Title: A look into the inner workings of pandemic prediction models

Author:

Steve Simon, PhD

University of Missouri Kansas City School of Medicine

Department of Biomedical and Health informatics

2411 Holmes Street

Kansas City MO 64108

Corresponding Author

Key Words:

**Abstract**

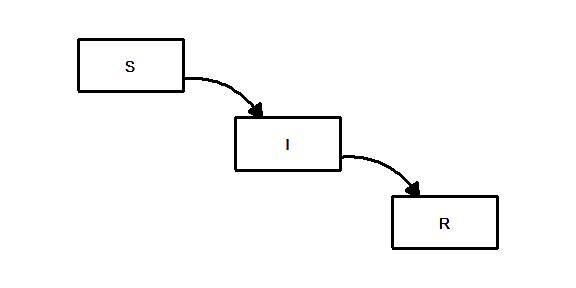
**Introduction**

From January through March, when people were learning how serious the COVID-19 crisis could be, a host of statistical models were being developed to predict how many people would become infected and how many people would die from COVID-19.

How are these prediction models developed? There are many approaches, but the most common uses some variation of the Susceptible-Infected-Recovered (SIR) compartment model.

Think of a compartment as a bucket of water with a hole in the side of the bucket. Water spills out at a certain rate, depending on the size and shape of the hole. The water that spills out of this bucket goes into a second bucket, which also has a hole in the side. This water goes into a third bucket, and so forth.

The compartments represent various disease states. In the simplest model, there are three compartments representing patients that are susceptible (*S*), infected (*I*), and recovered (*R*). The most interesting of these is the *I* compartment. The level in that compartment depends on how quickly the compartment becomes filled from the *S* compartment and how quickly it drains into the *R* compartment. It is important that the *I* compartment does not become too full for many reasons, but most important, because a large number of infected individuals can overwhelm the available health care resources.



##### Figure 1. Susceptible/Infected/Recovered (SIR) model

Mathematically, the *I* compartment is defined as:

The derivative notation, informs that the model is explaining not the amount in the *I* compartment, but the rate at which the level in the *I* compartment changes. When the rate is postive, the compartment is filling faster than it is draining. When the rate is negative, then the compartment drains faster than it fills.

The term represents the rate at which the *I* compartment drains. This is a constant times the amount in the compartment *I*. If is small, the *I* compartment drains slowly and infectious people stay infectious for a very long time. If is large, then the *I* compartment drains quickly, meaning that infectious people recover very quickly.

For most infectious diseases, there is very little control of . γ is controlled by how quickly the body’s immune system can overcome the infection.

The term represents how quickly *I* compartment is filling. This is a constant times the filling rate, calculated as the product of the amounts in *S* compartment and *I* compartment.

If *I* is small, there are not a lot of people who will infect others. If *S* is small there are not a lot of people to become infected.

Typically, *I* is small early in the course of a pandemic and represents the early part of the exponential curve where the disease spreads slowly. Over time, however, *I* may grow and with it, the rate at which *I* grows will increase, leading to an exponential increase in infections.

Eventually as the *S* compartment drains into the *I* compartment, *S* will shrink enough so that the product becomes small again. This represents herd immunity, the setting where the number of people susceptible to infection becomes too small to keep the epidemic going. This is likened to a forest fire that burns itself out because there is nothing left to burn.

### Dynamics that avoid an exponential explosion of cases

Some infectious diseases never “catch fire and some burn with an exponential fury”. The difference between them and what has been seen with COVID-19 is influenced by how quickly the Infected compartment fills compared to how quickly it drains. This is controlled by the constants and .

Designate as the ratio, .

This ratio plays a key role in the spread of disease. It is interpreted as the average number of people infected by a single individual before that individual ceases to be infectious.

If this ratio is less than one, then the *I* compartment drains faster into the *R* compartment than it can get filled by the *S* component. If the ratio *R*o is larger than 1, then the *I* compartment fills faster than it can drain. Thus, a smaller *Ro* is optimal.

This ratio (*Ro*) is made smaller by decreasing the numerator, , or increasing the denominator, . There is little that can be done to influence . Strengthening the immune system to eliminate the infection faster would reduce γ. However, therapies to do this are rare.

Decreasing , on the other hand, can be accomplished, even if there are a large number of susceptible people and a large number of infected people. In the case of COVID-19, β can be reduced through public health policies like masking and social distancing.

### Herd immunity

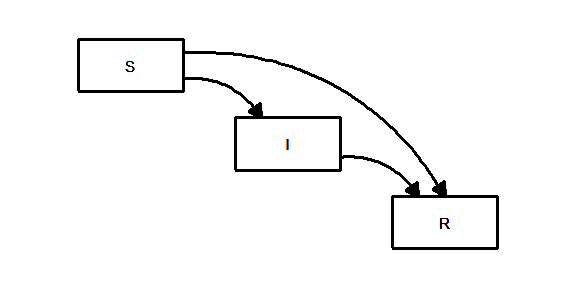
Even infections that burn with an exponential fury do not last forever. As the *S*usceptible compartment drains, the rate at which that *S-*compartment fills the *I*nfected compartment will slow. This occurrence can be achieved through “herd Immunity”. Herd immunity is when enough people become immune to a disease with the result being to stop the spread of the disease. The result is that all are protected, even those not immune to the disease. Herd immunity can be achieved through both infection and/or vaccination.

The proportion of patients who are still susceptible at time t, is designated as , Next, the change in reproductive number, , can be computed as

When herd immunity is reached, falls below 1.

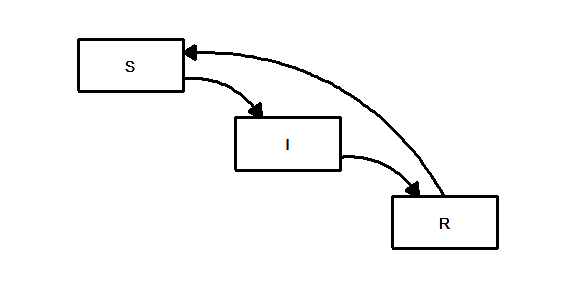
### Vaccination and re-infection models

The SIR model is quite flexible and even in its simplest form can accommodate two special cases: vaccination and re-infection. Vaccination can be incorporated into the SIR model by adding an additional path that goes directly from the *S*usceptible compartment to the *R*ecovered compartment, as shown in Figure 2. An estimate of the rate of vaccination is required for this model.



##### Figure 2. SIR model with vaccination

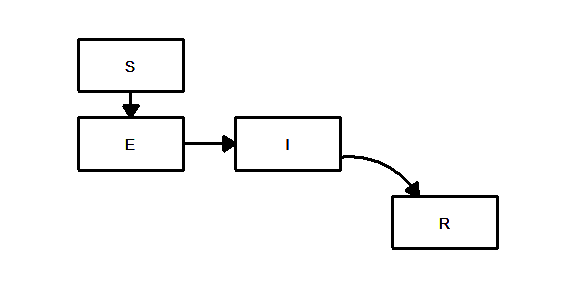
Re-infection is the scenario where people who recover from the infection can still migrate back to the Susceptible compartment. This could be due to mutations in the virus or an immune system that “forgets” about the infection over time. For COVID-19, cases of re-infection are rare, but simulations under this scenario show a future where waves of new infection returns repeatedly over time. (Figure 3)



##### Figure 3. SIR model with reinfection

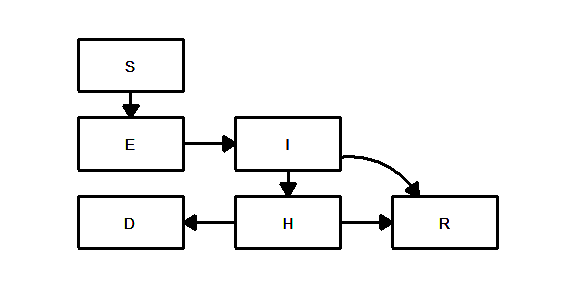
### Adding compartments to enhance the SIR model

The compartment model allows the addition of extra layers to more closely model the complex process of disease propogation. For example, an additional compartment can be added between the *S*usceptible and *I*nfected compartments. This would represent people who were exposed to the virus but who were not yet infectious (Figure 4).



##### Figure 4. Susceptible/Exposed/Infected/Recovered (SEIR) model

Multiple pathways out of the *I* compartment (Figure 5) can also be described by a model. Some of the individuals in these compartment be hospitalized (the *H* compartment) and from the hospital, some could recover and some could die (the *D* compartment).



##### Figure 5. Adding Hospitalized/Death compartments to the SEIR model

There are many other possible modifications to this model. The additional compartments allow for an increasing amount of sophistication in the predictions.

### How are these models used in practice?

The SIR model and various extensions models are helpful in two ways. First, infection data can be used to estimate the various rates in a statistical algorithm. This algorithm can then provide predictions about the future course of the infection. How well the model predicts is dependent on both the quality of the data that is entered into the model as well as how well the compartments of the model actually reflect the actual dynamics of the disease process.

The other value of the SIR model is in its ability to simulate settings that have not yet been observed. At the time that this article was written, there were hints that an effective vaccine might arrive soon for COVID-19. However, it is how many people will actually get vaccinated. A simulation, however, could show how much a vaccine would help in overcoming this virus. The simulation could vary the parameter associated with the vaccination rate to help understand what target rate would be needed to bring this COVID-19 under control.

### Concerns about the data sources

It is interesting to examine these models from a mathematical perspective, but predictions about the future course of any pandemic requires linking the theoretical models to actual data.

For COVID-19, data is available from various resources and registries. Thus, using real data, models can be built and tested.

There is concern, however, about the raw data. Tests for COVID-19 will sometimes produce false positive results and more often, false negative results. Also, there is a lag between when the test is done and when a positive result gets reported. More importantly, not everyone who has an infection will get tested.

A model can be built based on the number of deaths rather than the number of infections. While there are no false positive or negative results on an outcome like death, there is uncertainty about when a death should be classified as due to or associated with COVID-19.

While these are serious problems, they are not new to Epidemiologists. There are methods to adjust models for some types of over and under reporting. When adjustments are not available, there is often the ability to recognize the direction in which these problems will bias the model estimates.

### Concerns about heteorgeneity

A concern about compartment models is that they every individual the same. Some people, however, are less likely to get infected than others. Also, some parts of the country are more likely to see higher rates of infection than other parts of the country due to differences in population density. The models are also affected by important temporal changes such as business shutdowns and reopenings, increased testing, and mandates for masking. These variations lead to heterogeneity. If heterogeneity of individual, susceptibility and exposure are not accounted for, the accuracy of the model predictions will be impacted.

### Agent based models

An agent based model replaces buckets of water with individual particles. There particles move around at varying speeds and when an infected particle collides with a susceptible particle, the latter also becomes infected. This model can be likened to a giant simulated game of tag, with the infected people playing the role of “it”. Each particle can be controlled differently. With COVID-19, faster moving particles represent individuals who do not practice social distancing or masking well. Each particle can be assigned a different level of susceptibility based on social distancing or masking so that some become infected with very high probability after a collision and others become infected with a much lower probability. These models can include other sources of variation, such as changes in public health directives such as stay at home orders.

### Hierarchical models

A hierarchical model divides the country into regions (typically individual counties) and each region has a different SIR model. With 3,141 counties in the United States, this represents a huge increase in complexity, one that could potentially overwhelm even the fastest of computers. However, these hierarchical models will behave similarly in the dynamics of the SIR model for similar regions (counties).

### Time varying parameters

Time is one of the biggest sources of heterogeneity. The spread of an infectious disease changes over time as personal behavior changes and as government interventions change.

### Use of covariates

A covariate is a variable which is not of direct interest but which can help with prediction. If the covariate can account for some of the variation in the data, it can improve precision and reduce bias.

For example, considering population; A model can take the similarities and differences of indivdual counties into account using U.S. Census information (counties with very low population densities should have similar dynamics, for example).

Additionally, when considering heterogeneity of public health directives, different approached can be considered as a covariate for modeling.

### The Iowa State model

There are many prediction models available. A brief summary of the model developed at Iowa State (Wang et al. 2020) is explained below and in Table 1.

The Iowa State model has the three basic compartments, S for susceptible, I for infected, and R for recovered (Wang et al uses “removed” instead of “recovered”), plus a compartment D for deaths. The model predicts the rate of movement between these compartemts individually for each county.

The model uses a broad range of covariates (Table 1). The latitude and longitude allowed for counties that are close geographically to have similar infection patterns.

Table 1. Covariates in the Iowa State model

* Government declarations
* Emergency declarations
* School closures
* Bans on large gatherings
* Limits on bars, restaurants, etc.
* Travel restrictions
* Stay-at-home orders
* County demographics
  + Percent African American
  + Percent Hispanic or Latino
  + Percent aged 65 years
  + Male-to-female ratio
  + Population density
  + Socio-economic status
* Geographic features
  + Latitude
  + Longitude
  + Rural/urban classification
* County healthcare resources
  + Percent aged less than 65 years without health insurance
  + Government health care expenditures per capita
  + Total hospital beds per 1,000 population

### What to look for in a pandemic prediction model

While the mathematics for the typical prediction model are difficult to follow, even for experienced researchers, there are some easily discerned features that should be included.

First, does the model use standard Epidemiological principles based on the SIR model or some variation? The current COVID-19 pandemic is novel in a quantitative sense. The size and scope of this pandemic have no precedent other than perhaps the Spanish flu outbreak in 1918. Nevertheless, the mechanisms of transmission should not differ qualitatively from other diseases. Models that have predicted those past disease outbreaks well should be expected to perform reasonably well with the current pandemic.

Second, does the model include covariates that are expected to influence rates of transmission in the SIR model? A comprehensive set of covariates can account for the many sources of heterogeneity and will usually produce more precise and less biased predictions.

Third, does the model account for changes over time? A synamic model can capture changes that otherwise might be difficult to measure; such as COVID fatigue.

### Summary

Prediction is a perilous process. There is, however, great value in attempting to predict future cases and future deaths from an infectious disease like COVID-19. If one understands how statistical models work to make these predictions, one can assess the reliability of those estimates. However, regardless of the complexity of the model and methods used to build the model, one should always be cautious about limitations produced by heterogeneity and the quality of the data underlying these models. Nevertheless, the basic SIR model and extensions have been used to predict past epidemics, especially influenza, whose yearly appearances provide a wealth of data. When done well, these models provide useful predictions for hospitals and public health agencies.

### References

Note: These references are incomplete and not formatted properly just yet.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3935673/>

<https://www.washingtonpost.com/graphics/2020/health/coronavirus-how-epidemics-spread-and-end/>

<https://www.washingtonpost.com/graphics/2020/world/corona-simulator/>

<https://www.nature.com/articles/d41586-020-01003-6>

<https://www.cebm.net/covid-19/when-will-it-be-over-an-introduction-to-viral-reproduction-numbers-r0-and-re/>

Wang, Li, Guannan Wang, Lei Gao, Xinyi Li, Shan Yu, Myungjin Kim, Yueying Wang, and Zhiling Gu. 2020. “Spatiotemporal Dynamics, Nowcasting and Forecasting of Covid-19 in the United States.” <http://arxiv.org/abs/2004.14103>.