Journal of Data Science 19 (2), 219–242
April 2021
DOI: 10.6339/21-JDS1013
Philosophies of Data Science

Methods, Challenges, and Practical Issues of COVID-19 Projection: A Data Science Perspective

Myungjin Kim[1], Zhiling Gu[1], Shan Yu[2], Guannan Wang[3], and Li Wang[1,*]

[1]*Department of Statistics, Iowa State University, Ames, IA, 50011, USA*
[2]*Department of Statistics, University of Virginia, Charlottesville, VA, 22904, USA*
[3]*Department of Mathematics, College of William & Mary, Williamsburg, VA, 23187, USA*

Abstract

The coronavirus disease 2019 (COVID-19) pandemic caused by the novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has placed epidemic modeling at the center of attention of public policymaking. Predicting the severity and speed of transmission of COVID-19 is crucial to resource management and developing strategies to deal with this epidemic. Based on the available data from current and previous outbreaks, many efforts have been made to develop epidemiological models, including statistical models, computer simulations, mathematical representations of the virus and its impacts, and many more. Despite their usefulness, modeling and forecasting the spread of COVID-19 remains a challenge. In this article, we give an overview of the unique features and issues of COVID-19 data and how they impact epidemic modeling and projection. In addition, we illustrate how various models could be connected to each other. Moreover, we provide new data science perspectives on the challenges of COVID-19 forecasting, from data collection, curation, and validation to the limitations of models, as well as the uncertainty of the forecast. Finally, we discuss some data science practices that are crucial to more robust and accurate epidemic forecasting.

Keywords *COVID-19; disease spread; epidemic models; forecast; uncertainty*

1 Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was first reported in the city of Wuhan in Hubei Province in China in December 2019 and is still in progress globally. As of March 2021, more than one hundred ten million COVID-19 cases have been reported in 219 countries and territories, leading to more than two and a half million deaths. Although trends of COVID-19 cases and deaths have turned downward, new COVID-19 variants have been recently detected worldwide. They may cause further complexity and spread of COVID-19. In response to fast-spreading COVID-19, various research on epidemic models has been conducted to understand its dynamic nature in mechanistic and phenomenological ways.

To study the transmission of an infectious disease such as COVID-19, researchers build epidemiological models based on available data from current and previous outbreaks. The idea of epidemic modeling and forecasting is not new. According to Dietz and Heesterbeek (2002), the earliest disease transmission model dates back to Daniel Bernoulli. He developed an epidemiological model in 1760 and used it to demonstrate that life expectancy increased due to the use of inoculation against smallpox. In recent years, many methods have been developed to

*Corresponding author. Email: lilywang@iastate.edu.

Received April 21, 2021; Accepted April 22, 2021

investigate infectious disease forecasting (Brauer et al., 2019; Held et al., 2020). These epidemic models can be statistical models, computer simulations, or other mathematical representations of the virus and its impacts.

As described in Wang et al. (2021), the primary purposes of the epidemic models are: (i) to predict the number of beds and the amount of protective gear and equipment that will be needed to treat infected patients; (ii) to gain a better understanding of how the virus spreads; and (iii) to identify which factors contribute to the spread of the epidemic and help plan effective control policies for diseases. Epidemic projection is essential in the planning of the public health response to infectious disease outbreaks. For example, we can plan for the preventive steps needed and significantly diminish the impact of the disease if we can accurately forecast when the outbreak will hit its peak. As per World Health Organization (WHO), epidemic models are vital tools for successful control and hopeful elimination of measles; see SAGE Working Group on Measles and Rubella (2019).

Compared with other infectious diseases, COVID-19 brings unique challenges to data scientists when they are trying to predict the spread of the virus: (i) The spread of COVID-19 involves more human interventions, which boosts the prediction uncertainty. Compared with influenza, COVID-19 is much easier to spread and has a higher mortality rate, which requires more human actions to limit the disease spread and protect public health. (ii) Unlike other infectious diseases, we have limited historical data to study the trend of COVID-19 spread. For other infectious diseases, the availability of historical data and the recurring nature of seasonal epidemics provide promising opportunities for epidemic modeling.

The remainder of the article is organized as follows. Section 2 outlines key data features and issues, which pose significant challenges to modeling and projection. In epidemic modeling, there are three main types of models: (i) mechanistic models, (ii) phenomenological models, and (iii) hybrid models. Sections 3–5 discuss in more detail some representative forecasting methods in each type (see Figure 1) and how the data features and issues impact these forecasting methods. To help understand the factors that contribute to COVID-19 projection, Section 6 describes some essential data science practice that helps improve the accuracy of the prediction. Section 7 concludes the article with a discussion about the unique computational and methodological challenges brought by the uncertainties of pandemic and the limitations of models, as well as how to improve the data strategy and methods.

2 Data Features and Challenges

Data usually plays an essential role in epidemiology analysis since the epidemic data involve unique features not shared by traditional datasets. First of all, a new infection occurs usually due to an infectious contact with the infectious agent, and this contact could be person-to-person, such as droplets, coughing, sneezing, and wheezing. Secondly, typically, there is a latent period between the time of infection onset and the time the individual becomes infectious, and this latent period varies among different diseases. Third, when the infectious period ends, the individual either becomes susceptible again or becomes removed from the susceptible population. Here, a removal represents the individual who is immune or dead due to the infection. Many of these data features could affect the use of models and impact the accuracy and legitimacy of the forecasting. In this section, we discuss the main features and the resulting challenges of COVID-19 data.

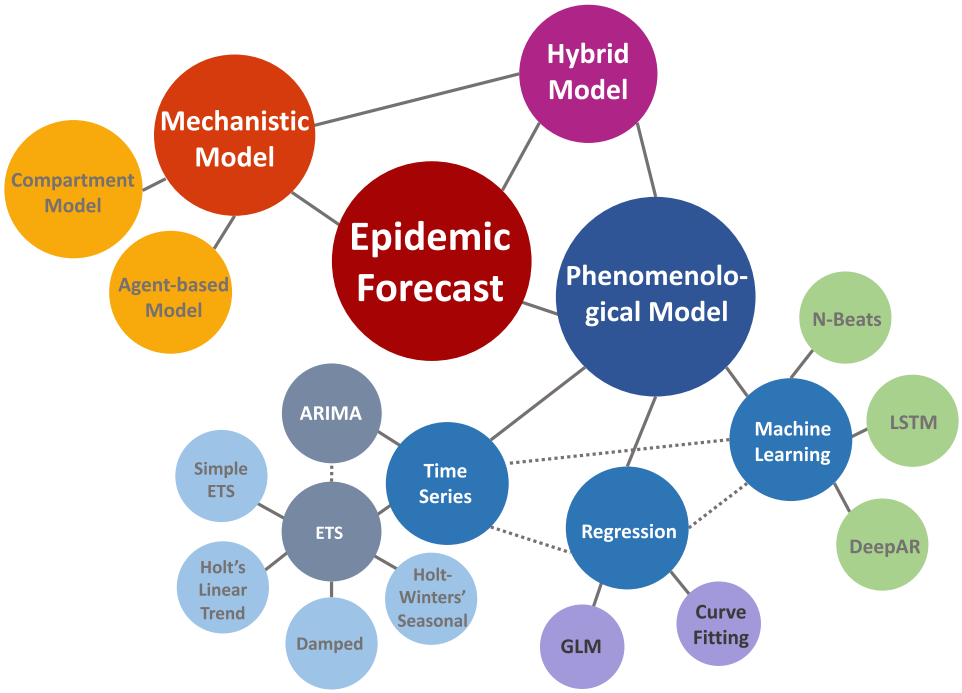

Figure 1: Various models for epidemic prediction, including (i) mechanistic models, (ii) phenomenological models, and (iii) hybrid models.

2.1 At the Early Stage

When a novel virus starts to spread, a lack of information in history imposes difficulties and uncertainties in the disease transmission projection. Specifically, it can induce the following problems.

Inflated Key Numbers for Theory-based Modeling As pointed out in Ioannidis et al. (2020), key parameters in theory-based epidemic modeling are difficult to determine. For example, people think that the coronavirus is commonly spread by close contact and sometimes by airborne transmission; however, the complete transmission modes are not fully clear. Moreover, COVID-19 produces various symptoms among infected individuals, and many cases are even asymptomatic. Therefore, without enough cases for reliable analysis, it is challenging to determine the length of the latent period. Furthermore, many other parameters, such as the fatality rate, infection fatality rate, and basic reproductive number, are inflated at the early stage of the outbreak. All of these would lead to more arduous work for epidemic modeling, especially for mechanistic models.

Lack of Historical Data Some of the well-studied infectious diseases, such as influenza, present strong seasonality. COVID-19 has only passed the one-year mark since the first case was reported. Thus, it is difficult to model and predict the seasonality effect due to the lack of historical data. Besides, without historical information on COVID-19, whether an individual will become immune after the infectious period is still unclear.

Other Effects The false-negative rate is exceptionally high in the early course of COVID-19 infection. Therefore, the infectious cases should not be ruled out solely based on the reverse transcriptase-polymerase chain reaction (RT-PCR) tests. As a result, it is important to embed clinical and epidemiologic information into the models (Kucirka et al., 2020). Since the beginning of the COVID-19 pandemic, several studies (Wang et al., 2020; Hoffman, 2021; Cramer et al., 2021; Castro et al., 2020) have illustrated that the day-of-week (DOW) effect, also referred to as or related to a seven-day cycle and the weekend effect, is present in epidemic data and can be very significant for some regions. When interpreting the data and forecasting the spread, it is crucial to consider these effects, especially in the short-term forecast. However, at the early stage of the pandemic, these effects are difficult to capture due to the sparsity of the data.

2.2 Throughout the Outbreak

Some data challenges remain throughout the outbreak, and we are also facing some new challenges as the pandemic continues.

Inconsistency and Non-existence of Standard Definitions Throughout the pandemic, the definitions of various important measurements are not consistent across time and regions. For example, as pointed out by Tenforde et al. (2020), recovery from COVID-19 is still undefined. The prolonged symptoms after being discharged from the hospital lead to divergent opinions on the time of recovery.

Another example is probable cases. Both probable cases and confirmed cases are recommended to be reported according to the CDC recommendation (Council of State and Territorial Epidemiologists, 2020) and the WHO guidance for public health surveillance (World Health Organization, 2020). More specifically, according to WHO, the definition of a "probable case of COVID-19 includes someone meeting clinical criteria and who is a contact of a probable or confirmed case, or a suspected case with chest imaging findings suggestive of COVID-19." However, it is difficult to measure how such definitions are practically implemented.

Lack of Uniform Data Reporting Mechanism The epidemic data vary in geo-temporal and demographic resolution, availability, and timeliness of release (Brooks, 2020). In general, for epidemic data, there are four popular open sources, including (i) the NYT (New York Times, 2021), (ii) the COVID Tracking Project by the Atlantic (Atlantic, 2021), (iii) the JHU (Johns Hopkins University Center for Systems Science and Engineering, 2021), and (iv) the USAFacts (USAFacts, 2021). As pointed out in Wang et al. (2020), these sources provide the data with different precision and focus. The NYT, JHU, and USAFacts released daily data at the national, state, and county levels, while the Atlantic releases daily national and state-level data along with testing, hospitalization, and recovery information. Recently, the Atlantic has stopped the daily data updating. Furthermore, different data sources implement different strategies when assigning the cases and deaths to a place. In particular, in New York City, various jurisdictions release data that can lead to double counting. Some sources might directly combine the official reports while other sources implement adjustments to solve the possible double-counting. This results in recognizable discrepancies between sources. Different geographical rules have been used for various sources. For example, unlike the USAFacts, NYT and JHU combine data in Kings, Queens, Bronx, and Richmond counties with New York City's data. A more detailed comparison can be found in Wang et al. (2020). All in all, we need to clearly understand the processing behind the data source we use for analysis.

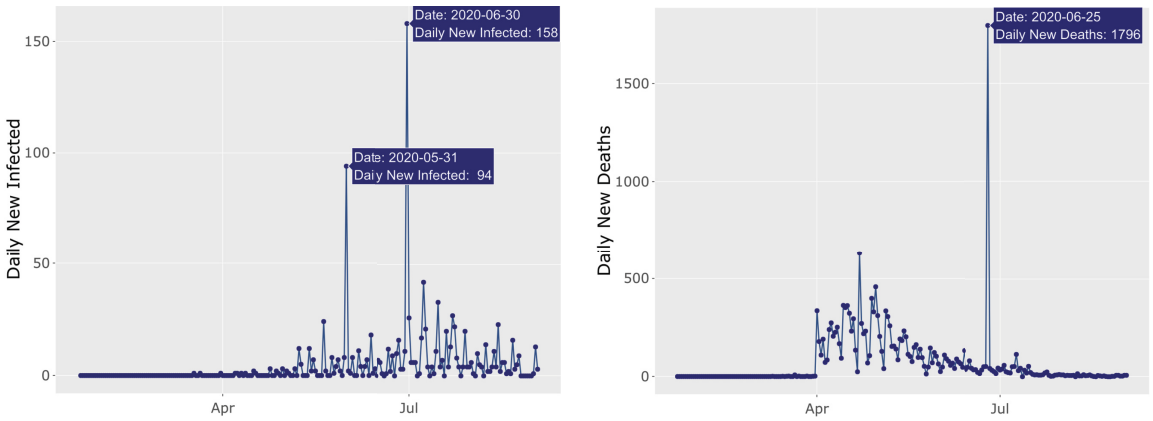

Figure 2: Spikes observed in the time series from January 23, 2020 to September 1, 2020: (a) the daily new infected count for the Grimes County, Texas, and (b) the daily new death count for the State of New Jersey. (Source: https://github.com/CSSEGISandData/COVID-19. Retrieved on December 12, 2020).

Integration Challenge Some other datasets, such as non-epidemiological spatial data, serological data, or pathogen genomic data, could also help people identify the significant factors of disease transmission. One of the essential pieces of information is the control policies placed. Since the beginning of the pandemic, many executive orders, such as emergency declarations, school closures, social distancing, have been issued by the federal, state, and local governments. Many research groups have made every effort to collect this information by consistently checking governmental websites, news, and press releases. Moreover, many other demographic characteristics, healthcare infrastructures, and socioeconomic features can be obtained from different sources, such as the American Community Survey. However, since the underlying process of data generation varies across different data sources, integrating them sensibly is a major challenge.

Data Privacy and Transparency The raw epidemic data might contain some personally identifiable information. However, the privacy issue does not warrant a delay or withholding public release. As suggested in Gardner et al. (2021), when aggregating the data, all of the personal information should be anonymized, and the aggregated data should be published promptly.

Ragged Data Availability For COVID-19, some data and information are reported with substantial delay (Funk et al., 2019). Backlogs lead to abrupt jumps as shown in Wang et al. (2020). For example, the spikes in Figure 2(a) might be related to the reported cases in the Texas Department of Criminal Justice (TDCJ) dating back to earlier in the outbreak. At the same time, backlogs and constant revision of data make the evaluation of forecasting extremely difficult, and the comparison between different sets of forecasting over-complicated.

These special characteristics pose significant challenges to data analysis and motivate the development of new epidemiological modeling and projection methods. Sections 3–5 below will describe how the data features and issues impact COVID-19 modeling and forecasting.

3 Mechanistic Models

Mechanistic models explicitly try to model the mechanisms of interaction among system components. There are two common frameworks in mechanistic models: compartment and agent-based models. These models show the commonalities in making explicit assumptions about the biological mechanisms that drive infection dynamics.

Compartment models are based on a dynamic system of equations with strong assumptions, such as a well-mixed population, homogeneity of the population, and large population (Huppert and Katriel, 2013). From their flexibility of modifications, compartment models are well-suited, especially for epidemic diseases that evolve further and require significant improvement in model specifications over time. Given initial conditions, they also easily and intuitively characterize a course of disease well under a simple structure. However, due to their unrealistic simplicity, they often lead to inaccurate results by themselves. Agent-based models use more detailed specifications of disease states and/or individual characteristics and behavior, which cannot be easily simplified into a compartmental form. They allow for exploring complex systems and capturing relationships among individuals and heterogeneity in their attributes (Rahmandad and Sterman, 2008). Also, they can model experiments that may be impossible or unethical to conduct in the real world and guide public health interventions. However, these models require intensive computational burdens, constraining sensitivity analysis, which is critical to ensure robust results. Besides, it is hard to obtain the contact networks and the distribution of individual attributes.

3.1 Compartment Modeling

A compartment model is characterized by a set of mathematical equations representing how individuals move from one compartment to another or interact among compartments. A majority of the compartment models in epidemiology are based on the Susceptible-infected-removed (SIR) model (Kermack and McKendrick, 1927) and its variants to conceptualize the dynamic changes of the population.

The conventional SIR model divides a population into three compartments: the susceptible class (S) indicates those who can become infected, the infected class (I) represents those who are infectious, and the removed class (R) corresponds to those with permanent infection-acquired immunity. Due to its simplicity with a small number of parameters, the SIR model has become popular and has been widely applied to track and forecast the trajectory of epidemic disease. For example, Chen et al. (2021) considered the SIR model for COVID-19 in Canada. The estimates and bounds of model parameters were extracted by minimizing the squared prediction error of daily infection numbers. Based on those values, the prediction bounds for S, I, R, and daily infection numbers can be obtained. However, the SIR model often oversimplifies complex processes of disease.

For many epidemic diseases, if an exposed period (infected but not yet infectious) is long enough or significant, its absence in the model could affect predictions (Brauer, 2008). The SEIR model additionally includes an exposed (E) compartment for the latent period. Efimov and Ushirobira (2021) designed the interval predictors for all compartments of the SEIR model for the epidemic course of COVID-19 by considering uncertainties of the model parameters. In addition, Tang et al. (2020) reviewed features and extensions of multi-compartment models such as SIR and SEIR models in various perspectives, and illustrated their specification, estimation and prediction with illustrative examples for the COVID-19 pandemic.

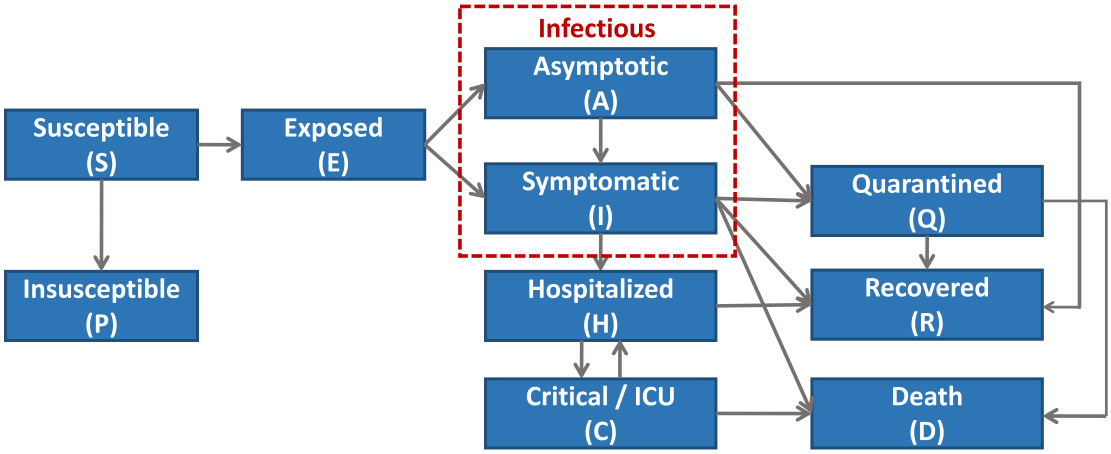

Figure 3: The epidemic compartment models for COVID-19.

As illustrated in Figure 3, some studies have extended compartment models to track transmission mechanism of interest and improve the predictability of the model by illustrating a broad category of classes in multiple directions. For example, hospital beds and Intensive Care Units (ICU) capacity is essential, especially for mortality or recovery. Singh et al. (2020) proposed the SEAIHCRD model with the addition of asymptomatic infectious (A), critical (C), death (D), hospitalized (H), and recovered (R) compartments. They focused on the basic reproduction number, the number of required hospital and ICU beds, and the case fatality rate for COVID-19 in Brazil, India, Mexico, Russia, South Africa, and the US. Another compartment model forecasting the future utilization of hospital and ICU beds is the SUEIHCDR (Neto et al., 2020) by adding critical (C), dead (D), hospitalized (H), recovered (R), and unsusceptible (U). They adopted the fourth-order Runge-Kutta numerical method when solving a system of ordinary differential equations. Besides, a strict quarantine policy for infected people is the distinct feature of the epidemic model for highly contagious diseases. To consider the quarantine period, Peng et al. (2020) and Godio et al. (2020) considered the SPEIQRD model, which additionally includes death cases (D), insusceptible (P), quarantined (Q), and recovered (R). They implemented the simulated annealing algorithm and the particle swarm optimization algorithm for model parameters, respectively, and simulate the progression of COVID-19.

3.2 Agent-based Methods

Agent-based models also referred to as individual-based models, are powerful simulation modeling techniques with interacting autonomous agents. Agents are artificial decision-making entities and are programmed to interact with other agents and the environment in specific ways. In epidemic models, agents can stay either at the state of susceptibility, the state of infected, or any pre-defined states. Then, these agents interact with each other based on a specific environment, which is usually defined through a social contact network. Agents can transfer from a state to another. A typical agent-based model consists of three key elements: (i) a population with the demographic characteristics of the studied region, (ii) a social contact network among the agents (individuals) in the population, and (iii) a disease model that translates the edge weights in the social contact network into infection probability (Hoertel et al., 2020). Agent-based models are helpful when data or the underlying dynamics are well suited for the network structure.

By simulating a set of simple agents' interactions, they can generate complex global patterns (behaviors) and visualize main properties from a global perspective.

This approach has been applied to COVID-19 data in different regions. A good example comes from a team from Imperial College London (Ghandi et al., 2020) who applied agent-based modeling to UK data. The authors predicted the spread of COVID-19 after including many non-pharmaceutical interventions (NPI). Hoertel et al. (2020) considered a stochastic agent-based model to predict the COVID-19 incidence cases in France. A team of researchers from various universities proposed a Global Epidemic and Mobility Model (GLEAM, https://covid19.gleamproject.org/) which employed an agent-based model to study the spatiotemporal dynamics of COVID-19 in the continental US.

4 Phenomenological Models

Phenomenological models, also known as statistical models, attempt to characterize and forecast the observed effects of epidemics without incorporating biological mechanisms and postulating conjectures that explain the observed phenomena. Phenomenological models can be simple; for example, in time series analysis, one often uses the most recent observation as a forecast or other information that may also be relevant such as the changes of control policy, effects of holidays, demographics, socioeconomic status, and other local features. Phenomenological models can also be highly complex, such as the neural network in machine learning and nonlinear systems of simultaneous equations.

4.1 Time Series Analysis

In time series analysis, we focus on using the past observations of a random variable to capture the underlying patterns and predict the future values. There are two popular methods for forecasting in time series analysis: exponential smoothing (ETS) and autoregressive integrated moving average (ARIMA) models. Given univariate time series data, the ETS forecasting method describes the data with a systematic trend or seasonal component. Meanwhile, the ARIMA method focuses on describing the autocorrelations.

4.1.1 Exponential Smoothing (ETS)

The ETS was proposed by Holt (1957), Brown (1959), and Winters (1960) in the late 1950s. There are several variations for the ETS method. In the simple ETS, exponential functions are used to assign exponentially decreasing weights as the observations get older. Therefore, the forecasts produced by ETS methods can be considered as weighted averages of past observations. Holt's linear trend method (Holt, 1957) allows the forecasting of data with a trend. Since the constant trend usually makes the long-term forecast either indefinitely increasing or decreasing into the future, this method tends to over-forecast by empirical evidence, especially for longer forecast horizons (Hyndman and Athanasopoulos, 2018). To handle this issue, Gardner and Mckenzie (1985) proposed a "damped" method that can "dampen" the trend to a flat line in the future. To analyze the time series with seasonality, Holt (1957) and Winters (1960) proposed the Holt-Winters' seasonal method. For time series with roughly constant seasonal variations, the additive Holt-Winters' seasonal method is preferred. Meanwhile, if the seasonal variations of time series change proportional to their level, the multiplicative Holt-Winters' seasonal method is often more valuable.

In general, the ETS method can cope with trends and seasonal variations simultaneously, and thus it is a simple yet powerful tool for time series forecasting. As discussed in Hyndman and Athanasopoulos (2018), when there is no clear trend or seasonal pattern, the simple ETS method performs well for forecasting since it can generate reliable forecasts quickly. When there is a trend in the time series, the "damped" method is one of the most popular time-series forecasting methods. When there is a seasonality pattern, Holt-Winters' seasonal method is usually preferred. When implementing the ETS method, there are several practical issues, including (i) the selection of the initial value, (ii) the sensitivity to outliers, (iii) the selection of the smoothing parameters, and (iv) the normalization of seasonal indices.

4.1.2 ARIMA Models

Another approach for time series forecasting is the ARIMA model. It combines the autoregression (AR) and moving average (MA) models. In an ARIMA model, linear correlations among the time-series are assumed, then the ARIMA model exploits these linear dependencies to explore the local patterns and denoise the data.

The ARIMA approach has three clear benefits. First of all, the interpretability level of the ARIMA model is very high. Therefore, researchers can better understand the relationship between the past and the current situations and explore the influence of some exogenous variables. Secondly, the ARIMA model has an automated way to maximize prediction accuracy, which can perform model selection efficiently. Thirdly, the ARIMA models have a high accommodative ability due to the simplicity of model updates based on recent events. However, one drawback of the ARIMA models is that they cannot deal with nonlinear patterns or relationships.

As discussed in Hyndman and Athanasopoulos (2018), the ETS model describes how unobserved components of the data (error, trend, and seasonality) change over time, while ARIMA focuses on the autocorrelations in the data. Furthermore, the additive ETS models are all special cases of ARIMA models, while the non-additive ETS models do not have any equivalent ARIMA counterparts. On the other hand, many ARIMA models have no ETS counterparts. Thirdly, all ETS models are non-stationary, but some ARIMA models are stationary. When modeling and forecasting a time series, such as the number of the confirmed cases, we can choose the best time series method based on validation methods like out-of-sample validation or information criteria like AIC and BIC if the candidate models are within the same class of models. It is worth noting that ETS and ARIMA models are in different model classes, so we cannot use the information criteria to compare them since the likelihood is computed in different ways.

Notice that the policy changes could significantly affect the spread of COVID-19. Intervention analysis is a helpful tool in the modeling procedures for incorporating the effects of those exogenous forces in time series analysis. By assuming that the ARIMA structure stays the same both before and after the intervention, the intervention analysis studies the mean of a time series. Typically, there are four types of mean changes, including (i) a permanent shift, (ii) an instant change then back to the mean level, (iii) gradually change to a new mean level, and (iv) an initial change followed by gradually return to the original mean level.

4.2 Machine Learning

Machine learning methods are attractive in the COVID-19 prediction with their great flexibility to capture disease spread patterns. There are two major categories of machine learning methods in the prediction of COVID-19. The first kind of methods are based on epidemic models and

trained by machine learning algorithms. Specifically, Zou et al. (2020) propose a variant of the SEIR model accounting for the untested/unreported cases of COVID-19, and the model is estimated by the standard gradient-based optimizer. Arik et al. (2020) integrate machine learning into compartmental disease modeling, which uses learning mechanisms, such as masked supervision from partial observations and Partial teacher forcing, to improve model estimation with limited training data.

The second kind of method considers COVID-19 reported cases as time series data and applies machine learning methods that are well-suited to it. Below, we describe some popular machine learning methods for forecasting epidemics. The long-short term memory (LSTM), proposed by Hochreiter and Schmidhuber (1997), is a recurrent neural network (RNN) architecture used in deep learning. Like standard RNN, LSTM connects nodes from a directed graph along a temporal sequence with wide applications in time series predictions. In Arora et al. (2020), LSTM is used to predict the number of COVID-19 reported cases for the state-level data in India. Devaraj et al. (2021) found that the LSTM model outperformed the ARIMA for predicting both short-term and medium-term infected cases.

Papastefanopoulos et al. (2020) considered DeepAR (Salinas et al., 2020) and N-BEATS (Oreshkin et al., 2019) in COVID-19 prediction. DeepAR is a forecasting method based on autoregressive RNN for probabilistic forecasting. It combines the likelihood methods and deep learning methods and applies them to the forecasting problem. There are two advantages in DeepAR. First, it produces probabilistic predictions by generating Monte Carlo samples that can be utilized to calculate consistent quantile estimates and address the uncertainty of the predictions. Second, as a deep-learning-based method, it can capture complex patterns such as seasonality and holiday effect over time. N-BEATS is a neural basis expansion analysis for interpretable time series forecasting. It is a predictive model with a deep neural architecture consisting of forwarding and backwarding residual links and a deep stack of fully connected layers. The method has several advantages, such as good interpretability, easy application to a wide array of target domains, and fast training.

In summary, machine learning methods provide four advantages. First of all, deep learning-based models are data-driven methods. They can cope with large-scale and noisy data, and discover many heterogeneous signals without laborious feature engineering. Secondly, machine learning methods usually learn a global model from related time series, whereas classical time series methods, such as ARIMA or ETS, fit a single model to each time series. Thirdly, in the COVID-19 prediction, researchers utilize ensemble methods that combine a set of candidate models (Ray et al., 2020). Many of the research groups are performing mechanistic modeling from compartment models. Deep learning-based techniques bring a unique perspective to how to detect signals from data with minimal assumptions. Fourthly, deep learning-based models provide excellent short-term forecasting, which is useful for guiding intervention and allocating resources (Rodriguez et al., 2020).

However, fitting machine learning models usually involves tuning a large number of unknown hyperparameters, such as the number of hidden layers in the deep learning-based method. Selecting a proper hyperparameter is computationally intensive and requires a fairly large amount of data. In the COVID-19 prediction, especially at the beginning of this pandemic, we have limited data sources. One needs to carefully check model complexity and avoid the over-fitting problem.

Also, in the spread of COVID-19, the transmission pattern depends on many factors such as different control measures and varies with location and time. Predictive models need to tune their (hyper)parameters over time to avoid issues such as under-fitting and over-fitting. Besides, when

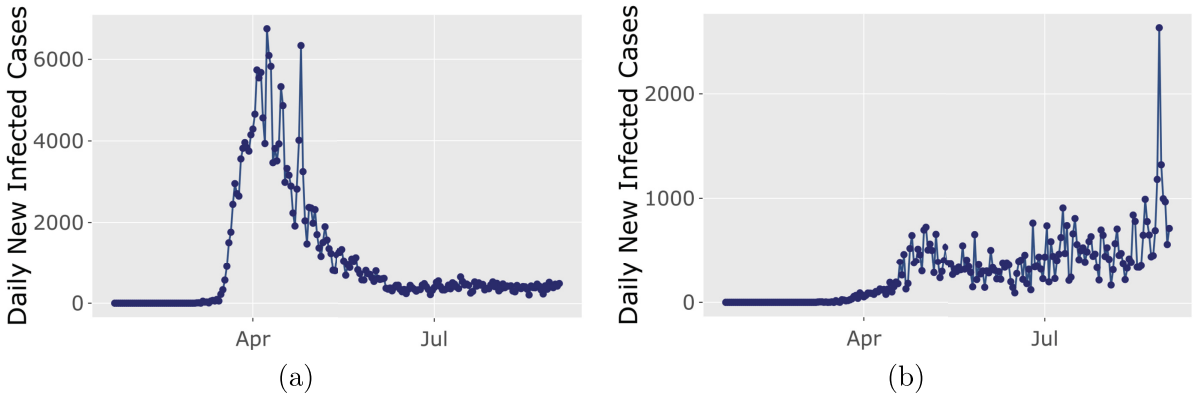

Figure 4: Daily new infected cases in the states of (a) New York and (b) Iowa from January 23, 2020 to September 1, 2020 (Source: https://github.com/CSSEGISandData/COVID-19. Retrieved on December 12, 2020).

we apply these models to different regions, it is important to check changes in models' behavior. For example, the model trained in New York may not be applied to Iowa. Figure 4(a)–(b) present time series plots of daily new cases in the states of New York and Iowa from January 23, 2020, to September 1, 2020, indicating different patterns of disease spread. One can tell there is one peak of daily new cases in the state of New York in April, 2020. In Iowa, there is one major peak in August 2020, with several small peaks in May and July 2020 (Santosh, 2020).

Machine learning methods are also closely connected with many other methods. For example, the simplest networks without any hidden layers are equivalent to linear regressions. A feed-forward network with p lagged inputs and no node in the hidden layer is similar to an ARIMA$(p, 0, 0)$ model but without restrictions on the parameters to ensure stationarity. Actually, the time series techniques are specifically designed for various types of time series data, while the many machine learning methods, such as the RNNs, are designed to handle sequence data in general. Typically, the ARIMA models are parametric, and they are fitted to and used to predict an individual time series. The RNNs are non-parametric models. With sufficient data, one can train a model and use the well-trained model for future prediction.

With time series techniques, one can typically build up a model with acceptable accuracy for the problem in minutes. However, as mentioned above, machine learning techniques are tedious and a much bigger hassle overall. A benefit of using machine learning models is that the well-tuned (hyper)parameters can help obtain optimal performance on most forecasting tasks. However, the settings of these advanced architectures are not straightforward and take much experimentation to get right.

4.3 Regression Methods

In epidemic modeling, regression methods are also popularly applied to estimate future prevalence or targets of interest.

Regression analysis can be used to evaluate the effect of nonpharmaceutical interventions (NPIs). For epidemic forecasting, it is often necessary to model interventions when using regression methods. Some types of regression methods, such as the segmented regression analysis, are powerful statistical tools for estimating intervention effects in interrupted time series stud-

ies. They also allow us to assess how much an intervention changed the dynamics of the disease spread, immediately and over time, instantly or with delay, transiently or long-term. The change in the level and trend from pre- to post-intervention can be estimated and used to test hypotheses about the intervention. We may also include other useful predictor variables that may have affected the transmission and fatality of the disease, such as local characteristics of the region; see Cao et al. (2021).

Regression analysis can also help investigate whether factors other than the NPI could explain the spread pattern change. For example, when analyzing the confirmed cases and deaths of COVID-19, Wang et al. (2020) considers demographics, socioeconomic status, mobility, and other county-level features in the regression models in addition to the control policies.

Altieri et al. (2021) considered various regression models for modeling the death count. The death count is assumed to follow a linear or exponential relationship with time or the current death count. Moreover, regression models could also make different assumptions on whether different counties share the aforementioned relationship or not, leading to the so-called "separate-county" models and "shared-county" models. In addition, regression models can also embed demographic features as explanatory variables.

Regression models are flexible to include useful predictors, such as the day-of-week (DOW) effect and effects of holidays. For example, the Los Alamos National Laboratory (LANL) has proposed a regression method called COVID-19 Forecasts using Fast Evaluations and Estimation (COFFEE), which includes the days of the week as dummy variables in the regression; see (Castro et al., 2020).

4.4 Model Comparisons and Connections

As illustrated in Figure 5, regression methods, time series analysis, and machine learning represent different cultures in statistics. Regression methods and time series analysis often assume that the data are generated by a given stochastic data model. On the other hand, algorithmic modeling approaches such as machine learning treat the data mechanism as unknown. General machine learning methods focus on approximating the non-linear underlying functions in a highly flexible non-parametric fashion to describe the variations in the algorithmic models.

Time series analysis accounts for the internal structure (such as autocorrelation, trend, or seasonality) of the data observed over time and extrapolates the patterns into the future. Compared to the time series approaches, the regression models allow for the inclusion of a lot of relevant information from predictor (external) variables. Time series models are very useful models when the data are accurate and correlated, the future is similar to the past, and directly modeling underlying processes is difficult. However, if we have enough information about the underlying transmission mechanism to simulate the actions and interactions of autonomous agents, physical or agent-based models usually work better than time series models.

Machine learning methods have flexible structures with mild model assumptions and can be easily generalized to many problems. They can automatically discover useful data patterns without specific feature engineering. However, the success of the machine learning method relies on large data sets. In the COVID-19 prediction, finding large and reliable data sets are vital for machine learning methods to generate accurate predictions.

In summary, when data is limited, but the underlying mechanism is well-studied or explicit, a deductive mechanistic model is more appropriate than other models. On the other hand, if there is a lack of knowledge on how a mechanism works, but a reasonable amount of data, it is appropriate to use an inductive approach with the many statistical or machine learning methods

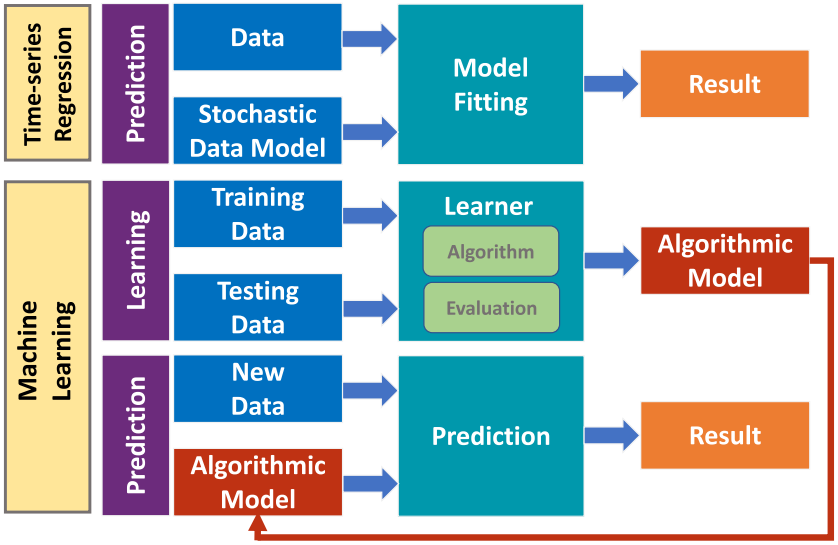

Figure 5: Two statistical cultures.

we have at our disposal. As for COVID-19 projection, researchers also use a combination of mechanistic models and phenomenological models, make updates, and improve predictions; see the descriptions in Section 5.

5 Hybrid Models and Ensemble Methods

Hybrid modeling often refers to the approaches where part of a model is formulated on the basis of mechanistic principles and part of the model has to be inferred from data because of a lack of understanding of the mechanistic details.

5.1 Integration of Different Types of Models

In the past two decades, statistical and machine learning-inspired time series methods have been successfully implemented in studying the spread dynamics of infectious diseases such as seasonal influenza. In COVID-19 studies, we have seen a growing number of hybrid methods combining characteristics of mechanistic models and phenomenological models; see two examples below.

The Institute for Health Metrics and Evaluation (IHME) proposed using a hybrid model for COVID-19 forecasting (IHME COVID-19 Forecasting Team, 2021). This model incorporates elements of curve-fitting in statistical models and compartment models. Due to various biases related to reported deaths, cumulative deaths are smoothed. Using estimated infection fatality ratio, death-by-age patterns, and infection to death duration, past/current daily new infections are then computed according to the estimates of past/current daily deaths. Based on those corrected infections and deaths, the IHME uses the SEIR model but with an additional pre-symptomatic compartment from infectious compartments for forecasting future infections and deaths. Time variation in the model is captured by allowing the time-varying transmission intensity (TI), which is associated with various time-varying covariates and time-invariant covariates. To determine the strength of the association between the time-varying TI and the covariates, a log-linear mixed-effects regression is performed across all locations. Estimations of future TI

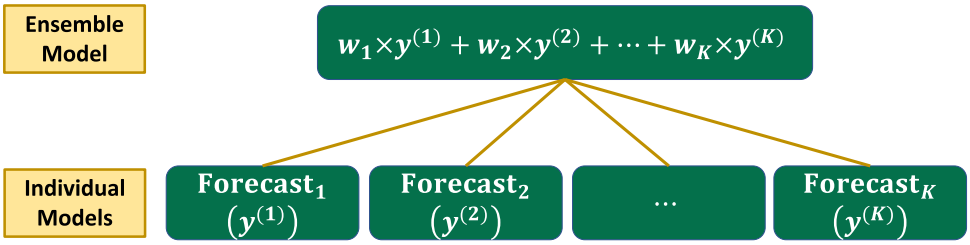

Figure 6: An illustration of ensemble method.

from the fitted regression are then adjusted for the final future TI based on the average fit over a window of time in the recent past. The final TI is used to predict future transmission through the corrected SEIR model for simplifications of model assumptions, such as the potential for importation and well-mixed population.

Borrowing the strength from both the compartment models and regression techniques, Wang et al. (2020) developed a novel hybrid spatiotemporal epidemic model (STEM). Traditional compartment models assume that the transition rate is proportional to the number of infectious individuals, which is commonly referred to as bilinear incidence rates. To relax this strong assumption, they considered the nonlinear incidence rates model by Liu et al. (1987). Note that the data observed are heterogeneous and time-varying. Thus, a "global" model may not be able to capture the spatial and temporal nonstationarity. Wang et al. (2020) incorporated more flexible regression techniques to comprise many local features and explore the spatially varying structure.

5.2 Ensemble Methods

Another hybrid method widely used in forecasting is the ensemble method, which combines multiple forecasting algorithms to improve predictive performance. The basic idea of ensemble methods is that when there is much uncertainty in finding the best model, combining different algorithms could improve prediction accuracy by reducing the instability of the forecast. A simple hybrid method employs a linear combination of the individual forecasts from various individual contributing models to generate the combined/a better forecast. Below, we introduce two examples of ensemble forecasting for COVID-19.

Based on the idea of the ensemble method, Altieri et al. (2021) proposed an ensemble forecast. The proposed combined linear and exponential predictors (CLEP) used various combinations of the five predictors discussed in Section 4.3, so we can consider it as a weighted average of the individual predictor.

Another example is the Reich Lab COVID-19 ensemble forecast, and one of the tasks is to predict the number of Americans who will die due to COVID-19. The Reich Lab (Ray et al., 2020; Cramer et al., 2021) collects and aggregates probabilistic forecasts from multiple models, and generates both point and interval predictions, referred to as "ensemble forecasts". The ensemble forecasts were first generated by averaging all eligible forecasts with equal weights, and later by computing the median across all eligible forecasts at each quantile level.

As demonstrated in Cramer et al. (2021), while the evaluation shows a highly variable performance of all individual models, the ensemble forecast achieves the best probabilistic accuracy in terms of the weighted interval score over the evaluation period. The rankings of models are highly variable, and it is very difficult to improve the median ensemble approach by using

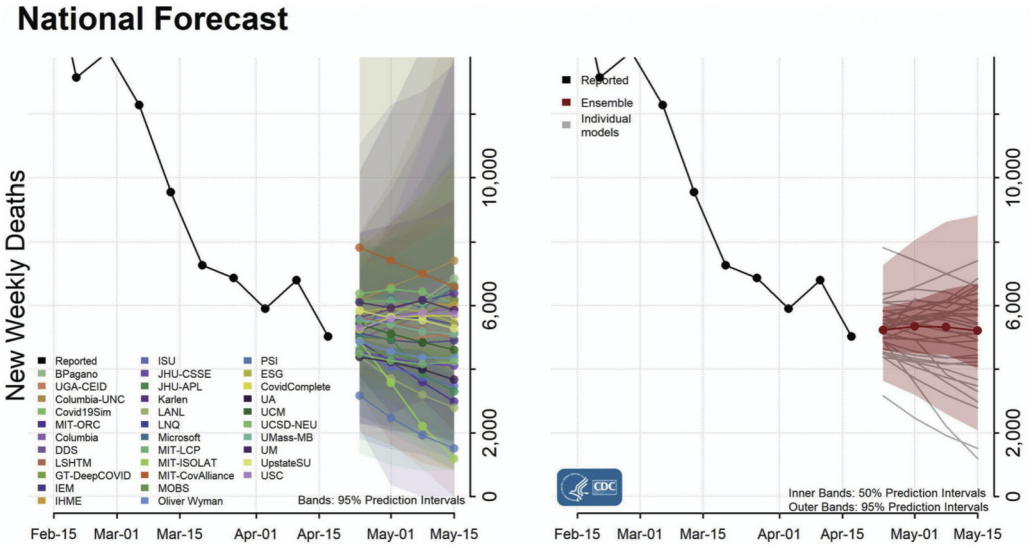

Figure 7: National COVID-19 forecasts (Source: https://www.cdc.gov/coronavirus/2019-ncov/covid-data/forecasting-us.html. Retrieved on April 21, 2021).

"trained" weights for component models. Last but not least, as the horizon becomes longer, the accuracy and calibration of the forecasts degrade as expected.

With the assistance of the COVID-19 Forecast Hub, the US Centers for Disease Control and Prevention (CDC) provides the forecast of deaths related to COVID-19 in the US; see Figure 7. There are more than 30 models being considered eligible to be included to produce an ensemble forecast. Each color represents a specific model which provides the point forecast (the connected dotted lines) and the interval forecast (the shaded bands).

6 COVID-19 Forecasting: A Data Science Procedure

In the information explosion era, the challenge of epidemic forecasting is not the lack of data but rather how to identify the most relevant high-quality data for meaningful results and how to integrate the data from various sources that might not have a standard format or be interoperable. Figure 8 summarizes the procedure of how the COVID-19 data are collected, curated, visualized, and fed to predictive models, which facilitates effective communication between data sources, scientists, and decision-makers, and provides important guidelines for the public. However, collecting clean, reliable, and timely data is challenging for most epidemic researchers. Infectious disease learning requires multiple iterations between data manipulation, visualization, and modeling. Communication is another critical part of data science procedure. It is imperative to have cross-collaboration with epidemiologists when making inferences, projections, and drawing conclusions.

As one of the individual teams for CDC's ensemble forecast, we have found it useful to make a variety of data cleaning and effects adjustments to better predict the potential spread of the COVID-19 pandemic. Both data wrangling and effects adjustments empower a better estimation of the underlying trend.

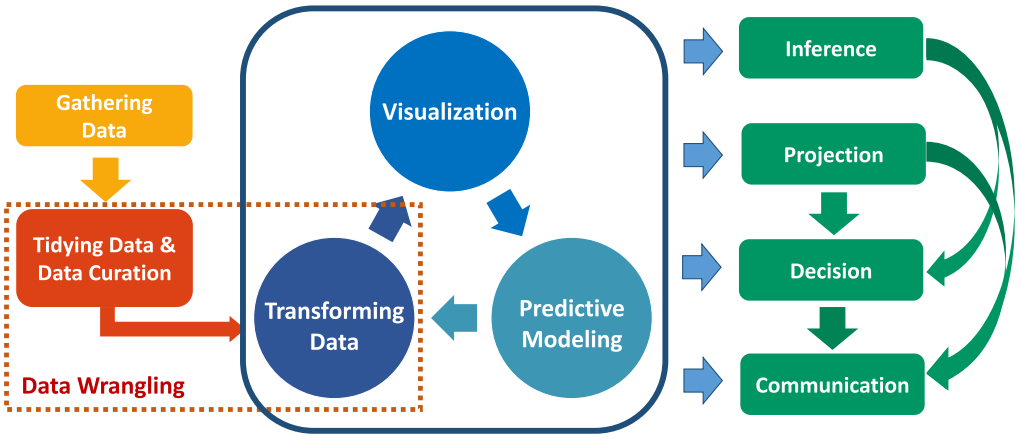

Figure 8: A data science procedure for COVID-19.

6.1 Data Wrangling

Several data wrangling steps are essential components in the first step of the COVID-19 data analysis due to the data issues mentioned above. For data cleaning, we mainly focus on the following four issues: (a) missing values, (b) anomalies, (c) abrupt level changes in the time series history, and (d) potential transformations which might help improve forecast performance. These data-cleaning strategies help us to better conform to the assumptions of forecasting models in the later analysis.

Anomalies typically present as a dramatic spike or drop, which cannot be explained by a trend, seasonality effect, or holiday impact. The time series with these anomalies usually returns to its previous level at trend after a short period. The first task when handling the time series is the detection of potential anomalies. There are several valuable methods in the time series framework, and most of them work on the remainders after removing the seasonal and trend components by some conventional methods such as the ETS, ARIMA, and machine learning. Once the remainders are obtained and have the desired characteristics, one can consider detecting the anomalies based on the interquartile range (IQR) and the generalized extreme studentized deviate (GESD) test (Rosner, 1983). The IQR method is usually faster compared with the GESD method; however, it may not be as accurate as the GESD method due to less resistance to the high leverage points of the GESD method.

Traditionally, these anomalies may occur due to some measurement error, and thus they can be excluded from the dataset for analysis. However, the anomalies in COVID-19 epidemic data typically arise from some specific reasons, such as the release of the result of a large batch of tests (Wang et al., 2020). For example, as illustrated in Figure 2(a), there is a spike in the daily new infected count for Grime County, Texas, on May 31, 2020; we later found that at least 80% of the active cases were from the Texas Department of Criminal Justice. Without adjustments, these data points can disrupt the seasonality, holiday, and trend estimation and thus adversely impact forecasting performance. Due to the specific reason for these anomalies, we cannot simply exclude these anomalies in an analysis. Wang et al. (2020) proposed an algorithm that can automate a procedure for "correcting" the history. In the first step, a reasonable estimate of the anomaly point and its corresponding residual are generated by any model discussed in Sections 3 – 5. Next, the causes and the problematic period are investigated, and the residual is distributed

proportionally to the value of the historical time series within the problematic period.

In contrast to anomalies, level changes represent a more permanent change of the level of time series. They may result from the change of reporting standard, such as some states starting to report probable cases during the pandemic. For example, the spike in Figure 2(b) is because New Jersey started reporting probable deaths and added 1854 probable deaths, which could date back to the earlier outbreak. We can adjust for level changes in a longer effective time series to tackle this issue, and often this strategy can make trend estimation easier.

Transformation is a prevalent tool in data analysis. It can help not only with meeting model assumptions but also with eventual forecast accuracy. The transformation techniques are often built in modern data analysis toolbox, such as the Box-Cox transformation in time series analysis.

6.2 Effect Adjustments

Effect adjustments are also useful in epidemic model fitting and prediction. While the cleaning adjustments are usually conducted in the data's pre-processing, the effects adjustments can be made in both the pre-processing and the modeling period. Several typical effects we noticed from COVID-19 are holiday effects, seasonality effects, and day-of-week effects. For example, in some time series models, we usually try to quantify the effect first, then remove the effect from the time series, make a forecast of the time series without the effect, and finally reapply the effect to the forecast.

6.3 Model Validation

As discussed in Held et al. (2020), in epidemic forecasting, there are several quantities which are of great interest, including (i) timing of the peak of incidence (Ray et al., 2017), (ii) onset timing (Pei et al., 2018), (iii) cumulative incidence (Lega and Brown, 2016), (iv) weekly new incidence (Paul and Held, 2011; Reich et al., 2016), (v) size and duration of the epidemic (Farrington et al., 2003), and (vi) curve of the epidemic (Jiang et al., 2009).

For a continuous or discrete time series, a point forecast is usually preferred, such as the number of daily confirmed cases. To evaluate the quality of these point forecasts, several measures have been proposed, see Gneiting (2011) for a comprehensive discussion. Some popular evaluation scoring functions that can be considered are (i) the absolute error (AE), $S(\hat{y}, y) = |\hat{y} - y|$, (ii) the squared error (SE), $S(\hat{y}, y) = (\hat{y} - y)^2$, (iii) the absolute percentage error (APE), $S(\hat{y}, y) = |\hat{y} - y|/y$, and (iv) the relative error (RE), $S(\hat{y}, y) = |\hat{y} - y|/\hat{y}$. In the infectious disease analysis, there are several other measures, but most of them are defined based on the four abovementioned scoring functions, for example the MAE used in Ray et al. (2020).

Probabilistic forecasts assign a probability to each of the different outcomes. Held et al. (2017) and Funk et al. (2019) have taken up the probabilistic forecasts in the infectious disease. To evaluate the probabilistic forecasts, we need to measure the "distance" between true observation y and the distribution function of the forecast, F. For binary data, there are two widely used scoring functions: (i) the logarithmic score (LS) function and (ii) the Brier score (BS) function. For count data, the logarithmic score can be defined as $\mathrm{LS}(F, y) = \log(\pi_j)$, where π_j represents the predicted probability of $y = j$. The ranked probability score can be considered as an extension of the BS, and it can be used for count data. The Dawid-Sebastiani Score (DDS) is another score that is usually preferred when the first two predictive moments are easy to derive compared to the entire distribution. Meanwhile, as pointed out by Held et al. (2020), depending

on whether the forecast distributions are analytically known or generated based on simulations, different scores might be computed. In practice, evaluation of several scores is usually recommended since it can demonstrate a more robust comparison of predictive performance.

Predictive methods consider the generalization performance and choose the model that best predicts future values. However, when fitting the model, especially for long-term forecasting, future values are not available. These methods assume that the behavior of observations from the future process is similar to that from the current or past process.

7 Discussion

In this article, we discuss the data science aspects of the COVID-19 projection. We selectively overview some features, issues, and challenges brought by COVID-19 data, and provide a sample of approaches for COVID-19 projection, focusing on mechanistic models, phenomenological models and hybrid models. Besides the challenges of data, there are several other challenges that are worth equal attention.

7.1 The Limitations of Models

Although epidemic models are very useful, they are constrained by what we know and what we assume.

Mechanistic models with some important inputs can easily simulate the dynamic spread of the disease, forecast behavioral changes in the population, and provide a simple but clear picture of the disease transmission. Due to the flexibility and ease of extensibility of mechanistic models, they are useful in investigating the effect of various factors, such as protection, control measures, quarantine, and hospitalization. However, those mechanistic models may contain various assumptions about the dynamics of disease transmission and a large number of parameters, making their results are sensitive to a small change of model assumptions and parameters (Singh et al., 2020; Wang et al., 2020). Hence, they have a high level of uncertainty, which is difficult to predict. In addition, constructing prediction intervals is challenging since mechanistic models are based on the deterministic process of disease. Proper data for the corresponding parts in the model is often not accurately available in a well-structured format. Mechanistic models also have difficulty in predicting the possibility of the second or third wave (Sun et al., 2020). The change of mechanical characteristics may enormously degrade the accuracy and predictability of the models.

Phenomenological models are useful to extract information from data. They focus on making inferences, such as prediction intervals and pursuing accurate predictions of target quantities. However, if the training data is not well-representative of future values, the phenomenological models usually overgeneralize the training data and lead to poor prediction. The performance of trained models highly depends on feature selection as irrelevant features may deteriorate the explainability of good features. There are usually numerous intermediate-tuning parameters involved in the model that are difficult to interpret. Thus, after we find a model with good predictability using many hidden parameters, it is challenging to distinguish the effects of main factors in phenomenological models. Since phenomenological models do not take into account how transmission progresses, they have difficulty in effectively understanding the whole process of disease in the future. Besides, phenomenological models are not well-suited for long-term forecasting as they rely on current or previous observations, and the future dynamics of disease transmission could change.

7.2 Understanding and Embracing Uncertainty in the Forecast

There are three fundamental sources of uncertainty in the forecast of disease.

First, the biological features of viruses are underlying factors in their transmission. The mutations of viruses make a system of their transmission variable; for example, newly discovered variants of COVID 19, such as B.1.1.7, B.1.351 and P.1, appeared in the UK, South Africa and Brazil, respectively, may result in cases and deaths of COVID-19 with different transmission rates and severity. In addition, the different pathogens have various incubation periods, which also vary with age, sex and location over time.

Second, the influence of forecasting on disease spread on human behavior generates further complexity, ambiguity, and uncertainty. For example, if forecasts of infections or deaths get serious, the government may announce strong restrictions on social contact and mandate that people wear masks in public. People also tend to cancel travel plans and spend more time at home as they feel anxious and frightened. These changes in human behavior can further affect the future progression of the epidemic by reducing the reproduction number and the number of infections (Zhang et al., 2020; Jarvis et al., 2020).

Last but not least, there are several data incorporated in epidemic models. A variety of data may be inaccurately measured or sampled. Missing and incomplete data often occur since individual counties and states provide different sources of information. Unassigned or under-reported cases lead to the primary source of bias. Those uncertainties from data may affect the forecast itself and also make its validation challenging.

7.3 Validation of Models and Complex Ground Truth

Model validation is a process to check whether a model, combined with its assumptions, provides an accurate representation given a sufficient accuracy level (Carson, 2002; Sargent, 2011). Unfortunately, all of those aforementioned data problems and uncertainties make model validation a "wicked problem". As stated in Rittel and Webber (1973), "a wicked problem is a problem that is difficult or impossible to solve because of incomplete, contradictory, and changing requirements that are often difficult to recognize."

Ideally, researchers assume to have full access to the desired measurements of disease prevalence from its entire history when forecasting future measurements. The correct measurements have significant implications for proper forecast, since the errors in the data may result in bias in the estimates of model parameters and a carry-over effect on the prediction. Under a pandemic such as COVID-19, correct data are crucial to enable decision-makers to quickly assess and respond to a situation. However, the exact data is not immediately accessible in practice. Moreover, accurate measurements may take weeks or even years to be completed.

Therefore, it would be beneficial to have a sequence of provisional estimates of each complete measurement, with later versions available with more overall accuracy. One popular strategy is the "modeling surveillance data revisions". Various methods can be used in this data revision process, such as the nonparametric one-ahead backcasting and forecasting methods.

7.4 Improving Data Strategy and Data Science Practice

There is much space for improving the existing data strategy and data science practice.

First of all, there are many data incorporated in the various models, and some of those data may be inaccurately measured. Poor data input might lead to unsatisfactory data-based forecasting. Investments should be made in the collection, cleaning, curation, and validation of

data. Many modeling teams, like us, have to spend much time implementing data curation before the prediction. It is critical to have a data validation process in place to guide the modelers. A uniform data reporting system can improve the consistency in data reporting protocols, and thus it benefits the data collection and even future disease analysis.

Secondly, forecasting the patterns of transmission is based on various assumptions. Models' assumptions and limitations must be appraised openly and honestly, even though it can be challenging to validate some assumptions within a short time period.

Thirdly, as discussed in Ioannidis et al. (2020), many methods used by the policymakers were not disclosed and never formally peer-reviewed. It is crucial that all of the data and code associated with the methods are open source to enable rapid integration across multiple research groups and government agencies in the future.

Last but not least, additional funding is needed. It would be great to see more funding opportunities for data science-related research of COVID-19 and other infectious diseases. Currently, most of the funding is on the medical side, but data science is also essential in aiding decision-making and facilitating scientific research in an effective and timely manner and therefore should be given attention.

Acknowledgement

This work was supported in part by National Science Foundation awards DMS-1916204, CCF-1934884, and the Laurence H. Baker Center for Bioinformatics and Biological Statistics. We would like to thank Dr. Glen Wright Colopy for his thoughtful comments and efforts towards improving our manuscript.

References

Altieri N, Barter RL, Duncan J, Dwivedi R, Kumbier K, Li X, et al. (2021). Curating a COVID-19 data repository and forecasting county-level death counts in the United States. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.1d4e0dae

Arik SO, Li CL, Yoon J, Sinha R, Epshteyn A, Le LT, et al. (2020). Interpretable sequence learning for COVID-19 forecasting. arXiv preprint: https://arxiv.org/abs/2008.00646.

Arora P, Kumar H, Panigrahi BK (2020). Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India. *Chaos, Solitons & Fractals*, 139: 110017. https://doi.org/10.1016/j.chaos.2020.110017

Atlantic (2021). The COVID tracking project. https://covidtracking.com

Brauer F (2008). Compartmental models in epidemiology. In: Brauer F, van den Driessche P and Wu J (eds.), *Mathematical Epidemiology*, 19–79. Springer.

Brauer F, Castillo-Chavez C, Feng Z (2019). *Mathematical Models in Epidemiology*, Texts in Applied Mathematics, volume 32. Springer, New York.

Brooks L (2020). Pancasting: forecasting epidemics from provisional data, Ph.D. thesis, Centers for Disease Control and Prevention.

Brown RG (1959). *Statistical Forecasting for Inventory Control*. McGraw-Hill, New York.

Cao W, Chen C, Li M, Nie R, Lu Q, Song D, et al. (2021). Important factors affecting COVID-19 transmission and fatality in metropolises. *Public Health*, 190: e21.

Carson JS (2002). Model verification and validation. In: Yücesan E, Chen C-H, Snowdon JL and Charnes JM (eds.), *Proceedings of the Winter Simulation Conference*, volume 1, 52–58.

Castro L, Fairchild G, Michaud I, Osthus D (2020). COFFEE: COVID-19 forecasts using fast evaluations and estimation. https://covid-19.bsvgateway.org/static/COFFEE-methodology.pdf

Chen LP, Zhang Q, Yi GY, He W (2021). Model-based forecasting for Canadian COVID-19 data. *PLoS One*, 16(1): e0244536.

Council of State and Territorial Epidemiologists (2020). Standardized surveillance case definition and national notification for 2019 novel coronavirus disease (COVID-19). https://cdn.ymaws.com/www.cste.org/resource/resmgr/2020ps/Interim-20-ID-01_covid-19.pdf

Cramer EY, Ray EL, Lopez VK, Bracher J, Brennen A, Rivadeneira AJC, et al. (2021). Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the US. medRxiv preprint: https://www.medrxiv.org/content/10.1101/2021.02.03.21250974v1.

Devaraj J, Madurai Elavarasan R, Pugazhendhi R, Shafiullah GM, Ganesan S, Jeysree AK, et al. (2021). Forecasting of COVID-19 cases using deep learning models: Is it reliable and practically significant? *Results in Physics*, 21: 103817.

Dietz K, Heesterbeek JA (2002). Daniel Bernoulli's epidemiological model revisited. *Mathematical Biosciences*, 180(1–2): 1–21.

Efimov D, Ushirobira R (2021). On an interval prediction of COVID-19 development based on a SEIR epidemic model. *Annual Reviews in Control*, in press. https://doi.org/10.1016/j.arcontrol.2021.01.006

Farrington CP, Kanaan MN, Gay NJ (2003). Branching process models for surveillance of infectious diseases controlled by mass vaccination. *Biostatistics*, 4(2): 279–295.

Funk S, Camacho A, Kucharski AJ, Lowe R, Eggo RM, Edmunds WJ (2019). Assessing the performance of real-time epidemic forecasts: A case study of Ebola in the Western Area region of Sierra Leone, 2014–15. PLoS. *Computational Biology*, 15(2): e1006785.

Gardner L, Ratcliff J, Dong E, Katz A (2021). A need for open public data standards and sharing in light of COVID-19. *The Lancet Infectious Diseases*, 21(4): e80.

Gardner ES Jr, Mckenzie E (1985). Forecasting trends in time series. *Management Science*, 31(10): 1237–1246.

Gneiting T (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494): 746–762.

Godio A, Pace F, Vergnano A (2020). SEIR modeling of the Italian epidemic of SARS-CoV-2 using computational swarm intelligence. *International Journal of Environmental Research and Public Health*, 17(10): 3535.

Held L, Hens N, O'Neill P, Wallinga J (2020). *Handbook of Infectious Disease Data Analysis*. Chapman and Hall/CRC.

Held L, Meyer S, Bracher J (2017). Probabilistic forecasting in infectious disease epidemiology: the 13th Armitage lecture. *Statistics in Medicine*, 36(22): 3443–3460.

Hochreiter S, Schmidhuber J (1997). Long short-term memory. *Neural Computation*, 9(8): 1735–1780.

Hoertel N, Blachier M, Blanco C, Olfson M, Massetti M, Sánchez-Rico M, et al. (2020). A stochastic agent-based model of the SARS-CoV-2 epidemic in France. *Nature Medicine*, 26(9): 1417–1421.

Hoffman H (2021). How day-of-week effects impact COVID-19 data. https://covidtracking.com/analysis-updates/how-day-of-week-effects-impact-covid-19-data.

Holt CC (1957). Forecasting seasonals and trends by exponentially weighted moving averages. Office of Naval Research Memorandum, Carnegie Institute of Technology.

Huppert A, Katriel G (2013). Mathematical modelling and prediction in infectious disease epidemiology. *Clinical Microbiology and Infection*, 19(11): 999–1005.

Hyndman RJ, Athanasopoulos G (2018). *Forecasting: Principles and Practice*. OTexts, Melbourne.

IHME COVID-19 Forecasting Team (2021). Modeling COVID-19 scenarios for the United States. *Nature Medicine*, 27(1): 94.

Ioannidis JPA, Cripps S, Tanner MA (2020). Forecasting for COVID-19 has failed. *International Journal of Forecasting*, in press. https://doi.org/10.1016/j.ijforecast.2020.08.004

Jarvis CI, Van Zandvoort K, Gimma A, Prem K, Klepac P, Rubin GJ, et al. (2020). Quantifying the impact of physical distance measures on the transmission of COVID-19 in the UK. *BMC Medicine*, 18: 1–10.

Jiang X, Wallstrom G, Cooper GF, Wagner MM (2009). Bayesian prediction of an epidemic curve. *Journal of Biomedical Informatics*, 42(1): 90–99.

Johns Hopkins University Center for Systems Science and Engineering (2021). COVID-19 data repository. https://github.com/CSSEGISandData/COVID-19

Kermack WO, McKendrick AG (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 115(772): 700–721.

KRR G, KVR M, SSP PR, Casella F (2020). Non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality. SSRN preprint: https://doi.org/10.2139/ssrn.3560688

Kucirka LM, Lauer SA, Laeyendecker O, Boon D, Lessler J (2020). Variation in false-negative rate of reverse transcriptase polymerase chain reaction–based SARS-CoV-2 tests by time since exposure. *Annals of Internal Medicine*, 173(4): 262–267.

Lega J, Brown HE (2016). Data-driven outbreak forecasting with a simple nonlinear growth model. *Epidemics*, 17: 19–26.

Liu W, Hethcote HW, Levin SA (1987). Dynamical behavior of epidemiological models with nonlinear incidence rates. *Journal of Mathematical Biology*, 25: 359–380.

Neto OP, Reis JC, Brizzi ACB, Zambrano GJ, de Souza JM, Pedroso W, et al. (2020). Compartmentalized mathematical model to predict future number of active cases and deaths of COVID-19. *Research on Biomedical Engineering*, in press. https://doi.org/10.1007/s42600-020-00084-6

New York Times (2021). Coronavirus (COVID-19) data in the United States. https://github.com/nytimes/covid-19-data.

Oreshkin BN, Carpov D, Chapados N, Bengio Y (2019). N-BEATS: neural basis expansion analysis for interpretable time series forecasting. arXiv preprint: https://arxiv.org/abs/1905.10437.

Papastefanopoulos V, Linardatos P, Kotsiantis S (2020). COVID-19: A comparison of time series methods to forecast percentage of active cases per population. *Applied Sciences*, 10(11): 3880.

Paul M, Held L (2011). Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts. *Statistics in Medicine*, 30(10): 1118–1136.

Pei S, Kandula S, Yang W, Shaman J (2018). Forecasting the spatial transmission of influenza in the United States. *Proceedings of the National Academy of Sciences*, 115(11): 2752–2757.

Peng L, Yang W, Zhang D, Zhuge C, Hong L (2020). Epidemic analysis of COVID-19 in China by dynamical modeling. arXiv preprint: https://arxiv.org/abs/2002.06563.

Rahmandad H, Sterman J (2008). Heterogeneity and network structure in the dynamics of diffusion: comparing agent-based and differential equation models. *Management Science*, 54(5): 998–1014.

Ray EL, Sakrejda K, Lauer SA, Johansson MA, Reich NG (2017). Infectious disease prediction with kernel conditional density estimation. *Statistics in Medicine*, 36(30): 4908–4929.

Ray EL, Wattanachit N, Niemi J, Kanji AH, House K, Cramer EY, et al. (2020). Ensemble forecasts of coronavirus disease 2019 (COVID-19) in the U.S. medRxiv preprint: https://www.medrxiv.org/content/10.1101/2020.08.19.20177493v1.

Reich NG, Lauer SA, Sakrejda K, Iamsirithaworn S, Hinjoy S, Suangtho P, et al. (2016). Challenges in real-time prediction of infectious disease: a case study of dengue in Thailand. *PLoS Neglected Tropical Diseases*, 10(6): e0004761.

Rittel HWJ, Webber MM (1973). Dilemmas in a general theory of planning. *Policy Sciences*, 4(2): 155–169.

Rodriguez A, Tabassum A, Cui J, Xie J, Ho J, Agarwal P, et al. (2020). DeepCOVID: An operational deep learning-driven framework for explainable real-time COVID-19 forecasting. medRxiv preprint: https://www.medrxiv.org/content/10.1101/2020.09.28.20203109v2.

Rosner B (1983). Percentage points for a generalized ESD many-outlier procedure. *Technometrics*, 25(2): 165–172.

SAGE Working Group on Measles and Rubella (2019). Feasibility assessment of measles and rubella eradication. https://www.who.int/immunization/sage/meetings/2019/october/Feasibility_Assessment_of_Measles_and_Rubella_Eradication_final.pdf.

Salinas D, Flunkert V, Gasthaus J, Januschowski T (2020). Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36: 1181–1191.

Santosh KC (2020). COVID-19 prediction models and unexploited data. *Journal of Medical Systems*, 44(9): 1–4.

Sargent RG (2011). Verification and validation of simulation models. In: Jain S, Creasey RR, Himmelspach J and White KP (eds.), *Proceedings of the 2011 Winter Simulation Conference (WSC)*, 183–198.

Singh A, Bajpai MK, Gupta SL (2020). A time-dependent mathematical model for COVID-19 transmission dynamics and analysis of critical and hospitalized cases with bed requirements. medRxiv preprint: https://www.medrxiv.org/content/10.1101/2020.10.28.20221721v1.full.

Sun J, Chen X, Zhang Z, Lai S, Zhao B, Liu H, et al. (2020). Forecasting the long-term trend of COVID-19 epidemic using a dynamic model. *Scientific Reports*, 10(1): 1–10.

Tang L, Zhou Y, Wang L, Purkayastha S, Zhang L, He J, et al. (2020). A review of multi-compartment infectious disease models. *International Statistical Review*, 88(2): 462–513.

Tenforde MW, Kim SS, Lindsell CJ, Rose EB, Shapiro NI, Files DC, et al. (2020). Symptom duration and risk factors for delayed return to usual health among outpatients with COVID-19 in a multistate health care systems network. *Morbidity and Mortality Weekly Report*, 69(30): 993–998.

USAFacts (2021). Coronavirus locations: COVID-19 map by county and state. https://usafacts.org/visualizations/coronavirus-covid-19-spread-map

Wang G, Gu Z, Li X, Yu S, Kim M, Wang Y, et al. (2020). Comparing and integrating us COVID-19 data from multiple sources with anomaly detection and repairing. arXiv preprint: https://arxiv.org/abs/2006.01333.

Wang L, Wang G, Li X, Yu S, Kim M, Wang Y, et al. (2021). Modeling and forecasting COVID-19. *Notices of the American Mathematical Society*, 68(4): 585–595.

Wang Q, Xie S, Wang Y, Zeng D (2020). Survival-convolution models for predicting COVID-19 cases and assessing effects of mitigation strategies. *Frontiers in Public Health*, 8: 325.

Winters PR (1960). Forecasting sales by exponentially weighted moving averages. *Management Science*, 6(3): 324–342.

World Health Organization (2020). Public health surveillance for COVID-19: interim guidance, 16 December 2020. https://www.who.int/publications/i/item/who-2019-nCoV-surveillanceguidance-2020.8

Zhang N, Jia W, Lei H, Wang P, Zhao P, Guo Y, et al. (2020). Effects of human behaviour changes during the COVID-19 pandemic on influenza spread in Hong Kong. *Clinical Infectious Diseases*, in press. https://doi.org/10.1093/cid/ciaa1818

Zou D, Wang L, Xu P, Chen J, Zhang W, Gu Q (2020). Epidemic model guided machine learning for COVID-19 forecasts in the United States. medRxiv preprint: https://www.medrxiv.org/content/10.1101/2020.05.24.20111989v1.