A Method to Model Outbreaks of New Infectious Diseases with Pandemic Potential such as COVID-19

Willem G. Odendaal

Abstract

The emergence of the novel coronavirus (a.k.a. COVID-19, SARS-CoV-2) out of Wuhan, Hubei Province, China caught the world by surprise. As the outbreak began to spread outside of China, too little was known about the virus to model its transmission with any acceptable accuracy. World governments responded to rampant minformation about the virus leading to collatoral disasters, such as plunging financial markets, that could have been avoided if better models of the outbreak had been available.

This is an engineering approach to model the spread of a new infectious disease from sparse data when little is known about the infectious agent itself. The most sought after estimate is the projected number of people who may be contagious, but still be within the incubation period thus remaining asymptomatic and undetected. This parameter is useful for estimating the extent of outbreak among a population over time and what to expect. It is thus important for common sense decision making by governments, organizations, and even individuals to contain, prepare, prevent, react, and safeguard. A Monte Carlo approach is implemented by using incubation period and a variable, called *testing efficiency* as variables.

The Coronavirus outbreak in USA will be used to illustrate the implementation of this modeling approach. Among others it is demonstrated that imposing early travel restrictions from infected countries slowed down the outbreak in the USA by about 26 days.

I. Introduction

EVEN before the coronavirus was officially declared a *pandemic*, it was already being heavily politicized in the USA already in the midst of a presidential election year. Coloring the outbreak with political slants in the media naturally exacerbated the transmission of misinformation faster than the virus itself. This lead to extreme over- and under reactions in society leading to collatoral calamities including: a crash in the stock market, unnecessary quarantine measures often involving harsh conditions, a blow to the travel industries, shortages and price gouging of medical supplies, and unnessecary panic in general born out of uncertain expectations.

If better modeling tools had been utilized and prediction widely communicated to the media and social media well in advance, these ramifications could have been avoided, or at least lessened. The problem is that health professionals and organizations speicializing in epidemics and infectious disease, were hard at work in laboratories focused on researching and gathering information about this new strain of which very little was known, and working on cures. Fortunately, the People's Republic of China had shared their sequencing so that other countries could also develop tests for detecting infections.

The challenge being addressed herein, is to model the spread of a new infectious disease knowing very little, other than incoming world data of confirmed infections and confirmed fatalities. The model can then be used to model, predict, and present likely trends of the outbreak.

II. Core Ingredients of the Model

This model is characterized by the following:

- Purpose
- Assumptions
- Mathematical approximation
- Presentation

A. Purpose

The main purpose of the model is to predict the spread of a new infectious disease knowing very little, other than sparse collected data such as incoming world data of confirmed infections and confirmed fatalities. Likely scenarios can then be presented based on known and unknown parameters that are clearly defined for societal decision making. Purpose is closely tied to presentation. One of the most important figures for which estimates can be projected using this model includes the unknown number of people who have already been infected on any given day in the future, especially if asymptomatic infections are also contagious during the incubation period.

B. Assumptions

It is very important to identify, define, and declare all assumptions in a model. For coronavirus, very little was known, so the following assumptions were made:

1) Unknowns: It is assumed that nearly nothing is known about the disease, except that it is an infectious disease, specifically a betavirus, that may or may not have similar characteristics to other known viruses. Unknown characteristics, parameters, variables, statistics, etc. may include among others:

- average incubation period, latent period, etc.
- means of transmission
- the level of asymptomatic transmission, if at all
- resilience to temperature, humidity, etc.
- survival rate on objects and surfaces
- testing accuracy (The percentage of infected people with symptoms tested, who tested positive.)
- testing efficiency (The percentage of all infected people with symptoms, who tested positive.)
- reliability of sample data

2) Reactionary Fine-Tuning: As with any model, it should be fine-tuned as more information is gathered. It should also be adjustable to external events that will influence the rate of the outbreak, or dramatic changes in the incoming statistical data.

3) The Law of Averages: For illustration purposes the law of averages will be used. The method can then easily be extended by implementing statistical theory with probabilities and uncertainties.

4) Valid Data Groups from global Data Population: For coronavirus the data population consists of case statistics from all over the world. These numbers were being dominated by case numbers from China. However, when plotting the cumulative number of confirmed cases on the planet as a function of time, the curve is artificially erratic. To establish a more reliable trend, the data from the PRC was therefore subtracted from the data collected from the rest of the world. It is then assumed that the rate at which the outbreak spreads in most countries should, on average, roughly follow the average rate in the world excluding China.

5) Mathematical Form: The cumulative rate of infections is assumed to be an exponential of the form:

$$F(t) = A\mathrm{e}^{b(t)} \tag{1}$$

where A and b are variables to be determined and t is time. The number e is the base of the natural logarithm equal to 2.71828.

This first-order equation is not the only mathematical formula that can be used. However, equation (1) has the unique property that it's time derivative and time integral results in the same function, except scaled by an amount. In other words, the virus is spreading at a rate that accelerates, that the accelerated rate itself accelerates, and so on. Another unique property of this function is that neither linear scaling of it, nor a constant shift in time, changes the time dependent exponent. Thus, time delays and taking percentages are interchangeable and can be transformed one into the other. In fact, derivates, integrals, time shifts, and linear scalling all result in a function of the same form and changes only the coefficient, A.

C. Discretizing the Model

Since data and graphs are best presented by date, time will be quantized as days in i, dates by $d =$*day number*, and *time periods* such as time delays also in days, by $p =$*period in days*. The difference between these two is that d might, for example, represent a date in the future, while p might represent an intrinsic parameter such as the incubation period, or the number of days that the average world spread lagged or lead the first introduction of the virus in a specific country.

The function is then rewritten as follows:

$$\begin{aligned} F(d,k) &= A\,\mathrm{e}^{b(d+p)} \\ &= A\,\mathrm{e}^{bd}\mathrm{e}^{bp} \end{aligned} \tag{2}$$

If $F(d,k)$ represents the number of new cases on a particular day, then its time integral gives the cumulative cases over a period:

$$\begin{aligned} G(d,p) &= \sum_{i=1}^{d} F(d,p) \\ &= \sum_{i=1}^{d} A\,\mathrm{e}^{b(d+p)} \\ &= \frac{A}{b}\,\mathrm{e}^{b(d+p)} \end{aligned} \tag{3}$$

If F represents a cumulative figure, then its value at a particular time is:

$$\begin{aligned} H(d,p) &= \frac{\Delta F(d,p)}{\Delta d} \\ &= F(d,p) - F(d-1,p) \\ &= Ab\,\mathrm{e}^{b(d+p)} \end{aligned} \tag{4}$$

where $\Delta d = 1$ day.

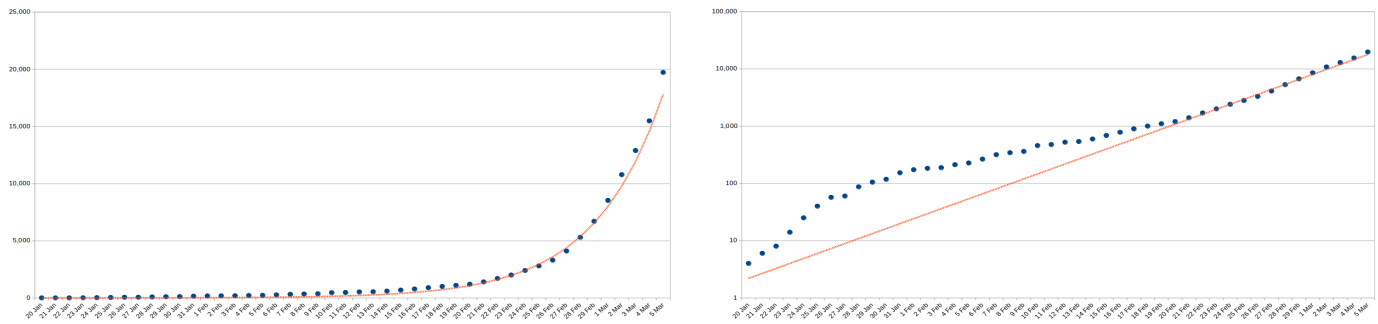
D. Implementing the Model for Novel Coronavirus

1) Modeling Cases for the World excluding China: To model the coronavirus outbreak in the world outside of China, a curve-fit is performed on the data collected.

The resulting equation for cumulative cases has the coefficients $b = 0.201$ and $A = 1.9$:

$$F(d)_{\text{World-China}} = 1.9\,\mathrm{e}^{0.201\,d} \tag{5}$$

The following figures show the correlation between the model and the case data for the world besides China:



(a) Linear Scale (b) Logarithmic Scale

Fig. 1. Cumulative number of confirmed coronavirus cases and the model for the world outside of China.

2) Practical Assumptions to Model Novel Coronavirus: To simplify the model and frame the modeling space, the following assumptions are made:

1) Asymptomatic infections are contagious some time after infection.
2) The average person who has been infected will show symptoms after the average incubation period.
3) If testing could be 100% efficient, the average infected person will become a confirmed case statistic exactly upon expiration of the average incubation period since infection.
4) The model must react sensibly to external events.

Using these assumptions:

- The cumulative number of people infected on a given day are then estimated by shifting the curve into the past by the average incubation period.
- Inefficient testing percentages are accounted for by dividing the result by the testing percentage.

Note that these two operations can be performed in any order.

E. Modeling the Outbreak in the USA

1) Initial Trend: Up until the end of February, the cumulative confirmed cases in the USA followed a function of the form:

$$F(d) = 1.1\,\mathrm{e}^{0.1\,d} \tag{6}$$

The rate of infection was initially significantly lower than the world average (outside of China). This can mainly be attributed to (i) imposing early flight restrictions on 31 January 2020 from heavily infected countries resulting in almost all active cases being isolated, and (ii) testing had only been done case-by-case on people who showed symptoms and had entered the country from heavily infected areas. At this point, the outbreak was largely contained.

2) Influence of Travel Restrictions: The imposition of early travel restrictions from China into the USA slowed down the virus by containing it mainly to individuals who had been to infected areas. The model indicates that the early flight restrictions slowed the virus down by 26 days, before it became ”community-spread” in the USA. (Here, the term *community-spread* refers to a combination of person-to-person and surface-to-person transmission modes.)

The following figure summarizes the two trendlines before and after the outbreak in the USA became community-spread. After it enters community-spread mode, a trend identical to the world’s trend (excluding China) is assumed.

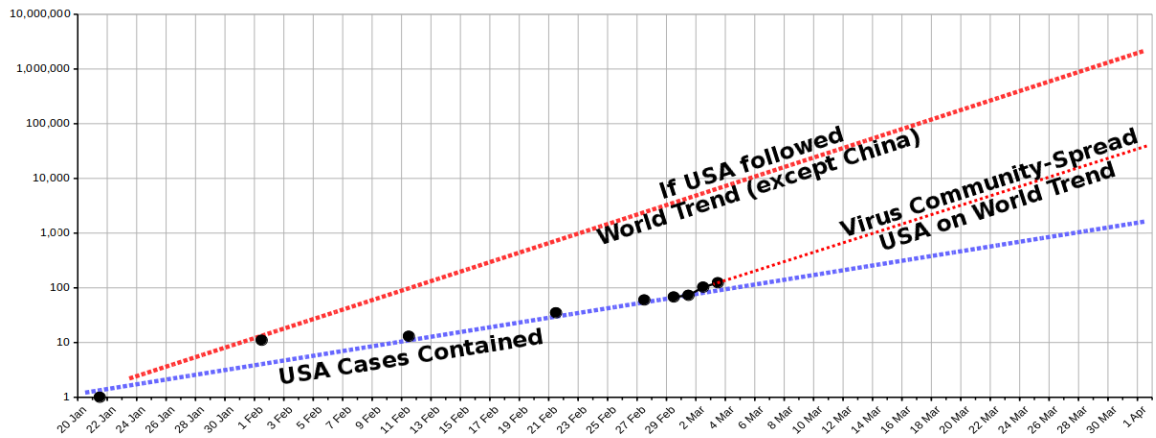


Fig. 2. USA confirmed cases, the USA models before and after *community-spread* mode, and the model for the world outside of China.

3) Influence of Testing Policies: During the first week of March 2020, the USA started expanding testing beyond their initial requirements. Testing was approved for people showing symptoms who did not have direct contact with people who had entered the country from infected areas. Production of testing kits in mass numbers began for distribution to the state laboratories. By this time, the virus had already left containment, It is reasonable to assume that the number of cases being confirmed were lagging the actual number of people becoming ill. As tests were being carried out to catch up with the rate at which infections were becoming symptomatic and being reported, a surge in data could be expected. Therefore, the model was *not* adjusted for the surge. The model for the world outside of China was still used after 29 February even though the rate at which actual new cases were being confirmed in USA exceeded the rate predicted by the world model. This ensured that the projected estimates of unreported infections weren’t artificially being inflated due to the lack of initial test data.

F. Introducing Testing Efficiency

Testing efficiency, a percentage, was previously defined as follows:

$$\eta = \frac{\text{total confirmed cases}}{\text{total number of symptomatic infections}} \times 100 \tag{7}$$

In other words, it represents the percentage people who have been tested positive out of all people who should have tested positive. It incorporates other parameters such as testing accuracy.

1) Estimating the Testing Efficiency in USA: It was elsewhere reported that less than 5% of the infected were tested in Wuhan, China following the outbreak. Since the USA's CDC and FDA limited testing just to individuals who were symptomatic and had entered the country from infected areas, the available sample data is extremely sparse and the actual testing efficiency in USA was anyone's guess. To improve the estimate, the fatality rate may be a better indicator as it is likely to be more precise than the statistic for confirmed cases, which is also dependent on self-reporting. Table II-F1 contains the fatality percentages of USA, the world, and a few other countries collected on 10 March 2020. The 2018-2019 figures for the common flu are also included for comparison.

	Confirmed Cases	Deaths	Fatality %	US ratio %
World	119,215	4,298	3.61%	122%
World outside China	38,437	1,140	2.97%	100%
South Korea	7,755	61	0.79%	27%
Germany	1,908	3	0.157%	5%
Italy	12,462	827	6.64%	223%
USA	1,010	30	2.97%	100%
Influenza (USA '18-'19)	44.8 million	61,099	0.136%	5%

If it is assumed that the testing efficiency for most countries was still improving on 10 March 2020, then it's also reasonable to assume that the true fatality rate of the coronavirus as a percentage will get lower with time after that date. Germany's fatality rate appears to be artificially low and close to the fatality rate of the common flu, while Italy's appears to be inflated. At the time of this sample the USA's fatality rate was tracking that of the world average (excluding China). The world average (outside of China) should be higher than the actual average value due to defficient testing. South Korea's low fatality rate is interesting, because they tested large numbers of people. Their fatality rate is likely slightly above the actual value, which probably lies somewhere between 0.3% and 2.5%.

The last column of the table takes the ratio of fatality rates for the respective region vs. the fatality rate of the USA. If the fatality rate is any indication, then the current testing efficiency of the USA lies somewhere beneath the ratio for South Korea, which is 27%. If we further assume that South Korea has a high testing efficiency of, say 50%, then the current US testing efficiency can be estimated as $27\% \times 50\% = 13.5\%$.

Once an appropriate range of values has been established for η, the total number of symptomatic infections can then be obtained as:

$$\text{total number of symptomatic infections} \quad = \quad \text{total confirmed cases} \times \frac{100}{\eta} \tag{8}$$

This excludes incubation period and means that only η out of every 100 new cases are being tested positive.

Using our model, let S now represent the cumulative number of symptomatic infections in USA. Then:

$$\begin{aligned} S(d,\eta) &= \frac{100A}{\eta}\,\mathrm{e}^{bd} \\ &= \frac{190}{\eta}\,\mathrm{e}^{0.201\,d} \end{aligned} \tag{9}$$

Figure 3 shows this result for $\eta = 2\%, 5\%, 12\%$, and 30%.

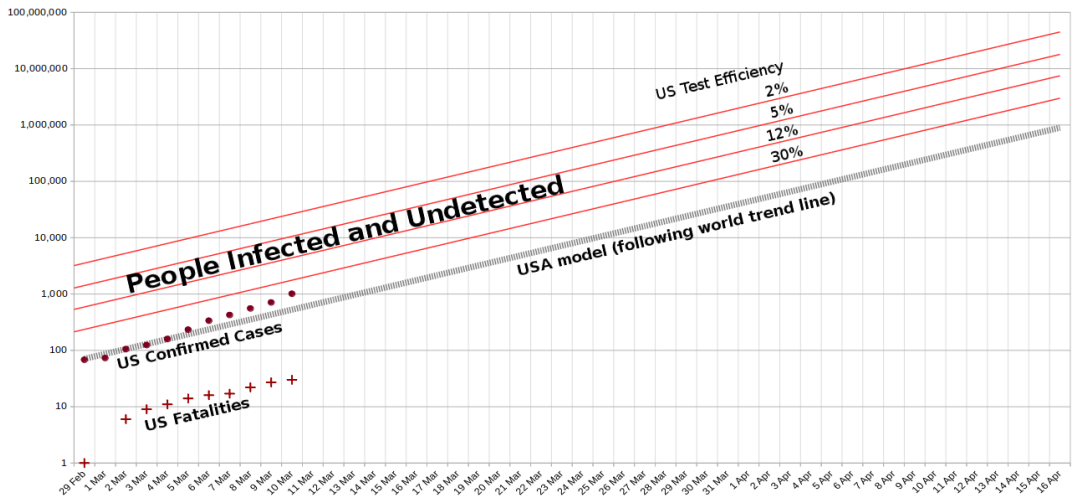


Fig. 3. The total number of people who have been infected, but remain undetected in USA as function of US testing efficiency.

2) Introducing Incubation Period: Setting η aside for the moment, let's now consider incubation period, which we've defined as the average period (in days) between a person becoming infected to the onset of symptoms of illness. For all practical purposes it is therefore defined as a characteristic of the virus itself.

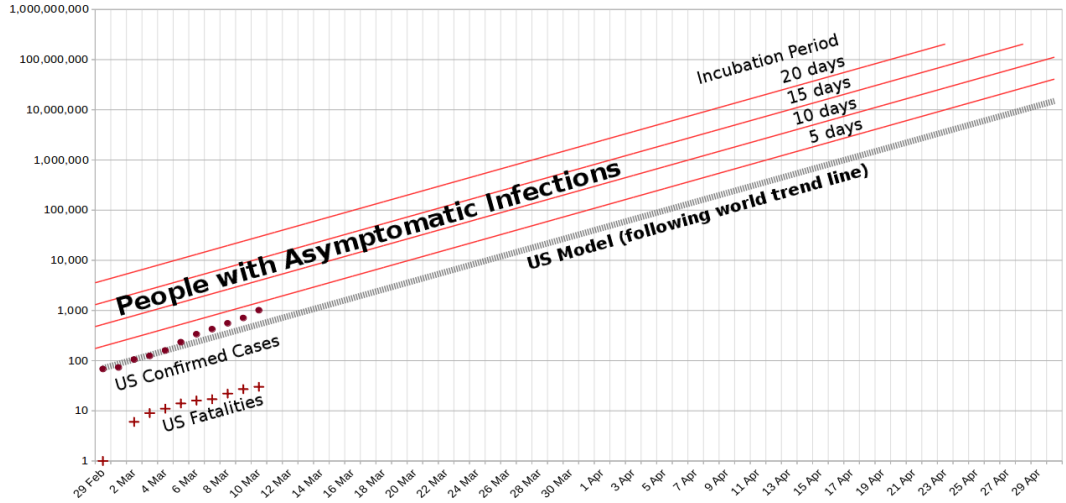


Fig. 4. Total number of infected people who are still asymptomatic, as a function of incubation period.

To incorporate it into the model, let I represent the cumulative number of asymptomatic infections, and p the incubation period. Then:

$$\begin{aligned} I(d,p) &= A\,\mathrm{e}^{b(d+p)} = A\,\mathrm{e}^{bd}\,\mathrm{e}^{bp} \\ &= 1.9\,\mathrm{e}^{0.201\,d}\,\mathrm{e}^{0.201\,p} \end{aligned} \tag{10}$$

In a perfect world of testing, and if all cases operate at the average, then every case added on a particular day will become a confirmed case after the incubation period. The curve for new cases confirmed would be precisely the same as the curve for the new infections, except lagging it by the incubation period. The cumulative rates of new infections would then also lead the cumulative cases confirmed by the average incubation period. This is illustrated in Figure 4, showing the total number of asymptomatic infections for average incubation periods of 5, 10, 15, and 20 days.

G. Incorporating both Incubation Periods and Testing Efficiencies

Combining the adjustment for the incubation period and the testing efficiency (in any order) into the model, the total number of infected individuals can be estimated. Let P now represent the cumulative number of infections. Then:

$$\begin{aligned} P(d,p,\eta) &= \frac{100}{\eta}\, A\, \mathrm{e}^{bd}\, \mathrm{e}^{bp} \\ &= \frac{190}{\eta}\, \mathrm{e}^{0.201\,p}\, \mathrm{e}^{0.201\,d} \end{aligned} \tag{11}$$

An incubation period of 4 days is assumed in Figure 5, which plots estimates of the total number of infected in the USA as a function of $\eta = 2\%, 5\%, 12\%$, and 30%.

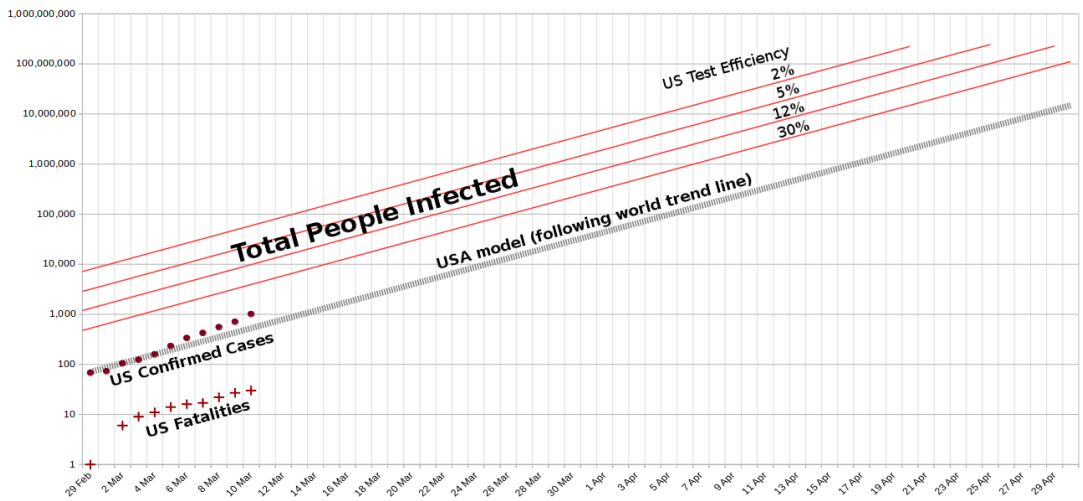


Fig. 5. Total number of people infected in the USA as a function of testing efficiency and assuming an incubation period of 4 days.

Figure 6 assumes an incubation period of 14 days.

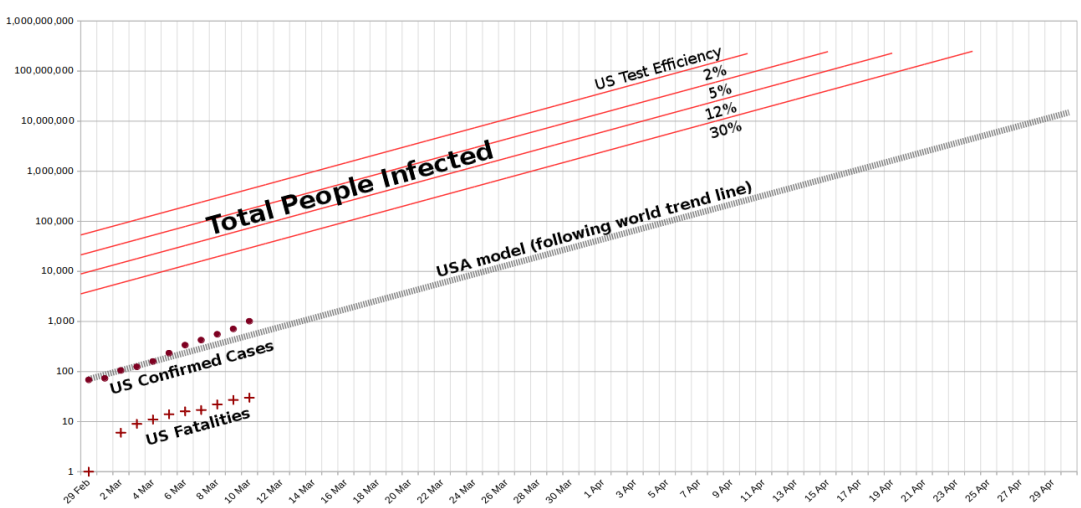


Fig. 6. Total number of people infected in the USA as a function of testing efficiency and assuming an incubation period of 14 days.

These plots for USA show how rapidly the virus outbreak spreads. It estimates that hundreds of millions may be infected by the end of the month unless the outbreak can be slowed down.