**Technical paper on Dalberg’s CoVID-19 Model**

Various experts, statisticians, and businesses are working on a wide range of epidemiological models to gain deeper understanding of CoVID-19. Tools derived from such models can predict broader impact of the disease in a geography, identify right policy interventions and enable better allocation of resources.

The Dalberg model is unique due to two main reasons:

1. We have used a modified version of the standard SEIRS model to build our projection engine. These modifications allow the model to reflect **unique features of CoVID-19 disease**, such as being infectious without exhibiting any symptoms, and different disease reproduction rates for people with and without symptoms
2. By applying a machine-learning on top of the simulation engine, the model is capable of identifying ‘dynamic’ parameters of the disease in near real-time and adjust the projections accordingly. This **helps create a ‘sandbox mode’ for policy interventions** and observe their impact within few days of implementing an intervention

There are four steps involved in developing this model:

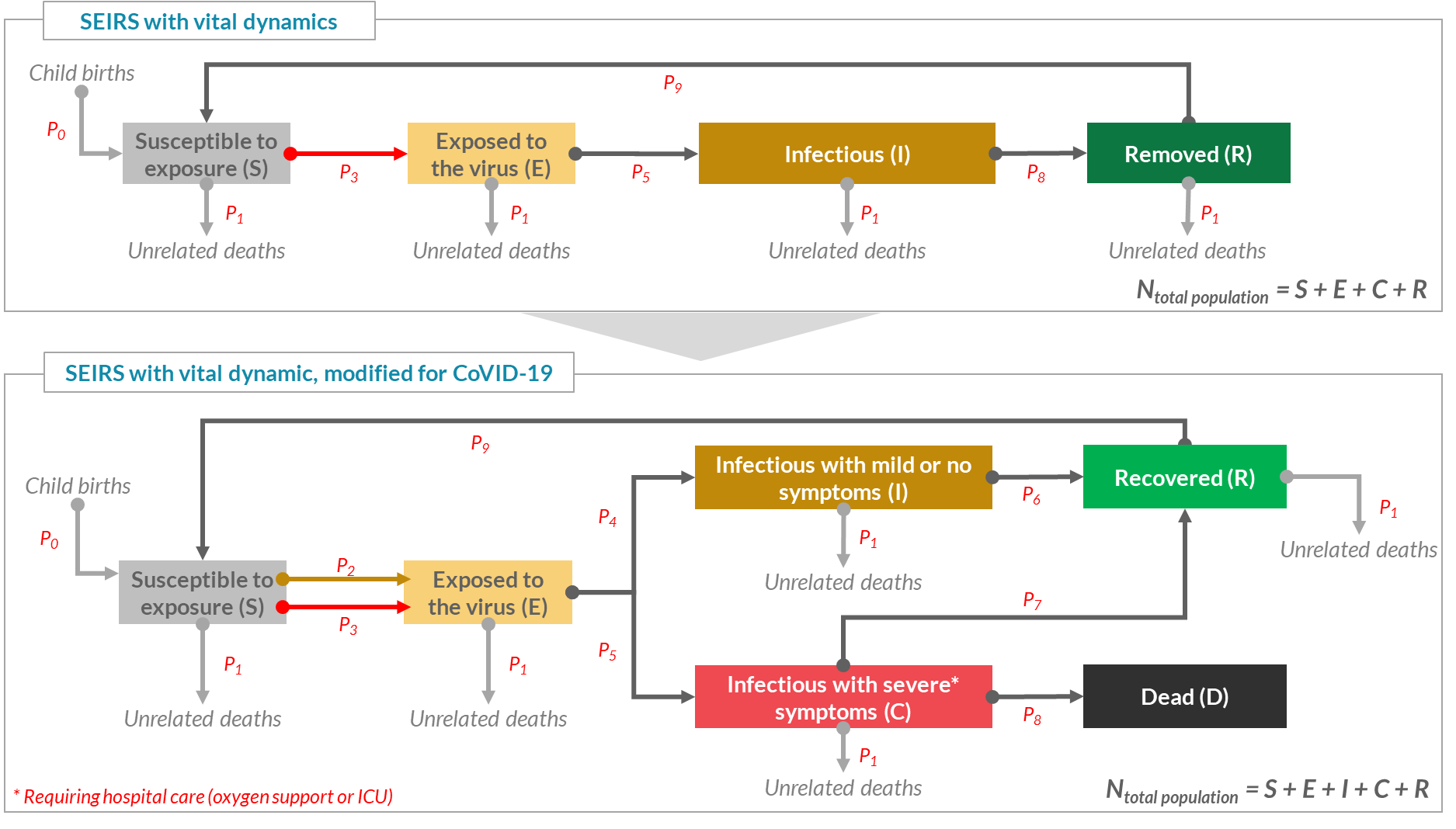
* **Step I**: Selecting the epidemiological model
* **Step II**: Developing differential equations governing shift of population through the disease cycle
* **Step III**: Understanding the nature of disease features
* **Step IV**: Building the simulation engine
* **Step V**: Building the Machine Learning layer for real time prediction

**Step I: Selecting the epidemiological model**

We picked a generalised SEIRS epidemiological model with vital dynamics, and made two specific modifications to it (Figure 1).

* We consider two categories of people who can be infectious – those who do not display any symptoms or display only minor symptoms, and those who display severe symptoms[[1]](#footnote-1). The lack of obvious symptoms in infectious individuals is an important characteristic of the CoVID-19, also making it a very insidious disease and hence important to model
* We have split ‘Removed’ into ‘Recovered’ and ‘Dead’ as these statistics are important for understanding the extent of casualty and for reliably using the ML algorithm, as we will explore.

Figure 1: Schematic representing shift of population through SEIRS model of disease cycle, and subsequent CoVID-19 modifications



Please note that for any short-term projection, the vital dynamics components (i.e. birth and death rates) and disease recurrence rate, if any for CoVID-19, will have no reasonable impact. Additionally, the model assumes a closed system with no movement of people in and out of the system, except due to birth/death.

**Step II: Developing differential equations governing shift of population through the disease cycle**

Aligned with the above schematic, we have following differential equations

* *N = S + E + I + C + R*
* *P0\*N + P9\*R - P1\*S - P2\*S\*(I/N) - P3\*S\*(C/N)*
* *P2\*S\*(I/N) + P3\*S\*(C/N) - P4\*E - P5\*E - P1\*E*
* *P4\*E - P1\*I - P6\*I*
* *P5\*E - P1\*C - P7\*C - P8\*C*
* *P7\*C + P6\*I - P1\*R - P9\*R*
* *P8\*C*

Where, at any given time,

* **N** is the total population
* **S** is part of the population that is susceptible to catching the disease
* **E** is part of the population that has been exposed to the virus, but hasn’t become infectious yet
* **I** is part of the population that is infectious, but showing mild or no symptoms
* **C** is part of the population that is infectious with severe symptoms, requiring hospitalisation
* **R** is part of the population that has recovered from the disease
* **D** is population that has died due to the disease and is not a part of the population

And the parameters (described in terms of disease or demographic features) are,

* **P0 =** Births per day as a percentage of total population

= Annual birth rate per thousand x (1000/365)

* **P1 =** Non infection related deaths per day as a percentage of total population

**=** Annual death rate per thousand x (1000/365)

* **P2 =** Number of virus exposures caused per day by each infectious, non-symptomatic individual

**=**

* **P3** **=** Number of virus exposures caused per day by each individual with severe symptoms

**=**

* **P4 =** Percentage of exposed who become infectious per day, with mild or no symptoms

**=**

* **P5:** Percentage of exposed who become infectious per day, with severe symptoms

**=**

* **P6:** Percentage of infectious people with no or mild symptoms, who recover every day

**=**

* **P7**: Percentage of infectious people with severe symptoms, who recover every day

**=**

* **P8:** Percentage of infectious people with severe symptoms, who die every day

**=**

* **P9:** Percentage of recovered individuals who may re-contract the virus per day

**=**

**Step III: Understanding the nature of disease features:**

Based on above description, the model requires 9 disease features as inputs. These features can be classified into two categories: six **static features** and three **dynamic features**.

* **Static features** are features of the disease that are mainly inherent to the disease itself and can remain largely unchanged across communities. Static nature of these features allows us to pick their values from global studies. However, it should be noted that these features may mutate in the future with the virus itself:

|  |  |
| --- | --- |
| **Static disease features** | **Value** |
| ***Percentage of exposed who become infectious with severe symptoms*** | 19% |
| ***Percentage of recovered population who may re-contract the virus*** | 0% (unused) |
| ***Days an exposed person takes to become infectious*** | 5 |
| ***Days for which a person with mild or no symptoms remains infectious before recovery*** | 20 |
| ***Days for which a person with severe symptoms remains infectious before recovery or death*** | 20 |
| ***Days before a recovered person re-contracts the virus*** | 30 (unused) |

* **Dynamic features** are those disease features that also depend on community circumstances and can be heavily influenced by the way in which communities and governments respond to CoVID-19. We identify 3 such features:
  + ***Number of exposures caused by infectious people with mild or no symptoms:*** This is disease’s basic reproduction rate for people showing mild symptoms, and can be influenced by community customs, population density, lockdowns, social-distancing, personal hygiene and usage of masks
  + ***Number of exposures caused by infectious people with severe symptoms:*** People showing severe symptoms may naturally have a higher disease reproduction rate, however they may also be easily identifiable. With a good quarantine mechanism in place, such people can be easily separated from the population, lowering their disease reproduction rate
  + ***Percentage of infectious people with severe symptoms who recover:*** This will depend on a community’s comorbidity factors (e.g. high diabetes prevalence) and access to healthcare

Dynamic nature of these features along with model’s high sensitivity towards them[[2]](#footnote-2) implies that values estimated in one location may not be used in another, and only measurements made in local and current context can provide reliable forecasts.

**Step IV: Building the simulation engine**

Using the model equations from Step II and values of *static features* from step III, we built a deterministic model in Python (WIP) and excel (Completed). Figure 2 shows time variance of susceptible (S), Exposed (E), Infectious with mild symptoms (I), Infectious with severe symptoms (C), Recovered (R), and Dead (D) over the next 2-years for India, using three set of values for dynamic features. As indicated earlier, the projections of peak hospital requirement and total number of fatalities vary significantly across the three set

*Figure 2: Simulation model to be updated*

**Step V: Building the Machine Learning layer for real time prediction**

Once we have the underlying simulation engine, in the next step we built a machine learning layer to figure out what set of values for the dynamic parameters best explains the real-world data. In other words, we try to find values of the dynamic features which best fits the projection curve on the real-world data.

Recognising lack of any standard R or Python library to run regression on custom models, such as those built using a modified SIERS, we have achieved the curve-fitting using first principle approach: We compare modelled values against the real-world data, estimate the overall error term, and allow disease features to vary within certain constraints with an objective to minimise this error term. This approach allows us to convert an ML problem into an Optimisation problem, where we try to find the minima of the error term in the *dynamic features* plane. In excel, we implement this optimisation using the *solver add-on* with GHG-Non-Liner algorithm. In Python (WIP), we take a brute-force approach. With just 3 dynamic features that can vary in small ranges, efficiency of calculation with brute-force was not found to be an issue. However, there remains a scope to make optimisation algorithm in python significantly more efficient.

For this curve-fitting exercise, we select two most commonly available daily data feeds for most countries (note that other real-world data feeds, such as number of hospitalised CoVID-19 cases, if it reliably reflects all severe cases, can also be used for this exercise):

1. Daily number of CoVID-19 deaths
2. Daily identified cases of CoVID-19

*A note on the overall Error term*

**Including multiple data feeds:** While, reported number of CoVID-19 deaths can be one of the most reliable data feeds, in a country such as Singapore with only 4 deaths[[3]](#footnote-3), it may also lack any statistical power. On the other hand, large number of daily identified cases can provide the required statistical power but may also be under-reported. To solve for this, we estimate overall error terms by assigning pre-specified weights to the normalised version of respective errors in the two data feeds. These weights reflect our preference (and trust) for the data feed to be used in fitting the curve.

Where,

represents number of days before today for the data point; is square error for the ith previousday from today in number of deaths; is square error for the ith previousday from today in number of cases; is weight assigned to the number of deaths data feed; is weight assigned to the number of cases data feed

**Building a recency bias:** Additionally, Given that the situation is rapidly evolving, including awareness amongst people toward social distancing and personal hygiene, and governments’ efforts, we also wanted to ensure that the ML model gives more weightage to recent data points compared to older data points. For this we have implemented a slow exponential reduction, where each previous day has 90% weightage than the subsequent day. Using this factor, the weight reduces to 60% for 5th previous day, 37% for 10th previous day, and 5% for 30th previous day compared to today.

**Results (so far):**

**For Singapore: TBU**

**For India: TBU**

**Next steps:**

* Finetuning the values of static features using global data
* Extensive testing and feedback
* Develop deployable versions that can be picked by local and state governments to test interventions

1. Severe symptoms: We define this group as those who require hospital care for oxygen or ICU support [↑](#footnote-ref-1)
2. Note on sensitivity of disease prediction against reproduction rates [↑](#footnote-ref-2)
3. As on 2nd April 2020 [↑](#footnote-ref-3)