Worldwide COVID-19 Case and Hospitalization Predictions

Jiying Zou*, Jongho Kim*, Keanu Spies*, Kofi Adu*, Praveen Pallegar*, Veer Shah*

(author names listed in alphabetical order)

with help from Prof. James Zou*, Irena Hwang*, Marco Giannitrapani**, Ryan Copping**, and Julie Ta**

*Stanford University, in collaboration with **Genentech, Inc.

1. Introduction

The global COVID-19 pandemic presents many challenges for society and our institutions. Arguably, one of the greatest issues is understanding and predicting the development of so-called "hot spots," regions where COVID-19 cases spread rapidly and put strain on local healthcare organizations and available resources. A variety of treatments have been proposed to treat COVID-19. Roche-Genentech are currently in the process of developing tests, treatments, and diagnostics for COVID-19, thus predicting these hot spots will be vital support in the decision-making process for supply, production, and distribution of tests and medicines. More generally, understanding case predictions on a country-level basis is crucial to efficient resource allocation and social protective policy establishment.

2. Problem statement

We propose the **SELICRD model**, based on the epidemiological SIR model, to predict the number of severe case hospitalizations over time and by country. Since severe-case patients utilize the most medical resources, they are likely to benefit most from effective treatment allocation.

3. Data Sources

Our models make use of several data sources: demographic data, death case data (daily), ICU bed data, and severe hospitalization case data (daily). Our data concerns the following countries: **Brazil, Canada, China, India, Italy, Netherlands, South Korea, Spain, UK, and USA**. We choose these countries for their large population size and consequently the large impact our work could have on the deployment of resources to patients in need.

3.1 Population by Age Group

The Demographic Statistics Database of the United Nations Statistics Department publishes population counts by age and age groups per country, accessible from the UN databank. We summarise these numbers into redefined age groups buckets of 10 years each (i.e. ages 0-9, 10-19, … 99-100) and retain counts from the most recent year available (*UNdata: Population by age, sex and urban/rural residence, 2019*) . To get estimates for 2020, we assume that age group distributions have not changed recently and proportionally scale their counts to each country's estimated 2020 population totals (*UNdata: Total population, both sexes combined, 2019*). These rescaled numbers are one of the main anchor inputs into our model.

3.2 Hospital Beds Data

Numbers for each country's hospital beds per 10,000 residents are provided by the World Health Organization (*UNdata: Hospital beds*). Most data come from 2008-2010, with the oldest estimate being for South Korea from 2002. According to this data source, most countries have not seen a dramatic change in hospital bed capacity over the 2000s, so we assume that these counts still hold valid for 2020. Although this may not be a solid assumption given the rapid demand for beds during this pandemic, we keep the number of hospital beds as a constant in our model for simplicity.

3.3 Confirmed Cases and Deaths Data

In general, we use the cumulative confirmed case and death count per country as provided by the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (*CSSEGISandData, 2020*). This has stood out as one of the most reliable and comprehensive data sources publicly available, compiling sources from official sources such as the WHO, ECDC, various CDC and Departments of Health, amongst others.

3.4 Severe Hospitalization Case Data

The severe hospitalization (ICU) data come from a variety of sources, ranging from government resources to reputable scientific repositories. We breakdown the sources below:

- Canada: Esri Canada in conjunction with provincial ministry reports (*Provincial Daily Totals, 2020*)
- China: National Health Commission daily reports (*Tracking the Epidemic, 2020*)
- Italy: Department of Civil Protection (*Rosini, 2020*)
- Netherlands: Coronavirus Hospitalized Patients (*Statista Research Department, 2020*)
- Spain: Investigative outlet Datadista reports, compiling data from the Spanish Ministry of Health, Department of Homeland Security, BOE, and Ministry of Transport (*Mobility and Urban Agenda*) (*Cámara, 2020*)
- US: The COVID Tracking Project by the Atlantic (*The Atlantic, 2020*)

We are not yet aware of well-documented severe case data for India, Brazil, Germany, and the UK. South Korea does not publicly release COVID-related hospitalization data for privacy reasons. We employ these data in the late stage of our model, where we use our predictions for cases of death per country to extrapolate cases of severe hospitalization (explained later in the **Model** section).

3.5 Data Caveats

For the sources we found, evaluating data accuracy proved difficult. Across various official sources, the data do not always agree. This is due to issues including, but not limited to, differences in disease-state definitions and administrative reporting redundancies, errors, and delays. For example, some regions may count deaths on date-of-occurrence while others wait for official death certificate issues and/or notice, creating a couple days' lag. Regarding definitions, "COVID-19 cases" could mean all COVID-like illnesses or only COVID-positive cases, and similarly "COVID-19 deaths" could include only COVID-19-induced ARDS deaths or also deaths from other comorbidities occuring during COVID-19 illness. A general lack of testing causes underreporting in both infection volume and rates, and what type of test results count differ between regions. While some data sources acknowledge these potential discrepancies, most do not post the necessary warnings, and each country experiences these challenges to

different extents over time. Even with the appropriate awareness, it is difficult to properly adjust the data, and so we choose to use them in their original reported form.

General data accuracy issues aside, a few blatant anomalies warrant explanation. On April 25th, Spain announced that they would count deaths on the date of occurrence rather than date of official notification, causing a 2000-count drop in its death data (*New York Times, 2020*). Around a week earlier, China announced a 1300-person addition to its death-toll who were not counted before (*Givetash et al., 2020*). South Korea has not publicly released any hospitalization-related data, and a lack of testing resources drives down India and Brazil's published numbers (*Phillips, 2020*). These are merely samples of the issues facing basic COVID-related data collection on an official level across the world.

4. Models

4.1 SELICRD Model

We modified and expanded on the SIR model, which is a popular compartment-based epidemiological model used for estimating infectious disease progression amongst a closed population. Each compartment represents a state of disease progression (e.g. "Susceptible", "Infected", "Recovered"). At any time point, each compartment has the following values:

1. A proportion of the total population in that compartment, defined in orange in Fig 3.
2. Probabilities associated with moving into and out of this compartment from neighboring compartments (e.g. how likely an infected case becomes critical), defined in blue in Fig 3.
3. Rates of entry from and to each of connected compartments (e.g. how quickly an infected case becomes critical, given that it does become critical), defined in red in Fig 3.

To estimate the numbers for any given compartment, we sum up the product of compartment size, chance of transition, and rate of transition from all possible other incoming compartments.

Our expanded model is called the **SELICRD Model** and involves the following disease-stage compartments: 1) Susceptible (not yet infected), 2) Exposed (carrying the virus but not yet infectious), 3) Latent (asymptomatic but can still transmit the virus), 4) Infectious, 5) Critical (in critical hospitalization or ICU care), 6) Recovered, and 7) Death. Building upon previous existing work (*Froese, 2020*), we added a compartment for critical-case patients to achieve our goal of modeling hospital resource needs. The Latent compartment is also a novel addition, added to more granularly model disease progression and achieve better prediction accuracy.

We define the model flow via a disease-progression dependency graph, representable by a system of ordinary differential equations (Fig 2). These can then be passed into a numerical solver to solve for parameter values and make estimates for each compartment. The model takes into account the specific hospital bed capacity per country and a time-dependent *R0* value to reflect the effects of implemented social distancing policies and other decreases in infection rate over time.

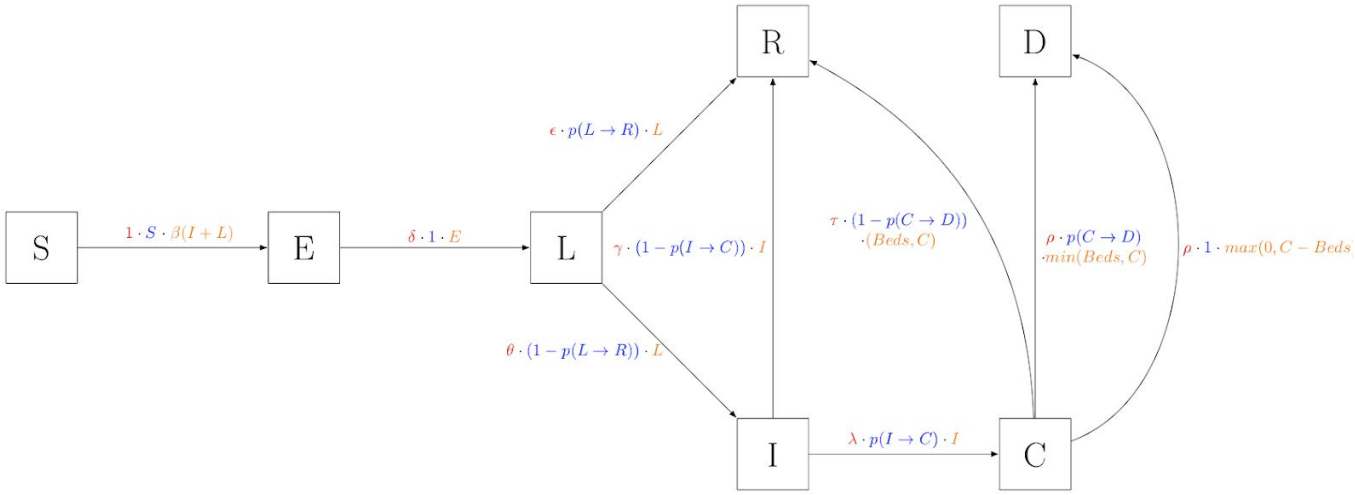


Fig 1. SELICRD model flow

$$\frac{\delta S}{\delta t} = -S * \beta * (I + L)$$
$$\frac{\delta E}{\delta t} = S * \beta * (I + L) - \delta E$$
$$\frac{\delta L}{\delta t} = \delta E - \epsilon p_{L \to R} L - \theta(1 - p_{L \to R}) L$$
$$\frac{\delta I}{\delta t} = \theta(1 - p_{L \to R}) L - \gamma(1 - p_{I \to C}) I - \lambda p_{I \to C} I$$
$$\frac{\delta C}{\delta t} = \lambda p_{I \to C} I - \tau \cdot (1 - p_{C \to D}) \cdot min(beds(t), C) - \rho \cdot p_{C \to D} \cdot \min(beds(t), C) - \max(0, C - beds(t))$$
$$\frac{\delta R}{\delta t} = \gamma(1 - p_{I \to R}) I + \tau(1 - p_{C \to D}) \cdot \min(beds(t), C) + \epsilon p_{L \to R} L$$
$$\frac{\delta D}{\delta t} = \rho \cdot p_{C \to D} \cdot min(beds(t), C) + max(0, C - beds(t))$$

Fig 2. ODEs describing the SELICRD Model

With such an expansive model, there are many parameters to solve for. These parameters fall into three classes: fixed parameters (requiring background research), fitted parameters via least squares regression, and R0 (fit using a separate model for decay, having sub-parameters of its own).

4.2 Fixed Parameters

All parameters regarding movement rate from one compartment to the next are treated as fixed parameters. Parameter values are shown below in Table 1. Note that rate is defined as 1 / # days on average to transition from one compartment to the next. Our parameter choices are explained below.

Parameter	Description (Rates)	Value (in 1/days)
δ	exposed → latent	1 / 2.5
	latent → recovered	1 / 14
θ	latent → infected	1 / 2
λ	infected → critical	1 / 11
γ	infected → recovered	1 / 17.5
τ	critical → recovered	1 / 11.5
ρ	critical → death	1 / 7.5

Table 1. Hyperparameters for the SELICRD Model

A number of factors affect γ, the recovery rate, with the most salient one being whether the infected patient develops Acute Respiratory Distress Syndrome (ARDS). Researchers indicate that most COVID-19 patients who do not develop severe symptoms take about 14 days to recover. However complications associated with more severe outcomes of the disease typically add another two weeks to recovery time for a total of 28 days. To represent $1/\gamma$, we take a weighted average of the time it takes a person with a mild form of disease and a severe form of the disease to recover. The proportion of people who eventually develop ARDS is a function of the size of a particular countries' elderly population and the percentage of people with underlying conditions (*Koma et al., 2020*). However, on average about 20% of a given population develops a severe illness due to the Sars-Cov-2 virus while the rest either have a mild case of the disease or show no symptoms at all. We make the simplifying assumption that this is a reasonable estimate for the majority of countries we analyze.

Information from the CDC indicates that COVID-19's incubation period is approximately 4.5 days, which we assume is consistent across countries. We can split this period into an exposed period and a latent period (corresponding to δ and θ, respectively). When a person has been exposed but does not yet have a high enough viral load to be contagious, we regard the person to be in the exposed category. The latent category is the subset of people who are contagious but have not yet developed symptoms. According to CDC Director Dr. Robert Redfield, exposed people are shedding the virus up to 48 hours before they show symptoms (*Whitehead et al., 2020*). Therefore we estimate the exposed period to be 2.5 days and the latent period to be 2 days for a total incubation period of 4.5 days.

We find from analysis of COVID-19 patient clinical progression data that patients with severe symptoms take about 11 days from first exposure to become critically ill (*Management of Patients with Confirmed 2019-nCoV, 2020*). We use this to estimate λ. For critically ill patients, on average death occurs in approximately 7.5 days or recovery happens in about 11.5 days ($1/\tau$). There have been reports of some patients rapidly deteriorating after initially experiencing mild symptoms, but we treat these as outliers.

Lastly, asymptomatic spreaders, people who have become exposed to the virus and are contagious but never develop symptoms are represented through . There is currently not a substantial amount of data about this group. However, initial evidence suggests that there is no reason to assume that the recovery period is significantly different from symptomatic carriers given the fact that people who do not show symptoms still have the capability to spread the virus. Since this group generally never becomes critical cases, we estimate their recovery to be within 14 days.

4.3 Fitted Parameters (via LS)

Since some parameters are difficult to research and define ourselves, we allowed them to be estimated via a LS fit, following the approach of Froese's models (*Froese, 2020*). We choose the parameter values minimizing MSE when predicting death count, given that death counts are widely available for countries and are more likely to reflect actual numbers (as compared to cases, which relies heavily on testing resources). We fit the following parameters:

- P(L to R): probability of recovering from the latent state
- P(I to C): probability of becoming critical from the infectious state
- P(C to D): probability of dying while in critical state
- R0 specific parameters: these define how R0 should decay over time (see next section)

4.4 $R0$ Model (Arctan Curve)

In epidemiology, the rate of infection is the probability of transmitting disease between a susceptible and an infectious individual. This is widely known as $R0$ (*Fisher, 2020*). Although there is no consensus for measurement, estimating *R0* is important to many epidemiological models. *R0* for an infectious disease over time is commonly modeled as some sort of reverse-S curve. Inspired by the recent DELPHI model (*Li, 2020*), model *R0* using an arctan function:

$$R_0(t) = \alpha * \gamma(t)$$
$$\gamma(t) = \frac{2}{\pi} \arctan\left(\frac{-(t-a)}{b}\right) + 1$$

In this model, α represents the initial infection rate and $\gamma(t)$ represents the government response at time t. The parameters are a and b which reflect the start of government response and the strength of such response respectively. Our model assumes that as the government starts responding to an epidemic (e.g. a shelter-in-place order), the rate of infection decreases. Therefore, we model $R0$ at each time t by multiplying an initial infection rate (α) with an arctan curve which represents governmental response. According to the "Overview of DELPHI Model" V2.0 (*Li, 2020*), this model is able to capture three phases of government response: 1) initial policy implementation phase, 2) active policy phase, and 3) saturation phase. For more detailed information, please refer to the paper (*Li, 2020)* and the explanation therein. As we discussed in the **Fitted Parameters** section, we fit a and b via LS as well.

4.5 Severe Case Prediction from Projected Deaths

The number of severe cases existing within a country per day are estimated using predicted deaths via a simple SLSC (scaled least-squares coefficient) model. We assume that all deaths come from severe cases, and that the death rate from severe cases is constant over time. This entails that when daily existing severe

cases are on the rise, deaths should be increasing at a rising rate. The greatest point of increase should correspond with peak severe case count, and thereafter the rise in deaths should be decelerating as severe case count decreases. Since we are using cumulative deaths, the number will never decrease over time, meaning the slope from a LS-fit can only be non-negative. Similarly, the daily existing severe case count is non-negative. Following this logic, we establish the following moel:

$$S_{i-lagdays} = c \cdot \beta_{D_{i-windowsize}, \ldots, D_i}$$

The interpretation is that the predicted number of existing severe cases on day (*i - lagdays)* is the LS slope estimate from regressing cumulative deaths (over a time window) against time (in days), scaled by a country-specific constant *c*. This can also be seen as a locally-weighted linear regression with an asymmetric neighborhood/sliding window of size *windowsize* and with constant kernel weights *c*. For example, the LS-slope calculated over days 2-17 with 3 lag days is assigned to day 14 as its severe case prediction. Generally, a sliding window of 15 days and lag of 3 days works well, but the country-specific constant should be tuned on a case-by-case basis.

We define severe cases as the number of patients in critical care or the ICU per day in all hospitals across the country. Sometimes critical/ICU case data is not available, so total hospitalizations per day could be used as a substitute, with appropriate adjustments for the constant *c*. In doing so, be wary that hospitalization admission policy and tallying procedures could vary by region and over time, and that not all hospitalizations result in severe cases.

5. Experiments

We used backtesting to validate our model performance and evaluate the amount of data needed to make reasonable predictions. To do so, we train our SELICRD model on a portion of the training data, and test our model on the remaining left out portion. Because our data is a time series, we do not randomly sample points for the training set; instead, we train on data from a given data point up to ***x*** number of consecutive days (e.g. 70%, 80%, 90% and 95% of whole available data set), and let the test set be on the remaining data. The starting date is defined as the first day that a given country hits ≥100 COVID-19 cases. Models are judged using mean absolute prediction error (MAPE) on the test set, and we select the model with lowest test error. Figure 3 illustrates examples for the US and Brazil.

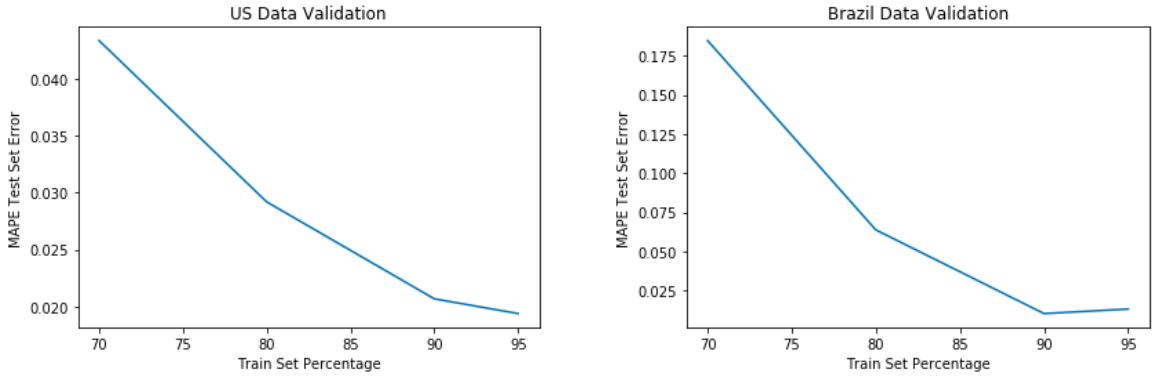


Fig 3. Validation results on SELICRD model using US (left) and Brazil (right) data

6. Results

For each of our target countries (United States, China, Brazil, India, Germany, Italy, Spain, United Kingdom, France, Canada, Netherlands and South Korea), we generate forecasts for the total number of deaths in the next 60 days. Figure 4 shows each country's projected deaths, and Figure 5 shows predicted and actual deaths and severe cases. Severe cases are predicted using predicted deaths data. Severe case predictions are not available for countries with missing hospitalizations/critical care/ICU data.

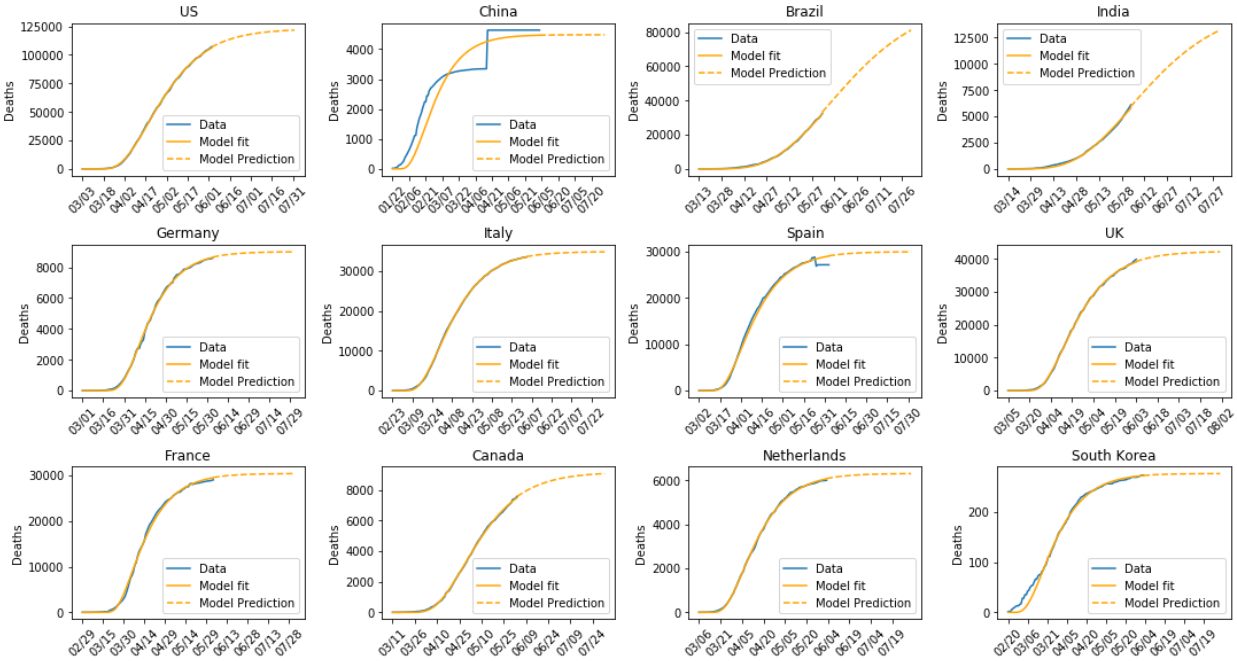
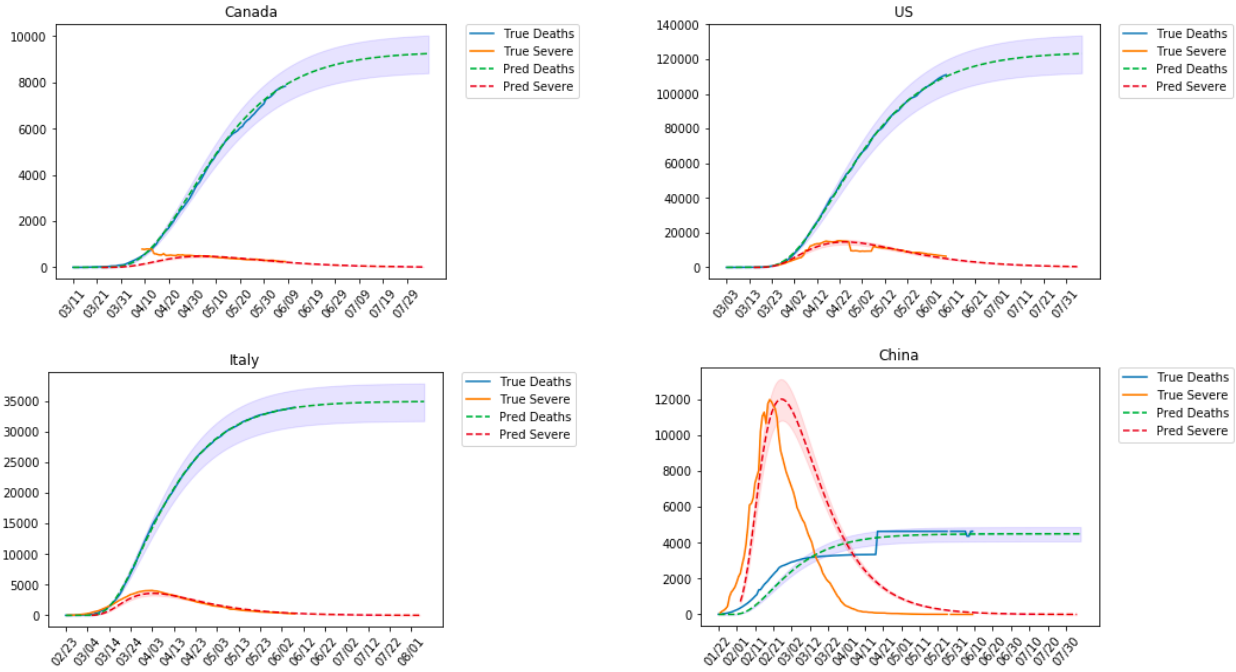


Fig 4. Death count projections for various countries from individually fitted SELICRD models



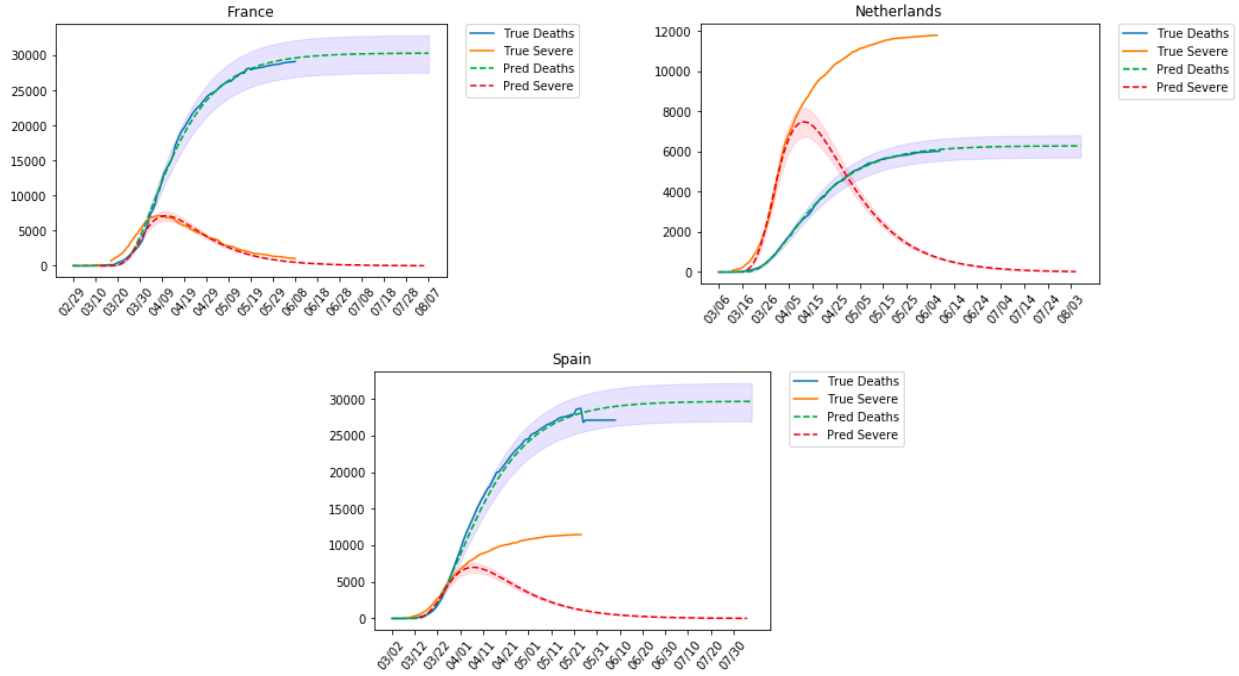


Fig 5. Death and severe hospitalization predictions with prediction uncertainty bands

In addition to the number of death forecasts on countries, we generated upper/lower prediction bounds by re-running our models with +/- 10% on the P(C to D) parameter estimate. These form the intervals shown in Fig 5. The main reason for providing intervals is that although best fitted parameters are determined in our model, it is still possible for these parameters to deviate from cases in the real world. In order to provide more robust results, we generated upper/lower bounds on each forecast and believe this may assist researchers/users in terms of interpretation and analysis on current epidemic phenomena.

Fig 5 also shows the model prediction on the number of severe cases. More specifically, for the US, Canada, and Italy, models predict For France and China, our model predicts the daily total critical care patients. For Spain and the Netherlands, models predict cumulative hospitalizations over time, which performs poorly due to the different-shaped curves. Note that these predictions depend on data availability.

7. Discussion

7.1 Strengths

The SELICRD model with an arctan *R0* submodel and LS-fitted parameters generally fits well on existing data and makes reasonable future projections. It considers government intervention effects (via a time-changing *R0* value) and resource disparities (specifically bed capacity per country). To avoid extrapolation error, the model should be re-run and parameters updated as more data becomes available.

7.2 Limitations

The model significantly depends on the data, so it is sensitive to data anomalies (as we discussed in **Data Sources**). Due to the complex nature of the model, and the relatively small dataset for each country (~120 days of data), it is likely that the model is overfitting to the data, which explains its sensitivity to changes in the data. Future work can be designing an improved model which is more robust to data anomalies and noise.

Furthermore, a model is only as good as the data supporting it. We have also discussed a variety of challenges with the credibility of our data despite being from official sources. For countries without data on severe hospitalizations, it may be possible to use proxy information, like the percentage of hospitalized COVID-19 patients who die, to extrapolate the severe cases from our cases of death. We have not yet found a reliable pattern to anchor predictions on. Improved data quality will no doubt produce better forecasts.

7.3 Future Directions

It may be possible to achieve similar performance using differently-specified disease states or a different *R0* model. While we have explored some possibilities, our chosen model is by no means the best model of its type. We invite future researchers to experiment with similar methods involving fewer parameters.

We made prediction bands by toggling values of P(C to D) within reasonable bounds. There are several other options for creating prediction bands, for example by varying multiple parameters by some reasonable magnitude, or by finding some way to calculate standard errors with statistical support.

This model can be applied to make projections for deaths and severe case hospitalizations for other countries or regions as well, given the appropriate type and amount of data. The challenge is that finding accurate and substantive data can be difficult. It may be possible to extrapolate across regions. Suppose that we are interested in studying the development of the virus in region *Y*. If there is no data, or suspected unreliable data for that country, we could find region *X*, which has a similar demographic age distribution and similar healthcare capacity and infrastructure for which there is substantial data. After training the model on region *X*, we can apply the model parameters to region *Y* to obtain a "transfer-learning" guess for region *Y*'s expected deaths and critical cases. It is important to note that such results cannot be easily validated, so any policy and / or hospital resource decisions based on these results should also consider factors outside of these projections.

A final interesting question is how accurate these projections prove to be over time. We are yet unsure how often the model needs to be updated (although we recommend at least weekly), or for what types of countries are better predictions made. While we can visually assess fit by country, the broader scope of applicability (e.g. to county, province, continent levels) remains in question.

Resources

UNdata | record view | Population by age, sex and urban/rural residence. (n.d.). Retrieved from http://data.un.org/Data.aspx?d=POP&f=tableCode:22

UNdata | record view | Total population, both sexes combined (thousands). (n.d.). Retrieved from https://data.un.org/Data.aspx?d=PopDiv&f=variableID:12

UNdata | record view | Hospital beds (per 10 000 population). (n.d.). Retrieved from http://data.un.org/Data.aspx?q=beds&d=WHO&f=MEASURE_CODE:WHS6_102

CSSEGISandData. (2020, June 7). CSSEGISandData/COVID-19. Retrieved from https://github.com/CSSEGISandData/COVID-19

Provincial Daily Totals. (n.d.). Retrieved from https://resources-covid19canada.hub.arcgis.com/datasets/provincial-daily-totals/data?page=12&selectedAttribute=DailyICU

Tracking the Epidemic. (n.d.). Retrieved from http://weekly.chinacdc.cn/news/TrackingtheEpidemic.htm

Rosini, U. (n.d.). pcm-dpc/COVID-19. Retrieved from https://github.com/pcm-dpc/COVID-19/blob/master/dati-andamento-nazionale/dpc-covid19-ita-andamento-nazionale.csv

Statista Research Department. (2020, June 8). Netherlands: coronavirus hospitalized patients 2020. Retrieved from https://www.statista.com/statistics/1109441/coronavirus-cumulative-hospital-cases-in-netherlands/

Cámara, C. (n.d.). datadista/datasets. Retrieved from https://github.com/datadista/datasets/blob/master/COVID 19/ccaa_covid19_uci_long.csv

The Covid Tracking Project Data API. (2020). Retrieved from https://covidtracking.com/api

The New York Times. (2020, May 4). Spain Coronavirus Map and Case Count. Retrieved from https://www.nytimes.com/interactive/2020/world/europe/spain-coronavirus-cases.html

Givetash, L., Chen, L., & Liu, D. (2020, April 17). Coronavirus: China updates death toll in Wuhan, adds 1,300 to total. Retrieved from https://www.nbcnews.com/news/world/coronavirus-china-announces-jump-death-toll-wuhan-china-n1186006

Phillips, D. (2020, June 7). Brazil stops releasing Covid-19 death toll and wipes data from official site. Retrieved from

https://www.theguardian.com/world/2020/jun/07/brazil-stops-releasing-covid-19-death-toll-and-wipes-data-from-official-site

Froese, H. (2020, April 22). Infectious Disease Modelling: Fit Your Model to Coronavirus Data. Retrieved from https://towardsdatascience.com/infectious-disease-modelling-fit-your-model-to-coronavirus-data-2568e672dbc7

Koma, W., Neuman, T., Claxton, G., Rae Follow, M., Kates, J., & Michaud, J. (2020, April 23). How Many Adults Are at Risk of Serious Illness If Infected with Coronavirus? Updated Data. Retrieved from https://www.kff.org/global-health-policy/issue-brief/how-many-adults-are-at-risk-of-serious-illness-if-infected-with-coronavirus/

Whitehead, S., & Feibel, C. (2020, March 31). CDC Director On Models For The Months To Come: 'This Virus Is Going To Be With Us'. Retrieved from https://www.npr.org/sections/health-shots/2020/03/31/824155179/cdc-director-on-models-for-the-months-to-come-this-virus-is-going-to-be-with-us

Management of Patients with Confirmed 2019-nCoV. (2020, May 20). Retrieved from https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-guidance-management-patients.html#:~:text=The incubation period for COVID,CoV-2 infection

Fisher, M. (2020, April 23). R0, the Messy Metric That May Soon Shape Our Lives, Explained. Retrieved from https://www.nytimes.com/2020/04/23/world/europe/coronavirus-R0-explainer.html

Li, M. (April 2020). Overview of DELPHI Model V2.0. Retrieved from https://www.covidanalytics.io/DELPHI_documentation_pdf