



Join 200,000 top engineers

Take the Quiz



Triplebyte Blog

We help engineers join great companies
Try our coding quiz

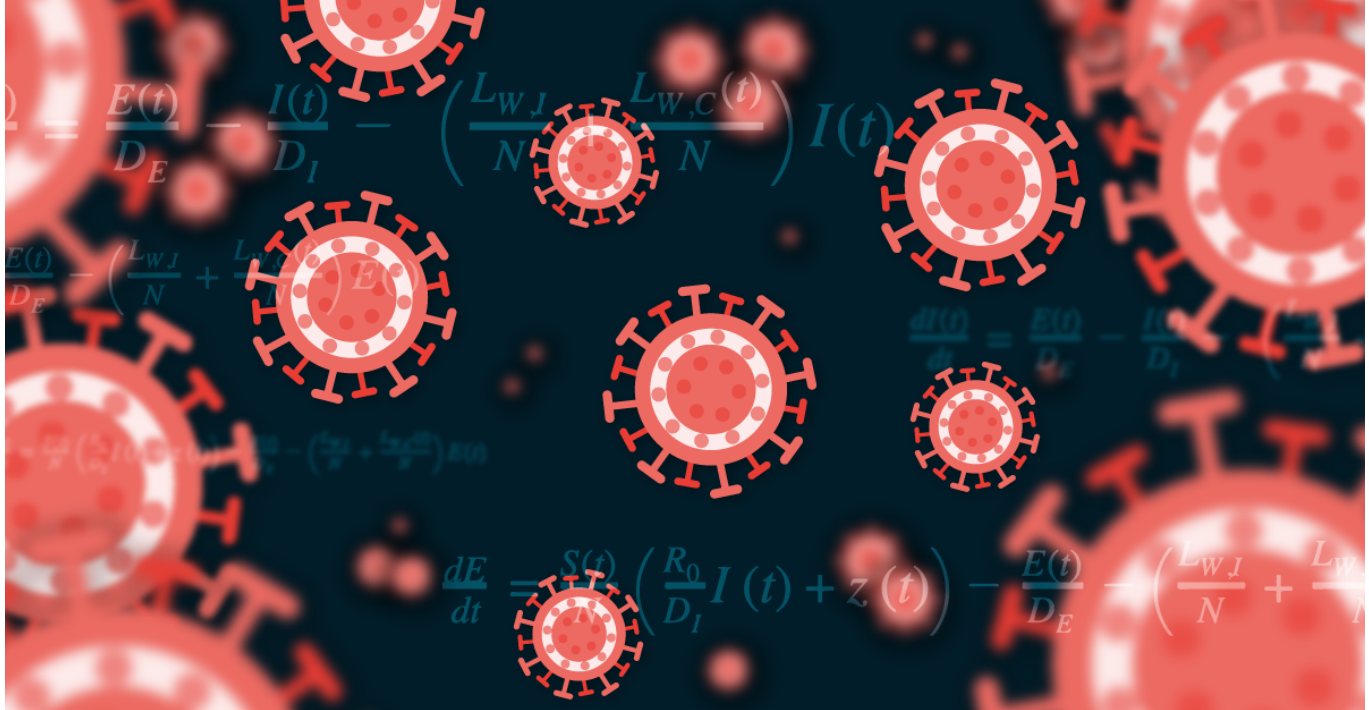
← Back to Blog

Share this article



How COVID-19 and Other Infectious Diseases Spread: Mathematical Modeling

BY JENNIFER CIARACHI ON MAR 12, 2020



On December 31, 2019, the Chinese city of Wuhan reported an outbreak of a novel coronavirus (COVID-19) that has since killed over 4,200 people. As of March 11, 2020, over 118,000 infections—spanning 113 countries—have been confirmed by the World Health Organization (WHO). The WHO is now describing the outbreak as a pandemic.

The virus is spreading rapidly. In the first month, the number of confirmed infections increased by over 1,000,000%. If the disease continued to spread at this rate, the entire global population could be infected before April. Fortunately, this is not how diseases actually spread, nor is it how they are modeled.

In the case of COVID-19, 90% of the confirmed infections are still restricted to four countries: China (67%), Italy (9%), Iran (7%), and South Korea (7%). However, these numbers are changing quickly. For example, China accounted for 74% of global cases on March 10.

The WHO, the Centers for Disease Control and Prevention (CDC), and governments within and outside of China are scrambling to minimize the spread of COVID-19. Infectious disease modeling is an essential part of this effort. A well-designed disease model can help predict the likely course of an epidemic, and reveal the most promising and realistic strategies for containing it.

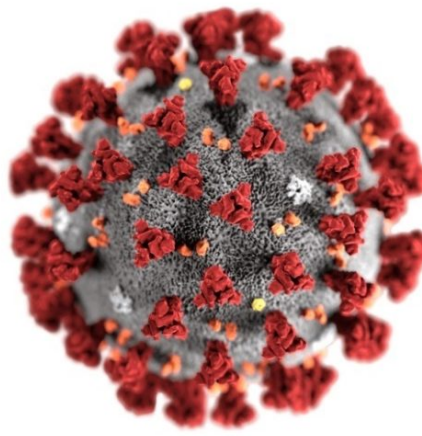


Illustration showing the structure of COVID-19. CDC/Alissa Eckert, MS

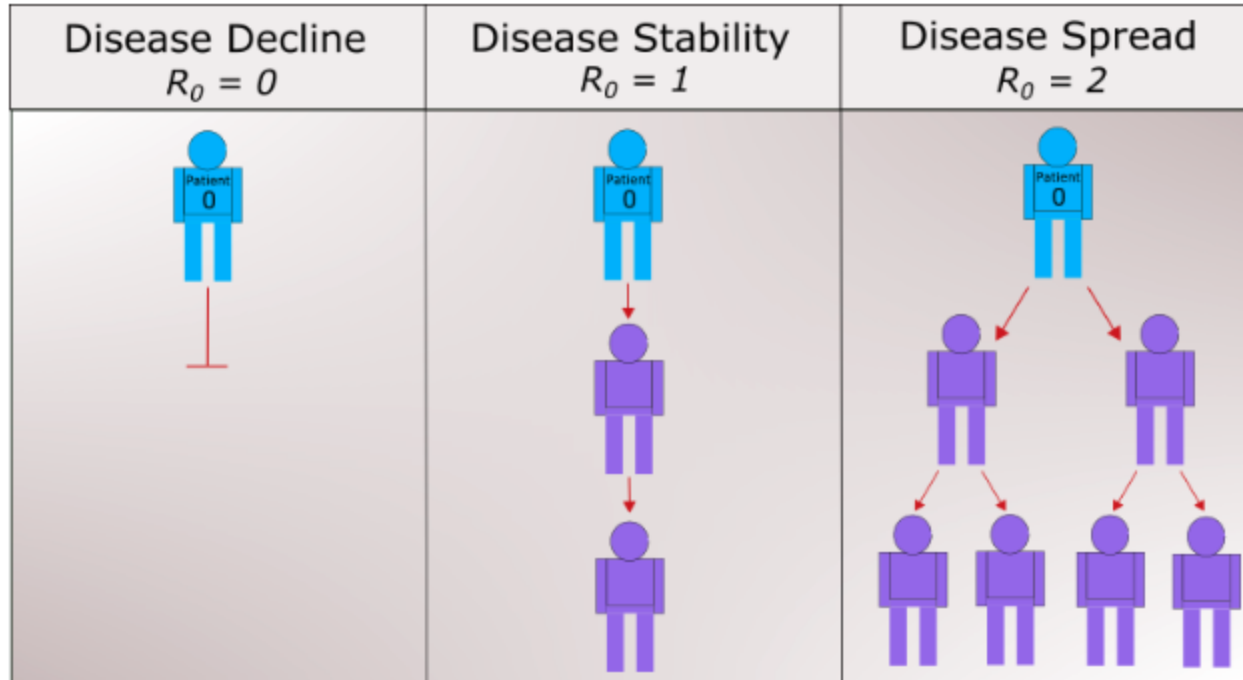
COVID-19 is a previously unencountered (i.e., “novel”) virus, so there are important unknowns that make simulating its spread particularly challenging. Ironically (but understandably), disease models often get the most public and media attention when they are the least reliable: at early stages of an outbreak, when critical data is sparse.

This article unpacks infectious disease models, exploring how the WHO and other groups are characterizing and forecasting the COVID-19 epidemic.

What’s Up with R_0 ?

One of the most important quantities in disease modeling is R_0 (pronounced R-nought), also known as the **basic reproductive number**. Determining R_0 is a fundamental goal of epidemiologists studying a new disease—like COVID-19—but what makes this quantity so useful?

R_0 is essentially a metric of how contagious a disease is. Simply put, R_0 is the average number of people in a susceptible population that a single infected person will spread the disease to over the course of their infection. R_0 can capture three basic scenarios:



If $R_0 < 1$ (left),

On average, an infected person infects less than one person.

The disease is expected to stop spreading.

If $R_0 = 1$ (middle),

An infected person infects an average of one person.

The disease spread is stable, or endemic, and the number of infections is not expected to increase or decrease.

If $R_0 > 1$ (right),

On average, an infected person infects more than one person.

The disease is expected to increasingly spread in the absence of intervention.

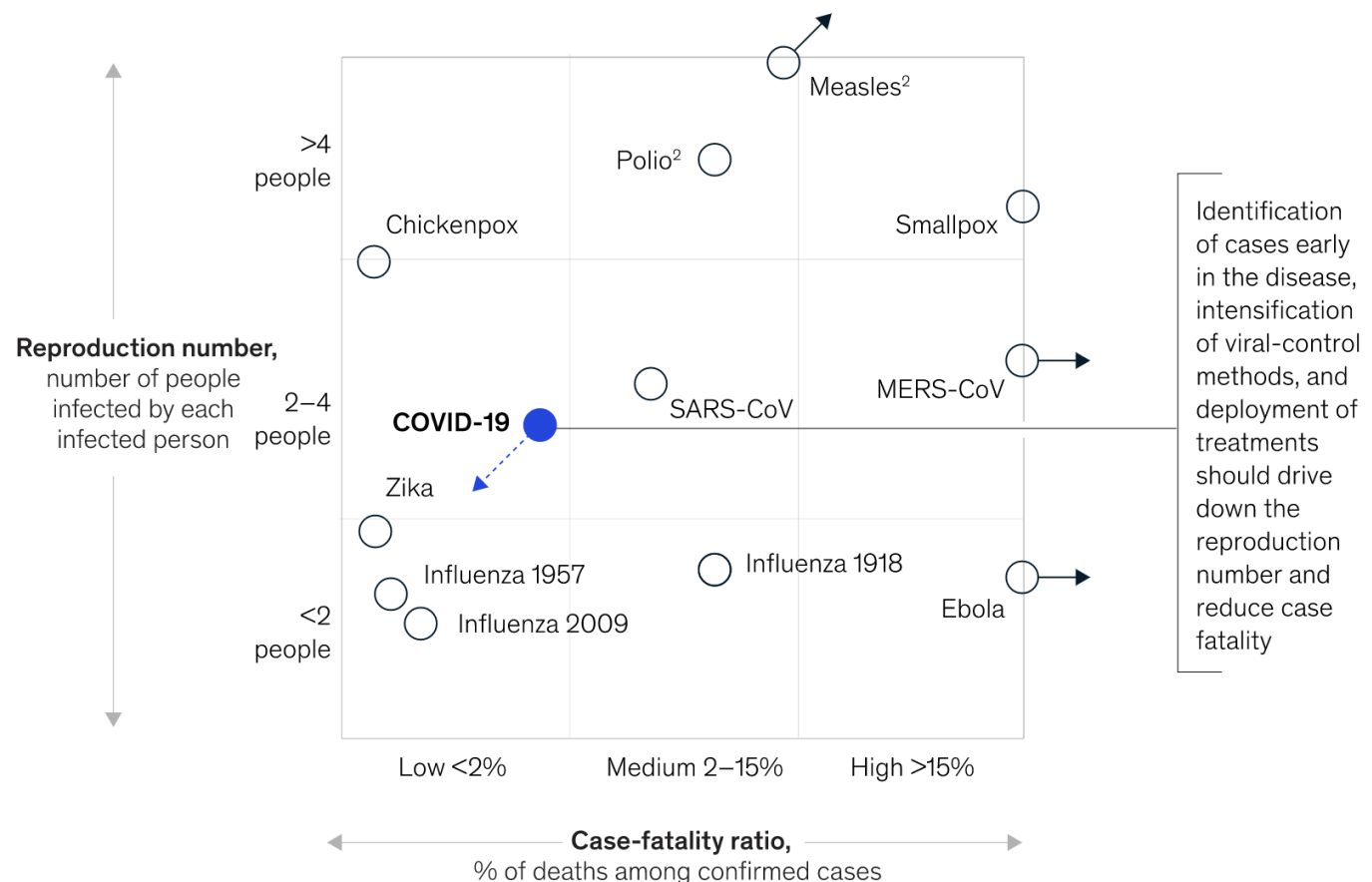
While early estimates of R_0 for COVID-19 vary, many hover in the range of 2-3. Pinpointing R_0 can help answer one of the most critical questions about an epidemic: under what conditions will the disease stop spreading?

R_0 is also tied to how much vaccination coverage is needed to control an outbreak. Measles, for example, is highly contagious, with a reported R_0 of up to 18. Assuming an R_0 of 18, stabilizing a measles outbreak (i.e., lowering its R_0 to 1) would require vaccinating about 94% of the population [$1 - (\text{goal } R_0 / \text{current } R_0)$, or $1 - 1/18 = 17/18$]. It is worth noting, however, that the often-cited R_0

range of 12–18 for measles is based on data from the 1900s, and more recent estimates indicate a much wider range of values that vary across regions and settings.

Importantly, R_0 measures a disease's potential for transmission, *not* how fast the disease will *actually* spread. Consider the ubiquitous nature of flu viruses, which have an R_0 of only around 1.3. A large R_0 is a cause for careful concern, but not a reason to panic.

Reproduction¹ and fatality² for selected human viruses



¹As determined at the beginning of an outbreak; can be reduced by effective intervention.

²Case-fatality numbers are reflective of the outbreak setting and depend on a number of factors, including patient's age, community immunity, health-system capabilities, etc. This graphic aims to offer a broad comparison.

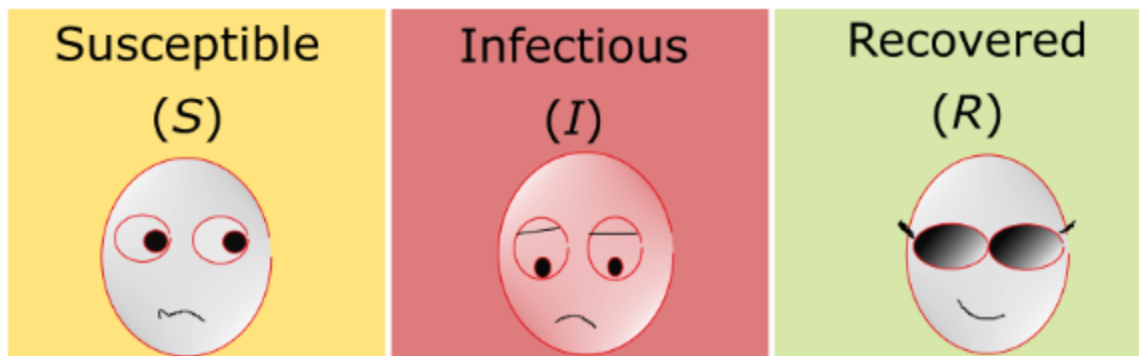
Source: World Health Organization; McKinsey analysis

R_0 is an average, so it can be skewed by factors like **super-spreader** events. A super-spreader is an infected individual who infects an unexpectedly large number of people. Super-spreader events occurred during outbreaks of SARS and MERS, other coronaviruses. Such events are not necessarily a bad sign, because they can indicate that fewer people are perpetuating an epidemic. Super-spreaders may also be easier to identify and contain, since their symptoms are likely to be more severe.

In short, R_0 is a moving target. Tracking every case and transmission of a disease is extremely difficult, so estimating R_0 is complex and challenging. Estimates often change as new data becomes available.

How Are Diseases Modeled in Populations?

Mathematical models can simulate the effects of a disease at many levels, ranging from how the disease influences the interactions between cells in a single patient (within-host models) to how it spreads across several geographically separated populations (metapopulation models). Models simulating disease spread within and among populations, such as those used to forecast the COVID-19 outbreak, are typically based on the **Susceptible – Infectious – Recovered (SIR)** framework.



SIR models are **compartmental disease models**. “Susceptible”, “Infectious”, and “Recovered” are compartments, and each individual in the population (N) is assigned to one of these compartments. To unpack this a bit further:

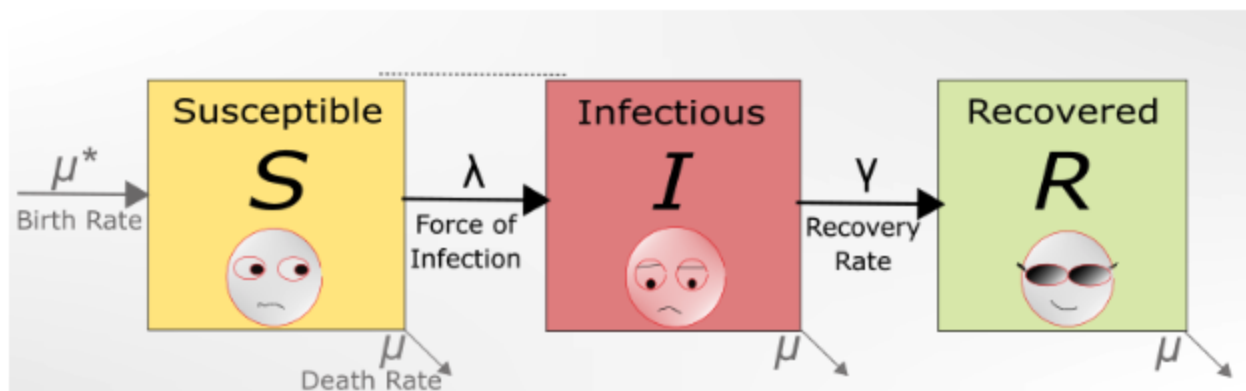
Susceptible individuals have no immunity to the disease (immunity can come from prior exposure, vaccination, or a mutation that confers resistance). Therefore, they can become infected. Susceptible individuals can move into the “Infectious” compartment through contact with an infectious person.

Infectious people have the disease and can spread it to others. Infectious individuals can move into the “Recovered” compartment by recovering from the illness.

Recovered individuals can no longer become infected, typically because they have immunity from a prior exposure. Many SIR-based models assume that a recovered person remains immune, which is often appropriate if immunity is long-lasting (e.g., chicken pox) or the disease is being modeled over a relatively short time period.

Because people can move between compartments, the number of people in each compartment changes over time. The SIR model captures population changes in each compartment with a system of **ordinary differential equations (ODEs)** to model the progression of a disease.

The standard SIR model can be schematically represented as:



Here:

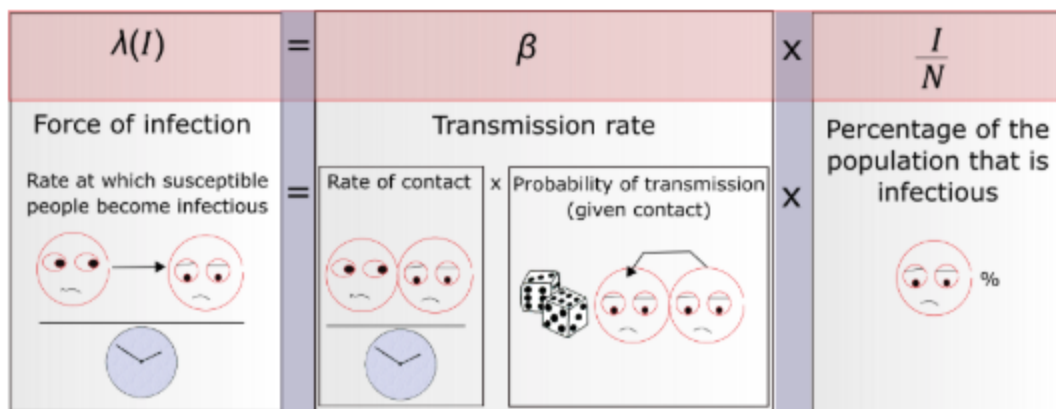
λ is the rate at which susceptible individuals become infectious—called the force of infection.

γ is the recovery rate, the rate at which people recover from infection.

The dashed line indicates that contact with an infectious individual is needed for a susceptible individual to move into the “Infectious” compartment.







Note that λ is not a constant, but a function of the size of the “Infectious” compartment. λ is also proportional to β , the **transmission rate**—the product of the rate of contact and the probability of transmission given contact:

$$\lambda(I) = \beta \frac{I}{N} .$$













Temporarily ignoring natural birth and death rates, the SIR model can be represented by the following system of ODEs:







$$\frac{dS}{dt} = -\lambda(I)S$$

$\frac{dS}{dt}$	= -	$\lambda(I)$	X	S
Rate of change of the number of susceptible people	= -	Force of infection The rate at which susceptible people become infectious	X	Number of susceptible people
#  — 		 →  — 		# 

$$\frac{dI}{dt} = \lambda(I)S - \gamma I$$

$\frac{dI}{dt}$	=	$\lambda(I)$	X	S	-	γ	X	I
Rate of change of the number of infectious people	=	Force of infection Rate at which susceptible people become infectious	X	Number of susceptible people	-	Recovery Rate Rate at which infectious people recover	X	Number of infectious people
#  — 		 →  — 		# 		 →  — 		# 

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

$\frac{dR}{dt}$	=	γ	X	I
Rate of change of the number of recovered people	=	Recovery rate Rate at which infectious people recover	X	Number of infectious people
#  — 		 →  — 		# 

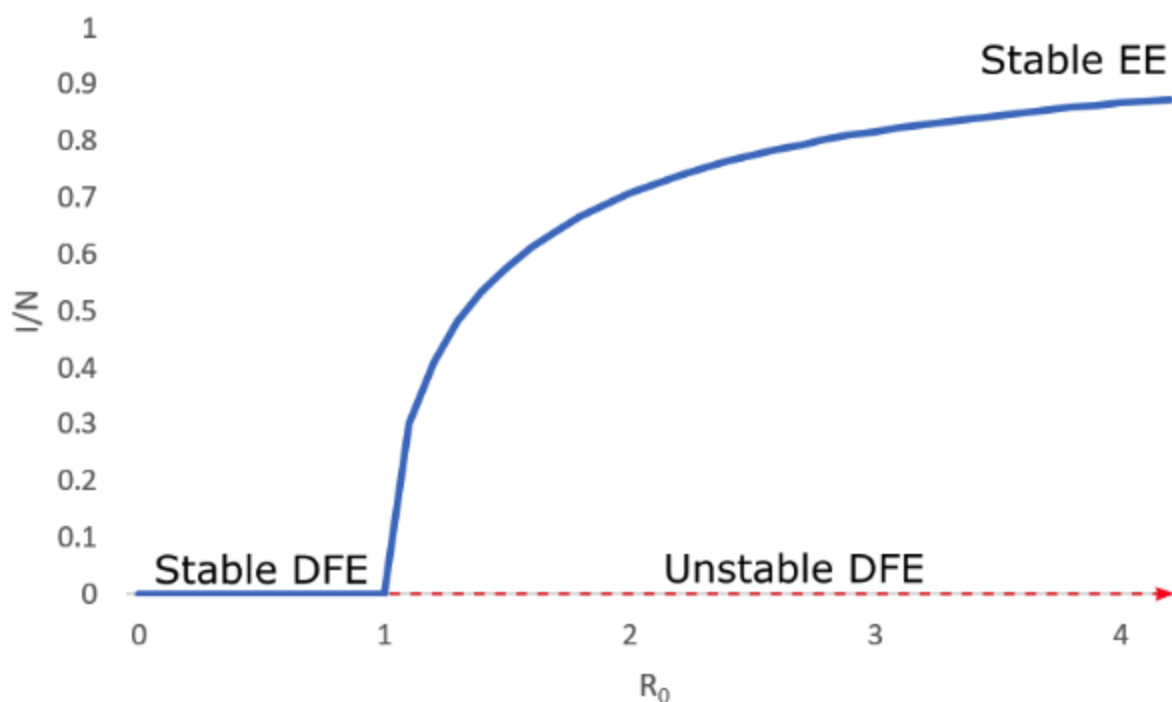
Putting It All Together: Equilibria and R_0

An important step in analyzing a system of ODEs is determining the equilibria, which is the same as setting all of the time derivatives equal to 0. In other words, if people are entering and exiting the compartment at the same rate, the compartment is in equilibrium.

Two equilibria are particularly important in epidemiological models: **disease-free equilibrium (DFE)** and endemic equilibrium. In DFE, there are no infectious individuals in the population.

In **endemic equilibrium (EE)**, there are always infectious individuals. EE requires a steady supply of susceptible individuals, for example, through birth or a loss of immunity. Otherwise, the number of infectious individuals will return to 0 once the epidemic has run its course.

Beyond determining equilibria, it is important to consider their stability. In other words, if the epidemic is near one of these equilibrium states, is it likely to move toward or away from that equilibrium? Additionally, under what conditions are DFE and EE stable and unstable? In fact, stability is closely related to R_0 : If $R_0 < 1$, DFE is stable. If $R_0 > 1$, DFE is unstable and EE is stable.

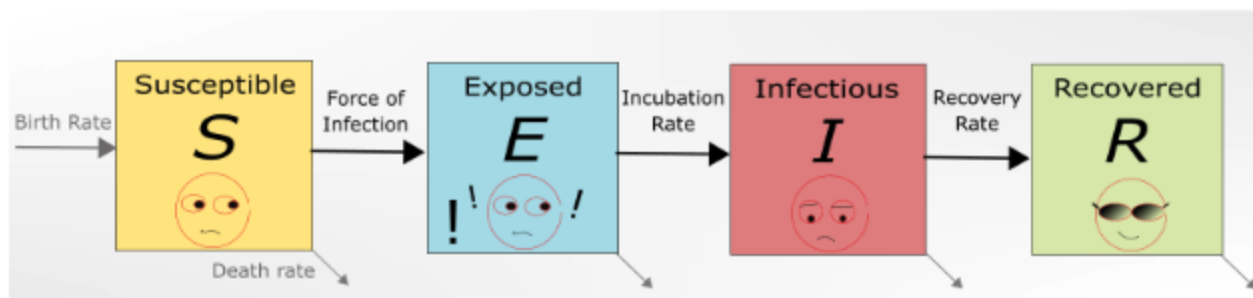


Compartmental Modeling of COVID-19

Most diseases also have a latent (or incubation) period, during which an infected individual cannot infect others. This additional compartment—E (Exposed)—is captured by an extension of the SIR model called SEIR.

The WHO used SEIR models to characterize and forecast the early stages of the COVID-19 outbreak in Wuhan.









SEIR models can be schematically represented by:











The addition here is the **incubation rate**, the rate at which exposed people become infectious.

The SEIR model published by the WHO on January 31 is presented below^[1], along with a table defining the parameters used. Examining the model, it quickly becomes clear that travel data is critical, because it directly affects transmission. The researchers incorporated flight booking information from the Official Aviation Guide (data from January-February 2019), passenger volume data from Tencent location-based services for over 300 cities, and travel data from the Wuhan Municipal Transportation Management Bureau.






$$\frac{dS(t)}{dt} = -\frac{S(t)}{N} \left(\frac{R_0}{D_I} I(t) + z(t) \right) + L_{I,W} + L_{C,W}(t) - \left(\frac{L_{W,I}}{N} + \frac{L_{W,C}(t)}{N} \right) S(t)$$

$\frac{dS(t)}{dt}$	=	$-\frac{S(t)}{N}$	x	$\left(\frac{R_0}{D_I} \right)$	x	$I(t)$	+	$z(t)$	+	$L_{I,W} + L_{C,W}(t)$	-	$\left(\frac{L_{W,I}}{N} + \frac{L_{W,C}(t)}{N} \right)$	x	$S(t)$
The rate of change of the number of susceptible people	=	The susceptible portion of the population	x	Average number of people infected by an infectious person over the average duration of infection	x	Number of people infected by infectious people	+	Number of people infected by zoonotic (animal) source	+	Travelers entering Wuhan	-	Percent of Wuhan population traveling out of Wuhan	x	Number of susceptible people
														

$$\frac{dE}{dt} = \frac{S(t)}{N} \left(\frac{R_0}{D_I} I(t) + z(t) \right) - \frac{E(t)}{D_E} - \left(\frac{L_{W,I}}{N} + \frac{L_{W,C}(t)}{N} \right) E(t)$$

$\frac{dE}{dt}$	=	$\frac{S(t)}{N}$	x	$\left(\frac{R_0}{D_I} \right)$	x	$I(t)$	+	$z(t)$	-	$\frac{E(t)}{D_E}$	-	$\left(\frac{L_{W,I}}{N} + \frac{L_{W,C}(t)}{N} \right)$	x	$E(t)$
The rate of change of the number of exposed people	=	The susceptible portion of the population	x	Average number of people infected by an infectious person over the average duration of infection	x	Number of people infected by infectious people	+	Number of people infected by zoonotic (animal) source	-	Number of exposed people over the average latency period	-	Percent of Wuhan population traveling out of Wuhan	x	Number of exposed people
														

$$\frac{dI(t)}{dt} = \frac{E(t)}{D_E} - \frac{I(t)}{D_I} - \left(\frac{L_{W,I}}{N} + \frac{L_{W,C}(t)}{N} \right) I(t)$$

$\frac{dI(t)}{dt}$	=	$\frac{E(t)}{D_E}$	-	$\frac{I(t)}{D_I}$	-	$\left(\frac{L_{W,I}}{N} + \frac{L_{W,C}(t)}{N} \right)$	x	$I(t)$
The rate of change of the number of infected people	=	Number of exposed people over the average latency period	-	Number of infected people over the average duration of infection	-	Percent of Wuhan population traveling out of Wuhan	x	Number of exposed people
								

Parameter	Definition
$S(t)$	Number of susceptible individuals at time point t
N	Total population
R_0	Basic reproductive number
D_I	Mean infectious period
$I(t)$	Number of infectious individuals at time point t
$z(t)$	Force of infection in baseline scenario
$L_{I,W}$	Average daily number of international inbound air passengers to Wuhan
$L_{C,W}$	Average daily number of domestic inbound travelers to Wuhan from mainland China
$L_{W,I}$	Average daily number of international outbound air passengers from Wuhan
$L_{W,C}$	Average daily number of domestic outbound travelers from Wuhan to mainland China
$E(t)$	Number of latent individuals at time point t
D_E	Mean latent period

The authors used this model to simulate the Wuhan epidemic since December 2019 and estimate the size of the outbreak. To simulate the spread of COVID-19 across China, they extended the model into an SEIR-metapopulation model, accounting for the effects of public health interventions by assuming different levels of reduced transmissibility (0%, 25%, and 50%) since the January 23 quarantine.

Assumptions of the model included:

An animal-based source of infection caused 86 cases at baseline (twice the number of confirmed cases).

The Wuhan population was 19 million (11 million residents + 8 million visitors).

Travel behavior was not influenced by the disease.

During the incubation period, infected individuals could not infect others (similar to SARS).

2019 travel data from China accurately reflected 2020 travel behavior (with the exception of Hong Kong, which was excluded due to the likely influence of recent social unrest on travel).

COVID-19 does not have strong seasonality in its transmission.

Stressing that the true size of epidemic remains unclear, the WHO estimated that, by January 25, 2020:

R_0 was 2.68, i.e., each infected person infected an average of 2-3 other people

The size of the epidemic doubled every 6.4 days

Up to 75,815 people in Wuhan may have been infected

Multiple major Chinese cities, including Guangzhou, Beijing, Shanghai, and Shenzhen, had already imported enough cases to spur local epidemics

Reducing transmissibility by 25-50% could substantially reduce the growth and size of local epidemics, while a reduction of more than 63% would cause the epidemics to fade out

Since the WHO study was published on January 31, much more data have become available. Nonetheless, the R_0 estimate of 2.68 remains squarely in the range of estimates reported in many other studies. Additionally, many of the Chinese cities that the study identified as susceptible to an epidemic—due to early imported cases—have since reported a high number of infections. For example, Guangzhou and Shenzhen are both in Guangdong province, which, as of March 11, is

second only to Hubei in the number of confirmed infections (granted, Guangdong is also a notably populous province). Wuhan is the capital of Hubei, where 67,773 cases have been confirmed as of March 11.

Many other researchers have since leveraged SEIR and similar models to study COVID-19. For example, researchers at Cedars-Sinai Medical Center and Peking University Health Science Center used an SEIR model to estimate the extent of the outbreak in the United States. Their study (which was posted on March 8 and has not yet been peer-reviewed) indicated that between 1,043 and 9,484 people in the United States would be infected by March 1, 2020. The lower end of that range—which is increasingly aligning with figures from the CDC and WHO—assumes that health interventions reduce transmission by 25%. The higher end of the range assumes no successful intervention.

Coping with Uncertainty

Although uncertainty is rarely featured in clickbait headlines, it is an important consideration in disease modeling, particularly during early stages of an outbreak.

For example, the WHO results characterizing the early outbreak in Wuhan were presented using 95% credible intervals (CrIs), which provide a range of values and a central point estimate. The central point estimate provides a “best guess”, while the range is an indicator of best- and worst-case scenarios.

Additionally, the researchers used a sensitivity analysis to project how the R_0 and behavior of the outbreak would change if the baseline number of cases was underestimated (by 50% or 100%). Spoiler alert: the R_0 is lower and the number of infections similarly reduced.

The new coronavirus is still pretty poorly understood, and some of the model assumptions reflect that uncertainty. For example, the model assumes that people with COVID-19 take the same amount of time to infect others as people with SARS.

AI-Powered Algorithms and COVID-19

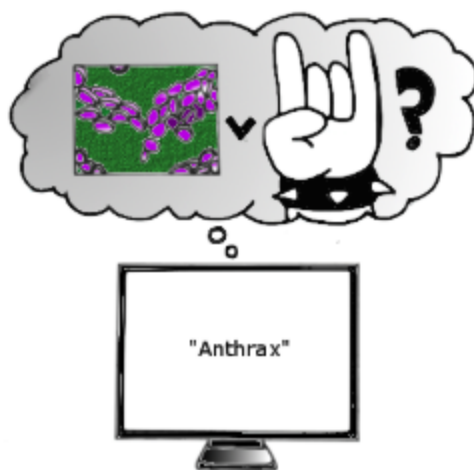
Other groups are also leveraging data-driven approaches to help predict the likely emergence of COVID-19 and other diseases.

Notably, the Canadian artificial intelligence company BlueDot, which launched in 2014, famously issued a warning to its customers about traveling to Wuhan on December 31, 2019—nine days before the WHO released a similar alert to the public.

BlueDot uses natural language processing and machine learning algorithms to analyze news reports, blog posts, and many other (non-social-media) sources, and compares this data to flight patterns to pinpoint possible outbreaks. The results are then screened and interpreted by human experts (i.e., epidemiologists) before being sent to BlueDot customers.

Like all methods, AI-powered algorithms have strengths and limitations. A core strength is their ability to quickly analyze enormous amounts of data. BlueDot's algorithms, for example, sift through 100,000 news reports in 65 languages per day. However, such algorithms are only as good as their data. Perhaps the most notorious illustration of this is Google's humbling experience with Google Flu Trends, which underestimated the spread of flu by 140% in 2013—then quietly disappeared.

By incorporating human experts, BlueDot likely circumvents some of the issues that plagued (pun intended) Google Flu Trends. In addition to its accurate COVID-19 warning, BlueDot also predicted the location of the South Florida Zika outbreak in 2016. In a recent interview with WIRED, Kamran Khan—founder and CEO of BlueDot—said, “What we have done is use natural language processing and machine learning to train this engine to recognize whether this is an outbreak of anthrax in Mongolia versus a reunion of the heavy metal band Anthrax”.



Other researchers, such as John Brownstein of Harvard Medical School and Boston Children's Hospital, are using machine learning to surveil social media posts, news reports, and health data for indications of COVID-19 outside of China. By scanning troves of data for relevant keywords (e.g., “respiratory”) and using natural language processing to determine their context (e.g., “I'm having respiratory problems”), researchers hope to determine where the virus might arise in time to contain its spread.

Closing Considerations: Don't Panic!

As you continue (intentionally or not) to come across news coverage of COVID-19, take necessary precautions, but don't panic!

Remember, even in the most affected countries, well below 1% of the population has a confirmed infection—and 70% of those infected in China have already recovered.

Nonetheless, it is prudent to abide by precautionary guidelines, such as those provided by the WHO. Although most infected people become mildly sick and recover, you can protect yourself and more vulnerable populations by:

- * Washing your hands frequently and thoroughly with soap and water or an alcohol-based hand product.
- * Staying at least 3 feet away from anyone who is coughing or sneezing.
- * Refraining from touching your face—particularly your eyes, nose, and mouth.
- * Covering your mouth and nose with tissue (or your own arm, if you're desperate) when you cough or sneeze, then immediately discarding the tissue.
- * Staying home until you recover if you start to feel sick with mild symptoms.
- * Seeking medical care early if you experience a fever, cough, and difficulty breathing.
- * Calling your health care provider in advance if you do seek medical care, and informing them about recent travel and contact with travelers.
- * Staying informed about the current state of the pandemic.
- * Following the advice of health care providers, health authorities, and your employer.

There are still many unknowns, and much of what is being announced today will likely change in the coming days and weeks as new data becomes available.

Have you worked on any ML models like the ones used to track COVID-19? Let me know at jen.ciarochi@gmail.com!

[1] R ($dR(t)/dt$) is not explicitly defined in the model. However, it can be determined using the other equations, as $N = (S + E + I + R)$. Additionally, disease spread does not depend on $R(t)$ (i.e., it is not a part of the other equations), so it can be omitted from the model. ↩

References

Blackwood, Julie C., and Lauren M. Childs. 2018. "An Introduction to Compartmental Modeling for the Budding Infectious Disease Modeler." *Letters in Biomathematics* 5 (1): 195–221. <https://doi.org/10.1080/23737867.2018.1509026>.

"Coronavirus Disease 2019 (COVID-19) in the U.S." 2020. Centers for Disease Control and Prevention. Centers for Disease Control and Prevention. March 10, 2020.

<https://www.cdc.gov/coronavirus/2019-ncov/cases-in-us.html>.

“Coronavirus Disease 2019 (COVID-19) Situation Report –50.” 2020. World Health Organization (WHO). https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200310-sitrep-50-covid-19.pdf?sfvrsn=55e904fb_2.

Desai, Rishi. n.d. “Understanding R Nought.” Khan Academy. <https://www.khanacademy.org/science/health-and-medicine/current-issues-in-health-and-medicine/ebola-outbreak/v/understanding-r-nought?modal=1>.

Fitzpatrick, Meagan C., Chris T. Bauch, Jeffrey P. Townsend, and Alison P. Galvani. 2019. “Modelling Microbial Infection to Address Global Health Challenges.” *Nature Microbiology* 4 (10): 1612–19. <https://doi.org/10.1038/s41564-019-0565-8>.

Knight, Will. 2020. “How AI Is Tracking the Coronavirus Outbreak.” WIRED, February 8, 2020. <https://www.wired.com/story/how-ai-tracking-coronavirus-outbreak/>.

Lanese, Nicoletta. 2020. “How Far Could the New Coronavirus Spread?” *Live Science*, January 31, 2020. <https://www.livescience.com/how-far-will-coronavirus-spread.html>.

Li, Dalin, Jun Lv, Gregory Botwin, Jonathan Braun, Weihua Cao, Liming Li, and Dermot P.b. MCGovern. 2020. “Estimating the Scale of COVID-19 Epidemic in the United States: Simulations Based on Air Traffic Directly from Wuhan, China,” August. <https://doi.org/10.1101/2020.03.06.20031880>.

Marill, Michele Cohen. 2020. “Wuhan Coronavirus 'Super-Spreaders' Could Be Wildcards.” WIRED, February 1, 2020. <https://www.wired.com/story/wuhan-coronavirus-super-spreaders-could-be-wildcards/>.

Munz, Philip, Ioan Hudea, Joe Imad, and Robert J. Smith. 2009. “When Zombies Attack!: Mathematical Modelling of an Outbreak of Zombie Infection.” *Infectious Disease Modelling Research Progress*: 133–150.

Niiler, Eric. 2020. “An AI Epidemiologist Sent the First Warnings of the Wuhan Virus.” WIRED, January 25, 2020. <https://www.wired.com/story/ai-epidemiologist-wuhan-public-health-warnings/>.

“Novel Coronavirus (COVID-19) Situation Dashboard.” 2020. World Health Organization (WHO). March 2020. <https://experience.arcgis.com/experience/685d0ace521648f8a5beeeee1b9125cd>.

Prosser, Marc. 2020. "How AI Helped Predict the Coronavirus Outbreak Before It Happened." SingularityHub, February 5, 2020. <https://singularityhub.com/2020/02/05/how-ai-helped-predict-the-coronavirus-outbreak-before-it-happened/>.

Wu, Joseph T, Kathy Leung, and Gabriel M Leung. 2020. "Nowcasting and Forecasting the Potential Domestic and International Spread of the 2019-NCoV Outbreak Originating in Wuhan, China: a Modelling Study." The Lancet. [https://doi.org/10.1016/s0140-6736\(20\)30260-9](https://doi.org/10.1016/s0140-6736(20)30260-9).

Yong, Ed. 2020. "The Deceptively Simple Number Sparking Coronavirus Fears." The Atlantic, January 28, 2020. <https://www.theatlantic.com/science/archive/2020/01/how-fast-and-far-will-new-coronavirus-spread/605632/>.

[Back to Blog](#)



Get offers from top tech companies

[Take our coding quiz](#)

Liked what you read? Here are some of our other popular posts...

How to Interview Engineers

By Ammon Bartram on Jun 26, 2017

We do a lot of interviewing at Triplebyte. Indeed, over the last 2 years, I've interviewed just over 900 engineers.

Whether this was a good use of my time can be debated! ...

[Read More](#)

Bootcamps vs. College

By Ammon Bartram on May 19, 2016

Programming bootcamps seem to make an impossible claim. Instead of spending four years in university, they say, you can learn how to be a software engineer in a three...

[Read More](#)

How to pass a programming interview

By Ammon Bartram on Mar 8, 2016

Being a good programmer has a surprisingly small role in passing programming interviews. To be a productive programmer, you need to be able to solve large, sprawlin...

[Read More](#)

BLOG

CONTACT

CAREERS

[Privacy Center](#)

[Privacy Policy](#)

[Terms and Conditions](#)

