

## The math behind epidemics

*A few simple metrics characterize outbreaks like COVID-19, but calculating them correctly is surprisingly tricky.*

Alison L. Hill



*Physics Today* **73** (11), 28–34 (2020);  
<https://doi.org/10.1063/PT.3.4614>



View  
Online



Export  
Citation

CrossMark



# THE MATH

## behind epidemics

Alison L. Hill

A few simple metrics characterize outbreaks like COVID-19, but calculating them correctly is surprisingly tricky.

ADAM VILMEK/SHUTTERSTOCK.COM

**Alison Hill** is an assistant professor in the Institute for Computational Medicine and the infectious disease dynamics group at Johns Hopkins University in Baltimore, Maryland. She is also a visiting scholar at Harvard University in Cambridge, Massachusetts.



# T

he year 2020 has been defined by the COVID-19 pandemic: The novel coronavirus responsible for it has infected millions of people and caused more than a million deaths. Like HIV, Zika, Ebola, and many influenza strains, the coronavirus made

the evolutionary jump from animals to humans before wreaking widespread havoc. The battle to control it continues.

When a disease outbreak is identified—usually through an anomalous spike in cases with similar symptoms—scientists rush to understand the new illness. What type of microbe causes the infection? Where did it come from? How does the infection spread? What are its symptoms? What drugs could treat it? In the current epidemic, science has proceeded at a frenetic pace. Virus genomes are quickly sequenced and analyzed, case and death numbers are visualized daily, and hundreds of preprints are shared every day.

Some scientists rush for their microscopes and lab coats to study a new infection; others leap for their calculators and code. A handful of metrics can characterize a new outbreak, guide public health responses, and inform complex models that can forecast the epidemic's trajectory. Infectious disease epidemiologists, mathematical biologists, biostatisticians, and others with similar expertise try to answer several questions: How quickly is the infection spreading? What fraction of transmission must be blocked to control the spread? How long is someone infectious? How likely are infected people to be hospitalized or die?

Physics is often considered the most mathematical science, but theory and rigorous mathematical analysis also underlie ecology, evolutionary biology, and epidemiology.<sup>1</sup> Ideas and people constantly flow between physics and those fields. In fact, the idea of using mathematics to understand infectious disease spread is older than germ theory itself. Daniel Bernoulli of fluid-mechanics fame devised a model to predict the benefit of smallpox inoculations<sup>2</sup> in 1760, and Nobel Prize-winning physician Ronald Ross created mathematical models to encourage the use of mosquito control to reduce malaria transmission.<sup>3</sup> Some of today's most prolific infectious disease modelers originally trained as physicists, including Neil Ferguson of Imperial College London, an adviser to the UK government on its COVID-19 response, and Vittoria Colizza of Sorbonne University in Paris, a leader in network modeling of disease spread.

This article introduces the essential mathematical quantities that characterize an outbreak, summarizes how scientists calculate those numbers, and clarifies the nuances in interpreting them.

For COVID-19, estimates of those quantities are being shared, debated, and updated daily. Physicists are used to distilling real-world complexity into meaningful, parsimonious models, and they can serve as allies in communicating those ideas to the public.

## Transmission dynamics

Few scientific fields have a single metric that both insiders and outsiders obsess over as much as infectious disease epidemiology's basic reproductive number,  $R_0$ . The unitless number is defined as the average number of new cases, or secondary infections, caused by a typical infected individual in a susceptible population.<sup>4</sup> It's a single quantity that describes how infectious a given pathogen is and how difficult it will be to control. (See box 1 for more about how models incorporate  $R_0$ .)

Infectious disease dynamics almost always display criticality or threshold behavior, whereby spread only takes off under certain conditions. Absent those conditions, the outbreak fizzles out, similar to a nuclear chain reaction. The value of  $R_0$  determines which outcome occurs. If disease spread is modeled by continuous differential equations,  $R_0$  helps determine when an equilibrium condition is stable or unstable. If the spread is instead captured as a series of stochastic reactions,  $R_0$  affects whether extinction or establishment is more likely.

Roughly speaking,  $R_0$  depends on the product of three factors: the contact rate, or the number of people an infected individual interacts with each day; the transmissibility, or the probability per unit time that any given contact results in transmission; and the infection duration. The goal of most infectious disease control efforts is to reduce  $R_0$  by altering one or more of those components. For example, the contact rate can be reduced by limiting an infected individual's connections through general social distancing or targeted isolation. The transmissibility can be reduced by limiting the chance of infection during each interaction through measures such as mask wearing. (For more on the physics of respiratory infection spread, see the Quick Study by Stephane Poulain and Lydia Bourouiba, *PHYSICS TODAY*, May 2019, page 70.) The duration of an infection can often be reduced by microbe-clearing therapies, like antibiotics for strep throat, but such drugs aren't yet available for COVID-19. Another way to decrease  $R_0$  is to reduce the number of susceptible individuals, which a vaccine could eventually do.

As a metric,  $R_0$  has several important limitations. It doesn't say anything about a disease's virulence, which characterizes how deadly it is. Infections with small  $R_0$  values, like SARS



BOX 1. QUANTIFYING TRANSMISSION DYNAMICS

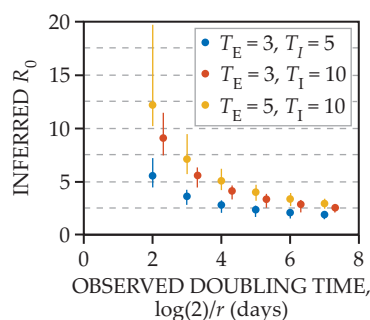
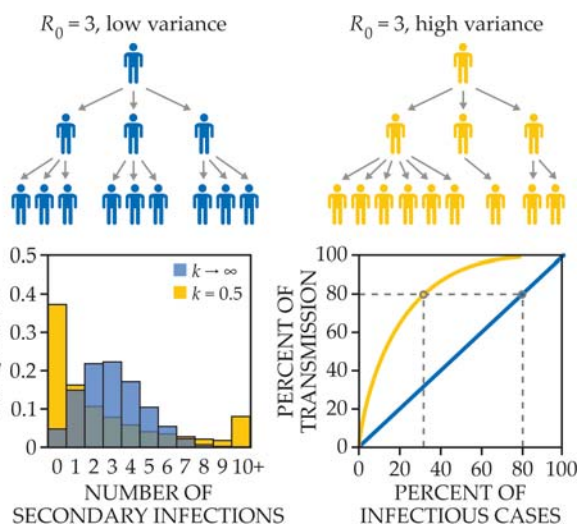
A disease’s basic reproductive number  $R_0$  describes the average number of secondary infections generated by a single infected individual introduced into a susceptible population. For an epidemic to take off,  $R_0$  must be greater than 1. An epidemic will tend to slow if the fraction  $f$  of the population that’s protected from infection is sufficiently large:  $f > 1 - 1/R_0$ .

The variance in secondary infections can be large and can lead to superspreading events.<sup>9</sup> The number of secondary infections is often summarized by a negative binomial distribution,

$P(x; R_0, k) = \Gamma(k+x) / \Gamma(k+1)\Gamma(x) p^k (1-p)^x$ , with mean  $R_0$ , where  $k$  parameterizes the dispersion of secondary infections,  $p = (1 + R_0/k)^{-1}$ , and  $\Gamma$  is the gamma function. If all individuals have the same intrinsic

infectiousness—that is, the variance is low (blue scenario on the right)—then the number of secondary infections is expected to have a Poisson distribution ( $k \rightarrow \infty$ ). If the infectiousness is heterogeneous, the distribution is said to be overdispersed and has a lower  $k$ . Overdispersion implies that a small number of individuals are responsible for a large percentage of secondary infections (dotted lines), whereas most others infect no one, which causes infection chains to go extinct. For COVID-19, a few studies have estimated  $k \approx 0.5$  (yellow at right), albeit with high uncertainty.

Estimating  $R_0$  directly is difficult. Instead, its value is usually inferred from the disease’s exponential growth rate  $r$  early in the epidemic and from the infection’s time scale<sup>10</sup> (figure 2). For example, if the average duration of the latent and infectious periods ( $T_E$  and  $T_I$ , respectively) are known and one assumes that



(severe acute respiratory syndrome), can be extremely lethal; others with high values, such as chicken pox, rarely lead to death. Some infections, like smallpox, have both a high  $R_0$  and a high risk of death. Also,  $R_0$  doesn’t reflect the time scale over which a disease spreads. The average number of new cases described by  $R_0$  could occur over a few days, as with the common cold, or many years, which is typical for HIV.

Contrary to popular belief,  $R_0$  is not an intrinsic property of an infection any more than the Reynolds number is a characteristic of a fluid. It is highly dependent on the population in which a disease spreads. The same infection could have a high  $R_0$  in a crowded population with poor hygiene and immune systems weakened by malnutrition but a much lower value in a population with better living conditions and general health. Even demographic details, such as the proportion of people in high-risk groups or patterns of social mixing, can influence  $R_0$ . The average number of secondary infections can also change dramatically over the course of an epidemic and is reflected by the effective reproductive number, which adjusts as individuals change their behavior to avoid infection.

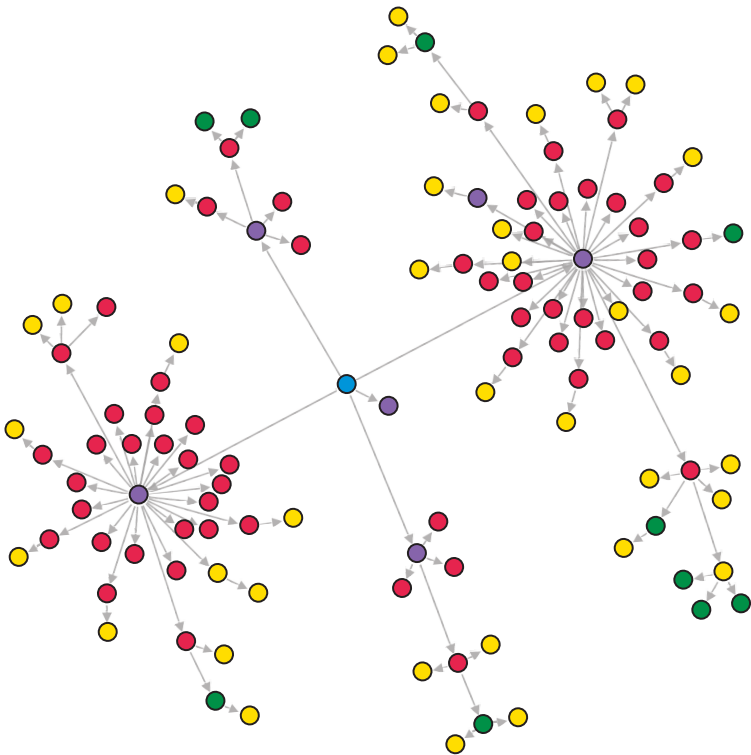
Despite the limitations, knowing a disease’s reproductive number is still useful in an outbreak. For example, Stephen Kissler and Christine Tedijanto at Harvard University found that with

the periods have exponential distributions, then  $R_0 = (1+rT_E)(1+rT_I)$  (dots on the lower graph). Other distribution shapes lead to different estimates for  $R_0$  (error bars). Country-level epidemic growth rates in the range of 0.1–0.4 per day have been observed for COVID-19, which corresponds to doubling times of 2–8 days. Estimates of  $R_0$  have generally been between 2 and 3, although they are sometimes much higher depending on the setting observed and the assumptions about the transmission intervals. (Images created using code from ref. 8.)

an estimated  $R_0$  of 2.2, people in the US would need to reduce their contacts by 60% through social distancing for at least 70% of the epidemic to avoid overflowing its current intensive care unit capacity. Luca Ferretti and Chris Wymant at Oxford University calculated that with their estimate of  $R_0 = 2.0$ , testing and contact tracing could control the epidemic only if 75% of confirmed and suspected cases were isolated within two days.

After estimating  $R_0 = 5.7$  early in the outbreak in Wuhan, China, Steven Sanche and Yen Ting Lin at Los Alamos National Laboratory calculated that control would require isolating 50% of infected people along with a 50% reduction in all contacts through social distancing. Huaiyu Tian at Beijing Normal University and colleagues estimated that early in the outbreak  $R_0$  was, on average, 3.1 in Chinese cities but that it quickly decreased to about 1 in cities that rapidly implemented control measures and further decreased to about 0.04 under more intense controls.

But where did those  $R_0$  values come from? Estimating  $R_0$  is notoriously difficult. A complete chain of transmission events starting from a single individual is rarely observed. That is often only possible when infection is still relatively rare, the symptoms are relatively unique, good diagnostic tests are available, and a high proportion of the population can be sampled (see figure 1). In contact-tracing studies, as soon as an individ-



**FIGURE 1. EXAMPLE TRANSMISSION NETWORK FOR COVID-19.** This transmission cluster was seeded by an unknown infected individual (blue) who attended a training course with other fitness instructors (purple). They in turn spread the infection to students in their exercise classes (red), to family (yellow), and to coworkers (green). (Created from data in ref. 7 using code from ref. 8.)

ual is diagnosed, public health professionals track down anyone that person might have contacted during his or her infectious period and test them for the disease; researchers use the data to estimate  $R_0$  for a single generation of infection.

But direct estimates of  $R_0$  can be biased. For example, outbreaks are more likely to be detected when many individuals were infected by a single source—a superspreading event—so estimates could be biased upwards. Alternatively, individuals enrolled in studies may be more likely than the average person to be diagnosed and isolated quickly, leading to underestimates of the true  $R_0$ . Indirect estimates are therefore more common and may give more representative values.

A common way of indirectly estimating  $R_0$  involves observing an epidemic's growth rate. Alone,  $R_0$  does not determine how quickly a disease spreads. It also depends on the time scale over which an individual's secondary infections occur. However, if the average time a person is infectious can be determined, then it's generally possible, with some mathematical tricks, to estimate a population's  $R_0$  from the rate of disease spread (see box 1).

## Time scale of infection

Exponential growth in the number of infections is a defining feature of epidemics early in their course. Estimates of the growth rate  $r$ , or alternatively the doubling time  $T_2 = \log(2)/r$ , can inform short-term projections of the epidemic.

Like  $R_0$ ,  $r$  is not an intrinsic property of an infection; it varies across regions and over time. Usually variation in  $r$  occurs for the same reasons as in  $R_0$ , such as changes in human behavior that reduce spread. But estimates of  $r$  are also subject to other factors. Dramatic changes in testing capacity that alter the proportion of cases detected and reported can lead to biased estimates of  $r$ , as can changes in reporting delays.

Observed exponential growth rates can be used to back out  $R_0$ , which has a more intuitive interpretation and is more directly connected to the underlying process of disease transmission. Researchers have derived mathematical equations to relate  $r$  to  $R_0$  under different assumptions about transmission (see box 1). In general, those formulas require knowing how long a typical individual is infectious and the delay between when someone is infected and when they become infectious, known as the latent period (see figure 2). A high observed exponential growth rate of infection implies a high  $R_0$  if either the latent period or infectious period is long, whereas it could imply a much smaller value if both those intervals are short.

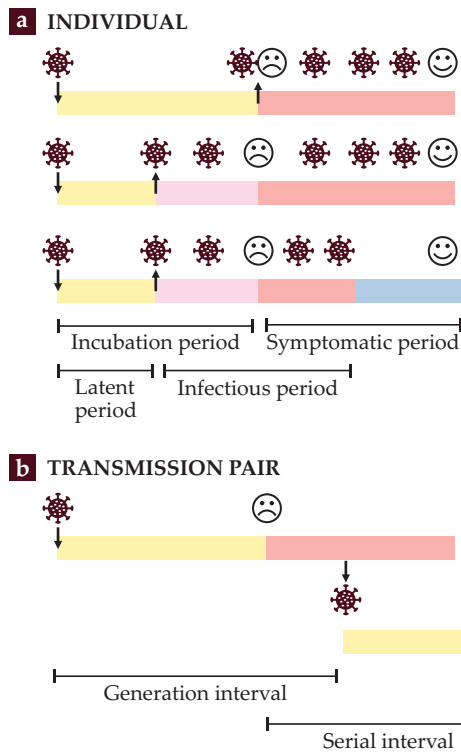
A disease's latent and infectious periods can be estimated by following individual patients with known infection exposure dates. But more than just the intervals' average durations is needed to determine the relationship between  $r$  and  $R_0$ : Enough patients must be studied to get a reasonable estimate of the full distribution.

For many infections, the latent and infectious periods are easily identified because they correspond with disease symptoms. However, for COVID-19 that is not the case: Individuals often shed the virus in their respiratory secretions and are highly infectious before they develop symptoms, such as a cough or fever. The incubation period—the time until symptoms develop—is therefore generally longer than the latent period (see figure 2). Furthermore, it appears that many of the symptoms of COVID-19 extend far beyond the infectious period. Epidemiological information, rather than symptom tracking, is therefore needed to estimate when someone was infectious.

Infectious disease epidemiologists often use observed transmission chains to determine the timing of infectiousness relative to the disease course. They do so by estimating either the generation interval, the time between when an individual was infected and when he or she infected a secondary case, or the serial interval, the time between when symptoms start in the first person and in the person they infected (see figure 2). Measuring the serial interval is more common because the onset of symptoms is generally easier to discern than the infection time.

The serial interval is a mathematical convolution of the incubation and infectious periods, so if one is known, the other can be calculated. Researchers have developed formulas that directly relate the serial-interval distribution to  $r$  and  $R_0$  without first recovering the individual periods. Those formulas have become the most common way to estimate  $R_0$ . However, the calculated  $R_0$  values are subject to biases in estimates of the serial interval. For example, individuals enrolled in research studies are often isolated shortly after diagnosis, which reduces the time they have to infect others.

Estimating the durations of infection stages provides information beyond  $R_0$  for epidemic control. The distribution of incubation periods indicates how long exposed individuals should be quarantined to safely rule out symptomatic infection. The



**FIGURE 2. TIME SCALES OF EPIDEMIC SPREAD.**

(a) The kinetics of infection spread depends on the timing of disease progression in infected individuals. The incubation period starts at infection (down arrows) and ends at symptom onset (frowny faces); the symptomatic period starts at symptom onset and ends at resolution (smiley faces). In contrast, the latent and infectious periods are determined by when an infected individual can transmit the disease to someone else (virus symbols), and the periods may or may not correlate with the timing of symptoms. Three possible scenarios are shown. Timeline labels are indicated for the third scenario, which is thought to describe COVID-19 infections.

(b) Transmission pairs—sets of individuals in which one infected the other—are often used to estimate individual time scales through either the generation or serial interval. (Figure by Alison Hill.)

distribution of infectious periods determines how long infected individuals should be isolated to prevent them from infecting others.

### How deadly is it?

So far we've characterized epidemics using the basic reproductive ratio  $R_0$ , which summarizes an infected person's transmission potential; the exponential growth rate  $r$ , which reveals how fast the epidemic is growing; and infection time intervals, which capture how the disease's course in one individual determines the time scale of infection at the population level. But those metrics miss a key feature: how deadly the disease is.

The lethality of an infectious disease is typically defined as the probability that an infected individual will eventually die of the disease and is commonly reported as the case fatality risk (CFR; see box 2). The CFR for COVID-19 has been hotly debated, and although scientists have generally converged on an estimate of around 1%, researchers, the press, and the general public continue to scrutinize that value. Some insist that COVID-19 is "just another flu," whereas others present evidence for total excess deaths far exceeding official reports. To understand the debates, it is important to understand the complications in estimating the CFR.

A common mistake in estimating a disease's CFR is to simply divide the cumulative number of deaths occurring up to a certain day by the cumulative number of cases diagnosed up to that same day. That ratio is a biased estimate of the likelihood of death given infection, especially during a rapidly growing epidemic. (See box 2 for a simple model illustrating that point.) To correctly ascertain the risk of death, researchers can use cohort studies in which they monitor a group of recently infected individuals until each one either recovers or dies. Performing such studies is difficult during an ongoing outbreak. Alternatively, simple death ratio measurements can be adjusted to ac-

count for epidemic growth and time to death.

Another complicating factor when calculating the CFR is determining who counts as a case. Definitions of the CFR in epidemiology literature make it clear that a case is someone diagnosed with infection, either by a specific test or at minimum based on symptoms. But that's a problem for infections like COVID-19, since true cases are underreported because of testing limitations and asymptomatic infections. If researchers truly want to estimate the probability of death given infection, then they need to correct for that undercounting.

The quantity being estimated is then more correctly termed the infection fatality risk (IFR). To estimate the degree of undercounting and calculate the IFR for COVID-19, epidemiologists either look at populations with near-universal testing or conduct random population-level testing to estimate the prevalence of current or past infection.

Other challenges, such as correctly identifying a cause of death, also affect estimates and interpretations of CFR and IFR values for COVID-19 and other infections. But more importantly, metrics like the CFR only count deaths; they don't include the many other harms that survivors suffer. The long-term complications of COVID-19 and the care required for serious cases, such as mechanical ventilation, are still under investigation, and simple metrics are unlikely to capture those effects.

### From description to prediction

Metrics like  $R_0$ ,  $r$ , and the CFR help classify and compare infections and quickly communicate risk. But their ability to predict the full burden of an epidemic is limited. For example, how many people an infection kills and the time scale over which that occurs depend not only on the CFR but also on how many people get infected, which itself depends on how easily the infection is transmitted, what fraction of the population is susceptible, and the efficacy of control measures. The number of new daily infections depends on the number of people currently infected and how long ago they were infected, which determines how many of them have already entered their infectious period. To put those ideas together and make informed predictions, mathematical models are needed.

Most dynamical models used to track infection spread in a population are compartmental models, in which individuals are classified into one of a few discrete states, such as susceptible, infectious, or recovered, based on their infection status<sup>5,6</sup> (see figure 3). The model tracks changes in the number of individuals in each state, usually with differential equations or discrete or continuous stochastic processes. The equations are inherently nonlinear because pairwise interactions between susceptible and infectious individuals generate new infections.

## BOX 2. ESTIMATING THE RISK OF DYING DURING AN EPIDEMIC

Epidemiologists use a disease's case fatality risk (CFR) to describe the percentage of individuals confirmed to be infected (red, top right) who will eventually die of the disease (black) rather than recover (blue). The true CFR can be accurately established only by following a cohort of infected individuals until their final outcome is observed.

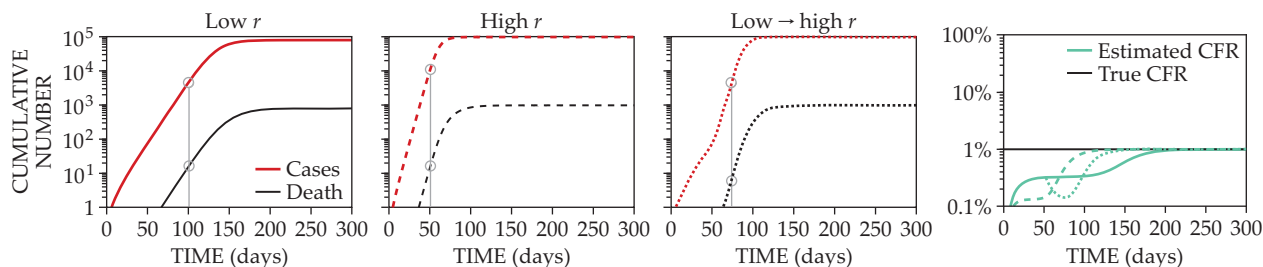
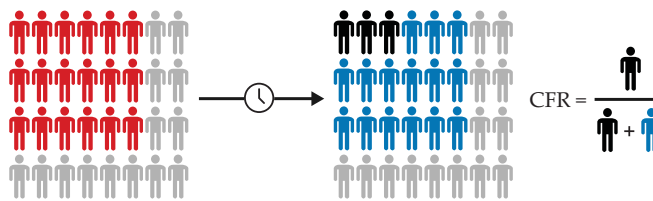
The ratio of the number of deaths observed up to a certain time to the number of cases reported up to that same day (gray circles on graphs below) can give a biased estimate of the true risk of death. The risk estimate is especially skewed when the epidemic has a high exponential growth rate  $r$ , when  $r$  changes rapidly, and when

a long delay exists between infection and death. That's because the pool of cases from which the observed deaths were drawn occurred in the past, when the epidemic was smaller.

In the simple infection model shown here, individuals are only infectious for about five days, but it may take an additional two weeks for them to die. The true CFR is 1%, which is dramatically underestimated by the ratio of deaths to cases early in the epidemic (right graph). In real data, the ratio can further be confounded by underreporting or reporting delays.

A note on terminology: The abbrevia-

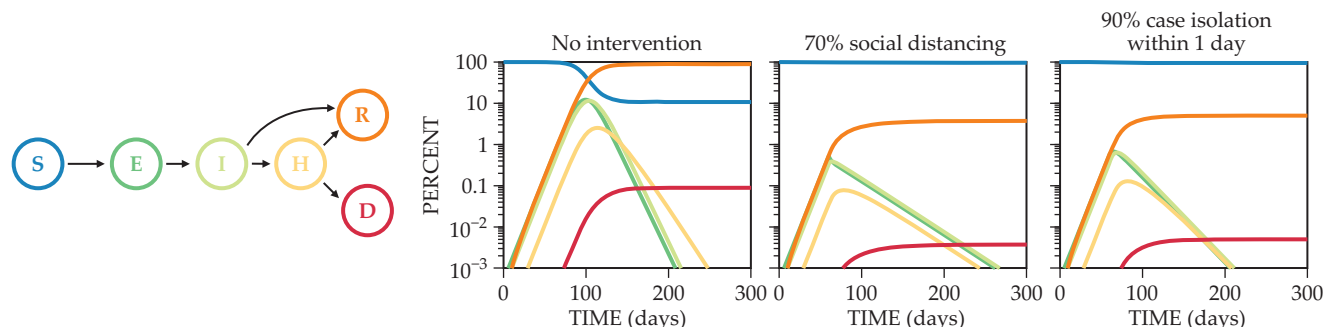
tion CFR is confusing because the R can stand for rate, ratio, and risk. In epidemiology, the three words have precise meanings. A rate generally implies a unit of inverse time and is rarely used to describe a short-term infection affecting only a portion of the population. A ratio compares two distinct populations. Only a risk metric describes a proportion in which individuals counted in the numerator are a subset of those in the denominator. That is what's needed to measure an infected person's chance of death. (Images created using code from ref. 8.)



Some simple standard epidemic models will be familiar to physicists from introductory dynamical systems courses and are named for the acronyms of their compartments. For example, the SIS model describes infections, like many sexually transmitted diseases, that don't produce long-term immunity: susceptible (S) individuals can become infectious (I) but then return to the susceptible state when they recover. In the SIR model, recovered (R) individuals are assumed to be permanently immune, a good approximation for many short-term viral infections like measles or yellow fever. (An online simulation tool that uses a compartmental SIR-type model to understand COVID-19 transmission is available at <https://alhill.shinyapps.io/COVID19seir>.)

Just like physicists, infectious disease researchers balance creating simple, understandable models with generating useful predictions. Compartmental models are always oversimplifications because in reality, the infection in one person's body is a continuum of states—the microbe multiplies and migrates between tissues, the immune system mounts a response, and symptoms develop. And the process of disease transmission can be much more complicated than the simple reaction-rate terms used in many equations. It depends on personal contacts and the highly structured nature of social networks (see box 3).

The level of detail needed for a model to be useful depends on its purpose. Some researchers modeling COVID-19 are interested in understanding the potential burden on the



**FIGURE 3. A MODEL OF COVID-19 SPREAD WITH AND WITHOUT INTERVENTIONS.** The compartmental model shown here classifies individuals as susceptible (S), exposed (E), infectious (I), hospitalized (H), recovered (R), or dead (D). At day 60, one of two interventions is implemented: either a general social distancing policy that reduces contact between infected and susceptible individuals by 70% or another case-based policy in which 90% of infectious individuals are isolated an average of one day after they become infectious. (Images created using code from ref. 8.)



BOX 3. SIMULATING INFECTION SPREAD IN NETWORKS

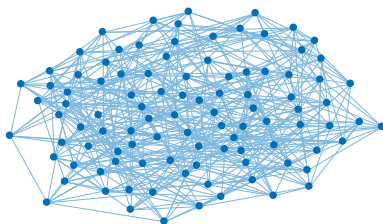
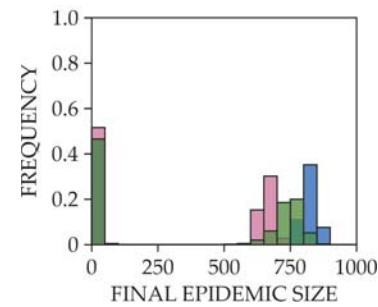
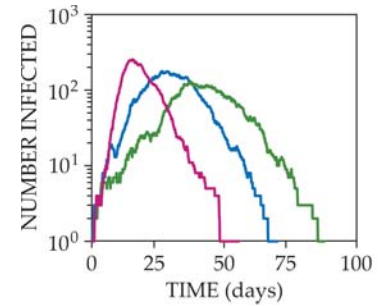
Human contacts are not random or uniformly distributed. They can be described by a contact network that determines which transmission paths are possible if an infection is introduced into the population. The structure of the network can heavily influence the extent of disease spread.<sup>11</sup>

Here, a simple, stochastically simulated susceptible-infected-recovered (SIR) model demonstrates that idea on three idealized networks: a uniform random network in which each individual is connected to 10 other randomly chosen people, a highly clustered small-world network that uses the Watts–Strogatz algorithm to preferentially connect individuals to 10 neighbors and then randomly rewires 10% of connections, and a heterogeneous network in which the number of neighbors for each individual is drawn

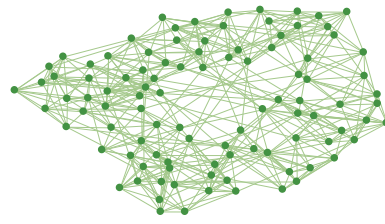
from a gamma distribution with mean and standard deviation of 10. All epidemics started with one infected individual.

Epidemic growth is fastest in the heterogeneous networks and boosted by highly connected superspreaders. It's slowest in the small-world network, where the high degree of interconnectiveness limits the susceptible contacts seen by an infected individual. The final epidemic size—the percentage of recovered individuals when the infection eventually dies out—varies across simulations because of stochastic effects, but it is generally highest in the uniform network and lowest in the heterogeneous network. (Images created using code from ref. 8.)

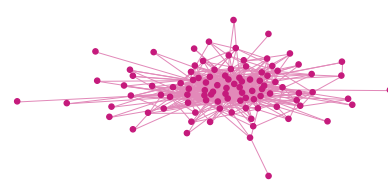
(For more details on the physics of networks, see the article by Mark Newman, *PHYSICS TODAY*, November 2008, page 33.)



Uniform random network



Small-world network



Heterogeneous network

healthcare system, so they extend SIR-type models to include advanced stages of infection that require hospitalization or admission into an intensive care unit. They also track the portion of individuals who die (see figure 3). In studies that make policy recommendations about social distancing strategies, modelers simulate detailed infection networks that describe individuals' interactions at home, school, and work and among friends. To understand the effectiveness of symptom-based isolation with or without additional quarantining of contacts, scientists augment basic models to track infectiousness over the disease's course.

Scientists continuously debate the relative merits and caveats of different modeling approaches for COVID-19. They refine models as their understanding of the disease changes and try to determine how to best communicate to the public the inherent uncertainty in model predictions. (For more on uncertainties in COVID-19 modeling, see *PHYSICS TODAY*, June 2020, page 25.)

Mathematical analysis and modeling are key tools in the study of infectious diseases and have been critical in our response to the COVID-19 pandemic. Estimating even seemingly simple metrics— $R_0$ , the CFR, and the incubation and infectious periods, among others—requires strict attention to nuances in the data and careful formulation of mathematical relationships. When designing complex models of epidemic dynamics,

modelers make trade-offs between keeping things simple enough to facilitate understanding and realistic enough to make accurate forecasts. Getting the numbers right is always a priority for scientists. During a public health crisis, the stakes are higher than ever.

Thanks to Michael Levy, Anjalika Nande, Andrei Gheorghe, Jean Yang, Norman Hill, and the reviewers for feedback on this article.

REFERENCES

1. S. P. Otto, T. Day, *A Biologist's Guide to Mathematical Modeling in Ecology and Evolution*, Princeton U. Press (2007).
2. N. Bacaër, in *A Short History of Mathematical Population Dynamics*, Springer (2011), p. 21.
3. S. Mandal, R. R. Sarkar, S. Sinha, *Malar. J.* **10**, 202 (2011).
4. J. M. Heffernan, R. J. Smith, L. M. Wahl, *J. R. Soc. Interface* **2**, 281 (2005).
5. M. J. Keeling, P. Rohani, *Modeling Infectious Diseases in Humans and Animals*, Princeton U. Press (2008).
6. E. Vynnycky, R. G. White, *An Introduction to Infectious Disease Modelling*, Oxford U. Press (2010).
7. S. Jang, S. H. Han, J.-Y. Rhee, *Emerg. Infect. Dis.* **26**, 1917 (2020).
8. All the code used to generate the results shown in this article and details about the methods used are publicly available at <https://github.com/alsnhll/PhysicsTodayCOVID19>.
9. J. O. Lloyd-Smith et al., *Nature* **438**, 355 (2005).
10. J. Wallinga, M. Lipsitch, *Proc. R. Soc. B.* **274**, 599 (2007).
11. J. C. Miller, *J. R. Soc. Interface* **6**, 1121 (2009).