9] These models have been applied to many emerging infectious diseases, for example, avian influenza, Ebola, HIV/AIDS and many others. The inclusion of different compartments is based on the nature of the diseases or the temporary stage of the epidemic.[10]

In order to be able to study the COVID-19 comprehensively, the removed compartment splits into two subcompartments, recovered (R) and deceased (D), as due to the noticeable case fatality rate (CFR) of the COVID-19, the number of deceased individuals becomes important for statistical analyses.[11] Delayed SIR-alike models are introduced to account for the time delays of transmissions of individuals between different classes.[12 13] Distribution fitting is the process of fitting a probability distribution to a series of data concerning the repeated measurement of a variable phenomenon. Distribution fitting aims to predict the probability or to forecast the frequency of occurrence of the magnitude of the phenomenon in a certain interval. There are many probability distributions of which some can be fitted more closely to the observed frequency of the data than others, depending on the characteristics of the phenomenon and of the distribution. The distribution giving a close fit is supposed to lead to good predictions. The conditions that need to be considered in the selection of probability distribution are the general trend, skewness (symmetry or asymmetry of the data), etc.[14]

The importance of estimation of the time averages to recovery and death, either via survival analysis or modelling, emerges where the facilities plan to provide the necessary supplies. Hence, in this study, we try to present a model to estimate these time averages using the process of curve fitting and confirm it by an optimisation problem model.

METHOD

Data preparation

Data were drawn from the Kaggle website comprised of the date of the observation, country-wise separation, segregation by state or province (if provided by the country) and cumulative infected, recovered and deceased population affected by COVID-19.[14] In order to mitigate the disjunction in the data, the data underwent a procedure that if there is a descent in the cumulative data, the data of the inconsistent day would be replaced by a weighted average of the data of the day before inconsistency and the next valid data. As another modification, for countries with state or province-wise separation of the data, we added up the data of these states or provinces for the same day to get a dataset for the whole country. If the data of a country were missing before or after a certain day, that day would be appointed as the onset or terminal day in dataset for that country. The mathematical method that is used for data preparation is provided in online supplemental information.

Patient and public involvement

It was not possible to involve patients or the public in the design, or conduct, or reporting, or dissemination plans of our research.

Data categorisation

In order to be able to study the COVID-19 comprehensively, the retrieved compartment of SIR model splits into two subcompartments, R and D, because due to the noticeable CFR of the COVID-19, the number of deceased individuals becomes important for statistical analyses.[11] Delayed SIR-alike models are taken to account for the time delays of transmissions of individuals between different classes.[12 13] In this study, we focus on the average time of incubation, recovery and death of individuals, according to a delayed susceptible–infected–recovered/deceased (SIRD) model given in system of equations (1)–(4).

$$\dot{S}(t) = -\alpha S(t)\, I(t-\tau_1) \tag{1}$$

$$\dot{I}(t) = \alpha S(t)\, I(t-\tau_1) - \beta R(t-\tau_2) - \gamma D(t-\tau_3) \tag{2}$$

$$\dot{R}(t) = \beta I(t-\tau_2) \tag{3}$$

$$\dot{D}(t) = \gamma I(t-\tau_3) \tag{4}$$

Where S, I, R and D stand for the susceptible, infected, recovered and deceased compartments, respectively.

α is the average incubation rate of an individual.

β is the average recovery rate of an individual.

γ is the average death rate of an individual.

τ_1 is the average incubation time of a susceptible individual.

τ_2 is the average recovery time of an infected individual.

τ_3 is the average death time of an infected individual.

Rezania A, *et al*. *BMJ Open* 2023;**13**:e065487. doi:10.1136/bmjopen-2022-065487

Curve fitting and goodness of fit (GoF)

After modification of the secondary data, the frequency distribution of each category of the secondary data was plotted against the number of days from the onset of the outbreak to determine the number of peaks to choose to fit a cumulative Gaussian distribution to each of these classes.

The validity of the fitted models was checked via three GoF indices, namely Pearson's reduced χ^2 statistics, root mean square error of approximation and standardised root mean square residual.

The cumulative Gaussian fitted to the modified secondary data for some countries could not pass the GoF criterion. So, by looking at the frequency distribution of these countries, it was decided to put skewed normal distribution and hence its cumulative function to use.

The steps to obtain time averages to recovery and death via curve fitting is explained in the following steps to be replicable:

1. Download the secondary data from the internet source.
2. List the countries in the dataset.
3. Pick a country to study.
4. If the data is given state wise or province wise, add up number of people in each category.
5. If the data (given cumulatively) show a descent, modify the data according to online supplemental equations s1 and s2.
6. Determine the number of waves according to the frequency plot of the three categories of the data (visually determined).
7. Determine the initial guesses of the parameters given in online supplemental equation s4 for each wave.
8. Minimise online supplemental equations s6–s8 simultaneously with regard to the constraints given in (online supplemental equations s9–s11.
9. Calculate the GoF indices and determine whether the fit is acceptable according to the defined criteria of the indices or not.
10. If the fit was acceptable proceed to step 14; if not acceptable, then minimise the skewed Gaussian error functions simultaneously with regard to the constraints (online supplemental equation s11, s20 and s21) (similar to step 8).
11. Repeat step 9. If the criteria are met, proceed to step 14; otherwise calculate the discrepancy (equation (5)) and step (equation (6)) indices.
12. Fit the last function fitted to the data but this time omit constraint (online supplemental equation s11) (checking for steps in the data).
13. Check whether the discrepancy and step indices pass the criteria to identify the incoherency in the data.
14. Subtract the mean of infected peak from the means of recovered and deceased peaks to attain the average time to recovery and time to death of the peak rendered by curve fitting method, respectively. Repeat for the number of waves determined in step 6.

All detailed procedure about Gaussian and skewed Gaussian distributions, and GoF can be found in online supplemental information.

SIRD compartmental model

A quite suitable compartmental model was selected for our study, by the process of elimination of irrelevant ones. SI and SIS models were eliminated because infected individuals either recover or die,[15] so the models excluding the removed compartment are unfit. The models involving exposed compartment are unsuitable for our study, as there is no data for a mediate class between susceptible and infected group in our dataset. Additionally, accounting for the exposed compartment requires predictions and estimations.[16] Since there is not enough evidence of maternal immunity and also lack of data,[17] the models involving maternal compartment would be eliminated. Finally, the models such as SIRS, in which the individual from the removed compartment, would be retransmitted to the susceptible compartment is ruled out, because there is not a strong claim for reinfection of population of the removed compartment in a single surge.[18] The SIR version of compartmental models seemed appropriate for our dataset, where the removed compartment was divided into two moieties, recovered and deceased.

The numerical solution method and the computation software are introduced in the online supplemental information.

Optimisation problem

The objective here is to find the optimum values of τ_2 and τ_3, and to compare these values with those achieved as explained in Gaussian fit and skewed Gaussian fit sections. For this purpose, two optimisation problems were defined (method I and method II) and discussed in the online supplemental information.

The steps to obtain time averages to recovery and death via optimisation problems are explained in the following steps:

1. Download the secondary data from the internet source.
2. List the countries in the dataset.
3. Pick a country in the dataset.
4. If the data is given state wise or province wise, add up number of people in each category.
5. If the data (given cumulatively) show a descent, modify the data according to equations (online supplemental equations s1 and s2).
6. If the number of waves is more than one, the time interval of each peak is found out by utilising equations (online supplemental equations s37 and s38).
7. Method I: Assign the initial value of t_3 (the initial value of t_3 is t_3^1).
8. Minimise equation (online supplemental equation s25) and obtain t_2^1 for each day (t) (t_3^1 is fixed). Note that $t+t_2^1$ cannot exceed the number of days from the onset of pandemic or the limit set in step 6.

9. Minimise equation (online supplemental equation s25) and obtain t_3^2 for each day (t) (the average of t_2^1 is calculated and fixed).
10. Minimise equation (online supplemental equation s25) and obtain t_2^2 for each day (t) (the average of t_3^2 is calculated and fixed).
11. Repeat steps 9 and 10 recursively until the criterion in equation (online supplemental equation s26) is met.
12. Rewrite t_3 as the product of t_2 (variable) and the ratio of the averages of t_3^{fin} to t_2^{fin} (parameter) as given in equation (online supplemental equation s29).
13. Minimise equation (online supplemental equation s28). The average of t_2 renders the value of τ_2 and likewise τ_3 of method I is achieved.
14. Method II: Divide the number of days of each wave to 3 (the initial number of divisions).
15. Assign the initial values of t_3.
16. Minimise equation (online supplemental equation s35) (Periodic Fatality Ratio (PFR) is fixed and t_3 is the variable to optimise).
17. Recalculate the values of PFR with regard to the new values of t_3 (PFR is a vector with the size equal to the number of divisions of days for each wave).
18. Minimise equation (online supplemental equation s35) again to obtain t_3^2.
19. Repeat steps 17 and 18 recursively until the criterion in equation (online supplemental equation s26) is met for t_3.
20. Minimise equation (online supplemental equation s36) to obtain t_2.
21. Calculate the SE of t_2 and t_3.
22. Increase the number of divisions of days of each wave by 1 and repeat steps 15–21.
23. Repeat step 22 until the number of days in each division is at least 2.
24. Calculate the average of t_2 and t_3 having minimum values of SE to achieve τ_2 and τ_3 of method II.

Compare the SE of t_2 and t_3 obtained from step 13 and step 24 (method I and method II). The smaller ones are selected to fill tables 1–4 for optimisation problem.

RESULTS

Gaussian curve fitting

The data of each country went through data preparation to assure consistency of the cumulative data (figure 1A), and then were subjected to the curve fitting procedure using sequential quadratic programming algorithm to find the ordinary least squares of errors defined in equations (online supplemental equations s6–s8) or their corresponding defined errors for skewed Gaussian fit. The results indicated an acceptable GoF for most of the countries and were able to point at the illogical jumps in the data and regulate it as well, as illustrated in figure 1B. The overall GoF was rejected for the cases (countries) with jumps in their data (steps) by our defined criteria, but the piecewise GoF before and after the steps satisfied the criteria. Yet, there existed some cases with too many steps, which could not satisfy any piecewise GoF in any arbitrary set of intervals as shown in figure 1C. The comparison between fitting power of Gaussian curve fitting and skewed Gaussian curve fitting is given as an example in figure 2. Although scaling the difference between the two fitting functions is not clearly visible on the cumulative data, it can be observed that skewed Gaussian curve fittings follow the trend of frequency distribution figures much better. The subtraction of the infected population mean from the mean of its corresponding recovered and dead peak renders average times to recovery and death (τ_2 and τ_3) derived by curve fitting technique, respectively.

In equations (online supplemental equation s3, s4 and s17, s18), the parameters a and σ represent population in each category and the timespan of a wave, respectively, and the parameters μ, and ζ are connected to the position of each wave in timespan and the relation between rise to decay of each wave, respectively.

Table 1 A comparison of time to recovery and death between the two techniques for 1-peak countries of COVID-19 pandemic

1-Peak country p=0.69	Curve fitting		Optimisation problem	
	τ_2	τ_3	τ_2	τ_3
Yemen*	19.80	3.28	22.67	5.00
Iraq	13.85	6.10	16.02	8.50
Georgia	16.73	12.93	17.70	11.12
India	11.46	10.71	12.08	9.01
Ukraine	35.09	12.00	32.95	12.44
Jordan*	16.55	2.21	17.21	8.33
Libya	24.62	10.61	32.55	10.33
Lithuania	39.68	13.70	30.09	12.25
Madagascar*	8.01	10.99	20.37	11.13
Belize	20.56	9.55	19.58	9.80
Tanzania†	5.94	2.62	N/A	N/A
Diamond Princess*	140.20	26.36	109.03	N/A
Central African Republic*	145.80	7.55	96.87	9.80

The conformity between valid data for countries with almost equal number of peaks is indicated by p value and reported in the table.
*The countries with step or discrepancy in their data are indicated as step.
†and/or discrepancy.

SIRD compartmental model

The improved SIRD model was implemented in two cases to model the COVID-19 breakdown for different countries. In Case 1, the model was directly fitted to the prepared data extracted for those countries that showed only one peak in our dataset, because of the inability of the SIRD models with the rate of incubation, recovery and death as time-invariant variables to fit to multiple peaks.

Table 2 A comparison of time to recovery and death between the two techniques for 2-peak countries of COVID-19 pandemic

2-Peak country p=0.51	Curve fitting				Optimisation problem			
	First peak		Second peak		First peak		Second peak	
	τ_2	τ_3	τ_2	τ_3	τ_2	τ_3	τ_2	τ_3
Kuwait*	13.89	8.14	4.73	36.42	13.90	12.81	12.56	33.74
Zimbabwe*	26.58	11.24	7.16	8.31	23.84	9.79	14.81	10.55
Suriname	11.88	6.59	13.86	7.61	12.52	9.54	10.69	7.71
Haiti*	52.93	12.50	59.10	36.62	45.56	13.65	78.75	33.32
Afghanistan*	54.75	24.30	24.04	5.79	42.45	17.43	52.39	14.35
Australia*	21.20	19.61	14.34	22.18	20.72	19.77	N/A	N/A
Brazil*	10.95	2.57	0	9.44	12.43	22.93	15.54	13.25
Bulgaria	31.95	8.75	33.25	11.29	30.02	15.31	33.87	13.26
Burma	16.04	9.93	20.44	8.29	15.02	8.31	15.36	7.20
Austria	16.00	10.80	14.43	19.62	16.15	10.35	11.69	20.10
Denmark	13.46	12.84	14.21	16.80	14.37	9.36	13.17	13.49
Canada*	49.83	11.69	1.84	12.51	28.46	14.07	10.65	16.40
Guatemala*	15.50	21.65	0	22.33	21.10	20.88	13.50	18.92
Chile*	0	0.5	21.50	0	9.59	9.73	7.33	10.46
Comoros	11.20	12.47	9.32	9.86	9.56	13.35	8.72	12.48
Switzerland†	15.38	4.17	11.42	18.96	17.12	12.13	17.79	27.33
Egypt†	54.27	8.90	11.44	28.01	37.01	12.69	18.02	14.95
Estonia	31.34	13.61	12.02	31.85	30.80	12.40	16.05	19.87
Eswatini	19.04	36.05	18.65	16.25	18.67	20.10	12.87	16.45
Finland*	19.92	0	7.00	15.14	21.09	15.40	27.14	4.66
Gambia*	24.50	9.20	0	9.27	31.63	11.13	5.95	10.50
Mexico*†	7.43	6.15	0	17.53	6.12	7.21	18.10	20.05
Germany*	17.81	15.49	10.96	30.58	15.33	24.83	16.26	24.64
Lebanon	31.39	7.75	31.45	10.45	29.24	8.54	30.33	10.53
Russia	23.89	15.95	23.91	19.93	25.82	16.37	25.98	21.49
Sudan†*	11.73	32.01	1.61	0.53	48.87	18.78	52.73	16.82
Nepal*	21.11	14.09	14.68	56.17	20.54	13.94	14.89	7.05
Singapore*	24.34	11.17	0	3.74	20.95	11.75	9.64	12.97
Thailand*	16.19	10.03	8.7979	0	15.35	10.60	15.64	13.54
Sweden (lack of data)	N/A	11.69	N/A	17.26	N/A	N/A	N/A	N/A
Pakistan	22.38	12.88	18.03	11.56	20.54	13.47	16.40	11.18
Kazakhstan*	13.56	18.64	17.35	0	18.63	5.13	31.62	6.90
Nigeria	34.15	11.22	16.97	18.72	31.21	10.97	17.63	18.54
Mali	29.68	8.05	39.05	15.33	28.75	8.51	41.21	15.31
Kenya*	39.73	17.84	52.73	0.14	25.09	18.30	N/A	N/A
Kosovo	22.17	8.29	24.31	18.30	18.81	13.98	21.57	19.59
Lesotho*†	28.43	49.52	0	20.80	40.58	36.21	N/A	N/A
Latvia*	26.92	3.47	28.30	1.82	40.83	16.49	35.41	14.14
Liberia*	13.33	0	21.65	18.03	16.72	45.73	26.76	17.02
Malawi†	29.72	18.74	6.51	4.97	30.49	20.16	44.28	15.39
Maldives	25.49	16.21	22.48	15.11	25.90	16.08	19.68	15.23

Continued

Table 2 Continued

2-Peak country p=0.51	Curve fitting				Optimisation problem			
	First peak		Second peak		First peak		Second peak	
	τ_2	τ_3	τ_2	τ_3	τ_2	τ_3	τ_2	τ_3
Mauritania	17.29	23.00	16.23	15.80	16.18	22.27	16.44	15.16
Poland*	29.41	11.08	26.59	10.88	29.81	13.26	22.52	10.83
Morocco	17.80	19.00	12.66	13.13	16.72	17.77	10.07	12.56
Namibia*	29.67	13.33	11.61	7.94	26.44	15.00	10.60	10.73
Mozambique *†	11.34	19.24	17.40	0.71	27.90	27.49	16.72	17.36
Argentina*	17.02	13.58	19.65	7.08	17.02	12.33	19.48	15.61
Armenia	21.99	17.72	22.39	22.68	22.78	17.19	20.90	20.09
Azerbaijan	15.78	15.15	15.28	14.14	15.15	15.19	15.00	14.95
Bahamas*	37.62	10.38	1.72	0	36.13	22.28	38.25	8.05
Belarus	28.08	19.75	12.83	19.64	28.86	20.78	11.29	20.09
Bolivia*†	34.53	29.20	41.68	9.41	41.85	10.33	41.31	16.60
Nicaragua*†	0.21	31.00	0	29.78	30.64	10.62	N/A	N/A
Cameroon*	17.86	40.67	0	94.91	15.93	26.76	20.01	13.31
Saint Vincent and the Grenadines	12.16	N/A	19.88	27.5651	9.33	N/A	20.50	11.11
Senegal	26.07	12.75	7.62	16.70	25.87	13.88	8.27	13.56
South Africa	16.24	11.33	22.25	9.71	16.73	12.59	20.62	9.43

The conformity between valid data for countries with almost equal number of peaks is indicated by p value and reported in the table.
*The countries with step or discrepancy in their data are indicated as step.
†and/or discrepancy.

In this model, the average incubation, recovery and death times are considered variable to obtain the optimal fit. In the second method, the improved SIRD model is fitted to the compartments derived by the Gaussian model. In this method, each surge is separated and the compartments are constructed as explained in the SIRD compartmental model subsection. In this case, the average incubation time is considered a variable and the average recuperation (recovery) and death times are considered parameters drawn from the difference between the mean of corresponding peaks (ie, the subtraction of the mean of the infected peak from the mean of the recovered peak which renders average time to recovery and similarly the average time to death is calculated). The data of countries with only one peak went through both cases of SIRD modelling as a means of comparison between the two methods as shown in figure 3A–C. The sum of absolute error is calculated for both cases with respect to the prepared data, as the observation (not the Gaussian model fitted to the observation) and the average and the SD of the ratio of the sum of absolute normalised error for case 1 (directly fitted to the prepared data) to case 2 (fitted to the Gaussian) for 10 countries with 1 peak in their data is calculated (1.3951±0.3275). The impediment of using SIRD model with invariant τ_2 and τ_3 to multiple-peaked data is solved in case 2 as shown in figure 3D–F.

Optimisation problems

The results of the optimisation problems are acquired after the comparison between the SE of t_2 and t_3 vectors of the two methods. This scenario rendered incoherent results for the countries with steps or discrepancy in their data.

The discrepancy between values of the two methods of estimation of the CFR as introduced in equations (online supplemental equations s31 and s32 will impede the correct calculations of the time averages to recovery and death via the optimisation problem; however, they can be obtained from the curve fitting procedure while neglecting the prior given in equation (online supplemental equation s11), because the error of curve fitting for these countries, as we put equation (online supplemental equation s11) in use, is relatively high (eg, where the value of reduced χ^2 GoF statistics is higher than 3), in such cases, the constraint of curve fitting presented in equation (online supplemental equation s11) is disregarded. As illustrated in figure 4A, as the number of the days from the onset of the pandemic grows, the CFR estimate approaches a horizontal asymptote. The intercept of this asymptote is equal to the estimation of the CFR, which is determined by applying equation (online supplemental equation s32) at the terminal data of the dataset. As mentioned earlier, equations (online supplemental

Table 3 A comparison of time to recovery and death between the two techniques for 3-peak countries of COVID-19 pandemic

3-Peak country p=0.48	Curve fitting						Optimisation problem					
	First peak		Second peak		Third peak		First peak		Second peak		Third peak	
	τ_2	τ_3	τ_2	τ_3	τ_2	τ_3	τ_2	τ_3	τ_2	τ_3	τ_2	τ_3
Iran	11.00	9.94	11.88	8.67	22.20	8.67	12.10	9.33	10.03	9.78	19.80	8.05
Chad*	13.56	41.31	11.71	2.77	3.31	17.73	12.66	46.05	10.41	11.35	11.05	18.62
Dominican Republic	24.04	21.48	35.24	11.29	40.69	10.80	21.86	22.98	33.77	11.37	44.42	15.18
Czech Republic	38.64	13.25	18.81	10.11	10.52	6.38	37.48	11.81	16.62	10.70	14.45	6.11
Italy†	36.06	19.97	11.00	4.06	14.32	5.61	38.46	20.08	30.22	15.82	34.36	16.88
Albania†	14.50	15.98	11.28	0	0	1.24	12.66	17.21	28.12	16.36	36.05	15.12
Honduras†	0.08	4.05	0	13.51	0	17.32	53.62	15.33	N/A	N/A	N/A	N/A
Bosnia and Herzegovina	18.89	28.07	24.06	16.84	26.33	13.00	18.40	27.13	23.84	16.72	25.60	10.87
USA (lack of data)	N/A	15.73	N/A	12.35	N/A	20.58	N/A	N/A	N/A	N/A	N/A	N/A
UK†	0	0	8.27	0	21.37	15.14	N/A	N/A	N/A	N/A	N/A	N/A
Costa Rica*	45.00	45.07	40.48	15.01	11.57	12.25	31.60	20.96	30.35	23.87	41.76	24.69
Croatia	17.03	20.95	12.71	14.09	6.71	16.03	18.35	20.36	11.57	13.56	6.72	14.79
Cuba	14.16	40.95	12.53	22.35	12.46	12.66	13.44	39.90	11.32	19.89	10.84	13.95
Bahrain	13.26	13.68	9.48	13.50	9.51	12.83	12.60	11.44	9.18	14.69	9.85	10.87
Ireland*†	10.66	13.03	0	6.57	0	22.06	N/A	N/A	N/A	N/A	N/A	N/A
South Korea	27.63	18.71	16.00	23.69	19.73	17.34	30.07	18.24	18.24	23.78	17.72	19.70
Malaysia	17.58	12.71	14.66	10.91	13.68	15.72	18.34	13.23	12.59	10.03	11.88	16.85
Vietnam*	14.06	N/A	15.68	29.62	5.95	N/A	16.34	17.63	21.65	16.24	28.70	23.12
Peru*	15.13	35.72	3.25	8.99	0	12.91	12.89	32.48	21.48	17.16	16.10	28.37
Japan	25.26	28.04	13.78	17.28	14.72	19.51	23.80	28.87	12.54	17.83	12.92	19.60
Spain*† (lack of data)	15.85	5.78	33.49	1.20	21.02	12.96	N/A	N/A	N/A	N/A	N/A	N/A
France†	8.50	17.14	14.67	1.63	16.69	0	N/A	N/A	N/A	N/A	N/A	N/A
Colombia	22.33	15.34	13.57	17.88	9.15	10.70	20.29	17.42	12.16	18.83	7.87	11.86
Netherlands†	0	0	17.95	4.10	14.79	16.80	N/A	N/A	N/A	N/A	N/A	N/A
Sri Lanka	19.34	21.45	13.56	12.62	15.38	15.51	21.73	21.17	15.33	12.82	13.35	14.93
Ghana*	44.48	17.50	9.00	14.27	9.00	13.30	17.21	17.73	8.50	17.38	9.00	14.36
Indonesia	27.29	24.32	18.86	14.94	17.19	10.50	26.40	20.83	16.62	12.45	15.51	11.93
Philippines*	29.53	14.81	0	26.59	1.02	0	38.94	26.72	20.68	12.06	17.98	17.48
United Arab Emirates	21.25	20.75	14.13	15.17	19.99	24.11	21.08	19.40	13.32	14.18	16.18	22.76
Taiwan*	27.01	33.21	27.33	5.72	27.31	11.10	29.43	31.87	N/A	N/A	N/A	N/A
Oman*	17.76	13.89	19.77	23.66	11.78	40.54	19.55	11.94	20.73	15.59	36.03	18.36
New Zealand	14.05	23.50	19.74	16.57	14.66	15.67	13.78	24.58	20.30	14.03	16.92	17.45
Luxembourg*	27.34	13.38	7.72	12.60	19.33	16.07	16.20	16.57	12.12	21.53	16.14	15.97
Uzbekistan	18.60	8.08	13.17	9.75	8.82	11.86	18.45	10.55	12.97	11.03	7.42	12.23
Norway*†	39.47	6.43	34.62	13.71	0.70	0	49.15	13.88	N/A	N/A	N/A	N/A
Jamaica†	43.04	19.31	4.67	0	14.18	10.85	24.93	12.45	37.67	8.16	34.10	12.29
Hungary†	33.47	40.48	35.77	7.18	0	18.40	30.68	28.41	31.36	12.06	N/A	N/A
Paraguay	26.25	7.94	19.20	20.17	28.35	17.42	25.28	9.29	16.78	18.14	31.01	14.68
Portugal	36.58	9.56	21.36	8.30	15.44	7.85	36.96	10.96	22.59	9.95	17.97	8.23
Tunisia*	34.04	7.16	4.95	10.59	14.92	8.72	26.57	10.81	26.93	10.91	20.14	11.43

Continued

Table 3 Continued

3-Peak country p=0.48	Curve fitting						Optimisation problem					
	First peak		Second peak		Third peak		First peak		Second peak		Third peak	
	τ_2	τ_3	τ_2	τ_3	τ_2	τ_3	τ_2	τ_3	τ_2	τ_3	τ_2	τ_3
Andorra	25.11	14.12	18.30	15.47	14.37	15.76	22.57	14.86	17.06	12.26	13.28	14.04
Angola	28.70	15.91	33.19	10.95	34.29	15.47	24.94	15.21	36.20	14.54	31.75	15.15
Bangladesh*	19.51	13.54	11.28	16.04	9.66	19.98	23.97	15.23	45.06	14.46	37.92	17.63
Niger	15.00	19.99	19.89	23.01	18.53	14.32	14.28	19.26	17.31	20.36	20.91	14.47
North Macedonia	20.63	27.11	28.40	10.56	22.75	11.45	18.47	29.56	26.16	11.46	23.67	13.65
Panama*	56.27	21.30	26.08	10.33	22.19	11.09	28.53	20.62	24.69	10.69	17.80	11.79
Rwanda	20.00	12.81	33.98	12.00	22.88	13.88	23.04	20.76	32.20	12.51	17.70	15.20
Saudi Arabia	17.18	17.66	11.00	13.00	17.88	12.40	17.90	20.67	14.88	15.20	18.60	15.96
Sierra Leone†	31.32	5.00	3.94	0	0	0	16.47	8.93	11.41	19.29	N/A	N/A
Slovenia*	40.35	22.12	15.43	14.53	1.93	0	32.07	13.64	14.23	4.98	14.07	7.72
Somalia*†	63.47	0	0	0	38.14	0	47.81	17.21	N/A	N/A	N/A	N/A
Syria†	1.53	37.63	30.67	0	9.79	5.54	20.60	12.83	4.61	7.89	N/A	N/A
Cyprus*†	37.56	40.58	18.83	0	17.58	12.03	56.18	23.97	34.43	31.37	N/A	N/A
Trinidad and Tobago	22.33	23.25	16.66	15.22	19.21	12.56	24.48	22.48	14.70	15.78	23.35	15.52
Uganda†	16.80	14.73	15.07	15.85	0	4.18	23.88	14.84	18.74	15.50	N/A	N/A

The conformity between valid data for countries with almost equal number of peaks is indicated by p value and reported in the table.
*The countries with step or discrepancy in their data are indicated as step.
†and/or discrepancy.

equations s31 and s32) can be used to estimate the CFR of countries; however, equation (online supplemental equation s32) is more likely to render a better estimate. It is predictable that in countries with consistent data, the estimation of the CFR from both equations (online supplemental equations s31 and s32 approach the same asymptote unlike figure 4B.

Hypothesis test

In order to investigate the nearness of the values of time averages obtained via the two techniques, a hypothesis test was performed on the data with acceptable values of the discrepancy index (DI) and the error of fit index (EI, equation (6)). DI and EI are defined in the next section. The null hypothesis is that the expected value of the normalised error between the values achieved via the two techniques is zero (H_0: $E((\tau_{j,i,op}-\tau_{j,i,cf})/\tau_{j,i,cf})=0$, where the subscript j can be 2 or 3, i is the peak number, and op and cf subscripts state that whether the average time is calculated by the optimisation problem of curve fitting, respectively) and the alternate hypothesis (H_1) is that the expected value of the mentioned variable is non-zero. The test is carried out for countries with identical number of peaks separately and their p values are given in tables 1–4. The result of the test clarified that the null hypothesis could not be rejected in a two-tailed test with 5% level of significance.

DISCUSSION

Since epidemics are a great concern for jeopardising the healthcare of humanity, it is important to develop more comprehensive models to assimilate their behaviour as thoroughly as possible. The ability to predict and estimate the average time to recovery and death of the infectious population of an epidemic, in addition to other factors, is imperative to maintaining and providing necessary stocks that healthcare institutes need for confronting the decease.

In the dataset, there were some countries which could not be fitted to either of Gaussian and skewed Gaussian model while applying the mentioned priors. After the examination of the data of such countries, it was found out that the reported numbers are not coherent (ie, addition of cumulative recovered and deceased population does not add up near to cumulative infectious data). In such cases, the prior remarked in equation (online supplemental equation s11) must be excluded so that the GoF criteria can be met. Nevertheless, the results of the fit of the data to these countries can be involved in the calculation of further statistical parameters, unless the value of DI (given in equation (5), is above 0.1. This error is calculated directly from the last day of our dataset and also from the amplitudes of the fitted Gaussian model for the discrepant countries. Another shortcoming for our dataset is that the data of one or more classes ceased to be

Table 4 A comparison of time to recovery and death between the two techniques for 4-peak countries of COVID-19 pandemic

4-Peak country p=0.13	Curve fitting							Optimisation problem								
	First peak		Second peak		Third peak		Fourth peak	First peak			Second peak		Third peak		Fourth peak	
	τ_2	τ_3	τ_2	τ_3	τ_2	τ_3	τ_2	τ_3	τ_2	τ_3	τ_2	τ_3	τ_2	τ_3	τ_2	τ_3
Slovakia	27.4	12.1	29.6	15.2	25.9	15.7	17.4	12.9	27.7	12.2	29.2	14.4	25.6	15.4	19.8	13.2
Ethiopia*	22.0	15.6	25.8	10.3	51.9	8.25	31.5	13.9	20.6	14.6	29.9	11.7	62.5	11.0	35.3	11.8
Hong Kong	20.9	13.1	13.4	12.2	16.3	12.7	13.5	11.3	19.7	12.3	15.1	11.6	16.2	13.2	15.1	11.4
Greece*†	19.6	20.6	17.2	6.41	31.6	0	23.8	25.8	54.2	6.24	32.3	15.8	0.10	0.10	N/A	N/A
Israel	21.2	13.4	22.2	17.2	12.8	11.3	17.0	11.0	19.7	14.2	23.4	17.1	12.7	11.8	17.5	11.0
Ivory Coast	14.2	15.5	22.0	15.9	16.4	11.1	7.27	21.7	15.4	16.2	21.3	15.4	16.6	13.2	8.58	20.2
Iceland	14.6	22.4	15.1	9.59	16.3	17.5	13.1	17.3	14.9	23.2	14.8	11.1	15.8	17.3	13.2	16.9
Mainland China*	19.5	79.0	80.0	26.9	5.70	19.0	63.6	4.50	22.8	17.7	N/A	N/A	N/A	N/A	N/A	N/A
Romania*	19.0	19.7	13.6	11.0	0	0	8.20	16.4	17.7	23.5	30.8	10.2	N/A	N/A	N/A	N/A
West Bank and Gaza	49.2	14.7	21.6	13.3	18.0	13.7	10.9	17.1	47.7	15.8	22.7	12.3	17.4	13.8	12.8	16.9
Algeria†	15.5	17.1	0.87	6.20	0	7.77	5.95	1.78	7.90	6.42	14.6	16.7	21.8	8.27	N/A	N/A
Malta	21.8	15.6	14.6	15.3	14.4	22.9	16.5	13.8	23.5	14.5	14.7	16.3	15.7	20.5	16.0	13.3
Montenegro	23.2	15.4	13.6	22.8	17.8	18.1	23.2	18.6	22.1	14.6	14.4	21.2	18.9	17.6	22.2	17.9

The conformity between valid data for countries with almost equal number of peaks is indicated by p value and reported in the table.
*The countries with step or discrepancy in their data are indicated as step.
†and/or discrepancy.

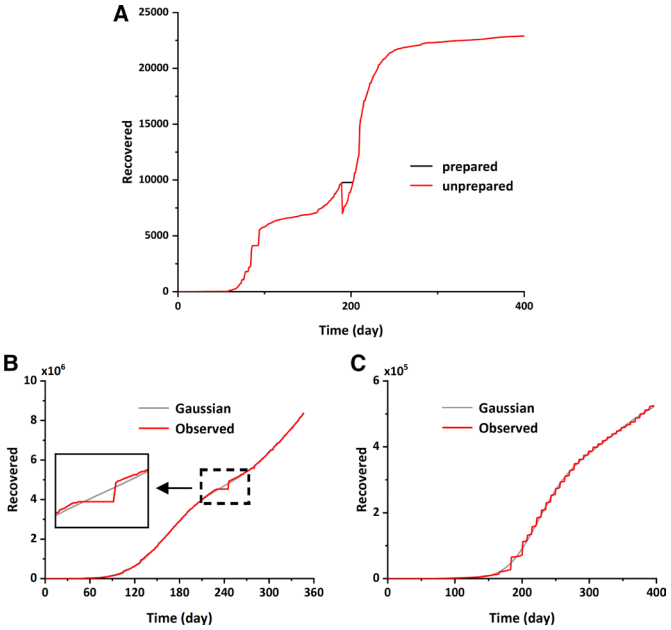

Figure 1 (A) Data preparation of recovered patients in Australia, (negative slope) the decline of cumulative data with advancing time were subjected to equation (online supplemental equation s1 and s2), which can be found in the online supplemental information, and revised accordingly. Piecewise GoF for recovered cumulative data of two different countries (B) Brazil with acceptable piecewise GoF, (C) Philippines with unacceptable piecewise GoF.

presented after a specific day, which compels us to prune the data of those countries up to that day or omit them in our calculations.

The countries with steps in their cumulative data do not hinder the implementation of the improved SIRD model, but the countries that have shown discrepancy in their data (having a DI greater than 0.1) do prevent carrying it out, because eventually, the infectious compartment will turn into either of recovered or deceased compartment, and existence of discrepancy makes an acceptable fit impossible. The results of the comparison between the two cases indicate that the model fitted by case 2, although it

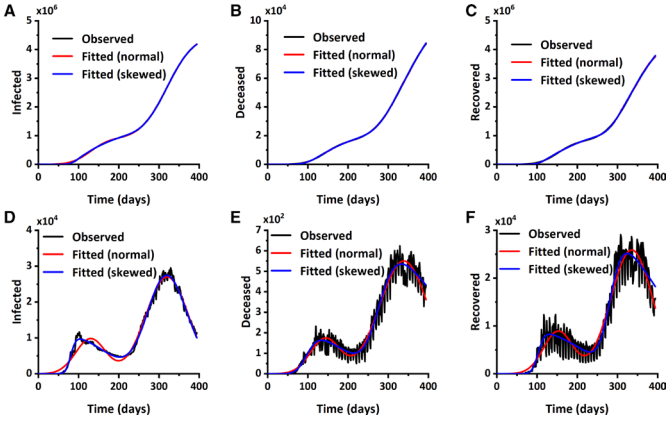

Figure 2 Comparison of curve fitting of the data via skewed and normal Gaussian distributions on cumulative (A–C) and frequency (D–F) data of Russia.

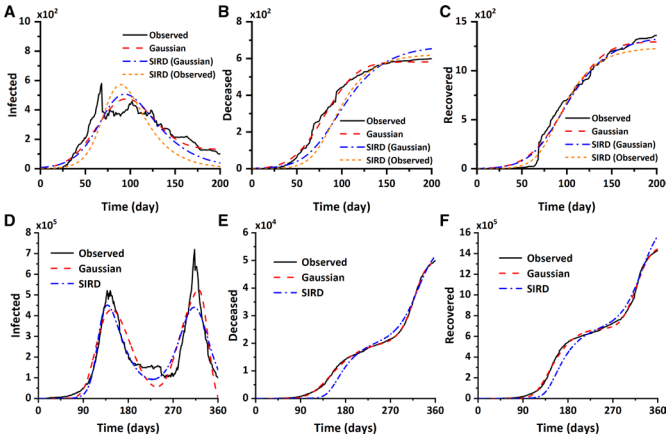

Figure 3 Comparison of both methods of SIRD fitting for (A) total infected, (B) cumulative deceased and (C) cumulative recovered patients for Yemen. Resolution of multipeaked data using SIRD model on (D) total infected, (E) cumulative deceased and (F) cumulative recovered patients with the help of case 2 for South Africa.

is not directly fitted to the observation, has a better fit for the most countries (ie, showed relatively a smaller sum of absolute error), even though case 1 has a greater degree of freedom (ie, also average time to recovery and death are variables of fit). This phenomenon can be attributed to the uncertainty of parameter estimation when fitting an epidemiological model to a noised dataset.[19] This proves that our suggested method is not only conducive to model a multimodal epidemic, but also is favourable to estimate the parameters of unimodal epidemiological models with.

The results of estimation of the CFR for some countries from our dataset were quite conflicting with what is reported from WHO as an average.[20] This can be attributed to either of the inability of the officials to account for all of the infected population for some reasons (eg, preference for home treatment) in contrast to 'the countability of the deceased population', or data manipulation or other errors. As a consequence, these results prevent researchers from performing credible multivariate analyses including the CFR factor, yet it can be used as a mean to roughly estimate the true infected population of countries. The convergence of the values of CFR estimated by the two methods confirms that equation (online supplemental equation s11) is applicable for the

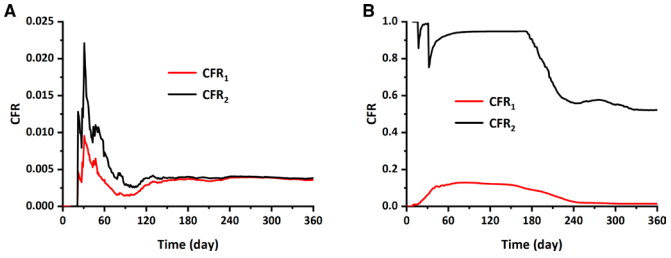

Figure 4 Instances of two kinds of case fatality rate (CFR) for (A) Bahrain with consistent data and (B) Netherlands with discrepant data.

considered country and the DI is within acceptable range, hence it can be applied for further statistical analyses.

Tables 1–4 provides the peak-wise time averages to recovery and death for countries obtained from the two techniques (curve fitting and optimisation problem) introduced in the model development section. The descending trend of the p values as the number of peaks grow can be attributed to the negative influence that overlapping of the peaks have on estimation accuracy of the two techniques. Contrary to the nearness of most of the average time values obtained by method I of the optimisation problem to those obtained by the curve fitting process, the deceased class exhibited relatively high values of SE, and it is caused by the difference in magnitude between recovered and deceased population; a small deviation in the number of recovered from its mean for a specific day causes drastic change in the calculation of the estimation of time to death for that day, causes the SE of t_3 to escalate.

Even though the corresponding average times to recovery and death drawn from the two techniques for most countries showed an acceptable agreement, there existed three types of inaccuracy which caused a disjunction in the average times, obtained by the two techniques or caused one or both techniques to malfunction and therefore not being able to produce a result. The first cause of inaccuracy is the lack of data on one or more categories of the people inflicted by the virus, this inaccuracy leads both techniques to produce invalid results. The second inaccuracy is when the difference in the two methods of estimation of the CFR is significant. This type of inaccuracy affects the optimisation problem in a way that it may not be able to render an output. The third inaccuracy occurs when there is an abrupt increment (steps) in the cumulative data of the three classes. This will cause a disjunction in the results of the two techniques for the specific peak where the step is observable or for more extreme cases causes the inability of optimisation problems.

The existence of these inaccuracies leads us to develop indices to quantify them. The DI is defined in equation (5):

$$\mathrm{DI} = 1 - 10^{-500\times(\mathrm{CFR}_2-\mathrm{CFR}_1)^2} \tag{5}$$

The DI is defined in a way that varies between 0 and 1 and it has a steep slope between the values 0.01 and 0.1 of $|\mathrm{CFR}_2-\mathrm{CFR}_1|$. The EI (briefly, the error index) is developed to quantify the frequency and amplitude of steps in the data by means of comparing the cumulative data with the fitted cumulative distribution to the data. This comparison is made via the reduced χ^2 GoF statistics. Since the reduced χ^2 GoF is calculated for each class of our data separately, EI includes those values of reduced χ^2 GoF that are above our defined acceptable range. If more than one of the classes of the data has a reduced χ^2 value above the acceptable range, the geometric mean of those values is applied to calculate the EI index. For countries with multiple peaks of the pandemic, the piecewise reduced χ^2 GoF is calculated for each peak whose bounds are determined by using equations (online supplemental s37 and s38). The *EI* is formulated in a way to vary between 0 and 1 too and is given in equation (6):

$$\mathrm{EI} = 1 - \mathrm{e}^{-0.02(\chi^2_{\mathrm{red}}-3)} \tag{6}$$

The DI of the countries with discrepancy in their data and the arithmetic mean of the ratios of the average times of the two techniques is given in table 5.

The 'time ratios' that are defined to correlate with the DI are formulated to be ≤1, for such purpose, the shorter τ_2 and τ_3 are divided by the longer one extracted by the two techniques, hence their average is between 0 and 1. The values of unavailable τ_2 and τ_3 for these calculations are deemed to be zero. The correlation between DI and the average of the ratios is −0.74 (figure 5A). Some of the causes of relatively large deviation of scattered data from the regression line are combination with the effect produced by the steps for most countries with discrepancy,

Table 5 The list of countries with discrepancy in reported data and their time ratio less than in their event data

Country	Time Ratio	Discrepancy index
Albania	0.3808	0.3228
Algeria	0.2819	0.1176
Bolivia	0.6842	0.1122
Cyprus	0.3009	0.7419
Egypt	0.6374	0.1504
France	0	1
Greece	0.2007	0.6563
Honduras	0.0134	0.7413
Hungary	0.5151	0.1171
Ireland	0	1
Italy	0.4965	0.3690
Jamaica	0.4411	0.1548
Lesotho	0.3579	0.8977
Malawi	0.5940	0.3906
Mexico	0.6377	0.1809
Mozambique	0.4829	0.2306
Netherlands	0	1
Nicaragua	0.0873	0.1645
Norway	0.2111	0.5024
Sierra Leone	0.2385	0.1067
Somalia	0.1255	0.4975
Spain	0	1
Sudan	0.2234	0.1993
Switzerland	0.4720	0.1638
Syria	0.0943	0.6301
Tanzania	0	0.9875
Uganda	0.5797	0.2036
UK	0	1

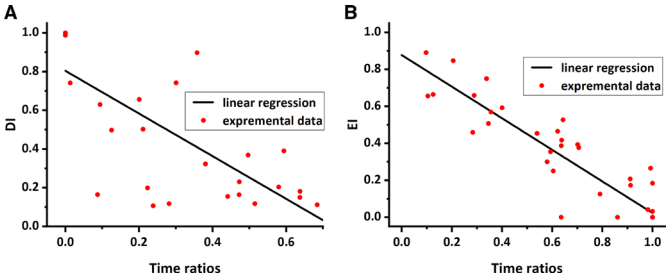

Figure 5 Correlation of experimental data and their linear regression and a function of 'time ratios' for (A) discrepancy index (DI) and (B) error index (EI).

inability to rule out the effect of a class and relate the DI to one of the time averages because the discrepancy is calculated using all three classes, and the inability in shrinking the multiple time averages of multiple peaks and not being able to determine which peak caused the discrepancy.

Three causes come to mind with regard to the relatively large deviation of scattered data from the regression line. The first one is combination with the effect produced by the steps for most countries with that of discrepancy. The second one is the inability to rule out the effect of a class and relate the DI to one of the time averages, because the discrepancy is calculated using all three classes. The last cause is the inability in shrinking the multiple time averages of multiple peaks and not being able to determine which peak caused the discrepancy. Nevertheless, the value of this correlation is large enough in magnitude to infer the existence of a negative correlation between the two variables.

The EI of the countries with reduced χ^2 values more than 3 and the geometric mean of the ratios of the time average of the specific peaks where the step is observed between the two techniques is given in table 6.

The defined 'time ratios' are formulated to be smaller than 1, like those of the DI, therefore their geometric mean is also between 0 and 1. Similar to the calculation of the DI, the values of unavailable τ_2 and τ_3 are deemed to be zero. When there is no step in recovered or deceased data, the τ_3 or τ_2 is ignored, respectively, and geometric mean is not used for the 'time ratios'. On the other hand, when there is a step in infected data and/or in both recovered and deceased data, the geometric

Table 6 The list of countries with one or more steps in their reported data and corresponding error of fit and their time ratio less than in their event data

Country	Error of fit (EF) index	Time ratios	Country	EF index	Time ratios
Afghanistan	0.2835	0.4589	Kosovo	0.6376	0.4170
Argentina	0.5397	0.4536	Kuwait	0.7057	0.3766
Australia	1	0	Latvia	0.6433	0.5269
Bangladesh	0.6042	0.2503	Lesotho	1	0
Bolivia	0.5931	0.3538	Liberia	1	0
Brazil	1	0	Luxembourg	0.4003	0.5925
Cameroon	1	0	Madagascar	0.7014	0.3932
Canada	0.9122	0.1728	Mainland China	1	0
Central African Republic	0.1252	0.6644	Mexico	1	0
Chad	0.58	0.2995	Mozambique	0.9814	0.0411
Chile	1	0	Namibia	0.0973	0.8911
Costa Rica	0.6218	0.4651	Nepal	0.7904	0.1255
Cyprus	1	0	Nicaragua	1	0
Ethiopia	0.3386	0.7500	Norway	1	0
Finland	1	0	Oman	0.2887	0.6589
Gambia	1	0	Panama	0.3457	0.5070
Germany	0.3557	0.5691	Peru	1	0
Ghana	0.6362	0.3869	Philippines	1	0
Greece	1	0	Poland	0.2052	0.8469
Guatemala	1	0	Romania	0.6361	0
Ireland	1	0	Singapore	1	0
Jordan	0.9922	0.2653	Slovenia	0.8603	0
Kazakhstan	1	0	Somalia	1	0
Kenya	1	0	Sudan	1	0.0315
			Yemen	0.1041	0.6560

mean is considered for the 'time ratios'. The correlation between EI and the geometric mean of the defined ratios is −0.93 (figure 5B). The causes of relatively small deviation of scatter data from the regression line are exactly opposite of the causes mentioned for the relative high deviation of the data of DI. For example, fewer countries with steps have manifested discrepancy effect. For some countries one of the recovered or deceased classes can be dispensed with in the calculation of the defined ratios for EI as long as it does not show a reduced χ^2 value over 3 but if the infected class contains a step in its data, no classes are excluded from calculation. Another constituent of this relatively large correlation is that in bimodal or multimodal distributions, the specific peaks, and consequently their EI via piecewise GoF, where steps have occurred are extractible. These correlations establish the validity of the two techniques to estimate the average times to recovery and death of COVID-19 statistically as a case study.

The data on infected and recovered population, despite the data of deceased population, is acquired via specific test of the infection and the accuracy of this data partly depends on the tendency of individuals to check on their health status and this might affect the time averages and moreover the CFR factor of the epidemiological disease. These observations pose threat to the validity of the resultant data, but if the dataset is well-labelled with respect to its validity or if there exists a prior knowledge of the epidemiology through precise statistical analyses, the true values of interest can be calculated. Further, the valid data of the time averages can be subjected to multivariate analyses and machine learning algorithms to investigate effective factors and predict these time averages in different conditions, which is out of scope of the current study.

Spatial stratified heterogeneity (SSH) of the time-to-event data (tables 1–4) is measured.[21 22] Since the study of heterogeneity in current study is not spatiotemporal, the average data of all the waves of the pandemic is allocated for measuring SSH. The mean of the data resulted by the two techniques is used for attributing a single time-to-event data to a country. For those countries where an average time-to-event via one technique was not available, the other technique is used to assign the time-to-event data to that country. The data are partitioned into six strata by minimising the within-strata variances and maximising the between-strata variance. The result shows extreme heterogeneity with q-statistic results being 0.9339 and 0.9397, and f-test results being 358.65 and 361.75 for average time to recovery and time to death of countries, respectively. The results of f-tests indicate that heterogeneity of the study population for the selected variables. These results are also projected in figure 6 in which spatial stratification of times to recovery and death for all countries is given. The number of countries with DI and EI larger than the critical values of their corresponding indices are about half of the available countries data, therefore the observed heterogeneity is partly caused by such phenomenon. This can intervene in the analysis of findings and deviate the epidemiological interpretations from their origin.

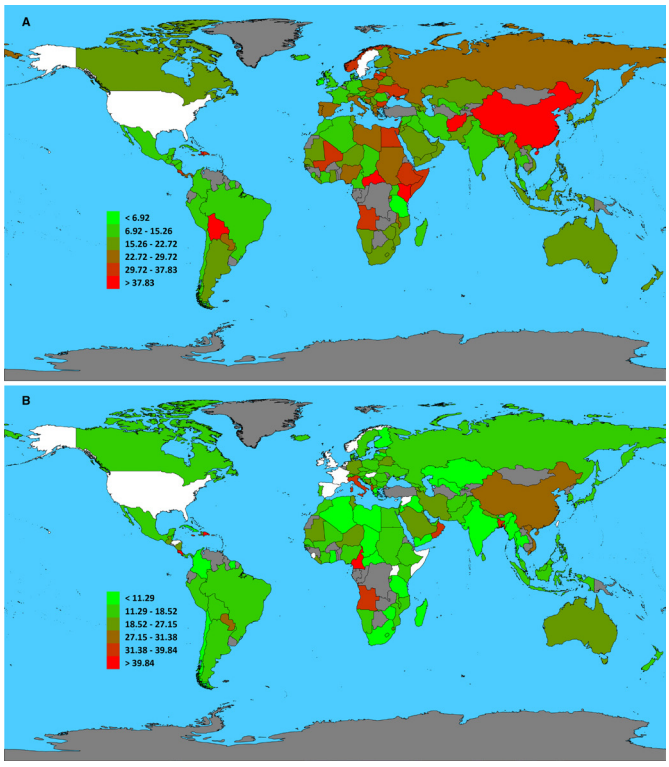

Figure 6 Spatial stratified heterogeneity of (A) time to recovery and (B) time to death of countries. Countries with white colour indicate that the times are undeterminable, and the grey colour shows that COVID-19 data are not available in open data resource Kaggle.

In the current study, two techniques, curve fitting and optimisation problem, are developed to calculate the recuperation and death average time of epidemics based on their observed data. The curve fitting process is a utility for fitting better unimodal SIRD model and are also able to generate SIRD models with two or more peaks. The result of the estimation of the time averages by the two techniques agreed to one another in absence of irregularities. It was also shown that the closer the peaks are to each other, the less is the conformity between the results of the two techniques. Two indices are defined to correlate the level of mismatch in the result of the two techniques for estimating the average times and both indicated strong negative correlations between intensity of irregularities and level of matching the results of the two techniques. Further to this study, the findings of this work can be subjected to statistical analyses to correlate them with determining factors (eg, socioeconomic parameters, geography, seasonal changes, etc) to achieve a better understanding of their underlying behaviour for COVID-19 pandemic.

Author affiliations
[1]Department of Polymer Engineering and Color Technology, Amirkabir University of Technology, Tehran, Iran
[2]Department of Polymer Engineering, School of Chemical Engineering, College of Engineering, University of Tehran, Tehran, Iran

[3]Social Determinants of Health Research Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[4]Department of Clinical Toxicology, Shohada-e-Tajrish Hospital, School of Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran

Acknowledgements We wish to thank Mrs. Akram Kia from Social Determinants of Health ResearchCenter for her great assistance in registration of our study.

Contributors AR, EG, DH-M, FF and HH-M provided a substantial contribution to the design and interpretation of the paper and revised drafts. AR initiated the project and wrote the initial draft. AR and HH-M are the guarantors for the study.

Funding Shahid Beheshti University of Medical Sciences, award no 33247.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request.

ORCID iD
Hossein Hassanian-Moghaddam http://orcid.org/0000-0003-4370-0544

REFERENCES

1 Moein S, Nickaeen N, Roointan A, *et al*. Inefficiency of sir models in forecasting COVID-19 epidemic: a case study of isfahan. *Sci Rep* 2021;11:4725.
2 WHO. HIV/AIDS. 2021. Available: https://www.who.int/data/gho/data/themes/hiv-aids [Accessed 13 Nov 2021].
3 Taubenberger JK, Morens DM. 1918 influenza: the mother of all pandemics. *Emerg Infect Dis* 2006;12:15–22.
4 Hui DS, Azhar EI, Memish ZA, *et al*. Human coronavirus infections—severe acute respiratory syndrome (SARS), middle east respiratory syndrome (MERS), and SARS-cov-2. *Encyclopedia of Respiratory Medicine* 2020;146–61.
5 Ivorra B, Martínez-López B, Sánchez-Vizcaíno JM, *et al*. Mathematical formulation and validation of the be-FAST model for classical swine fever virus spread between and within farms. *Ann Oper Res* 2014;219:25–47.
6 Ivorra B, Ferrández MR, Vela-Pérez M, *et al*. Mathematical modeling of the spread of the coronavirus disease 2019 (COVID-19) taking into account the undetected infections. The case of China. *Commun Nonlinear Sci Numer Simul* 2020;88:105303.
7 Brauer F. Compartmental models in epidemiology. *Math Epidemiol* 2008;1945:19.
8 Liu M, Thomadsen R, Yao S. Forecasting the spread of COVID-19 under different reopening strategies. *Sci Rep* 2020;10:20367.
9 Dissanayake D. Different types of compartmental models for understanding disease spread | by dinusha dissanayake | sep, 2021 | towards data science. 2021. Available: https://towardsdatascience.com/different-types-of-compartmental-models-for-understanding-disease-spread-29ac066d59dd [Accessed 14 Nov 2021].
10 I. of M. (US) F. *Infectious disease emergence: past, present, and future*. 2009.
11 Balabdaoui F, Mohr D. Age-Stratified discrete compartment model of the COVID-19 epidemic with application to Switzerland. *Sci Rep* 2020;10:21306:21306.:.
12 Pei L, Zhang M, Kuniya T. Long-Term predictions of COVID-19 in some countries by the SIRD model. *Complexity* 2021;2021:1–18.
13 Chen Y, Cheng J, Jiang Y, *et al*. A time delay dynamical model for outbreak of 2019-ncov and the parameter identification. *Journal of Inverse and Ill-Posed Problems* 2020;28:243–50.
14 McLaughlin MP. *A compendium of common probability distributions*. Michael P. McLaughlin, 2001.
15 Alvarez F, Argente D, Lippi F. *A simple planning problem for COVID-19 lockdown*. 2020.
16 Hamzah FAB, Lau CH, Nazri H, *et al*. CoronaTracker: world-wide COVID-19 outbreak data analysis and prediction. *nCoV* [Preprint] 2020.
17 Gee S, Chandiramani M, Seow J, *et al*. The legacy of maternal SARS-cov-2 infection on the immunology of the neonate. *Nat Immunol* 2021;22:1490–502.
18 Victor Okhuese A. Estimation of the probability of reinfection with COVID-19 by the susceptible-exposed-infectious-removed-undetectable-susceptible model. *JMIR Public Health Surveill* 2020;6:e19097.
19 Lloyd AL. Sensitivity of model-based epidemiological parameter estimation to model assumptions. In: *Mathematical and Statistical Estimation Approaches in Epidemiology*. Springer Netherlands, 2009: 123–41.
20 Wu Z, McGoogan JM. Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the Chinese center for disease control and prevention. *JAMA* 2020;323:1239–42.
21 Wang JF, Zhang TL, Fu BJ. A measure of spatial stratified heterogeneity. *Ecological Indicators* 2016;67:250–6.
22 Yin Q, Wang J, Ren Z, *et al*. Mapping the increased minimum mortality temperatures in the context of global climate change. *Nat Commun* 2019;10:4640.