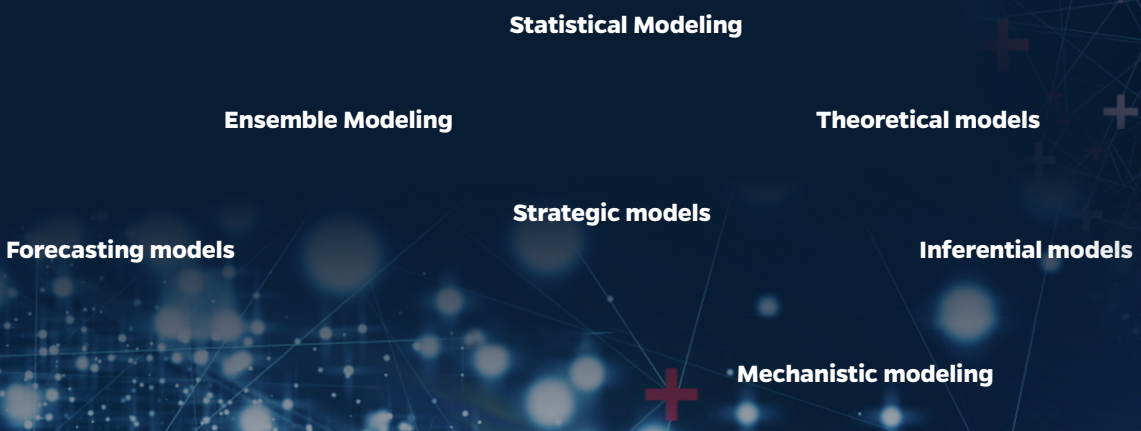
Infectious Disease Transmission Modeling Primer



JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

The increasing need for informed policy and decision-making to address the unprecedented nature of the COVID-19 pandemic have brought infectious disease modeling to the center stage of public health. Against the backdrop of many uncertainties and countervailing forces, models can be the best way to synthesize what is known, forecast future pandemic developments based on what is known, and understand disease risks under consideration of uncertainties. Accordingly, models related to COVID-19 have been used to inform policy decisions related to health care, public health, and beyond. While models can be valuable tools for understanding characteristics of the disease and informing policy responses, awareness of their strengths and weaknesses is crucial for optimizing their effectiveness as a tool for decision-making.

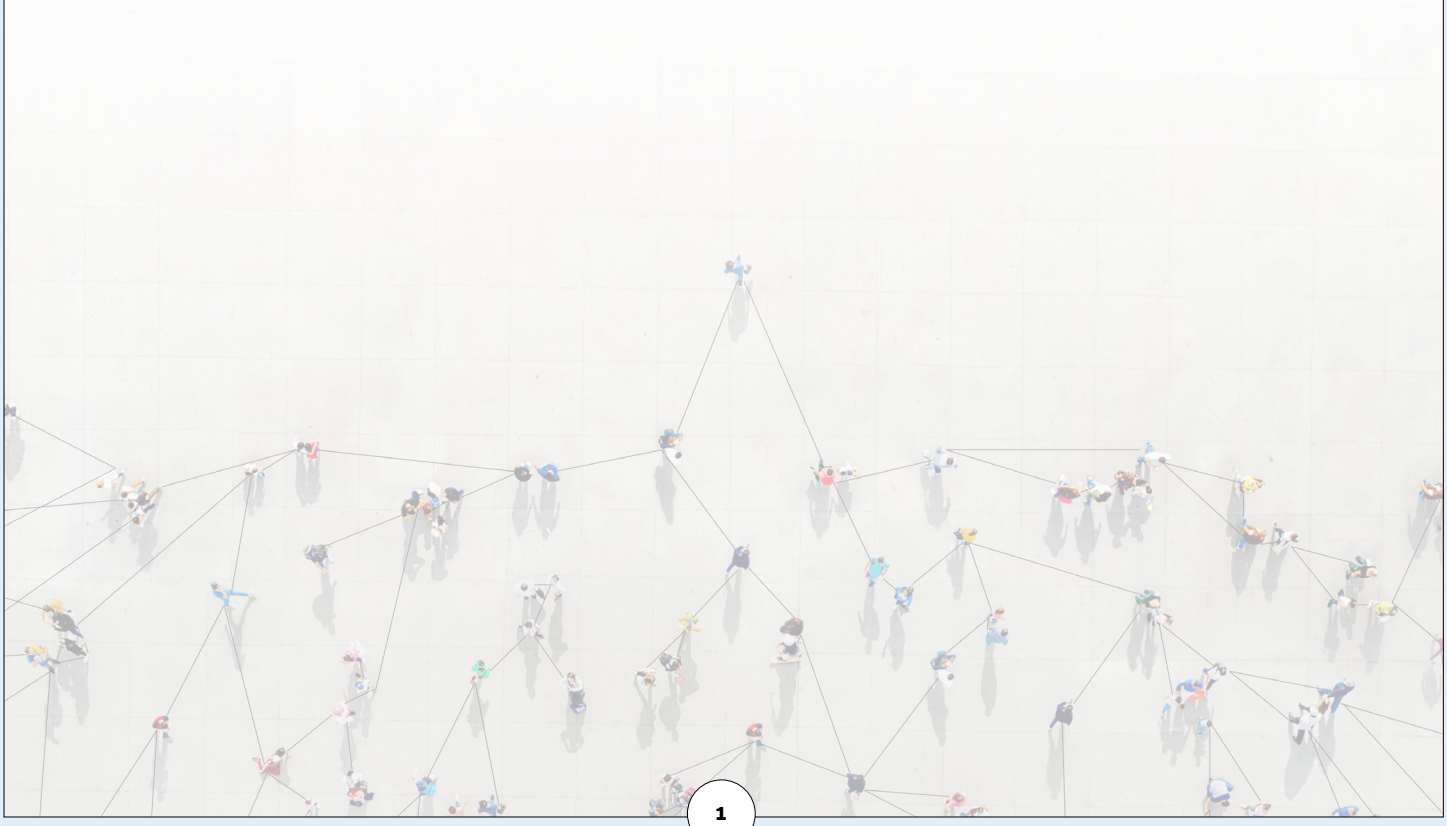
This primer introduces key concepts in and considerations for infectious disease transmission modeling and epidemiological data analysis relevant to modeling. This document was created for policymakers using non-technical language and provides answers to commonly asked questions. We've organized the primer into the following five sections, with each written to be read on their own. Please skip sections, begin in the middle, or read from the beginning to end, based on your interests and needs.

Purpose of this Primer **2**

1. Modeling 101: The Basics **3**

2. Modeling and Policy Decisions: An Overview **9**

3. Questions and Answers: What Policymakers are asking Modelers **11**



Purpose of this Primer

The COVID-19 pandemic is a public health crisis on a scale that most people alive today have never experienced. The importance of leadership and policymakers' access to accurate and timely information about COVID-19 to inform decision-making is critical. Equally important is their ability to assess the relevance and meaning of that information. Infectious disease transmission models (herein referred to as models) produce information that can inform policy decisions. Models describe how an epidemic may progress (e.g., number of people who may become infected and get sick, be hospitalized, and die) and how the trajectory of disease may change with different types of interventions (e.g., stay-at-home orders, partial business closures), in different populations, or over different time periods. A well-constructed model informed by accurate data can be a valuable decision tool if policymakers understand the model and the meaning and limitations of resulting estimates. This primer explains aspects of model development, the terms used to describe models, and how to interpret model output. We developed this primer as a ready reference for those working in government, business, and the health care community who are making decisions about COVID-19 interventions and responses at the local, state, and national levels.

Models are a part of our everyday lives. We rely on weather forecasts (a model output) to inform routine, personal decisions (what to wear) and public decisions with implications for health and safety (whether to close roads and schools, when to evacuate towns). Businesses rely on models to assess investment risks, understand market fluctuations, and inform insurance rates. These same familiar approaches have been used throughout the pandemic to apply the best available science to understanding how the pandemic will likely affect health and impact available resources.

Applications of Modeling for COVID-19 Policy and Practice

Models have been essential to shaping local, state and national COVID-19 responses. Their applications will continue to evolve as the pandemic progresses and the treatments and interventions evolve. Examples of their application are listed below.

- Monitor the growth and spread of COVID-19 pandemic at the local, state, and national levels
- Evaluate efforts to mitigate and control COVID-19 spread
- Identify trends in COVID-19 infections, hospitalizations, and deaths
- Guide purchase and allocation of health care resources
- Inform decisions about school closure and re-opening
- Inform decisions about business closure and re-opening
- Monitor the impact of seasonal influenza on COVID-19 infections, hospitalizations, and deaths
- Assess the impact of different vaccine distribution strategies

1. MODELING 101: THE BASICS

Infectious disease transmission models (herein referred to as models) are tools for using epidemiological, biological, statistical and mathematical techniques to describe how an epidemic may progress. They provide a precise framework to integrate these different types of information to develop scientifically informed guidance. Models are able to describe the trajectory of an epidemic and how the trajectory may change with different types of interventions, in different populations, or over different time periods. For example, models provide estimates using the best available data about the number of people likely to be infected, hospitalized, and die from a disease. Mathematical and statistical equations are used to produce estimates of the course of the COVID-19 pandemic to identify how quickly the virus may spread, how many people may become infected and get sick, be hospitalized, and die. Importantly for policymakers, models can be used to understand how policies may impact those estimates. The number of people a model estimates may get infected, require hospitalization, and die will be different depending on whether the government takes no action or implements effective policies that reduce the spread of disease. Therefore, a well-constructed model informed by accurate data can be a valuable tool for policymakers.

1.1. Model types by function

1. **Theoretical models:** Sometimes referred to as conceptual or illustrative models, these models are used to illustrate a key concept or test a hypothetical scenario by assessing how a disease system will likely behave under well-defined conditions. Theoretical models are sometimes simplistic, focusing on the impact of a single parameter or intervention on disease dynamics, where lack of complexity is considered a feature rather than a shortcoming.

   Example: Impact of waning immunity on population-level susceptibility. We do not know exactly how immunity to SARS-CoV-2 will wane and if this will vary across different demographic groups. Since we have limited understanding of these immunological dynamics, theoretical models can be developed to investigate how different assumptions about waning immunity (from the available literature and studies or other similar viruses) can impact the number of cases, hospitalizations, and deaths.

2. **Strategic models:** Also known as planning or scenario models, these models rely on the latest knowledge and best available evidence about how an infectious disease spreads and the effectiveness of interventions to reduce that spread to estimate the potential impact of different public health interventions on future epidemic trajectories and disease outcomes. These models are often developed with particular locations or settings in mind and the population structure is explicitly chosen to reflect that place. Planning models typically assess a range of plausible scenarios in the absence and presence of interventions over an extended time frame (months and years) and are therefore useful in informing policy decisions about interventions to interrupt the spread of disease.

   Example: The impact of different school re-opening strategies on cases. As schools begin to reopen, different school systems are choosing between different strategies (completely in-person, hybrid, completely online) and are balancing the risks of an outbreak. Strategic models can be used to investigate the effectiveness (in terms of reduced cases) of different strategies.

3. **Inferential models:** These models are typically used to test hypotheses about how a disease system works, infer key unknown parameters (e.g., the reproduction number, population susceptibility), or estimate the impact of interventions on observed disease trends. Inferential models rely heavily on the availability of data, including information on disease outcomes (e.g., case counts, deaths) as well as intervention timing and uptake in populations.

   Example: The impact of social distancing measures on transmission (as captured by the time-varying reproduction number, R_t). Using data on reductions in mobility patterns and estimates of transmission (time-varying Rt), you can see how much transmission reduced as a result of social distancing measures.

4. **Forecasting models.** These models rely on current and past behavior to forecast future disease trajectories and the impact of interventions to reduce COVID-19 disease spread. Forecasting models generally are most accurate over a shorter time scale (days and weeks) and are therefore useful in informing health care resource planning decisions in the short-term.

   Example: Forecasted estimates of case counts over the next two weeks across the US.

Susceptible-Exposed-Infected-Recovered (SEIR) — most standard mechanistic model used to model SARS-CoV-2 transmission

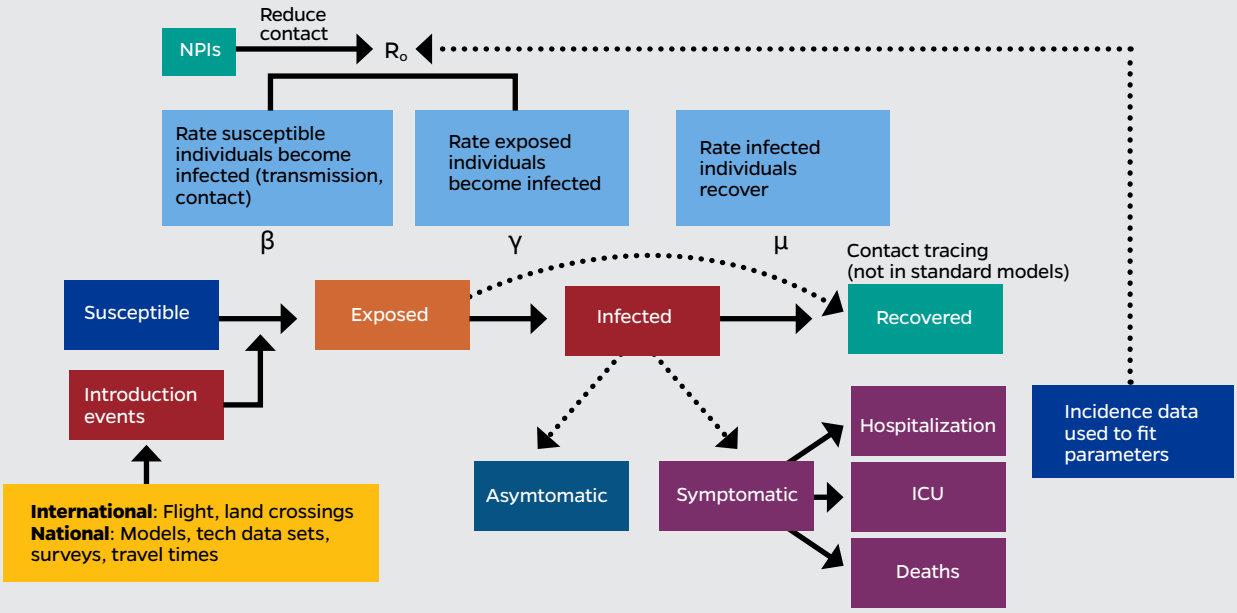
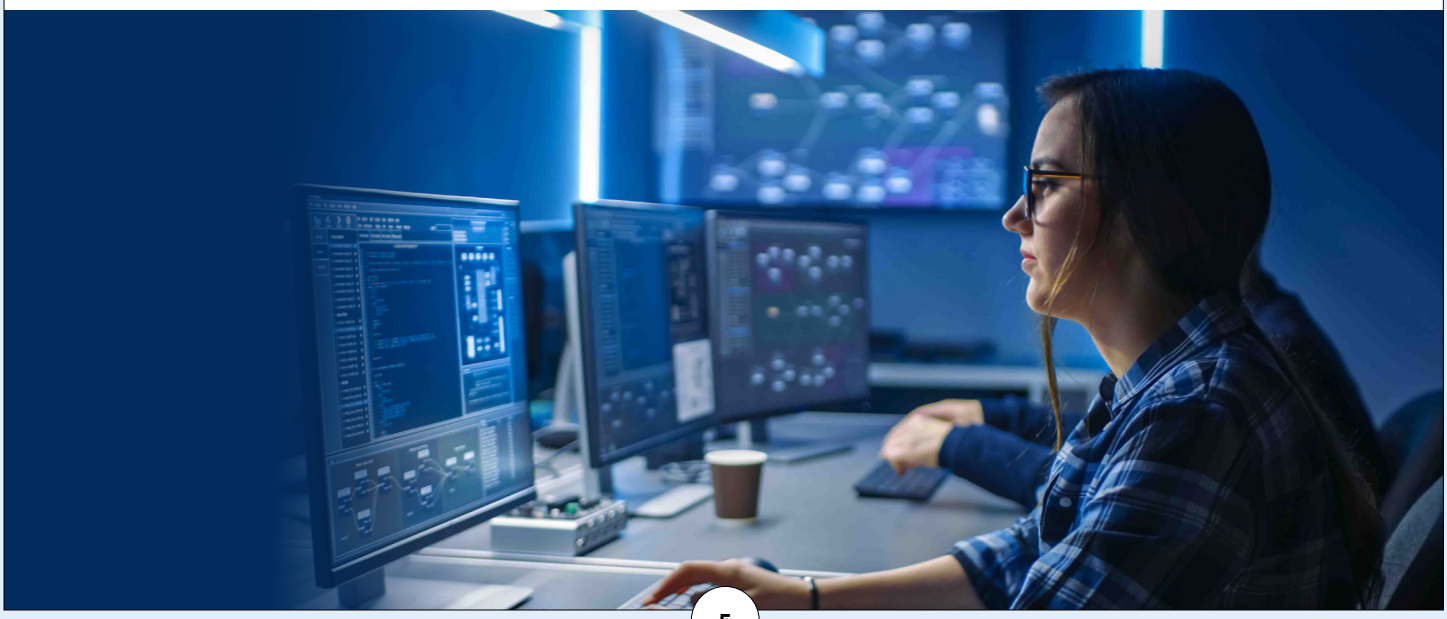


Figure 1. A typical SEIR model of SARS-COV-2 transmission. In an SEIR model, individuals start susceptible (S), then become exposed (E) following contact with an infected and infectious individual, after some time, exposed individuals become infected and infectious (I), then individuals become recovered (R). Each of these states are referred to as compartments. The rate at which individuals traverse between these different states β, γ, μ may be fixed values or drawn from a distribution of likely values, and these rates may be impacted by interventions. Although shown here as just a few compartments, each compartment can be further broken down by age or space to reflect different transmission patterns in different locations and/or by age groups. Infected individuals will become asymptomatic or symptomatic, and symptomatic individuals can then progress to hospitalization, ICU, or death. Incidence or death data can be used in inference models to fit some parameters (like R_0). Importantly, non-pharmaceutical interventions (NPIs) like social distancing and school closures reduce R_0. While this is not the only way to integrate NPIs and different NPIs may impact different parts of transmission (for example, travel restrictions may prevent introduction events from occurring), this is the most common method. Other NPIs like contact tracing have more variability in how they're integrated into models, however here is one example where contact tracing can remove some exposed individuals (the infected contacts of a known case) from infected others and hence those exposed individuals are not able to transmit to others since they move into the recovered class.

1.2. Types of modeling approaches

Infectious disease models can be constructed using a variety of methodological approaches, each with their own strengths and limitations.

1. **Mechanistic modeling:** This approach inherently takes into account biological and nonlinear feedback mechanisms, allowing models to incorporate how these processes dictate the occurrence of infections and transmission dynamics. The most common mechanistic models of SARS-CoV-2 transmission use the general framework of a Susceptible-Exposed-Infectious-Recovered (SEIR) compartmental model, where in the simplest case, a population of individuals is divided between 4 mutually exclusive disease-related states, referred to as compartments. (S,E,I and R; Figure 1). In this model (and other mechanistic models), there is a set of parameters that governs how individuals become infected, what happens once individuals are infected, and who is likely susceptible or immune. Probabilities are assigned (typically using real-world, empirical data) to determine the transition from one compartment into another. Because mechanistic models incorporate non-linear feedback from biological processes, as more people get infected the disease will spread faster.
2. **Statistical Modeling:** This approach uses statistical methods that fit functions (or curves) of data. The fitted functions are then used to extrapolate or predict into the future. Examples of statistical models include regression analyses (e.g. logistic or linear) and machine learning approaches. Statistical modeling typically represents highly data driven applications and is therefore reliant on data and data quality. However, unlike mechanistic models, these models usually do not incorporate biological mechanisms or transmission dynamics, or how diseases progress. Thus, they are also not well-suited for long-term projections and inferring differential impacts of multiple interventions.
3. **Ensemble Modeling:** This approach aggregates predictions from multiple diverse models, which can be both mechanistic and statistical models, to reduce errors in predictions. Thus, this approach uses the wisdom of multiple models from different research groups to produce a combined, improved prediction. The most popular ensemble model for COVID-19 is the COVID-19 Forecast Hub by the Reich Lab of the University of Massachusetts Amherst, which is also used by the CDC for national, state and county forecasts.



1.3. Key components of a COVID-19 model

Infectious disease models include 3 key components:

1. **Structure:** The parts of models that model designers deem important to generating the best estimates using available data. This is the formula of a model.

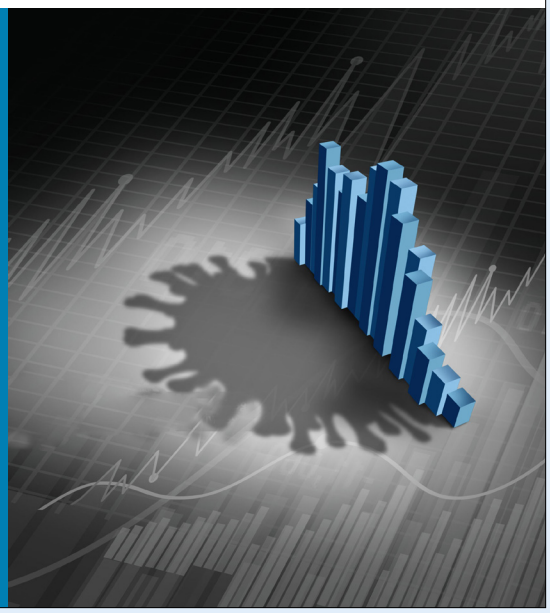
   The general structure of most mechanistic models of COVID-19 transmission has remained relatively the same during the pandemic. The majority of mechanistic models classify people as susceptible, exposed, infected, and then they either become sick and recover or die. However, models have evolved to better reflect reality by including geography, age, and more complex states of infection (asymptomatic and pre-symptomatic infections in addition to symptomatic infections). See Figure 1. This allows models to better describe the biology of the virus and its trajectory of impact as it moves through the population and what interventions may be used at what time and to what effect. For example, a model could estimate the impact of contact tracing and school closures on rates of COVID-19 infection, hospitalization, and death. Moreover, models are being developed to explore other emerging questions such as the projected impact of vaccination on the trajectory of COVID-19.

2. **Inputs and Parameters:** The variables and data used to generate estimates from models. Inputs and parameters will evolve as we learn more about the COVID-19 disease trajectory. As inputs become more precise and accurate, the resulting models will be more realistic representations of the impact of COVID-19.

   Initially, models for COVID-19 were based on transmission parameters (i.e, variables) that were derived from what we know about other coronaviruses or influenza viruses. However, increasing knowledge about SARS-CoV-2 and increasingly precise data on hospitalization rates and disease severity by age, for example, allow models to be directly informed by COVID-19 empirical data. More and improved data on COVID-19 also allows modelers to estimate the time-varying reproduction number (Rt), which enables models to reflect the true transmission patterns in different locations over time more accurately.

3. **Model Outputs:** The estimates of COVID-19 that models generate. This might include the number of infections, people hospitalized, and deaths that might be expected under different policy decisions over a given time frame. As inputs and parameters are refined, the resulting outputs will be more accurate. Modeled outputs can also be comparisons between different scenarios.

Overall, the types of outputs of most COVID-19 models have not substantially changed as most models still focus on estimating cases, hospitalizations, and deaths. However, better data lead to more refined and accurate model estimates, and incorporate more policy intervention scenarios. For example, models have been updated to stratify estimates by age groups to account for increasing disease severity with age and have been improved by using data on hospital bed capacity. Models are also now able to estimate the actual impact of policy interventions on disease transmission rates (such social distancing, stay-at-home orders, and contact tracing) as opposed to just simulating what their impact may be.



1.4. Characteristics of Good Models

Models should be reasonable, useful, and relevant. Here is how we think about these qualities in the current context of COVID-19.

1. **Reasonable:** Do the assumptions in the model (transmission rate, intervention effect, etc.) reflect the biological, behavioral, and population dynamics that are of interest? Will the model be able to simulate or reflect the true COVID-19 situation? Not all models need to include all aspects of the population or transmission dynamics to produce useful estimates. While it is not easy to judge the ideal model complexity needed to answer the questions of interest, models should be able to reproduce the most essential aspects of the modeled system.

2. **Useful:** Is this model asking (and answering) questions that are relevant for decision-makers and that advance policy discussions? Models are one source of information for policymakers. They compare how factors that influence disease transmission or different intervention scenarios affect estimates of health impacts (e.g., people infected, hospitalized, and who die). They can inform policy questions, but they do not provide answers as to the best policy action to take. That is for policymakers to decide.

3. **Relevant:** In comparison to other models or data, does this model meaningfully add to what we already know? Does this model incorporate new information? While it is important to continuously validate and improve models based on the available data, developing models for theoretical scenarios should focus on questions that remain unknown so as to maximize the impact of available resources.

1.5. Assessing Models: Key Questions to Ask

If you are using modeling to guide decision-making, you should always ask questions of the model. Asking the following key questions may help you quickly decide whether a COVID-19 model meets your needs and is ultimately useful.

1. **Is the model responsive to the question(s) you are asking?**
   Not all questions decision-makers have will be answerable through modeling. Be clear about the question(s) you are trying to answer and assess if the model and its assumptions are appropriate to answer these questions.

2. **How is uncertainty or chance incorporated into the model?**
   Our evidence and knowledge of infectious diseases, including COVID-19, is incomplete, so uncertainty is always present and needs to be incorporated into model estimates and described in conjunction with these estimates. Sources of uncertainty may include:
   - **parameters:** transmission (e.g. R0), the time it takes to recover, and the duration of immunity after infection
   - **structure:** modeled components (e.g. asymptomatic transmission)
   - **setting:** closed (isolated) populations with no importation of cases vs. importation of cases as a result of population mobility
   - **intervention effects and management processes:** assumed level of change in transmission, expected impact on case counts and deaths
   - **stochastic/ statistical uncertainty:** random error or chance since, for example, every case may not be counted

3. **Are you using the model to inform short (weeks) or long-term (months and years) decisions?**
   Modeling short- and long-term consequences usually requires different modeling approaches.

4. **Are multiple models available to inform decision making?** Decisions are best informed by predictions from multiple models (e.g., ensemble modeling). This may be challenging as there are often multiple models that are similar, but not the same. For reference, the COVID-19 Forecast Hub provides an often-referenced example of an ensemble forecast.

5. **How wide are the confidence intervals surrounding the point estimates?**
If model findings are presented as singular points without intervals that provide a range within which the true estimate lies, the model either does not account for uncertainty or the confidence intervals are not presented. Both explanations may result in overly confident interpretations, which should be an indication to question or doubt the model and its results. Understanding uncertainty is particularly important for long-term predictions as uncertainty increases with the length of the time period reflected in the estimates. Confidence intervals increase with greater uncertainty.

6. **Is the model representative of your population of interest?**
This refers to the setting modeled (e.g., geography, demography), the pandemic phase (e.g., how current is the data relative to the COVID-19 trajectory of disease), as well as the observed data (e.g, transmission rates and sources of estimated intervention impact). If a model refers to a population or setting different than the one you are interested in, you should try to assess to what extent the estimates are applicable to your population of interest. Consult an expert about the robustness or "generalizability" of the model estimates.

7. **What are the limitations in the data (both in terms of scope and quality)?**
Our understanding of the pandemic is evolving rapidly, as is our ability to collect complete and accurate data. Both our understanding of COVID-19 and our data collection abilities affect the strength of the models. It takes time to conduct sound science despite the urgency of the crisis and the need for evidence informed decision-making. Recognizing that the information used to create model estimates is changing, and that this will cause estimates to change, is important to assuring realistic expectations among those using model estimates to inform policy decisions.

8. **What additional information or data do you need to answer my question(s)?**
Optimal collaboration between modelers and policymakers should include bidirectional communication and feedback. As much as modelers want to help policymakers make informed decisions, they also want to know how they can improve their models to fit the needs of policymakers. Models are dependent on data input and policymakers can help improve modeling by supporting the collection and sharing of high quality data for use in models.

9. **What steps have been taken to validate the model?**
Have steps been taken to validate the model? For forecasting models, the best validation is usually past performance in prediction, but measures of model fit, "out of sample" predictive power, and performance on synthetic data are all methods that can help validate a model's performance. Performance should always be measured against a reasonable null/naive model (e.g., projecting that future weeks will have the same incidence as the week of a forecast) to see if they are actually adding value.

2. MODELING AND POLICY DECISIONS: AN OVERVIEW

Models can be valuable tools to guide policy decisions. For example, decisions about when and how to reopen after stay-at-home orders or how to allocate vaccines are two areas where policymakers have used models to inform decision-making. Models can be responsive to national questions, or fine-tuned to guide localized intervention strategies at the state or county level. Regardless of the geographic scope, access to data about the population of interest is key to yielding precise estimates of local COVID-19 trajectories in relation to different intervention options.

Models can help guide decisions for reopening schools and universities. However, balancing student and staff safety and the impact of continued school closures on learning, child development, and the economy requires a multi-disciplinary effort to assess the full range of economic, developmental, societal, and public health outcomes associated with these decisions.

As we roll out a coronavirus vaccine, modeling will be useful to those charged with making decisions about distribution and rollout. Questions about the optimal geographic areas and populations for vaccines and therapeutic drugs can aid in improving safety and effectiveness, and ensuring equitable access to vaccines and therapeutics. Modeling can guide policy and planning to optimize the reach, prioritize equity, and guard against systemic racism in all aspects of vaccine allocation, distribution, and administration.

Ideally, models provide decision-makers with balanced assessments of public health, economic, and social impacts associated with policy options for controlling COVID-19 and other infectious diseases.

2.1. Examples of Policy Decisions that can be Informed by Models

As we look to the challenges in the near term that will likely come before policymakers, several questions can be informed by modeling:

1. **How can we safely and responsibly reopen the economy and societal life?**
   - What are the benefits and drawbacks of partial or phased reopening and/or stricter social distancing?
   - Which sectors of the economy should be prioritized for reopening?

2. **How will decisions about reinstituting closures be made? How can we safely and responsibly reopen schools and universities?**
   - What are the economic, developmental, social, and public health impacts associated with virtual, hybrid, and in-person instruction decisions?
   - How will decisions about reinstituting closures be made?
   - Are there scientifically defensible reasons for different reopening decisions for different grades and for different types of schools (e.g., public, private, boarding)?

3. **What is the influence of seasonal influenza on COVID-19 impacts?**

4. **What are the scientifically optimal and equitable strategies for trialling vaccines and therapeutic drugs?**

5. **How can we maximize the public health impact of vaccine allocation, distribution, and administration which includes systems that are equitable and anti-racist?**

6. **What intervention strategies can be tailored to address state and county level needs?**

2.2. Incorporating Uncertainty from Model Estimates into Decision-Making

The major issue in modeling COVID-19 has been and continues to be that many aspects of this pandemic are poorly understood. COVID-19 is a new disease. There are limited data on the disease and inherent uncertainties regarding how people will behave in response to interventions to reduce disease spread (e.g., policies around wearing masks, social distancing). The extent to which people follow public health recommendations also impacts the effectiveness of those interventions that are being modeled.

Representing uncertainty

There are common ways to represent uncertainty in modeling approaches so that decision-makers can assess how much weight to give the model estimates.

Uncertainty in **statistical models** is primarily displayed through confidence (or prediction) intervals around a point estimate of the output. This is often referred to as "stochastic uncertainty" and represents the effect of random error or chance. Confidence intervals should widen as estimates project further into the future because uncertainty inherently increases with time.

In **mechanistic models** uncertainty is often reflected in the results of a sensitivity analysis. By using a range of values that are possible for any parameter included in the model, the sensitivity analysis qualifies resulting estimates so that decision-makers can better determine how to use the results. For example, some people may develop symptoms within 2 days, whereas others might develop them in 7 days. Thus, we can sample when they would develop symptoms from a range of values. This range can be informed by actual data (the ideal situation) or might represent a best guess based on the currently available evidence (less ideal but often the only option). Additionally, we can run models multiple times. Oftentimes, mechanistic models include stochasticity (randomness) and by running a model many times we can capture some of this uncertainty.

The biggest sources of uncertainty in COVID-19 modeling

Currently, the biggest sources of uncertainty in COVID-19 models include: **protective immunity**, or the level and duration of immunity among people who are infected and recover; **contact rates between susceptible and infected individuals**, or how much exposure is needed for someone to be infected with COVID-19; and the **contribution of asymptomatic/presymptomatic transmission to community spread**, particularly among children.

While improved testing of symptomatic cases, traced contacts, and population-wide and random testing continue to contribute to our understanding of infection rates and seroprevalence, more data on the COVID-19 immunity are needed to make predictions about long-term disease dynamics and pandemic trajectories. Additionally, the quality and population coverage of testing for SARS-CoV-2 varies by locality across the US and globally, resulting in substantial uncertainty regarding the reliability of testing data. We also have limited understanding of how social factors, such as occupation, education, and social class influence who becomes infected, what happens to their disease trajectory post infection, and how this varies across different areas or regions.

More data on the novel coronavirus immunity are needed to make predictions about long-term disease dynamics and pandemic trajectories.

3. ANSWERS TO COMMON QUESTIONS FROM DECISION-MAKERS

People keep saying 'mechanistic' models, what is the 'mechanism' behind these models?
Mechanistic models frequently consist of the structure "susceptible-exposed-infected-recovered" (SEIR). These models explicitly include biological processes for disease transmission and progression (which is the mechanism in this case). The mechanism in these models can include different progressions, for example infected individuals recovering or susceptible individuals becoming exposed or exposed individuals becoming infected. Thus mechanistic models incorporate disease transmission factors that impact the estimated outcomes. The biology of the novel coronavirus is important to understanding the trajectory and impact of COVID-19.

What is the difference between a statistical, mechanistic, and ensemble model?
Statistical models are more sensitive to underlying assumptions for the functional relationship between the observed data and future predictions or forecasts. Thus, their results may seem more optimistic if they are based on optimistic assumptions or data that indicates (a potentially fallacious) positive trend. Mechanistic models, on the other hand, simplify the underlying mechanism of disease progression by assigning members of a population to different compartments (e.g. susceptible-exposed-infected-recovered).

Some commentators have remarked that statistical models seem to be more optimistic than mechanistic models. However, the assumptions in mechanistic models are based on the best available knowledge about the disease process and include biological processes innate to disease spread. Changing this mechanism is more difficult as it involves making stronger assumptions, for example that susceptible and infected people do not have contact.

What is the difference between a forecast and projection?
Forecasts are most commonly based on statistical models that predict the number of expected cases, deaths, or other health-related outcomes. Forecast models are often used for short term predictions that are highly dependent on the available data. Statistical assumptions about how these data are related to infectious disease processes are used to extrapolate future cases, deaths, and hospitalizations.
Projections of the trajectory of the disease are normally an output from mechanistic transmission models. These are commonly used to compare different scenarios or suites of implemented interventions. They are also useful for exploring questions where we have very limited data or currently epidemiological or clinical understanding. For example, mechanistic models can be used to estimate the impact that different assumptions about waning immunity might have on possible hospitalized cases. This is a question where we currently have limited understanding, but it may be possible to look at various projections using mechanistic models.

Understanding model terminology

What does parameterizing a model mean?
Infectious disease models are mathematical constructs based on variables (parameters) that determine the likelihood or rate at which people are affected by a disease, such as getting infected. In compartmental models, parameters are used to determine the rate at which people progress from one compartment (susceptible, infected, etc.) to another. Parameterization is the process of defining or choosing the optimal parameters for a model. This is usually based on available data or assumptions about the most plausible values if there is scarce or no data. Parameters can be specific for different locations or age classes.

What do you mean when you say the estimation was 'unreliable'?
This usually refers to modeling in situations with sparse data on which to base the model. This is particularly true in the beginning of a pandemic when we have limited data making estimation more difficult. Estimations may also be unreliable based on reported data. For example, testing results may have an inherent delay and estimating transmission parameters without taking this into account could be incorrectly over or under estimating the transmission rate.

How is model accuracy determined?
In epidemiology, accuracy is defined as the degree to which a measurement or estimate represents the true value of what is being measured. Thus, accuracy indicates a model estimate's closeness to the "truth". However, the 'truth' may be difficult to define, measure, and quantify. Hence, accuracy and measures of the accuracy of a model are highly dependent on the availability of good-quality data. However, the accuracy of COVID-19 model predictions is mostly limited by our current knowledge about the virus and transmission dynamics in populations. Uncertainty regarding the biological characteristics of SARS-CoV-2 and the "true" number of infected individuals in the population represent two major issues limiting model accuracy.

Does adding more variables or information to a model always make it better?
Not necessarily. Making models more realistic, such as adding high-risk populations in long-term care facilities, may provide important policy guidance. However, adding in every possible variable will make the model more complex. Increasing complexity may make models less representative of natural disease transmission processes, thereby making them more difficult to interpret. For example, if there is no or very limited quality data for the included variables, this will add substantial uncertainty to the model and reduce its reliability. The model complexity always needs to be considered in the context of the question being asked, ability for data or epidemiological insight to inform the model, and usefulness in including those aspects. It will not always be true that more is better, sometimes it just adds complexity without gaining additional insight.

What is stochasticity and why does it matter?
In general, stochasticity refers to the role of random processes or chance represented in mathematical models. For example, every time a susceptible person comes in contact with an infected person, the susceptible person may or may not become infected. While we may expect the susceptible person to become infected based on experience or model calculations, there is always a possibility that it may not happen due to chance. In other instances, stochasticity often refers to random processes where an event may not happen due to chance, which results in uncertainty surrounding our estimates.

Making models more realistic and accurate

How well does a model reflect my population?
Representativeness is the extent to which a model (and its predictions) reflect the population of interest. This is influenced by how well the model is able to reflect underlying characteristics of the target population or geographic location, such as age structure, population demographics, and local transmission rates. Models can also include the effect of public health interventions implemented in a given place at a specific time to improve their representation of the real situation. However, all models will only be a representation of the true population of interest and will always be limited by our current understanding of the epidemiological situation and/or data from that location (or population).

How do models reflect different time points in the course of an epidemic or pandemic?
Optimally, the model should be calibrated (meaning fit to) to data reflective of the current phase or situation both in terms of transmission rates and disease frequency as well as implemented interventions. If the disease parameters do not reflect the current situation, the predictions will be unreliable. This was a major issue in early US pandemic models that relied on data and experience from other countries that did not reflect the characteristics of the pandemic in the US at that time.

What affects the accuracy of models?
The accuracy of model estimates can be improved by adding in more data or refining the model structure to better capture relevant aspects of disease transmission. For example, many COVID-19 models incorporate age as it is a key risk factor for COVID-19. However, there is a trade-off between increasing the complexity of a model by adding in more data and decreasing interpretability of complex model predictions.

In addition, model predictions depend on the availability and quality of the data used to inform the model. If the input data is biased towards underreporting of cases and we fit a model exactly to those data, then the resulting model will be inaccurately underestimating disease transmission even though it fits well with the data.

Interpreting modeling results

What do the bars (or lines) that surround the projected cases, deaths, etc. in plots mean?
These bars display the range of possible values around a midpoint value (usually the mean). When interpreting model results, it is important to consider the bars (or areas) surrounding the point estimates. Bars covering a large range of values signal there is substantial uncertainty in the model results. Small ranges indicate more precise estimates. In general, the width of the bar should increase over the time horizon of the projections as the distant future is always more uncertain. If this is not the case, you should be cautious about trusting these results.

Models show that the time-dependent reproduction number (Rt) went below one. Why do case counts still increase after this apparent reduction in transmission?
The time-dependent reproduction number (Rt) reflects the transmission potential of an infectious disease at a specific point in time. This transmission potential is dependent on the number of infected, susceptible, and immune individuals in the population at that time. As long as there is still contact between infected and susceptible individuals in a population, transmission will rebound and Rt will increase.

There are multiple models, which one should I trust?
You should compare estimates from different models developed by research groups representing diverse areas of expertise. This will allow you to get a more comprehensive answer to your questions. Additionally, "ensemble modeling" aggregates results from multiple diverse models into a combined estimate. This is the approach the CDC currently uses to forecast cumulative national death counts based on the COVID-19 Forecast Hub.

Informing policy decisions with models

How can you include aspects or factors that we do not know about yet into a model, for example interactions between tuberculosis and COVID-19 or interventions that have not been implemented?
For most factors that are included in models, we rely on the available data or evidence from the literature. For emerging topics like COVID-19, we can use surrogate data or information from other diseases that appear to be similar to COVID-19 or implemented interventions that may only be slightly different from the ones we want to model. For example, if we want to model the impact of school closures on COVID-19 spread, we could look at data for daycare closings since the schools and daycare share many similar traits relevant to SARS-CoV-2 transmission. In the absence of any information or data to guide our modeling assumptions, it is good practice to make these assumptions as simple as possible and to be very transparent about the data being used.

When you add non-pharmaceutical interventions (e.g., policies requiring masks or social distancing) into a model, how is this changing the model?
Non-pharmaceutical interventions (NPIs), such as quarantines, travel restrictions, and school closures, often seek to change human behaviors. Thus, they may be incorporated into models by reducing the contact rate between susceptible and infected individuals or reducing the transmission rate (represented by reductions in the reproduction number R_0). However, NPIs may also be integrated as a completely separate mechanism into the model.

What if I need a model to project out for at least 1 year or 5 years?

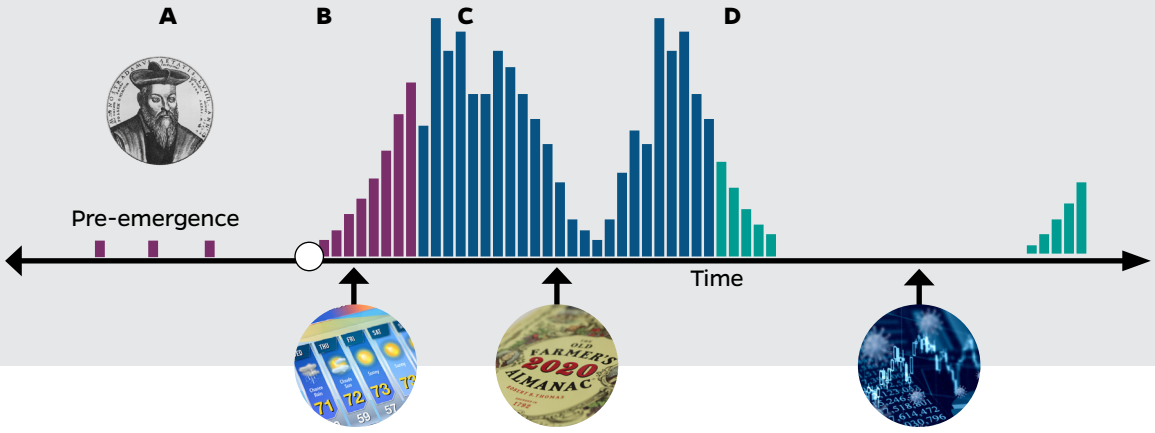
Different model types are better suited to making short- or long-term projections. If your aim is to project long-term effects, such as over multiple years, we recommend basing your decisions on models that incorporate a biological mechanism i.e.,a (mechanistic model) for disease transmission. It is important to remember that uncertainty increases over time and therefore all models become less reliable the further the projections reach into the future. This should be reflected by widening error bars around the model estimates over time. If this is not the case, you should be cautious about the model as it may not be accurate.

Can every question be informed by modeling?

No, models are not a universal solution to every problem. Models are limited by the availability of data and information to base our assumptions on. It is not a good idea to model a question that we do not know anything about. Conversely, it may not be useful to model a question for which we already have answers and abundant evidence. The main utility of models is to make complex relationships explicit and to guide decision-making in situations with substantial uncertainty. Do not trust anyone who says their model provides the definitive answer or solution.

Can anyone build a model?

Yes, technically everyone can build a model, but not everyone can build one that is informative and useful. A background in both technical aspects (mathematics, statistics, and computer science) along with public health and epidemiology are needed to develop realistic, sound, and useful models.



Model predictions and stages of an emerging epidemic. Early after emergence (B), and throughout the epidemic we can perform short term forecasts using both dynamic and epidemic models 2-4 weeks out with reasonable accuracy. An analogy for this period is your 10-day weather forecast, not perfect but reliable for planning. We are also able to, using the basic laws of epidemic spread, make some accurate predictions about the long-term fate of the epidemic (D) under reasonable sets of assumptions (e.g., the total percentage of people who will eventually be infected if there is not vaccine), akin to long term climate projections that show the earth will generally tend to warm if levels of greenhouse gases increase. Mid-term forecasts (C) at the range of months to a few years, are particularly difficult to model as they are driven by the complex interactions between disease dynamics and human behavior. The analogy here is the farmer's almanac, we can say something that might help some for planning, but the specifics are likely to be fairly inaccurate. The final period of interest is the pre-emergence period (A) where we just don't yet have the science to know which viruses are likely to emerge as threats to human health, and which will die out. The best analogy for forecasts here might be the prophecies of Nostradamus (adapted from https://science.sciencemag.org/content/357/6347/149.abstract).

What are some good resources to learn more about modeling?

- https://www.nejm.org/doi/full/10.1056/NEJMp2016822
- https://www.pnas.org/content/117/28/16092
- https://fivethirtyeight.com/features/a-comic-strip-tour-of-the-wild-world-of-pandemic-modeling/
- https://www.nationalacademies.org/news/2020/06/national-academies-release-covid-19-data-guide-for-decision-makers Imperial College https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/covid-19-planning-tools/
- https://www.nature.com/articles/d41586-020-01003-6
- https://www.coursera.org/specializations/infectious-disease-modelling
- https://www.jhsph.edu/covid-19/articles/10-tips-for-making-sense-of-covid-19-models-for-decision-making.html
- https://magazine.jhsph.edu/2020/making-sense-myriad-models
- https://www.jstor.org/stable/j.ctvcm4gk0

About the Authors

We are a cross disciplinary team from the Johns Hopkins Bloomberg School of Public Health that includes epidemiologists and infectious disease specialists, and people with expertise in computer science, economics, policy, and risk assessment who have worked within the United States and around the world to advance science informed decision-making for some of the most challenging public health issues.

Thank you for your work at this critical time in human history.

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH