A model for forecasting the COVID-19 pandemic and assessing responses to it

David I. Ketcheson*

April 18, 2020

Abstract

We describe a model for predicting the death toll over time due to the ongoing COVID-19 epidemic. The model is based on recorded numbers of deaths and makes predictions of daily numbers of new infected persons and deaths. The effect of non-pharmaceutical interventions (NPIs) is modelled empirically, based on recent trends in the death rate. The model can also be used to study counterfactual scenarios based on hypothetical NPI scenarios.

1 Overview of the model

The central part of the forecasting model is the SIR model of Kermack & Mckendrick [2]. The SIR model is a compartmental model that divides the population into three groups:

- S(t): Susceptible (those who have not yet been infected)
- I(t): Infected (those who can currently spread the disease)
- R(t): Recovered (those who are now immune and cannot contract or spread the disease)

The model takes the form of a system of three ordinary differential equations:

$$\frac{dS}{dt} = -\beta I \frac{S}{N} \tag{1}$$

$$\frac{dI}{dt} = \beta I \frac{S}{N} - \gamma I \tag{2}$$

$$\frac{dR}{dt} = \gamma I \tag{3}$$

Here N is the total population. Since R = N - S - I, the system can be modelled with just the first two equations.

*Computer, Electrical, and Mathematical Sciences & Engineering Division, King Abdullah University of Science and Technology, 4700 KAUST, Thuwal 23955, Saudi Arabia. (david.ketcheson@kaust.edu.sa)

We extend this model by incorporating a time-dependent intervention parameter $q(t)$, which models the change in the contact rate β due to social distancing, school and work closures, and other such measures. This extension takes the following form:

$$\frac{dS}{dt} = -(1 - q(t))\beta I \frac{S}{N} \tag{4}$$

$$\frac{dI}{dt} = (1 - q(t))\beta I \frac{S}{N} - \gamma I \tag{5}$$

$$\frac{dR}{dt} = \gamma I \tag{6}$$

A value $q(t) = 0$ corresponds to no intervention, while $q(t) = 1$ would correspond to a complete elimination of any human contact. This parameter can model both population-wide measures and those specifically targeting infected or suspected individuals.

1.1 Details of forecast algorithm

The model is applied to individual countries or states. For each region, it involves the following steps:

1. Get recorded numbers of deaths per day since the start of the epidemic, denoted D_i, either from the JHU data set or the NYT data set.

2. Infer the number of daily new infections I_i for dates in the past, based on an assumed distribution of the time from infection to death. Let $p(k)$ denote the probability that a newly infected person dies k days after infection. We have

$$\sum_k p(k) = F$$

where F is the infection fatality ratio, currently estimated to be about 0.006 for COVID-19. The quantities D_i and I_i are related by

$$D_i = \sum_{k=0}^{\infty} I_{i-k} p(k).$$

The problem of solving this equation for I_i given D_i is known as deconvolution, and it is ill-conditioned. Due to non-smoothness of real-world recorded death rates, the deconvolved signal is typically highly oscillatory and includes negative values. A good regularization for this problem is the subject of ongoing work. In the meantime, the model estimates a mean time to death m based on the recent death rate trend and then sets

$$I_{i-m} = D_i/F.$$

For a uniform (in time) distribution of deaths D_i the mean time would be $m = 17$ days, but the actual value of m at a given time and place may be shorter (if the epidemic is growing) or longer (if the epidemic is diminishing). This inference step

above gives values of I_i up to m days in the past. The cumulative number of recovered individuals at day i is then computed as

$$R(i) = \sum_{j=0}^{n} I_j(1 - e^{-\gamma(n-j)}).$$

The number of actively infected persons on day n is then

$$I(i) = (\sum_{0}^{n} I_i) - R_i^C.$$

The number of susceptibles on day i is

$$S(i) = N - R(i) - I(i).$$

3. Let $i = 0$ denote the present day. The SIR model is run starting from day $i = -m$ up to $i = 0$. The initial data is given by the inferred values above. The value of $q(t)$ is determined by fitting the model output to the recorded numbers of deaths. This fitting is done in two ways. In the first way, a constant value is assumed $q(t) = q_0$. In the second way, a linear function is assumed: $q(t) = q_1 + \delta t$. In the second fit, there is a penalty for using a non-zero slope; in both fits, $q(t)$ is constrained to lie in $[0, 1]$. Note that the second fit gives a value of q_1 for the intervention at the present day. The value q_1 will be used as the assumed intervention effectiveness in the next step, while the two fits are used to give a range of plausible intervention parameters

$$[\min(q_0, q_1) - 0.2, \max(q_0, q_1) + 0.2], \tag{7}$$

used to generate a range of model outcomes. This range is also constrained to lie in $[0, 1]$.

4. Finally, the output values of the last step at day $i = 0$ are used as initial data for the forecast. The primary forecast is generated using the intervention effectiveness value q_1, while the range is generated using the set of values (7).

Forecasts can also be generated using some assumed future intervention policy that changes in time, but at present we are not generating any such forecasts.

1.2 Details of counterfactual no-intervention scenario algorithm

The model can also be used to generate a hypothetical scenario starting at some time in the past. The main purpose of this is to examine what might have happened in a given region in the absence of intervention. This involves the following steps:

1. Determine a starting date for the model. This should be a date before substantial intervention started, but after a statistically significant number of deaths had occurred. For some regions (those that are very small or where the epidemic started very late) these two requirements may be contradictory, and no reasonable modeling can be done. At present, the model finds a the earliest day on which the cumulative death toll was greater than 50.

2. Determine initial numbers of susceptible, infected, and recovered for the starting date. The number of infected is assumed to be about $\alpha D_i/F$ where D_i/ifr is the number infected m days before the start date and $\alpha \approx 10$ accounts for growth of the epidemic during those m days. There is an extremely high level of inherent uncertainty in estimating this initial data, and the model outcome is highly sensitive to it.
3. This initial data is fed into the SIR model without intervention.

2 Data sources

Daily deaths:

- For countries: https://github.com/CSSEGISandData
- For states: https://github.com/nytimes

Population data: https://population.un.org/wpp/Download/Standard/Population/

3 Key parameter values

Here we discuss the key parameters and give the values used in the model. The discussion here is relatively brief; additional references used in determining parameter values for the model are listed for now at https://github.com/ketch/covid-blog-posts/wiki/Parameter-estimates.

The basic reproduction number σ_0. This is given by β/γ and is the average number of new infections that would be generated by one infected individual in a fully-susceptible population (in the absence of intervention). Early estimates of the basic reproduction number were low, with some even less than two [7], but more recent estimates are closer to four [5]. A survey of values for the basic reproduction number can be found in [3]. We take $\beta = 0.27$ and $\gamma = 0.07$, giving $\sigma_0 \approx 3.85$, similar to the value used in the March 30 Imperial College report [1].

The doubling time. This is the time it would take for the number of infected persons to double, and is given in the SIR model by $\log(2)/(\beta - \gamma)$; with our parameters this is about 3.47 days, which is consistent with observed patterns in many countries prior to intervention.

The infection fatality ratio F. This is the fraction of infected persons that die from COVID-19. There is no way to determine this directly because we don't know the total number of infected persons; we can instead calculate the case fatality ratio based on confirmed cases, which is much higher, especially in the case of COVID-19, due to the large number of mild and asymptomatic cases. Furthermore, the IFR for COVID-19 is known to vary strongly with age. We use age-specific IFR estimates from [6]; these have been reinforced by our own (unpublished) analysis of data from Spain and the USA. These are combined with UN demographic data for each country to determine an effective IFR:

$$F_{\text{eff}} = \frac{1}{N} \sum_j F_j N_j$$

where F_j is the IFR for age group j and N_j is the population in age group j. For regions in which we lack detailed demographic data, we use a value of 0.6% based on estimates from [4, 7, 6].

4 Sources of error and uncertainty

The adage that "all models are wrong, but some models are useful" is certainly applicable here. The difficulty of modeling an epidemic of a new disease while it is ongoing is high, and all results should be taken as rough estimates. Here we list some of the primary sources of error and uncertainty.

- Uncertainty in the key parameters β, γ, F. Published scientific estimates of each of these values vary by at least a factor of 2, and two of these parameters enter into the exponent of the exponential growth of the epidemic. Therefore, even small uncertainties in these parameters can lead to very large uncertainties in the model outputs.
- Inaccuracies in the input data. Our model intentionally disregards confirmed case numbers because they are a very poor indicator of real infections. Numbers of deaths are typically much more accurate, but still contain substantial errors. It may be expected that in some countries these numbers are deliberately manipulated for political purposes; in this case **it must be emphasized that the model outputs will be completely meaningless.** We believe this to be the case for instance in Iran and Russia, where evidence suggests the death toll has been drastically under-reported. But even in countries where the responsible parties try to report accurate numbers, real death rates are often significantly higher than those reported. For instance, comparison of excess mortality in Spain with published COVID-19 death numbers suggests that the death toll may be underestimated by about 1/3.
- Inaccurate model assumptions. The SIR model is one of the simplest epidemiological models. One of its key assumptions is that of homogeneous mixing among the population – i.e., that any two individuals are equally likely to have contact. This is of course far from true in the real world. More detailed models involving spatial structure and human networks can partially remedy this deficiency. We have stuck to the SIR model so far because it seems that the parameter uncertainties already listed above – which will impact any epidemiological model – are more signicant than this modeling error.
- Modeling the impact of intervention. Other models, such as that of [1], have attempted to explicitly model the effect of officially-announced intervention policies, lowering the contact rate by some amount starting when the policy is put in place. This requires an inherently uncertain estimation of the quantitative impact of a given policy, which will certainly vary in time and between different societies. To avoid this, we use an empirical assessment of intervention, choosing a contact rate that reproduces the data. This contact rate is somewhat sensitive to the noisy daily death rate data. Additionally, the model is slow to adapt to changes in intervention

policy, since their effects do not show up in the death rate until at least 2-3 weeks after they are implemented.

- Assumptions about future policy. Our primary forecast assumes that the current level of intervention will be maintained over the entire forecast period. For long-range forecasts, this is very unlikely. The model is not intended to show what will likely happen in reality, since future intervention policy is a matter of politics and not susceptible to detailed mathematical modeling. Rather, the forecast values should be interpreted as showing what will likely happen *if the current intervention stays in place.* We plan to release forecasts for other intervention scenarios in the future.

References

[1] Seth Flaxman, Swapnil Mishra, Axel Gandy, H Unwin, H Coupland, T Mellan, H Zhu, T Berah, J Eaton, P Perez Guzman, et al. Report 13: Estimating the number of infections and the impact of non-pharmaceutical interventions on covid-19 in 11 european countries. 2020.

[2] William Ogilvy Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721, 1927.

[3] Ying Liu, Albert A Gayle, Annelies Wilder-Smith, and Joacim Rocklöv. The reproductive number of covid-19 is higher compared to sars coronavirus. *Journal of travel medicine*, 2020.

[4] Timothy W Russell, Joel Hellewell, Christopher I Jarvis, Kevin Van-Zandvoort, Sam Abbott, Ruwan Ratnayake, Stefan Flasche, Rosalind M Eggo, Adam J Kucharski, CMMID nCov working group, et al. Estimating the infection and case fatality ratio for covid-19 using age-adjusted data from the outbreak on the diamond princess cruise ship. *medRxiv*, 2020.

[5] Marlena M Siwiak, Pawel Szczesny, and Marian P Siwiak. From a single host to global spread. the global mobility based modelling of the covid-19 pandemic implies higher infection and lower detection rates than current estimates. *The Global Mobility Based Modelling of the COVID-19 Pandemic Implies Higher Infection and Lower Detection Rates than Current Estimates (3/23/2020)*, 2020.

[6] Robert Verity, Lucy C Okell, Ilaria Dorigatti, Peter Winskill, Charles Whittaker, Natsuko Imai, Gina Cuomo-Dannenburg, Hayley Thompson, Patrick GT Walker, Han Fu, et al. Estimates of the severity of coronavirus disease 2019: a model-based analysis. *The Lancet Infectious Diseases*, 2020.

[7] Joseph T Wu, Kathy Leung, Mary Bushman, Nishant Kishore, Rene Niehus, Pablo M de Salazar, Benjamin J Cowling, Marc Lipsitch, and Gabriel M Leung. Estimating clinical severity of covid-19 from the transmission dynamics in wuhan, china. *Nature medicine*, pages 1–5, 2020.