Spring 2020

An Examination of COVID-19 Statistical Modeling

Shane Vaughan
ssv11@zips.uakron.edu

Senior Honors Research Project:

An Examination of COVID-19 Statistical Modeling

Shane S. Vaughan

Buchtel College of Arts and Sciences

The University of Akron

Abstract

The 2019 novel coronavirus, also known as COVID-19, is an infectious disease which was first reported in late 2019 and soon spread to become a global pandemic, prompting major action from world governments. Soon after, many institutions began attempts to analyze and predict the spread and severity of the disease via statistical modeling. Some information is not available for public consumption; however, a number of institutions have published the results of their analyses and some have made public repositories of the code used to build the models. This research paper attempts use these and other resources to examine the modeling techniques employed by several such institutions and assess their accuracy and efficacy. In addition, it will explore these techniques and discuss their implementations as well as what assumptions they rely upon and how small parameter changes affect output.

Senior Honors Research Project:

An Examination of COVID-19 Statistical Modeling

In December 2019, in Wuhan, China, a new respiratory disease emerged and was found to have been caused by a novel coronavirus which soon became known as COVID-19. The spread of the highly transmittable virus eventually prompted unprecedented societal upheaval as world governments scrambled to contain the pandemic. Self-quarantining, shelter in place and stay at home orders have left many individuals and families under drastically different circumstances than just a few short months ago. The state we find ourselves in now is highly uncertain and leaves the majority of the population asking the same questions; indeed, there has never been a more important time for investigation with the purpose of finding a solution to such a dire problem. Researchers around the world have been working tirelessly to answer questions related to the pandemic to better inform government entities and the public. Specifically, statisticians, data scientists, epidemiologists and mathematicians are attempting to model as many prescient attributes of the pandemic as possible. The Centers for Disease Control and World Health Organization, as well as government entities and, most prolifically, academic institutions, have employed several different modeling techniques from various branches of statistical analysis to answer pressing questions, with varying degrees of accuracy and success. In this research paper, we would like to present three different models from three different academic institutions, examine the statistical techniques with which they are constructed and discuss the motivation behind each technique, as well as the assumptions they are based on, the data they are describing, the challenges in their implementation and the interpretation of their results.

University of Bern

The first model we want to examine was built by the Institute of Social and Preventative Medicine at the University of Bern in Bern, Switzerland. This model attempts to describe the case fatality ratio (CFR) of COVID-19 based on data from China; it accounts for "underreporting of cases and the time delay to death" (Riou et al, 2020). The model is also age-specific, drawing different conclusions for different age groups. It was built by Julien Riou, Anthony Hauser, Michel J. Counotte and Christian L. Althaus, all from the University of Bern, and finalized on March 3, 2020 (Riou et al, 2020).

Technique

The authors of the model describe the modeling technique they used as "an age-stratified susceptible-exposed-infected-removed (SEIR) compartmental model" (Riou et al, 2020). A SEIR model is closely related to a SIR (susceptible-infected-removed) model; this is a technique that is also seeing widespread use in COVID-19 modeling. The major difference between the two modeling SEIR considers the transition from susceptible to exposed while SIR does not; this makes SEIR models particularly successful at explaining the effect of preventative intervention measures (Azose, 2013). Under ideal circumstances, SEIR models can be incredibly effective at modeling the spread of infectious disease.

Assumptions.

The following assumptions were noted by Riou and the other authors:

- People showing symptoms "had an age-specific case-fatality rate"
- Time from onset to death $\gamma \sim Lognormal(\mu = 20.2, \sigma = 11.6)$
- All deaths recorded
- "all symptomatic cases… 80 years and older were also recorded"

- "risk of transmission… was homogeneous by age"
- "49% of infections lead to symptoms"
- "younger individuals…have milder symptoms that decrease the probability of seeking care and being identified"
- "older individuals have more severe symptoms and are more likely to be identified"
- Surveillance bias (i.e. underreporting the number of cases/deaths)

Parameterization.

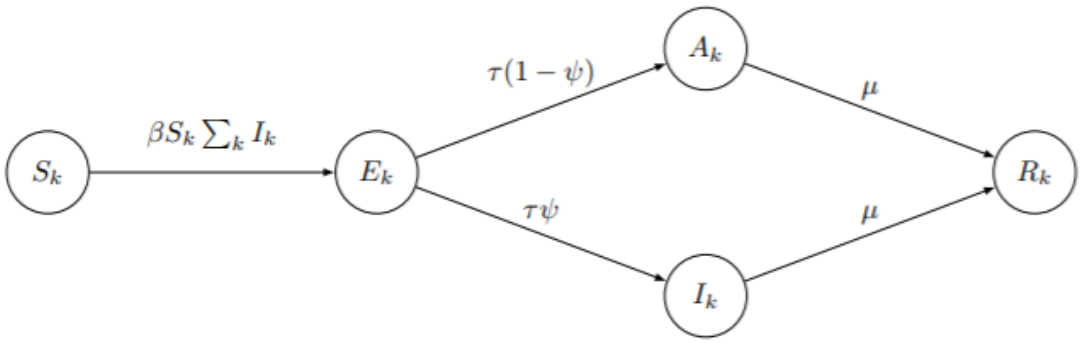

Figure 1: *This diagram from the authors shows the different states of the SEIR model and the model parameters. Interestingly, we see that they consider the susceptible, exposed, infected and removed states as well as a fifth state which they refer to as "asymptomatically infected". Individuals who are asymptomatically infected do not contribute to the spread of the disease.*

The following model parameters were indicated by Riou and the other authors:

- $\beta - transmission\ rate$
- $\tau - incubation\ rate$

- $\psi - probability\ of\ symptomatic\ infection$
- $\mu - removal\ rate$
- $\rho_k - reporting\ rate\ of\ symptomatics\ (by\ age\ group)$
- $\epsilon_k - probability\ of\ death\ among\ symptomatics\ (by\ age\ group)$
- $\gamma - delay\ from\ disease\ onset\ to\ death\ (discretized\ by\ day)$

Data

Riou and the other researchers from Bern had access to daily Chinese records of COVID-19 cases and deaths from January 1 to February 11. It is interesting to note that the researchers assumed that this data was affected by surveillance bias; this means that in building the model they accounted for an underreporting of cases and deaths. At the time this model was built, data from China, where the disease originated, was the best option; it was, however, a prescient decision by the researchers to assume some underreporting had occurred. Additionally, the researchers were able to fit the model to four different response variables: number of confirmed cases, number of deaths, age distribution of confirmed cases and age distribution of deaths (Riou et al, 2020).

Motivation

The motivation behind using SEIR to model COVID-19 is its ability to capture the effects of preventative measures on the overall progression of the disease. By differentiating between susceptible and exposed, we can see the effects that public health measures have on reducing exposure to the disease, whereas some other epidemiological models (including SIR models) can only explain how individuals move from a susceptible state to an infected state. The researchers took this brand of compartmentalization to a higher degree by including a state for individuals who become asymptomatically infected; the researchers estimated the proportion of symptomatic

infection via the “testing (of) every passenger on the Diamond Princess ship” (Riou et al, 2020). The Diamond Princess was a cruise ship which experienced an outbreak of COVID-19.

Implementation

The researchers implemented their model using the statistical programming language R coupled with Stan, a programming language which is often used in Bayesian modeling. More than likely, it was useful for the researchers to implement the model in such a fashion because of their distribution-based assumptions. A Bayesian approach is particularly well suited to a situation like COVID-19 modeling because they probabilistically explain inherent uncertainty effectively.

Results and Conclusions

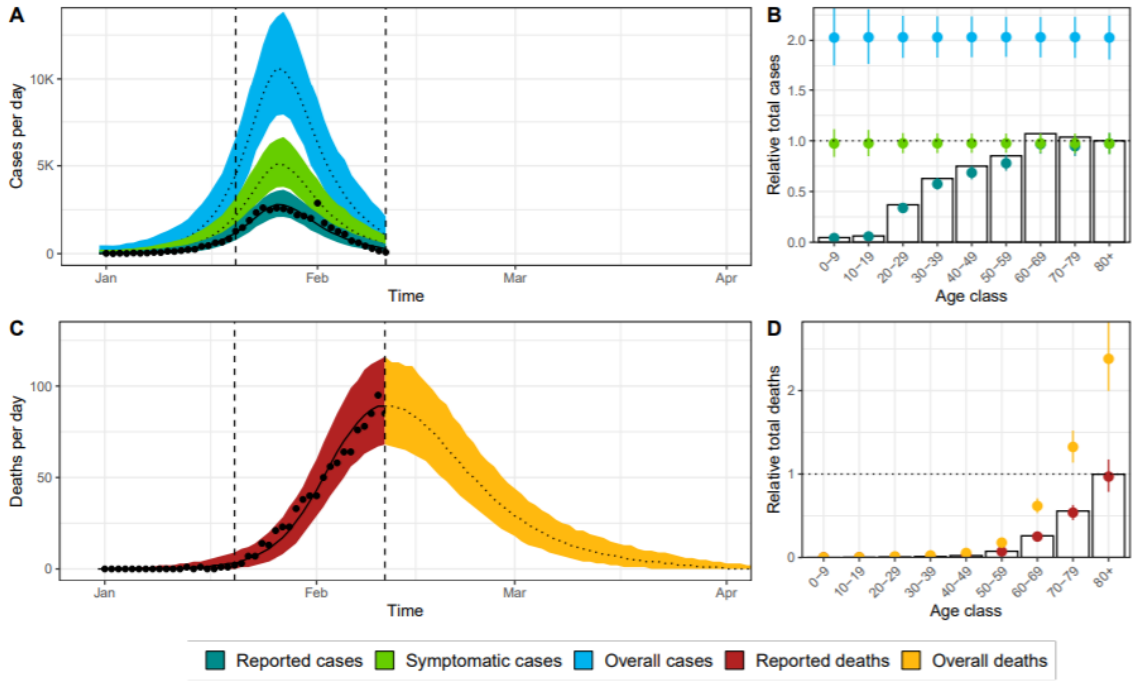

Figure 2: *Model fits from the Bern SEIR model (Riou et al, 2020). The four fits above represent the four response variables considered by the researchers.*

Riou and the other researchers from Bern were able to fit the model using their four chosen response variables, producing results that describe many different facets of the pandemic and how successful preventative measures have been. They estimated that COVID-19 had a case-fatality ratio of 1.6% in Hubei, China between January 1 and February 11 with a 95% confidence interval of (1.4%, 1.8%) (Riou et al, 2020); estimates of the true CFR vary wildly due to limited testing capacity and uncertainty in the data. They were also able to estimate several other aspects of the outbreak in Hubei.

Overall, this SEIR model produces many actionable insights about COVID-19 and the way the number of cases and fatality rates progress over time. The researchers note, however, that their work is heavily reliant on the key assumptions discussed earlier (Riou et al, 2020). One major problem presented in the current infancy of COVID-19 modeling is the vast amount of uncertainty in the data; the researchers note that varying estimates of the symptomatic infection rate are particularly problematic.

The University of Texas

Next, we examine a model built at The University of Texas at Austin COVID-19 Modeling Consortium. This model attempts to explain the effects that social distancing policies have on the daily death rate of the disease. Uniquely, the researchers from Texas were able to use mobile phone GPS data obtained from SafeGraph to quantify several different aspects of social distancing for use in the model. The model was built by Spencer Woody, Mauricio Tec, Maytal Dahan, Kelly Gaither, Michael Lachmann, Spencer J. Fox, Lauren Ancel Meyers and James Scott (Woody et al, 2020).

Technique

In a similar fashion to the researchers from Bern, the researchers from Texas take a Bayesian approach to the modeling task, but unlike Bern they do not work in the framework of an SEIR model. Instead, they use a GLM (generalized linear model) framework that is purely statistical and does not rely on the same assumptions or parameters as the Bern model. Within this framework, they use a negative binomial distribution to describe the log per-capita death rate and time-evolving Gaussian curves to approximate death rates locally and then regress on covariates related to social distancing and fit using Markov Chain Monte Carlo (Woody et al, 2020) The result is what the researchers describe as " a negative-binomial mixed-effects GLM for daily COVID-19 deaths" (Woody et al, 2020). It is worth noting that the researchers from Texas take great care to contrast their model with a similar model built in March 2020 by the Institute for Health Metrics and Evaluation (IHME) at the University of Washington. We reached out to the University of Washington for information on this model but received no response.

Assumptions.

In contrast to the Bern model, the Texas model makes relatively few assumptions. They assume, much like IHME, that "the expected daily death rate… can be locally approximated by a curve proportional to a Gaussian kernel" (Woody et al, 2020). Note that this is used in the fitting of the time-evolving Gaussian curves, which they assume can locally approximate daily death rates. The researchers from Texas also make a key assumption that "social-distancing behavior in a state remains unchanged from the average behavior over the seven most recent days of available data" (Woody et al, 2020). They also assume that the heteroskedasticity and overdispersion inherent in the data can be handled by incorporating the negative binomial portion of the model.

Parameterization.

The parameters used in the model are relatively complicated. Firstly, they parameterize the time-evolving Gaussian curves using maximum daily expected death rate, the day on which the expected death rate reaches its maximum value and a steepness parameter that describes the speed with which the death rate rises and falls (Woody et al, 2020). They then move to the log scale to obtain a vector of beta values which are the coefficients of "a locally quadratic regression on *t*, the number of elapsed days since deaths crossed the threshold value of 3 per 10 million" (Woody et al, 2020). This process is an example of a nonparametric regression technique called LOWESS, or locally weighted polynomial regression. Finally, they incorporate a vector of predictors based on mobile phone social distancing data which "are allowed to 'flatten the curve' by changing its shape" (Woody et al, 2020).

Data

One of the most unique aspects of the Texas model is its use of mobile phone GPS data obtained from SafeGraph which informs the model's social distancing parameters. SafeGraph is a data company which usually provides points-of-interest data on businesses, traffic, airports and other entities. However, during the pandemic they have turned their attention toward COVID-19 data, and the researchers from Texas were able to make use of their GPS data. Briefly, SafeGraph provides the researchers with real time mobile phone geolocation data which accurately quantifies the number of individuals in any given place. According to Woody and the other researchers, the "data source quantifies… changes in visitation patterns to public places" and "time spent at home versus at work" (Woody et al, 2020). This data has allowed the researchers to gain much clearer real time quantifications of social distancing practices to inform parameters in the model.

Motivation

The main motivation behind the Texas model seems to be expanding upon the original IHME model; the authors outline three key similarities and three differences between the IHME model and their own. The authors note that "ours is not an epidemiological model, in the sense that we do not try to model disease transmission, nor do we use or attempt to estimate underlying epidemiological parameters… our model is purely statistical" (Woody et al, 2020). Indeed, their work relies on relatively few assumptions and makes use of many different statistical techniques.

Implementation

Much like the researchers from Bern, the authors of the Texas model make use of Stan (integrated into R) to fit their model. The authors note that they fit their model using "Markov Chain Monte Carlo, sampling from the posterior distribution of all model parameters" (Woody et al, 2020). Markov Chain Monte Carlo methods are often used in a Bayesian framework such as this; the researchers need to estimate a posterior distribution (i.e. the distribution of daily COVID-19 deaths) by sampling from many prior distributions with varying degrees of informativeness. In describing this method of model fitting with regard to their model, the researchers note that they "use weakly informative priors on the fixed effects…, the second-stage regression coefficients…, and the covariance matrix… of the random effects" (Woody et al, 2020).

Results and Conclusions

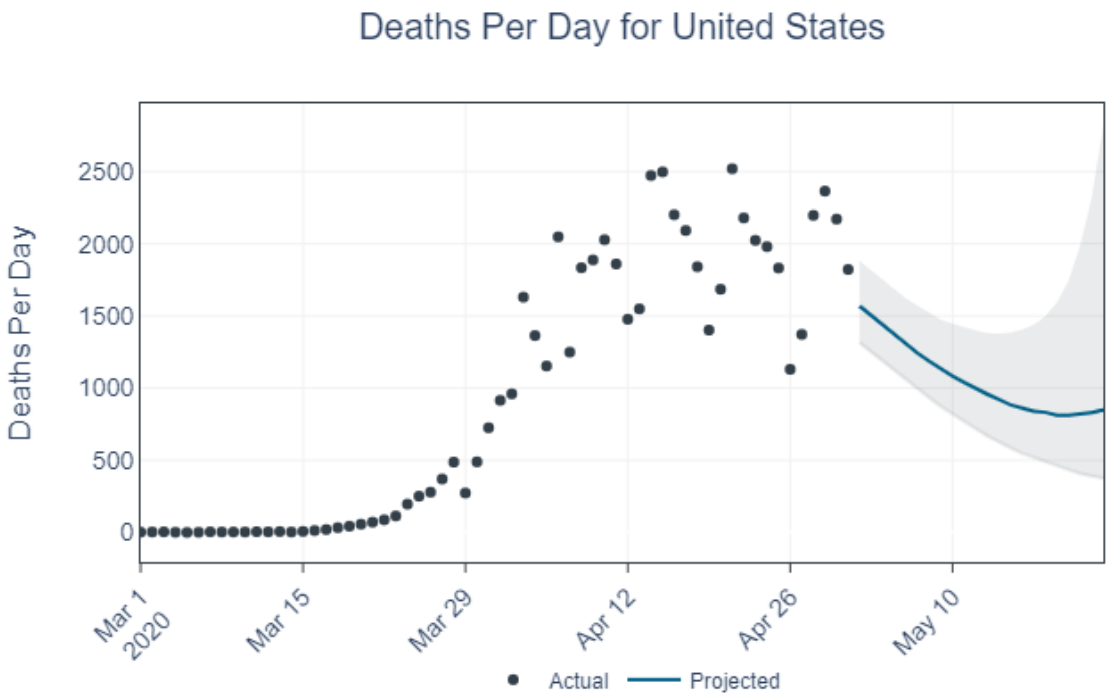

Figure 3: *Plot downloaded from https://covid-19.tacc.utexas.edu/projections/, the website which hosts the researchers' model. The plots and predictions are updated daily.*

The researchers from Texas report no results in their paper; instead, they update the website above with new results daily. Their model obtains new information constantly from SafeGraph which informs new results.

Overall, Woody and the other authors have produced a model which provides well informed predictions on death rate to the granularity of metropolitan areas across the United States. The authors note that, unlike the IHME model and many others, they did not use any data from China or Europe and instead focused all their attention on modeling COVID-19 daily death

rates in the United States. This level of focus allows the model to home in on results relevant to the American public which they can see in an accessible format, updated on a daily basis.

Humboldt University of Berlin

The final model we want to examine was built by researchers at Humboldt University of Berlin. This model is unique from the other two models we have analyzed in two major ways. Firstly, it attempts to predict the ICU load due to COVID-19 infections rather than the death or transmission rate. Secondly, in contrast with the previous models, this model does not exactly fit a pre-existing technique; rather, it is somewhat of a hybrid between a time series and an exponential regression model. The model was developed by Matthias Ritter, John-Dylan Haynes and Kerstin Ritter (Ritter et al, 2020).

Technique

The Berlin researchers' model describes how ICU admissions develop over time and describing that starts with describing how many people test positive for COVID-19 per day. Thus, the researchers work with a time series where the value on any given day is the number of new individuals who have tested positive for the virus on that day. Naturally, some individuals will need ICU care, and some will not, due to age, pre-existing conditions and other factors. The researchers then consider the amount of time between when an individual is tested positive for the disease and when that individual is admitted to an ICU. They then work with the "average number of days patients remain in an ICU" (Ritter et al, 2020). Once the model estimates its parameters, it then assumes that the number of infections grows exponentially with respect to those parameters (we will discuss this assumption further in the Assumptions section). The model finally uses these components to make predictions about ICU load in the future.

Assumptions.

Fundamental to the Berlin model is its assumption that ICU load grows exponentially along with the number of infections. However, Ritter and the other researchers make a modicum of other assumptions. Firstly, the model needs to assume a certain ICU rate, which the researchers note "has to be either estimated from the data or pre-specified" (Ritter et al, 2020). Ritter and the other authors apply their model to Germany, specifically Berlin, where this rate is not available, but they estimate an ICU rate of 5% based on information from other locations. The authors note, however, that "the ICU rate depends on age and the total number of tests because more testing will lead to a higher number of mild cases and thus a lower ICU rate" (Ritter et al, 2020). Secondly, the model assumes that the average time between a positive test and an ICU admission vary based on a number of factors, thus they must make an estimation based on the data in order to best fit the model. Finally, the researchers note that they must also assume a certain average number of days that a given individual needs to be in an ICU, which can also vary based on a number of factors (Ritter et al, 2020).

Parameterization.

$$\widehat{\mathrm{IC}}_t(K, \alpha, l^*) = \sum_{k=1}^{K} \alpha \, \Delta \mathrm{PT}_{t-l^*-k+1}$$

Figure 4: *The model equation as written by Ritter and the other researchers from Berlin.*

The model, as described by the researchers, is simple; it needs to estimate only three parameters with respect to time before fitting an exponential growth model. The model considers the ICU rate, denoted α (estimated at 5% as discussed in the Assumptions section), the average time lag between a positive test and ICU admission (which has been estimated to best explain the

data (Ritter et al, 2020)), denoted l*, and the average length in days of an ICU stay, denoted κ. These parameters are then coupled with ΔPT, the number of positive tests on a given day. The equation, using the properly estimated parameters and run over all days where observations are available, produces a time series of the number of patients who will be under ICU care on a given day. The researchers note that the model is "sensitive to the assumed or estimated parameters and the underlying data" (Ritter et al, 2020).

Data

As mentioned previously, the researchers designed their model to be directly applicable to Germany and specifically Berlin, so they used data from hospitals in Berlin to inform their model building process and the estimates of the parameters involved. The researchers make their data available online.

Motivation

There are certainly many motivations behind designing this model for use by Berlin hospitals and hospitals across the globe. In statistical modeling, often the simplest and most easily accessible and generalizable techniques either yield the best results or can be adapted to yield the best results, and that is certainly the case with the Berlin model. The Berlin researchers' work is described in their paper as a "simple statistical model", and indeed it is. The simplicity of a model increases its generalizability and reproducibility for use by the rest of the scientific community.

Implementation

$$\mathrm{RMSE}(K, \alpha, l^*) = \sqrt{\sum_{t=1}^{T} (\mathrm{IC}_t - \widehat{\mathrm{IC}}_t(K, \alpha, l^*))^2}$$

Figure 5: *Ritter, Haynes and Ritter's RMSE equation, minimized to optimize the model.*

In keeping with the theme of simplicity and reproducibility of their modeling process, the team from Berlin implemented their model in Excel. This is extremely unique; often, researchers find that their models can only be fit using advanced software, as we examined in the previous sections. We have seen that the Berlin model is built using a simple equation with few parameters. We also see in Figure 5 that the root mean squared error (RMSE) equation used to find the best model fit is relatively simple as well. These factors allowed the researchers to implement their model using arithmetic operations in Excel.

Results and Conclusions

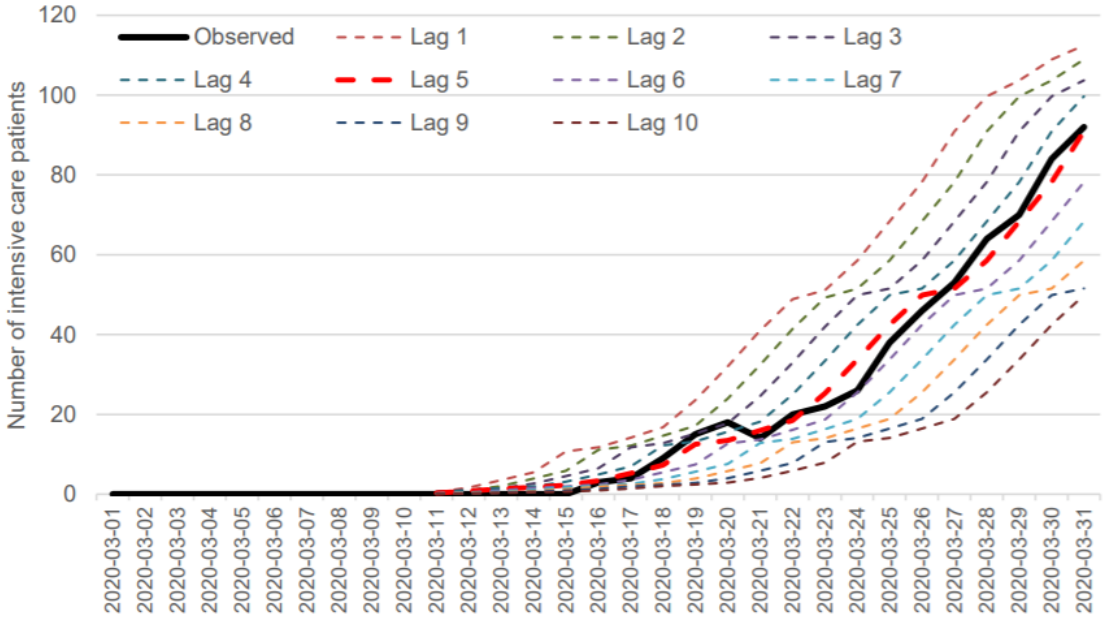

Figure 6: *The Berlin model's predictions compared to the observed values for the entirety of March 2020. The model was optimized at a lag of 5 and k-value (average length of ICU stays in days) of 14.*

Ritter and the other researchers present several different results from their model output, including the plot shown above in Figure 6. We see that the bands on the graph represent the

different lag values considered in the model building process; the black line represents observed values, and we see that the model fit at a lag of 5 days very closely approximate the observations over the month of March.

Overall, the researchers from Berlin accomplished what they set out to do: build a simple statistical model to help hospitals in Berlin predict ICU loads. The authors also intend for their model to help inform policy decisions, noting that “risk models are needed that allow policymakers to estimate the future ICU load to take appropriate measures” (Ritter et al, 2020).

References

Riou, J. et al (2020). Adjusted Age-Specific Case Fatality Ratio During the COVID-19 Epidemic in Hubei, China, January and February 2020. *Preprint (obtained from* https://www.medrxiv.org/content/10.1101/2020.03.04.20031104v1.full.pdf)

Azose, J. (2013). Statistical Inference in a Stochastic Epidemic SEIR Model with Control Intervention: Ebola as a Case Study. *University of Washington, courses*.

Woody, S. et al (2020). Projections for first-wave COVID-19 deaths across the US using social-distancing measures derived from mobile phones. *The University of Texas at Austin COVID-19 Modeling Consortium.*

Ritter, Haynes, Ritter (2020). Covid-19—A simple statistical model for predicting ICU load in exponential phases of the disease. *Obtained from* https://arxiv.org/ftp/arxiv/papers/2004/2004.03384.pdf