Data-driven Modeling of the COVID-19 Pandemic Using Penalized Linear Regression

by

Sabrina M. Corsetti

A thesis submitted in partial fulfillment
of the requirements for the degree of
Bachelor of Science with Honors
(Department of Physics)
at the University of Michigan
2021

Advisor:

Professor Tom Schwarz

A C K N O W L E D G M E N T S

I would like to express the deepest appreciation to my advisor, Professor Tom Schwarz, who spent more than two years encouraging my every academic endeavor and instilling in me the will to pursue research for the fun and love of science. His constant guidance and support of my research both at and beyond U-M has made me into the scientist I am today.

I would also like to thank Professor Marisa Eisenberg and Professor Emily Martin for their mentorship throughout this project. Their thoughtful commentary and insights have motivated the implementation of nearly every significant improvement to the model.

In addition, I am deeply indebted to all of the research advisors and mentors that I have had throughout my undergraduate career: Professor Wolfgang Lorenzon, Professor Junjie Zhu, Dr. Avi Purkayastha, and all of the numerous postdocs, research scientists, and graduate students that lent me their ears and their advice during times of need.

I also extend the sincerest gratitude to the students that have worked on this project with me: Ella McCauley, Thomas Baer, Robert Myers, Yitao Huang, and Karl Falb. Without their hard work and ingenuity, the model's contributions could never have grown to the extent that they have.

Outside of my research, I owe a huge thank-you to all of my friends back home and at Michigan that have kept me sane throughout the past four years. From encouraging each other to take breaks and enjoy the world around us to supporting each other during difficult times, the mutual support of my friends has been a constant motivating force for me.

Finally, I must thank my family with all of my heart for providing me with the resources to come to Michigan and pursue my research both in and beyond this work. It is through their unwavering support and encouragement that I have met the others acknowledged here and made so many years worth of incredible memories.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Data-driven Modeling of the COVID-19 Pandemic Using Penalized Linear Regression

by

Sabrina M. Corsetti

Chair: David Gerdes

Since early 2020, the push to subdue the COVID-19 pandemic has brought unprecedented levels of attention to disease modeling efforts. The U-M COVID-19 model presented in this thesis was developed in response to a demand for COVID-19 spread, hospitalization, and mortality predictions. The model makes regional, state, and national predictions for these three data categories using ridge regression, a machine learning algorithm rooted in penalized linear regression. To make its predictions, the model learns the relationship between consecutive sets of COVID-19 data points. Once the model has learned the necessary relationships, it can produce future predictions indefinitely, with uncertainties given by the bootstrapping method. As of March 2021, the model makes its predictions based on varying combinations of case, hospitalization, death, social distancing, and testing data. However, the model is highly flexible and capable of making predictions based on any combination of inputs. This study presents the underlying mathematics of the model, as well as its prediction performance for the United States, the state of Michigan, and the Grand Rapids region of Michigan.

CHAPTER 1

Introduction

In December 2019, the first cluster of the novel human coronavirus disease COVID-19 was detected in Wuhan, China. The disease, caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), quickly spread outside Wuhan's borders and became a global threat. In March 2020, COVID-19 was officially declared a global pandemic by the World Health Organization (WHO).

Since the early stages of the pandemic, numerous measures have been implemented to prevent the spread of COVID-19, such as city-wide lockdowns and social distancing regulations. As these measures have varied greatly between different regions, public officials and health systems have been tasked with monitoring their efficacy and adapting their strategies accordingly.

Forecasts of COVID-19 spread, hospitalizations, and mortality provide one avenue for health authorities to gauge the efficacy of lockdown measures and respond proactively to increased disease spread. While numerous mechanistic prediction methods exist, they tend to depend on assumptions regarding variables affecting the spread of the pandemic, such as mask compliance rates and population density. Machine learning algorithms offer an alternative paradigm for forecasting pandemic spread, in that machine learning algorithms can learn from trends in a dataset without users supplying them with explicit assumptions.

Machine learning algorithms are a popular tool set for researchers in a variety of fields, due to their ability to find subtle trends in datasets that would be immensely difficult to detect using traditional statistical methods. One popular application of machine learning algorithms is in the context of particle physics, in which algorithms such as decision tree classifiers can be used to probe the Standard Model with high precision and seek new physics in decay data [1].

Created by a group of U-M Physics researchers, the U-M COVID-19 model was developed in an effort to leverage particle physics-based machine learning expertise to create purely data-driven forecasts of COVID-19 cases, hospitalizations, and deaths. The model

is based on ridge regression, a form of machine learning related to ordinary least squares regression with provisions to protect from overfitting. It makes its predictions based on a combination of datasets containing historical COVID-19 case, hospitalization, and death counts, as well as relative social distancing measures.

Since its development in spring 2020, the model has been used for numerous applications. As of March 2021, it is primarily applied in three directions: national and state case and death predictions within the United States, regional case and death predictions within the state of Michigan, and hospitalization predictions within the Grand Rapids region of Michigan. State- and region-level predictions for Michigan are provided directly to the office of the governor, and Grand Rapids hospitalization predictions are provided directly to the Business Intelligence & Analytics department at Metro Health.

Unlike the other prediction types, our state and national predictions do not have direct recipients. Rather, we contribute them to the COVID-19 Forecast Hub - the official data source for the CDC COVID-19 Forecasting page [2],[3]. Building off of historical influenza forecasting efforts, the goal of the COVID-19 Forecast Hub is to aggregate numerous individual forecasts into a single, high-performing ensemble prediction. On a weekly basis, dozens of global research groups submit predictions from the national down to the county level for the United States. The Hub administrators then combine these predictions into the weekly ensemble, taking the medians of the submitted predictions as the ensemble's "point" predictions, and deriving confidence intervals from the spread of the submitted predictions. As outlined in [2], the ensemble has consistently outperformed nearly all of the individual contributing algorithms.

Our participation in the COVID-19 Forecast Hub has shaped certain aspects of our prediction-making and error analyses. Namely, in compliance with the Hub's guidelines, we make our predictions based on the ground truth datasets encouraged for universal contributor use. Likewise, we perform our error analyses based on the prediction interval and percentile reports required for Hub submission. All relevant Hub guidelines can be found in the Hub's technical README.

All of the code used to generate the U-M COVID-19 model's predictions is available in the model's git repository [4], and all data produced by the model is available within the COVID-19 Forecast Hub under a cc-by-4.0 license. Weekly predictions for the COVID-19 Forecast Hub are generated using batch jobs within the Great Lakes cluster, a U-M Advanced Research Computing resource, which allows for the creation of predictions for 51 regions (all 50 states, plus the national aggregate) simultaneously. The model is highly flexible with regards to data inputs and machine learning algorithms, which makes it a useful multi-purpose tool for regression-based modeling.

CHAPTER 2

Foundations and Related Works

2.1 Foundational Models

Within epidemiology, numerous models exist that shed light on the basic dynamics of disease propagation. While these models do not account for many of the complex factors that contribute to disease spread, such as population density and travel practices, they provide a basis for the construction of more sophisticated disease models.

One such simplified disease model is the SIR model - a system of ordinary differential equations representing three separate portions of a population during an epidemic:

$$\begin{aligned} S = S(t)&\text{: the number of } \textit{susceptible} \text{ individuals,} \\ I = I(t)&\text{: the number of } \textit{infectious} \text{ individuals, and} \\ R = R(t)&\text{: the number of } \textit{recovered} \text{ individuals,} \end{aligned} \tag{2.1}$$

where time t is measured in days, and both recoveries and deaths are treated as recoveries [5].

Under the assumptions of the baseline SIR model, all previously uninfected individuals are susceptible to infection, and all recovered individuals are fully immune to the illness. Thus, we can redefine the model in terms of fractions of an N-person population, where:

$$\begin{aligned} s = S(t)/N&\text{: the susceptible } \textit{fraction} \text{ of the population,} \\ i = I(t)/N&\text{: the infectious } \textit{fraction} \text{ of the population, and} \\ r = R(t)/N&\text{: the recovered } \textit{fraction} \text{ of the population,} \end{aligned} \tag{2.2}$$

and we now have $s + i + r = 1$ for all times t.

The baseline SIR model makes several further population assumptions for the sake of

problem simplification. Namely, births and immigration are disregarded, so the susceptible population never grows, and a fixed fraction k of the infected population is assumed to recover every day. Here, k is taken to be $1/\tau$, where τ is the average time people remain infectious.

Under these assumptions, we can derive the full set of SIR equations:

$$\begin{aligned} \frac{ds}{dt} &= -bs(t)i(t), \\ \frac{di}{dt} &= bs(t)i(t) - ki(t), \\ \frac{dr}{dt} &= ki(t), \end{aligned} \tag{2.3}$$

where b is the average number of contacts between an infectious individual and other individuals (susceptible or otherwise) per day. So, each infectious individual generates $bs(t)$ new infectious individuals per day.

While the standard SIR model is too simple to capture all of the complex dynamics underlying disease spread, especially in a socially distant population, it is frequently used as a foundation for more robust models. As a simple example, the SEIR model is an improved model that adds an extra component to the SIR model: E, the number of *exposed* individuals in a population [6]. This addition accounts for a non-zero incubation period, during which an individual has contracted a disease, and is thus no longer susceptible, but is not yet infectious.

2.2 Data-driven Forecasting

Beyond simple extensions like the SEIR model, extremely robust models can be developed based on the SIR model. For example, machine learning can be used to create SIR-based models capable of forecasting disease characteristics up to days or weeks ahead. One such model, developed in 2020 for the purpose of modeling COVID-19, used machine learning to forecast COVID-19's basic reproduction number $R_0(t)$ for upwards of two weeks ahead [7]. Here, the basic reproduction number represents the average number of additional people infected by a single infectious person. If a population experiencing an epidemic can maintain $R_0(t) < 1$, the epidemic will no longer be self-sustaining and will instead die out. As a result, monitoring and predicting R_0 can help health officials decide upon strategies for mitigating a disease's spread [8].

While the model eventually incorporated undetectable - or asymptomatic - cases in

its forecasts, it initially relied only on the basic SIR equations (2.3). To represent the relationship between current COVID-19 transmission and previous data, these equations were transformed into discrete time difference equations:

$$\begin{aligned} s(t+1) - s(t) &= -b(t)s(t)i(t), \\ i(t+1) - i(t) &= b(t)s(t)i(t) - k(t)i(t), \\ r(t+1) - r(t) &= k(t)i(t), \end{aligned} \tag{2.4}$$

where $b(t)$ and $k(t)$ are now time-dependent to account for changes in contact and recovery rates based on factors like social distancing and advancements in treatment.

Based on the early-pandemic assumption that most people in a region have not been infected, and thus $S \approx N$, the following relationships were determined for $k(t)$ and $b(t)$:

$$\begin{aligned} k(t) &= \frac{r(t+1) - r(t)}{i(t)}, \\ b(t) &= \frac{[i(t+1) - i(t)] + [r(t+1) - r(t)]}{i(t)}. \end{aligned} \tag{2.5}$$

From these equations, given historical data from a given period $\{i(t), r(t), 0 \leq t \leq T-1\}$, it is possible to measure $\{b(t), k(t), 0 \leq t \leq T-2\}$. Then, using machine learning it is possible to predict $\{b(t), k(t), t \geq T-1\}$ and in turn the basic reproduction number R_0, where $R_0 = b(t)/k(t)$.

Specifically, using ridge regression - a form of machine learning rooted in penalized linear regression - future values of $b(t)$ and $k(t)$ can be determined by taking linear combinations of previous values of $b(t)$ and $k(t)$. The computation of forecast values $\hat{b}(t)$ and $\hat{k}(t)$ follows as:

$$\begin{aligned} \hat{k}(t) &= \sum_{n=1}^{N} a_n k(t-n) + a_0, \\ \hat{b}(t) &= \sum_{m=1}^{M} c_m b(t-m) + c_0, \end{aligned} \tag{2.6}$$

where M and N are user-fixed values, and the coefficients a_n and c_m are optimized for the dataset through ridge regression's minimization of an internal objective function.

Once these coefficients have been determined, an iterative process can be used to gen-

erate successive $\hat{b}(t)$ and $\hat{k}(t)$ indefinitely. In other words, expectations can be defined for future transmission and recovery rates, and thus for $R_0 = b(t)/k(t)$.

2.3 The Model

While predicting COVID-19 transmission characteristics is useful for understanding the effects of containment efforts, direct predictions of COVID-19 cases, hospitalizations, and deaths can be more intuitively useful to the public and policy-makers. For example, if a sharp increase in cases is forecasted, policy-makers can implement more stringent social distancing measures in advance. Likewise, if a large uptick in hospitalizations is predicted for a hospital system, the system can impose a temporary halt on elective procedure scheduling to ensure the availability of resources for COVID-19 patients.

Merging the goal of predicting cases, hospitalizations, and deaths with the math underlying the ridge regression model from section 2.2 led to the construction of the U-M COVID-19 model. Specifically, we adapted the method from the ridge regression model above to create a direct prediction mechanism for COVID-19 cases, hospitalizations, and deaths over an extended timeline of four weeks. Building from equation set (2.6), we predict future cases, hospitalizations, or deaths $\hat{x}(t)$ from previous cases, hospitalizations, or deaths respectively using:

$$\hat{x}(t) = \sum_{n=1}^{N} a_n x(t-n) + a_0, \tag{2.7}$$

where coefficients a_n are again determined through ridge regression.

2.3.1 Ridge Regression

Ridge regression is a form of machine learning algorithm with roots in ordinary least squares (OLS) regression. Like OLS regression, ridge regression uses linear combinations of previous data points to generate predictions, as in equation set (2.7). However, as a protection against overfitting - a phenomenon in machine learning in which a model fits too closely to a particular set of data and may therefore fail to fit additional data [9] - ridge regression introduces a "penalty" on the magnitude of the linear combination coefficients. The effect of this penalty is realized through the minimization of the objective function [10]:

$$\min_{a_j} \sum_{t=N}^{T-1} (x(t) - \hat{x}(t))^2 + \lambda \sum_{n=0}^{N} a_n^2, \tag{2.8}$$

where $x(t)$ is the true dataset, $\hat{x}(t)$ is the set of generated predictions, a_n are the coefficients in (2.7), and λ is an arbitrary parameter.

In the $\lambda \to 0$ limit, the objective function (2.8) becomes the OLS regression objective function. In the $\lambda \to \infty$ limit, (2.8) is dominated by the second sum, and minimization requires $a_n \to 0$ for all $n = 0, 1, ..., N$. Thus, by fixing the finite parameter $\lambda > 0$, one encourages the ridge regression algorithm to minimize both the sum of square differences between the data and the fit, as well as the magnitude of the coefficient vector $\vec{a}$.

By encouraging the minimization of $||\vec{a}||$, the penalty λ works to prevent any single a_n from growing too large. Since each a_n ascribes a weight to the n^{th} input in equation (2.7), an abnormally large weight can indicate the presence of overfitting in response to a significant, yet non-generalizable feature in the data used to train the model. In the case of COVID-19 modeling, such a feature might be a large spike caused by a backfill of probable COVID-19 cases, as occurred for the state of Michigan on June 5, 2020 [11]. Since such spikes are typically several orders of magnitude larger than the surrounding data, OLS regression is likely to assign it a large weight in recognition of its significance. However, in the case of ridge regression, $\lambda > 0$ discourages such a large weight assignment.

As λ is an arbitrary parameter, an optimal λ can be chosen through the testing of several models, each with a different λ. Based on the predictive performance of each model over the training set, traditionally tested using leave-one-out cross validation scored by mean square error [12], an optimal model can be chosen and used to iteratively generate predictions $\{\hat{x}(t), t \geq T\}$.

2.3.2 Iterative Predictions

Equation (2.7) provides the formula by which predictions can be made for $t \geq T$. However, in order for equation (2.7) to be executed, an optimal coefficient vector $\vec{a}$ must be chosen through the minimization of the objective function in (2.8). For a ridge regressor to choose an optimal vector $\vec{a}$, it must be provided with training data consisting of sets of N consecutive days worth of case, hospitalization, or death data, paired with the number of cases, hospitalizations, or deaths on consecutive day $N + 1$. Then, the minimization of the expression in (2.8) can be attained with respect to the equation:

$$M\vec{a} = \vec{b}, \tag{2.9}$$

where M is a real-valued matrix, and the i^{th} rows of M and $\vec{b}$ correspond to the i^{th} $(N+1)$-day set of data points.

For example, suppose that we have 100 days worth of national COVID-19 case data for the United States. Further, suppose we wish to generate future predictions for national cases using 21 days worth of previous data to create each prediction. The equation used to train the involved ridge regressor becomes:

$$\begin{bmatrix} \text{cases}_1 & \text{cases}_2 & ... & \text{cases}_{21} & 1 \\ \text{cases}_2 & \text{cases}_3 & ... & \text{cases}_{22} & 1 \\ \text{cases}_3 & \text{cases}_4 & ... & \text{cases}_{23} & 1 \\ | & | & ... & | & | \\ \text{cases}_{79} & \text{cases}_{80} & ... & \text{cases}_{99} & 1 \end{bmatrix} * \vec{a} = \begin{bmatrix} \text{cases}_{22} \\ \text{cases}_{23} \\ \text{cases}_{24} \\ | \\ \text{cases}_{100} \end{bmatrix}, \tag{2.10}$$

where each cases_i represents national cases on day i.

In this example, if we take $\hat{x}_i = M_{i-21,1:22}\vec{a}$ for each $i = 22 : 100$, an optimal $\vec{a}$ can be determined using expression (2.8). Then, using $\vec{a}$, we can iteratively generate predictions $\hat{x}_i$ for $i > 100$ using:

$$\text{cases}_i = \begin{bmatrix} \text{cases}_{i-21} & \text{cases}_{i-20} & ... & \text{cases}_{i-1} & 1 \end{bmatrix} * \vec{a} \tag{2.11}$$

2.3.3 Additional Feature Incorporation

Case, hospitalization, and death predictions based on the scheme exemplified by equation (2.10) allow for ridge regressors to use information about the shape of the curve for $0 \leq t \leq T-1$ to generate trajectories for $t > T-1$. However, a ridge regressor trained using only a curve's shape for $0 \leq t \leq T-1$ will falter when external forces change the underlying dynamics of the pandemic.

Several prominent external forces can directly influence COVID-19 dynamics, such as social distancing regulations and vaccinated population counts. Likewise, certain COVID-19 statistics can lead other statistics, in the sense that an increase or decrease in one precipitates an increase or decrease in another. For example, as COVID-19 symptoms frequently take time to escalate from mild to severe, changes in a region's incident case rate typically lead corresponding changes in the death rate by 2-8 weeks [13].

In order to generate ridge regression predictions based on one or more of these factors, in addition to the shape of the curve for $0 \leq t \leq T-1$, one must simply modify M and $\vec{b}$ in equation (2.9) to include additional features. For example, should one develop a relative metric for social distancing in the United States and wish to incorporate it into the example outlined in (2.10), they need merely to expand (2.10) to:

$$\begin{bmatrix} \text{cases}_1 & \dots & \text{cases}_{21} & \text{social}_1 & \dots & \text{social}_{21} & 1 \\ \text{cases}_2 & \dots & \text{cases}_{22} & \text{social}_2 & \dots & \text{social}_{22} & 1 \\ | & | & \dots & | & | & & \\ \text{cases}_{79} & \dots & \text{cases}_{99} & \text{social}_{79} & \dots & \text{social}_{99} & 1 \end{bmatrix} * \vec{a} = \begin{bmatrix} \text{cases}_{22} \\ \text{cases}_{23} \\ \text{cases}_{24} \\ | \\ \text{cases}_{100} \end{bmatrix}, \tag{2.12}$$

where each social_i represents relative social distancing levels on day i, and $\vec{a}$ now has dimensions 43×1.

While successive U-M COVID-19 model predictions are made iteratively based on prior case, death, or hospitalization predictions, as in equation (2.11), we do not predict future values of metrics such as relative social distancing. Thus, in the context of the example presented in equation (2.12), successive prediction-making requires modification of (2.11):

$$\text{cases}_i = \begin{bmatrix} \text{cases}_{i-21} & \dots & \text{cases}_{i-1} & \text{social}_{S-21} & \dots & \text{social}_S & 1 \end{bmatrix} * \vec{a}, \tag{2.13}$$

where S is the size of the training dataset. Using this extension method, it is possible to incorporate any combination of inputs into a ridge regressor's predictions.

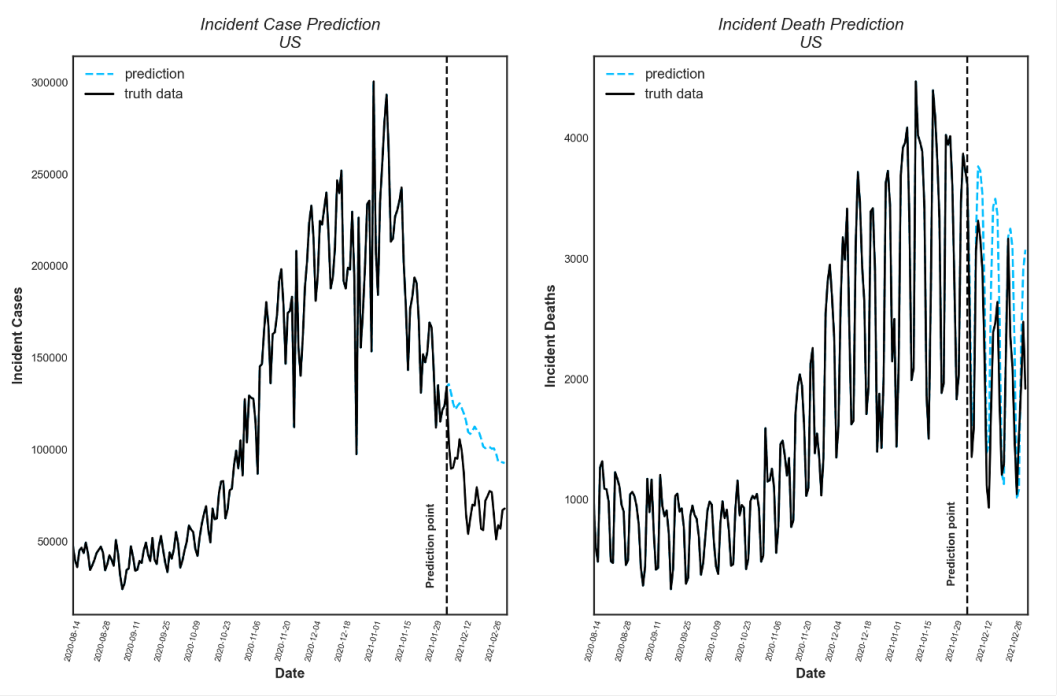

Figure 2.1: Four-week incident case and death predictions for the United States, beginning on Feb. 5, 2021. While the prediction data does not perfectly match the truth data, the use of 95% prediction intervals helps us capture this uncertainty, as demonstrated in Fig. 3.3.

A sample of incident case and death predictions for the United States is presented above in Fig. 2.1, with case predictions made using 35 previous days of case and social distancing transit data, and death predictions made using 35 days of case and death data.

CHAPTER 3

Data and Methods

3.1 Datasets

The U-M COVID-19 model is equipped to make forecasts for several different categories of COVID-19 data: case and death predictions for the U.S. at the region, state, and national levels, and hospitalization predictions for the Grand Rapids region of Michigan at the system and regional levels. Case predictions are made based on case and relative social distancing data, death predictions are made based on death and case data, and regional and system hospitalization predictions are made based on hospitalization and case data.

The datasets used by the model are regularly pulled from several sources. Case, hospitalization, and death data is pulled from a repository hosted by a group at Johns Hopkins University, social distancing data is pulled from a Google respository, and any remaining local-level data is pulled directly from healthcare providers for the purpose of generating predictions.

3.1.1 Johns Hopkins University CSSE COVID-19 Repository

As a contributor to the COVID-19 Forecast Hub, the U-M COVID-19 model makes its U.S. case, hospitalization, and death predictions based on the accepted "ground truth" data agreed upon by the Hub. While no ground truth datasets have been selected for metrics like relative social distancing and vaccination rates, all ground truth case, hospitalization, and death data is pulled from the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University [14]. The U-M COVID-19 model relies on data from the repository's csse_covid_19_daily_reports directory, which contains a host of COVID-19 data for nearly every region of the globe, reaching back as early as 1/1/2020. Sample case data for the state of Michigan is displayed in Fig. 3.1.

3.1.2 Google Mobility Reports

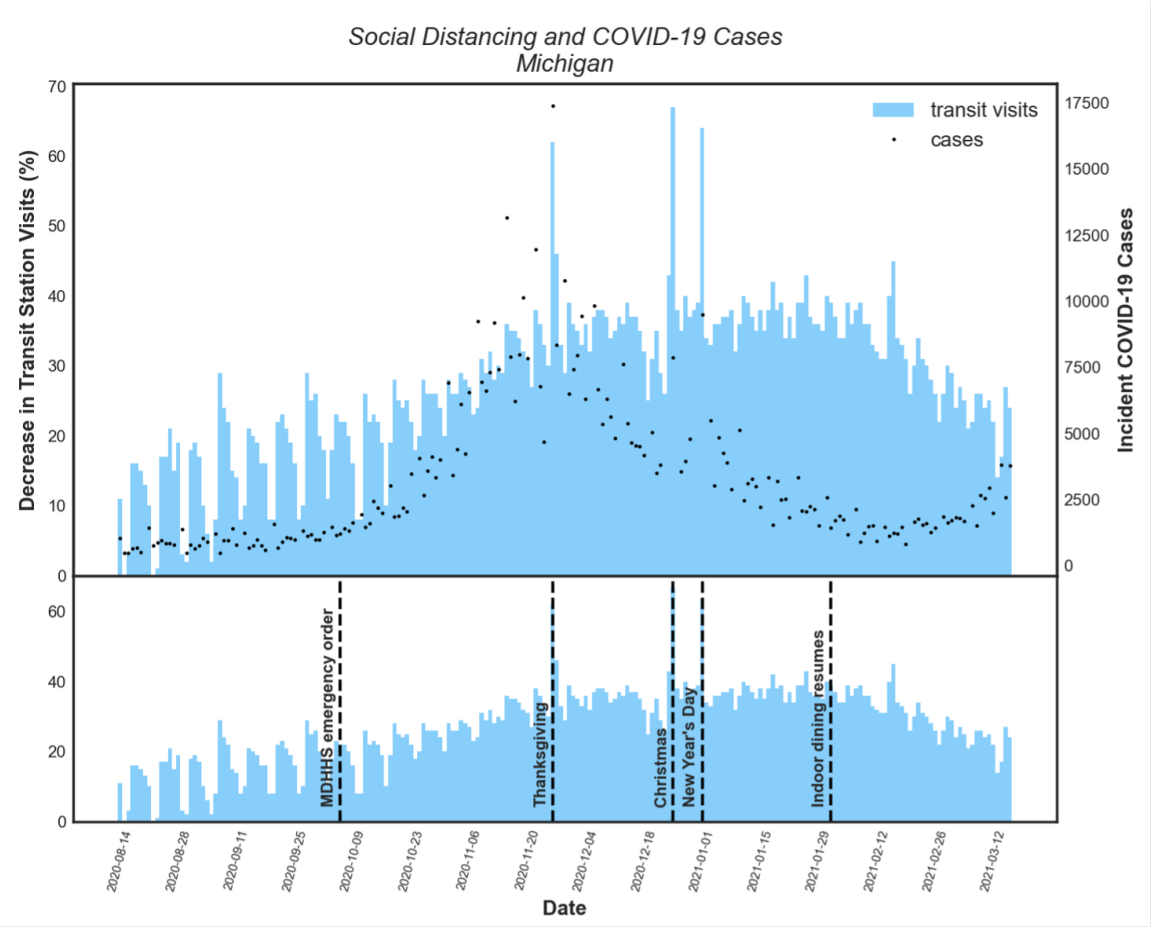

Figure 3.1: *Top:* Social distancing and incident case data for the state of Michigan, with social distancing represented by transit station visits in terms of percent decrease from baseline. *Bottom:* Same social distancing data, with markers for five significant events affecting social behaviors.

To aid COVID-19 research efforts, Google has developed the COVID-19 Community Mobility Reports resource [15]. The Mobility Reports site contains a list of csvs with relative social distancing data based on anonymized mobile device location data. The csvs provide data for most countries, provinces, states, and counties across the globe. The data consists of percentages representing people's change in social activity from a pre-pandemic baseline average. Each country or sub-region's data is divided into 6 categories, each representing different types of activity. For example, one category gives daily percentage changes in visits to transit stations, whereas another category gives changes in time spent in park areas. The other four categories measure relative time spent in residences, time spent in workplaces, visits to retail and recreation locations, and visits to grocery stores

and pharmacies. All data used by the U-M COVID-19 model comes from the transit and residential categories, as time spent at home is inversely correlated with all forms of in-store shopping and in-person work activity, and transit visits are positively correlated with non-essential travel activities. Sample transit data for the state of Michigan is displayed in Fig. 3.1.

3.1.3 Metro Health Proprietary Data

In preparing hospitalization predictions for Metro Health, a Grand Rapids region hospital system, we rely on two types of data: COVID-19 cases and hospitalizations. As we make regional predictions for Metro Health based on COVID-19 activity in the Grand Rapids region, and in Kent County specifically, we pull Kent County case data from the JHU CSSE COVID-19 Repository. However, hospitalization datasets compiled at the sub-state level are uncommon. So, we make our hospitalization predictions based on datasets provided directly by Metro Health, which contain data points regarding COVID-19 hospitalizations within the Grand Rapids region Metro, Spectrum, and Mercy Health systems.

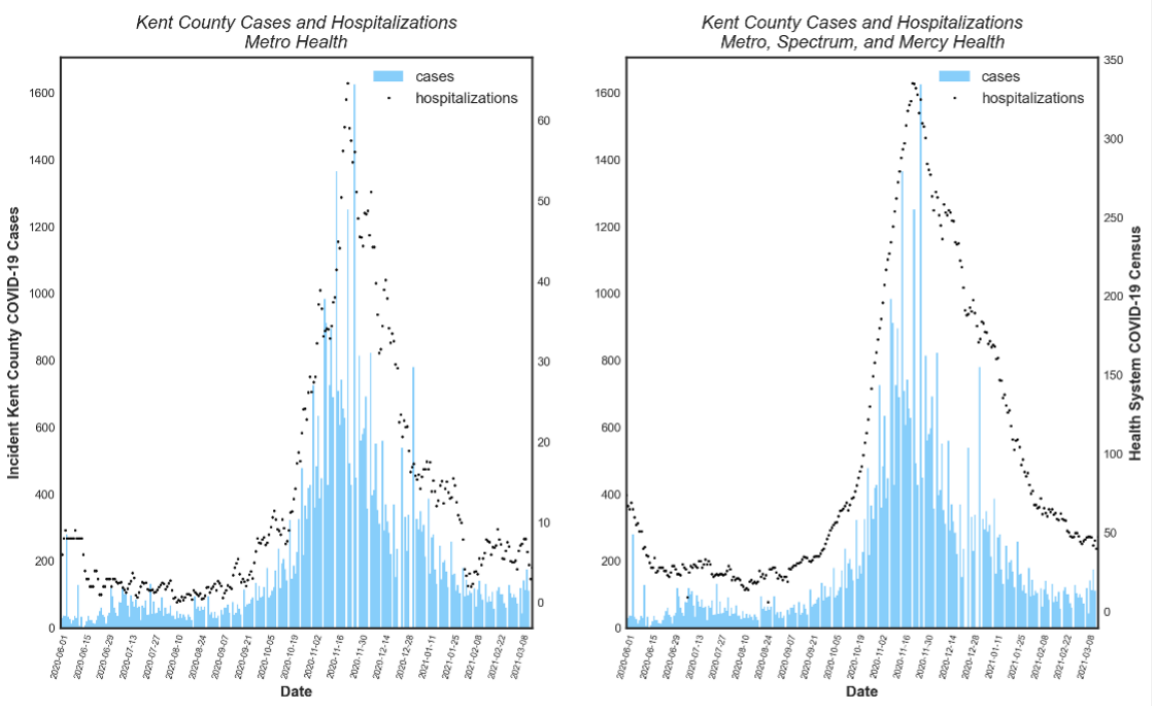

Figure 3.2: *Left:* Kent County cases and Metro Health COVID-19 hospitalizations. *Right:* Kent County cases and Grand Rapids area COVID-19 hospitalizations across Metro, Spectrum, and Mercy Health. Generally, we see peaks in cases lead peaks in both sets of hospitalizations by approximately one week.

While our goal in the context of Metro Health is to provide hospitalizations specific to the individual hospital system, Metro Health's COVID-19 patient intake is closely related to the intake of two other hospital systems in the Grand Rapids region: Spectrum Health and Mercy Health. Thus, when we provide predictions for Metro Health, we rely on two sets of data. The first dataset is specific to Metro Health, containing COVID-19 patient counts, or censuses, by date. The second dataset contains combined census values for all three systems, again by date. Plots of the two datasets in comparison to Kent County COVID-19 cases are presented in Fig. 3.2.

Using these datasets, we create two sets of predictions - one for Metro alone, and another for the combination of Metro, Spectrum, and Mercy. By studying both sets of predictions, the Metro analytics team can determine their expected COVID-19 patient intake based on both their own previous data, as well as their expected proportion of COVID-19 care with respect to the other systems.

3.2 Ridge Regressors

The U-M COVID-19 model makes use of two types of ridge regression algorithms implemented in the Python Scikit-learn library [16]. The first algorithm is a traditional ridge regression algorithm, whereas the second is an extension of the first. The second algorithm incorporates native cross-validation, which allows for the testing of multiple points in the algorithm's parameter space, leading to the determination of a single strongest model parameter set.

3.2.1 Scikit-learn Ridge

The first algorithm used by the U-M COVID-19 model is Scikit-learn's Ridge() model. Given inputs M and $\vec{b}$ as in equation (2.9), as well as a user-specified λ penalty, the model's fit() method minimizes the objective function in equation (2.8) to provide the coefficient vector $\vec{a}$. Once $\vec{a}$ has been determined, the model's predict() method can produce new predictions based on a set of inputs, as exemplified by equation (2.11).

3.2.2 Scikit-learn RidgeCV

The second algorithm used by the U-M model is the Scikit-learn RidgeCV() model. RidgeCV() is generally identical to Ridge(). However, it has an expanded parameter space that allows for automatic cross-validation over more than a single point. In the case of the

U-M COVID-19 model, this allows us to supply RidgeCV() with an array of potential λ penalties. The RidgeCV() can then use comparative leave-one-out cross validation scored by mean square error to determine an optimal λ. This optimal λ can then be fed into a lighter-weight Ridge() model for prediction-making. Typically, we provide RidgeCV with λ options of $\{1E-8, 2.5E-8, 5E-8, 7.5E-8, 1E-7, 2.5E-7, ..., 5E5, 7.5E5\}$.

3.3 Error Estimation

The question of how to generate proper prediction intervals for a ridge regression model is controversial. Since the imposition of a penalty λ introduces a deliberate bias into a ridge regression model in favor of lower variance, traditional standard error measurements fail to capture the model's true uncertainty [17]. In light of this, many penalized regression packages deliberately do not offer tools for uncertainty evaluation, as is the case with the package discussed in [17]. However, in the case of an application like COVID-19 modeling, prediction intervals can provide critical information about the trajectory of the pandemic, especially around local maxima and minima in cases and deaths.

3.3.1 Bootstrapping

Bootstrapping is a common method used to meet the demand for prediction intervals while accounting for uncertainties due to model bias [18]. As bootstrapping uses random sampling with replacement to assign measures of accuracy to sample estimates, the method allows for estimation of the sampling distribution of virtually any statistic.

The U-M COVID-19 model's bootstrapping code is adapted from a preexisting Python implementation [19]. The implementation accounts for 3 sources of prediction error, where for a model trained on a data sample of size n, given a new observation x_0:

$$y_0 := y(x_0) = \hat{y}_n(x_0) + \eta(x_0) + \eta_n(x_0) + \epsilon(x_0). \tag{3.1}$$

Here, $\hat{y}_n(x_0)$ is the model's prediction for input x_0, and $\eta(x_0)$, $\eta_n(x_0)$, and $\epsilon(x_0)$ are the model bias, model variance noise, and sample noise respectively.

To generate a prediction interval for any given point x_0, the COVID-19 model's training data is randomly sampled $B >> 0$ times. The model is then fit on each of the resulting subsets, and a prediction $\bar{y}_{b,n}(x_0)$ is generated for each $b \leq B$. Taking the variance of the set $\{\bar{y}_{b,n}(x_0),\, b \leq B\}$, we get our estimate for the first source of uncertainty - model variance noise.

Proceeding with $\bar{y}_{b,n}(x_0)$ as defined above, the bootstrapping code estimates the model bias and sample noise. To do this, it first computes the validation errors:

$$\textit{validation\_error}_{b,i} := y(x_i) - \bar{y}_{b,n}(x_i), \tag{3.2}$$

for every $b \leq B$ and $x_{i \leq n}$ which is not in the b^{th} random training sample. Taking the average of the errors gives us an estimate of the sum $\eta(x_i) + \epsilon(x_i)$. However, this estimate will tend to be too large due to the artificial weakening of bootstrapped predictions through the random sampling of the training set. Accounting for this inflation through the .632+ bootstrap estimate method ([20]), we receive a final estimate for the model bias and sample noise.

Using this bootstrapping method, we can retrieve prediction intervals for each predicted case, death, or hospitalization point, with each prediction interval corresponding to a user-determined percentile range. As the U-M COVID-19 model's predictions are generated chronologically, with each successive prediction dependent on previous predictions as in sample equation (2.11), the uncertainty of each predicted point must take the uncertainties of previous points into account.

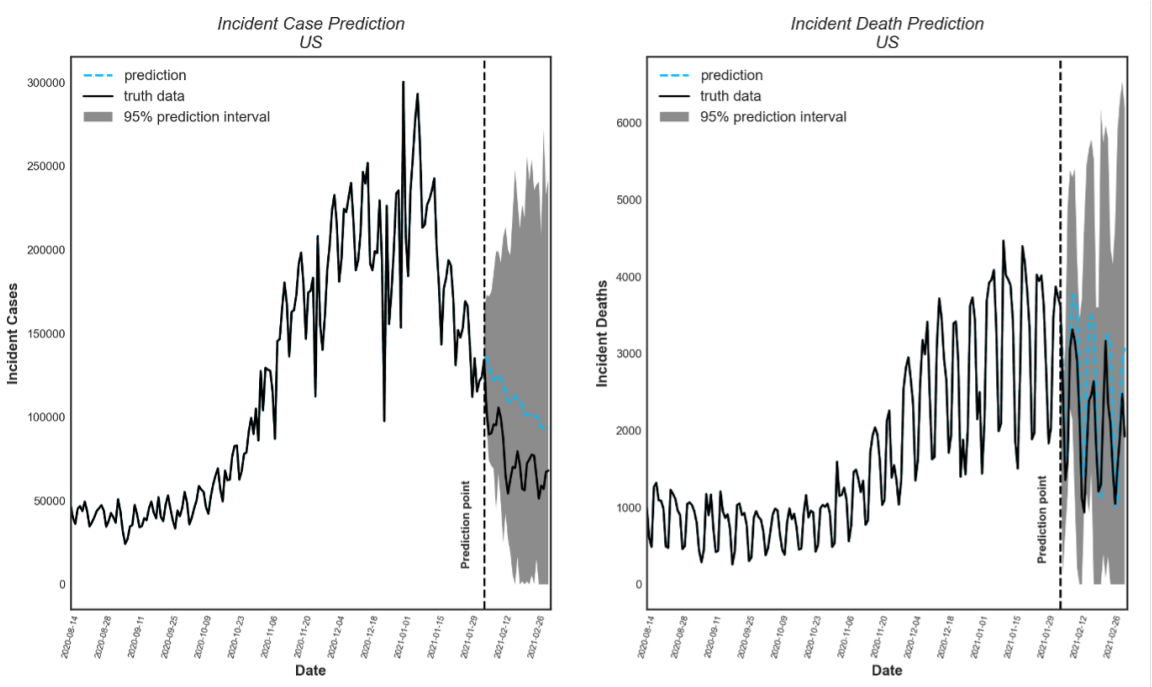

Figure 3.3: Four-week incident case and death predictions for the United States, beginning on Feb. 5, 2021. A replica of this plot without 95% prediction intervals is displayed in Fig. 2.1. While the point predictions displayed in blue do not perfectly align with the truth data, the 95% prediction intervals capture the truth data curves.

In deference to the COVID-19 Forecast Hub standard, we produce 7 percentiles $\{2.5, 10, 25, 50, 75, 90, 97.5\}$ for each case prediction and 23 percentiles $\{1, 2.5, 5, 10, ..., 90, 95, 97.5, 99\}$ for each death and hospitalization prediction. For a given point, we create a distribution of the model's propagating error by repeatedly sampling the percentile distributions of each previous point and adding the results. Then, by drawing new percentile bounds based on this broader distribution, we incorporate compounding uncertainties into our consecutive prediction intervals. A sample of case and death predictions for the United States is presented with uncertainties in Fig. 3.3 above. As in Fig. 2.1, the case predictions were made using 35 previous days of case and social distancing transit data, and the death predictions were made using 35 days of case and death data.

CHAPTER 4

Data Analysis

4.1 External Analyses

Analyzing the performance of long-term prediction methods is no small effort. Numerous factors can be taken into account, both in terms of a model's individual performance, as well as its performance in comparison to other models. In general, the most relevant metrics for individual Forecast Hub COVID-19 models are their accuracy and precision, both on individual days and over time.

In an effort to analyze the quality of its constituent models, both Forecasting Hub administrators and contributors have developed performance metrics for comparing and contrasting models. Taking one approach, the Hub administrators developed a Hub-specific metric used to compare models using a single value representing relative performance over the full prediction time frame. The ensemble metric ranks individual models in terms of their general performance against a naive baseline model.

Taking a different approach, Dr. Steve McConnell - a Forecast Hub contributor - developed a data center with comprehensive evaluations of each model across multiple metrics. The primary metrics considered are accuracy and precision for each set of predictions made by each model. While this information is crucial for an in-depth understanding of each model's strengths and weaknesses, it is difficult to summarize concisely for a full-scale Hub analysis, which is where the Hub-specific metric can serve as an effective complement.

4.1.1 Relative WIS Comparative Metric

In the most recent ensemble analysis, the Forecast Hub administrators used relative weighted interval scores (WIS) as a metric for each model's death prediction performance against a naive baseline model [2]. The weighted interval score is a method for evaluating predictive model performance that accounts for both prediction accuracy and interval coverage.

WIS values are computed as a linear combination of interval scores, where a single interval score for forecast distribution F, observation y, and uncertainty level α is taken as:

$$IS_\alpha(F, y) = (u - l) + \frac{2}{\alpha}(l - y) \times 1(y < l) + \frac{2}{\alpha}(y - u) \times 1(y > u), \tag{4.1}$$

for $1(\cdot)$ the indicator function and l and u the lower and upper quantiles of F.

Using a linear combination of 11 interval scores chosen to be equivalent to the quantile loss function, the Hub administrators computed a mean WIS for each model, where the resulting number described the average closeness of the model's distribution to the observed data, with units on the scale of the observations. Dividing each model's mean WIS by the Forecast Hub's naive baseline's mean WIS thus gave the relative WIS metric, where a relative WIS > 1 meant that the baseline model had superior performance.

At the time of the pre-print's writing in January 2021, the U-M COVID-19 model had a relative WIS of 1.51, with its largest contribution stemming not from inaccuracy, but from overly-tight prediction intervals. The results of this analysis inspired the introduction of the uncertainty propagation methods discussed in section 3.3.1, which have significantly improved our 95% prediction interval coverage. Based on preliminary results available at the CMU Delphi Group Forecast Evaluation Dashboard, our model has maintained an average relative WIS < 1 for deaths in the United States since the implementation of our new error propagation measures in early February 2021 [21].

4.1.2 Covid Complete Data Center

The Covid Complete Data Center, developed by Forecast Hub contributor and software engineer Steve McConnell, provides in-depth insights into the evolution of COVID-19 models over time, as well as the factors underlying their relative WIS values [22]. The data center's analyses are broken down into categories, starting from the highest level with comparative analyses versus individual model analyses. From there, the analyses are further broken down into weekly comparative accuracy and prediction interval coverage reports, as well as individual model assessments over the full time scale of the pandemic.

The Covid Complete Data Center has been an incredibly useful tool for gauging the performance of the U-M COVID-19 model. As an example, it was the Center's individual prediction interval reports, along with the model's relative WIS value, that prompted improvements to our model's uncertainty propagation methods going into 2021. While the Data Center offers comprehensive insights into a model's performance throughout the pandemic, its historical prediction analyses for the U-M COVID-19 model are based on early versions of the model, prior to the most recent algorithm and uncertainty improvements. In

addition, all of its analyses focus on death predictions, rather than case predictions. Thus, additional analyses of the model's performance are warranted based on retroactive predictions for both cases and deaths generated using the current, most robust version of the model.

4.2 Historical Accuracy

While the U-M COVID-19 model consistently makes predictions for more than 50 regions and subregions of the United States, it has three primary regional targets: the United States, the state of Michigan, and the combination of Metro, Spectrum, and Mercy Health. As these targets cover all three levels of the model, from the national down to the state and regional levels, they provide a strong case study for the model's performance given different datasets. For both the United States and Michigan, this section provides analyses of case and death prediction accuracy over the span of 86 days worth of predictions. Likewise, it provides analyses of 86 days worth of hospitalization predictions for both Metro Health, as well as the combination of Metro, Mercy, and Spectrum Health.

All of the predictions in this section were generated using the methods outlined in Chapter 2. Each set of historical predictions was generated using the most up-to-date version of the U-M COVID-19 model, as of March 2021. However, to ensure a fair assessment of the model's capabilities at every point in time, all of the historical predictions were generated using versions of the model trained only on the pandemic data available prior to that date. For example, predictions made for October 14, 2020, were generated using a model trained exclusively on pandemic data through October 13, 2020.

4.2.1 National Performance

As a general overview of case and death prediction performance in the United States, one can study the agreement between four-week cumulative case and death prediction trajectories and the corresponding truth data. By plotting these trajectories on top of each other, we can visually determine where the model most significantly strays from the truth data. In turn, we can determine the strength of the model's predictions given various local conditions, such as sharp increases or decreases in the truth data curve slope. Fig. 4.1 provides a trajectory comparison for both cases and deaths in the United States for a sample of weekly predictions made over the span of 86 days, from November 22, 2020, to February 16, 2021.

Qualitatively, Fig. 4.1 demonstrates the COVID-19 model's tendency to under-predict both cases and deaths at the beginning of sharp truth data slope increases. Likewise, the

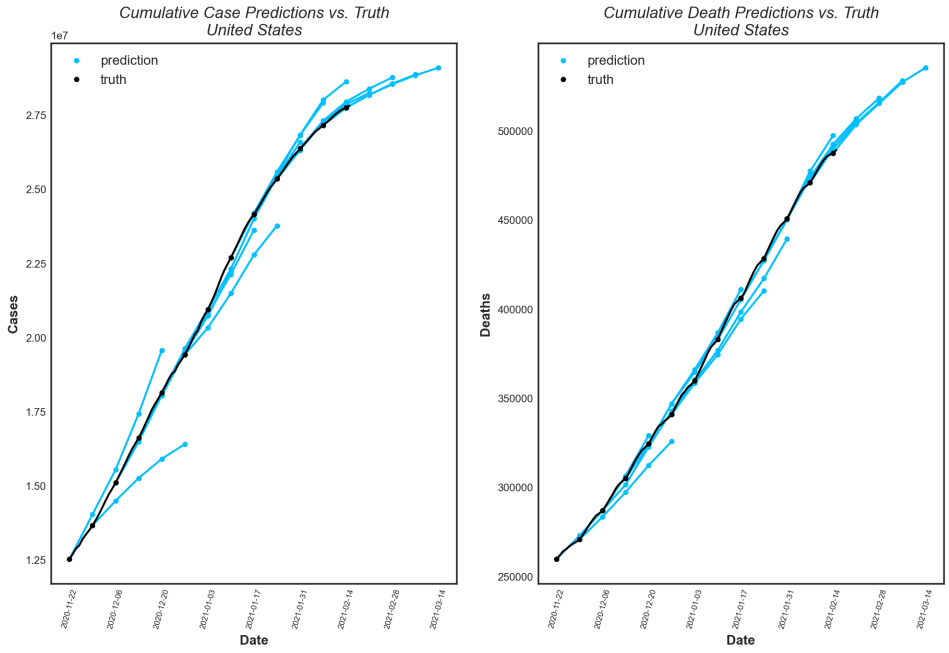

Figure 4.1: Four-week cumulative case and death prediction trajectories versus truth data for the United States. Regions where the predicted trajectories are covered by the truth data curve are the regions with the strongest predictions.

model tends to over-predict at the beginning of sharp slope decreases. This points to a feature of the model, in that it can take up to approximately two weeks for the model to fully catch on to significant changes in case and death trends.

To help us evaluate our handling of this uncertainty, we can gauge the model's performance in terms of accuracy and prediction interval coverage, evaluated using percent errors. For predictions at any level, an ideal model will consistently have errors as close to 0% as possible. As a loose benchmark for model performance, contributors to the COVID-19 Forecast Hub tend to aim for consistent error rates of approximately $\pm 25\%$ to $\pm 50\%$, up to 4 weeks ahead [22]. In addition, a model with ideal prediction interval coverage will have truth data fall within the corresponding 95% prediction interval range 95% of the time.

Fig.s 4.2 and 4.3 below show historical United States case and death prediction accuracies, with upper and lower bounds given by 95% prediction intervals. In general, we see superior performance for death predictions, in that nearly all death prediction errors fall within the 50% error bound. While we see higher average error rates for case predictions, especially up to four weeks ahead, we still see most predictions falling within 50% error and comparable performance compared to death predictions in the one- to two-week range.

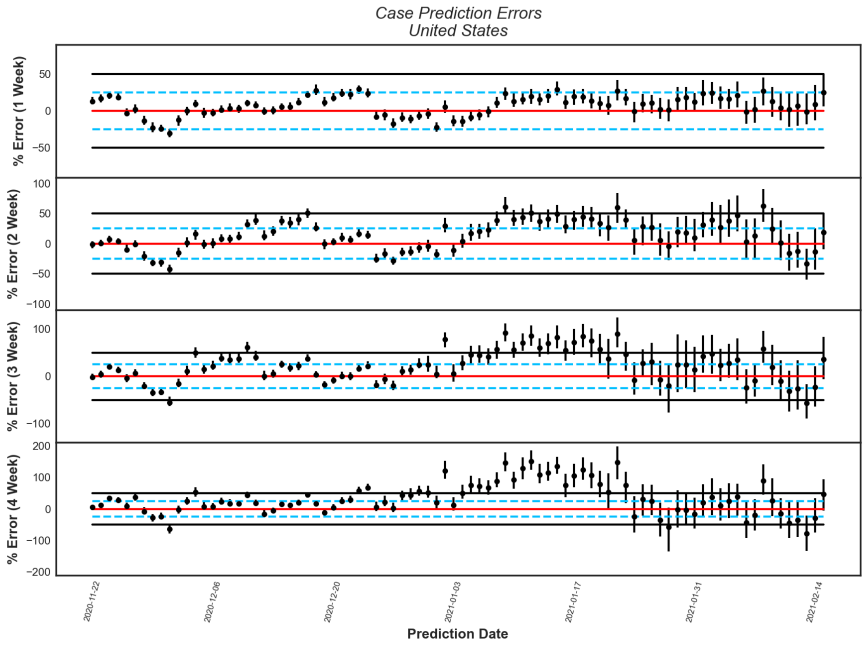

Figure 4.2: Daily one through four week ahead case prediction errors for the United States, by prediction date, with 95% prediction intervals. *Red solid line:* 0% error. *Blue dashed lines:* ±25% error. *Black solid lines:* ±50% error.

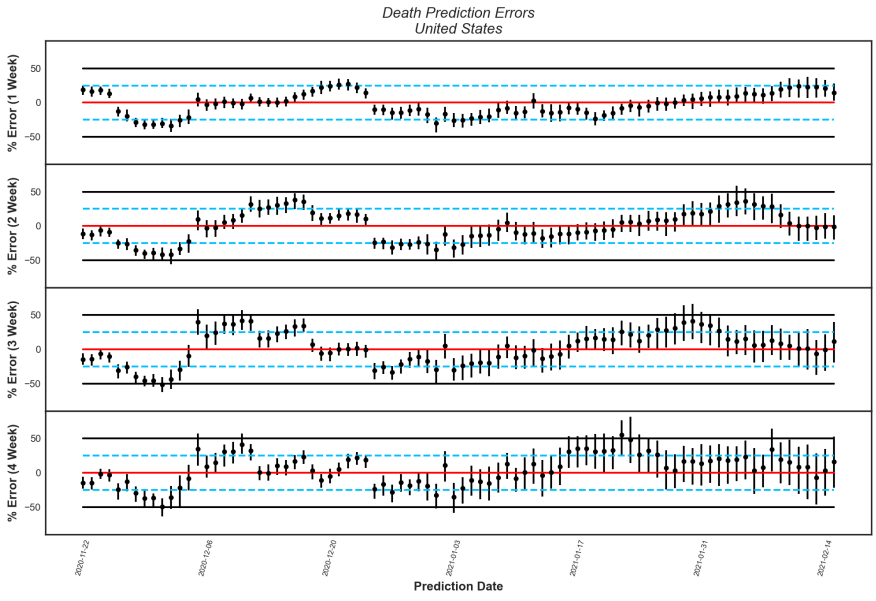

Figure 4.3: Daily one through four week ahead death prediction errors for the United States, by prediction date, with 95% prediction intervals. *Red solid line:* 0% error. *Blue dashed lines:* ±25% error. *Black solid lines:* ±50% error.

The enhanced performance for death predictions is likely the result of deaths being a lagging indicator for COVID-19 spread. As it typically takes several days or weeks for a COVID-19 patient to progress from mild to life-threatening symptoms, proportional incident death increases and decreases tend to lag behind case trajectory shifts. Thus, as we use incident cases as an input for death predictions, cases provide advanced notice of impending death curve shifts.

Beyond accuracy, Fig.s 4.2 and 4.3 depict a trend in 95% prediction interval coverage. As the size of the training dataset increases, or as predictions are made on later dates, prediction interval coverage tends to increase. This increase indicates that, in general, using more training data allows for the model to better gauge its own uncertainty. This points to the benefit of using large training sets to generate predictions, within the timespan of the pandemic. However, modellers must be cautious in choosing which early-pandemic data to include for training, as testing and data reporting inefficiences contributed to low signal-to-noise ratios in the pandemic's early stages.

The data from Fig.s 4.2 and 4.3 can be summarized as:

Table 4.1: Average National Prediction Errors

Prediction Type	1-Week [%]	2-Week [%]	3-Week [%]	4-Week [%]
Cases	12.74	22.83	32.20	45.42
Deaths	13.97	17.95	19.53	18.90

Table 4.2: National Prediction Interval Coverage Rates

Prediction Type	1-Week [%]	2-Week [%]	3-Week [%]	4-Week [%]
Cases	36.05	32.56	38.37	36.05
Deaths	33.72	43.02	45.35	52.33

From these tables, we can determine that the U-M COVID-19 model's predictions are highly accurate at the U.S. national level. For both cases and deaths, average prediction errors are $< 25\%$ up to two weeks ahead and $< 50\%$ up to four weeks ahead. Death error rates are typically even smaller, with average predictions errors $< 20\%$ up to four weeks ahead.

While the model boasts high accuracy for national predictions over this time frame, it exhibits relatively low prediction interval coverage rates. With rates varying from approximately 33% - 52%, case and death truth values fall outside of their corresponding prediction intervals in a majority of cases. While Fig.s 4.2 and 4.3 depict increasing interval lengths with larger training set sizes, training on larger datasets may not suffice for creating high-precision predictions in instances where precision is extremely important, such as hospitalization predictions. This phenomenon points to the usefulness of custom prediction interval definitions for specific targets, based on recipient needs. An example of such custom interval use is detailed below in section 4.2.3.

4.2.2 State Performance

Following suit from the national-level analysis, we can visually assess the COVID-19 model's state-level performance by plotting predicted and truth cumulative case and death trajectories. Analogues of the United States trajectory plots in Fig. 4.1 are provided for the state of Michigan below, in Fig. 4.4.

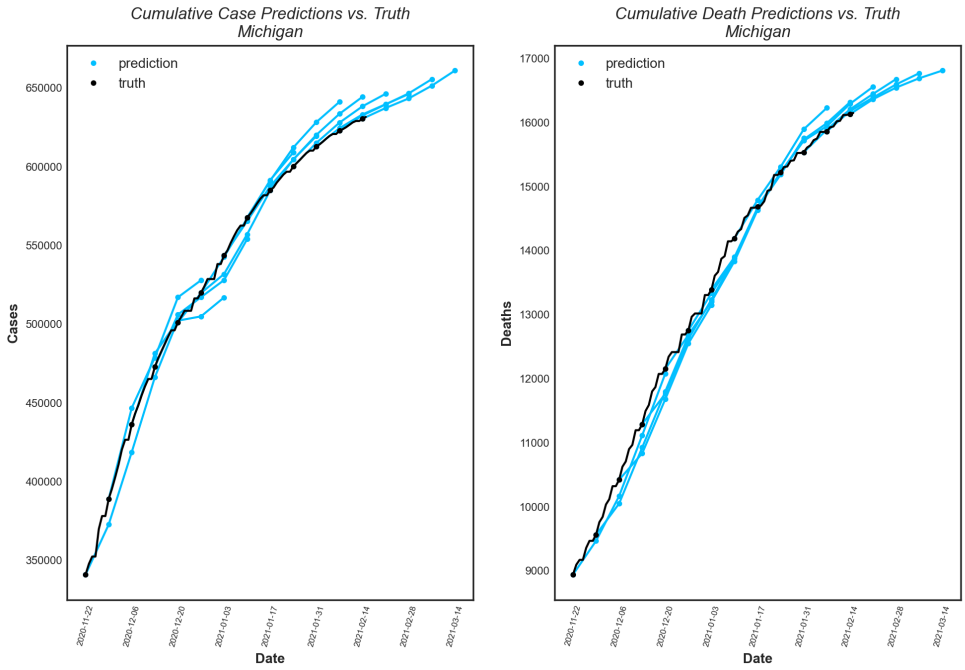

Figure 4.4: Four-week cumulative case and death prediction trajectories versus truth data for the state of Michigan. Regions where the predicted trajectories are covered by the truth data curve are the regions with the strongest predictions.

In comparing Fig.s 4.1 and 4.4, similarities are apparent in terms of the model's over- and under-prediction tendencies. Specifically, as in the case of national predictions, the state-level model tends to under-predict during times of rapid slope increases, and it tends to over-predict during times of rapid slope decreases.

Apart from these similarities, an important distinction exists between the truth curve noise levels. As represented by the jitters in the data between successive weekly truth datapoints, the Michigan data has a significantly higher proportion of noise than the national data. While both datasets experience weekly oscillatory behavior shaped by data reporting patterns, the U.S. curves are smoothed by the contributions of multiple states and territories, each with unique reporting patterns [23]. On the other hand, Michigan's data is shaped exclusively by the state's own reporting patterns, leading to higher reporting-induced noise.

As evidenced by Fig.s 4.5 and 4.6 below, modeled after Fig.s 4.2 and 4.3, this increased noise results in significantly larger 95% prediction intervals and error spectra:

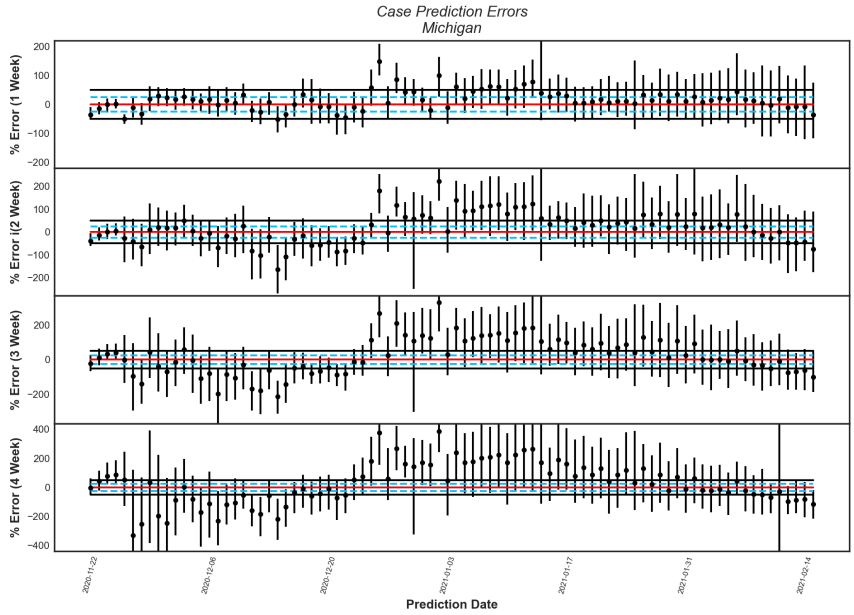

Figure 4.5: Daily one through four week ahead case prediction errors for the state of Michigan, by prediction date, with 95% prediction intervals. *Red solid line:* 0% error. *Blue dashed lines:* ±25% error. *Black solid lines:* ±50% error.

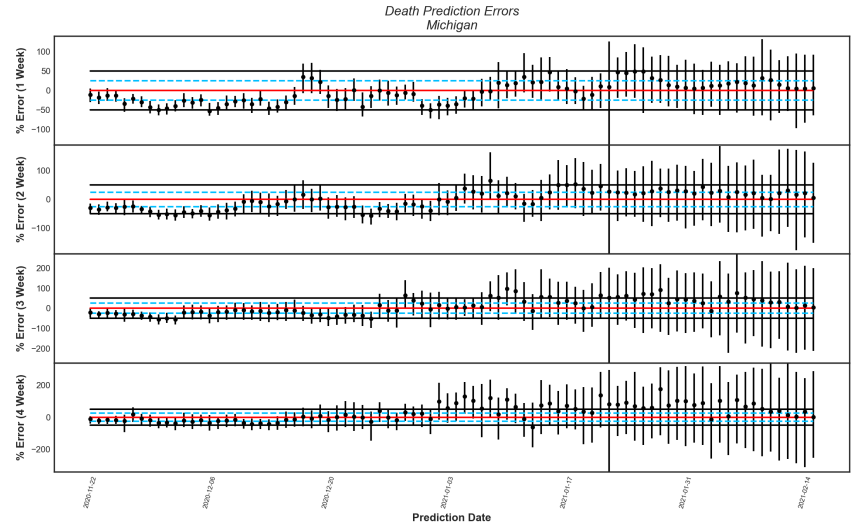

Figure 4.6: Daily one through four week ahead death prediction errors for the state of Michigan, by prediction date, with 95% prediction intervals. *Red solid line:* 0% error. *Blue dashed lines:* ±25% error. *Black solid lines:* ±50% error.

While we see decreased accuracy for state of Michigan predictions, especially in terms of cases, the increased width of the 95% prediction intervals frequently accounts for this uncertainty, in that the intervals often capture the truth data. In addition, while Michigan case prediction errors often stray far outside the 50% error margin up to four weeks ahead, most one-week-ahead case predictions remain within these bounds.

As in the case of the United States predictions, the Michigan death predictions benefit from the information provided by case trajectory changes. Thus, even up to four weeks ahead, most predictions continue to fall within the 50% prediction error margin.

Numerically, these observations can be summarized as:

Table 4.3: Average Michigan Prediction Errors

Prediction Type	1-Week [%]	2-Week [%]	3-Week [%]	4-Week [%]
Cases	27.00	53.91	85.39	116.26
Deaths	22.96	28.41	40.38	55.215

Table 4.4: Michigan Prediction Interval Coverage Rates

Prediction Type	1-Week [%]	2-Week [%]	3-Week [%]	4-Week [%]
Cases	90.70	84.88	77.91	79.07
Deaths	72.09	70.93	77.91	77.91

As observed in Fig.s 4.5 and 4.6, the data in tables 4.3 and 4.4 exhibits a strong contrast with the U.S. national prediction data in tables 4.1 and 4.2. While death predictions for the state of Michigan remain highly accurate, with error rates $< 56\%$ up to four weeks ahead, we see average error rates as high as $\approx 116\%$ up to four weeks ahead for case predictions.

While Michigan case and death predictions are noticeably less accurate than United States case and death predictions, Michigan prediction interval coverage rates are significantly higher, ranging from 70.93% to 90.70%, as opposed to 32.56% to 52.33%. Even further, as exemplified in Fig.s 4.5 and 4.6, prediction interval lengths again tend to increase with increased training set sizes. Thus, interval coverage rates for both cases and deaths increase to nearly 100% around January 1, 2021. This suggests that in the case of predictions based on high-noise state data, the use of large training datasets can help prediction intervals capture uncertainties induced by extra noise.

4.2.3 Health System Performance

Evaluating prediction performance for the Metro, Mercy, and Spectrum Health systems requires a slightly different analysis than the national- and state-level analyses. This is for three reasons. The first is that for health systems, we make COVID-19 census - or patient count - predictions, rather than incident or cumulative hospitalization predictions. While census predictions are particularly important for health system administrators to anticipate COVID-19 resource requirements, they are not conducive to trajectory plots such as Fig.s 4.1 and 4.4. The second reason is that, as we are not directly predicting incident or cumulative hospitalizations, we cannot evaluate prediction accuracy based on weekly incident hospitalization counts the way that we could with cases and deaths.

The third and final reason is slightly more involved, in that in an effort to provide predictions with a broader uncertainty margin, we prepare uncertainties for our health system hospitalization predictions differently than our case and death uncertainties. Namely, rather than using bootstrapping, we provide prediction intervals based on our weakest historical model performance, with intervals given by $\pm 50\%$ up to two weeks ahead and $\pm 100\%$ between two and four weeks ahead. While these intervals are clearly much broader than our case and death intervals, they provide an enhanced opportunity for Metro Health to determine the full spectrum of possible resource demands.

In response to these differences, we can perform a similar accuracy-based analysis as we performed for the national and state levels. However, we can modify it slightly to demonstrate key factors in the predictions. Namely, rather than plotting weekly incident prediction errors, we can plot weekly mean percent errors, where each week's mean percent error (MPE) is computed as:

$$\text{MPE} = \sum_{i=1}^{7} \frac{|\text{prediction}_i - \text{truth}_i|}{7 \times \text{truth}_i}, \tag{4.2}$$

where each prediction_i and truth_i respectively represents the i^{th} daily predicted and truth values for that week.

Now, we seek two indicators of strong model performance: near-0% MPEs and prediction intervals with minima at 0%. If a given week's MPE interval reaches 0%, that means that each individual day's prediction interval included the corresponding truth data. Thus, an ideal model will have all interval minima at 0%. Fig.s 4.7 and 4.8 provide MPE plots for predictions created over the span of 86 days, from November 22, 2020, through February 16, 2021. As in the national and state analyses, all predictions were made using versions of the model trained only on data available prior to the prediction date.

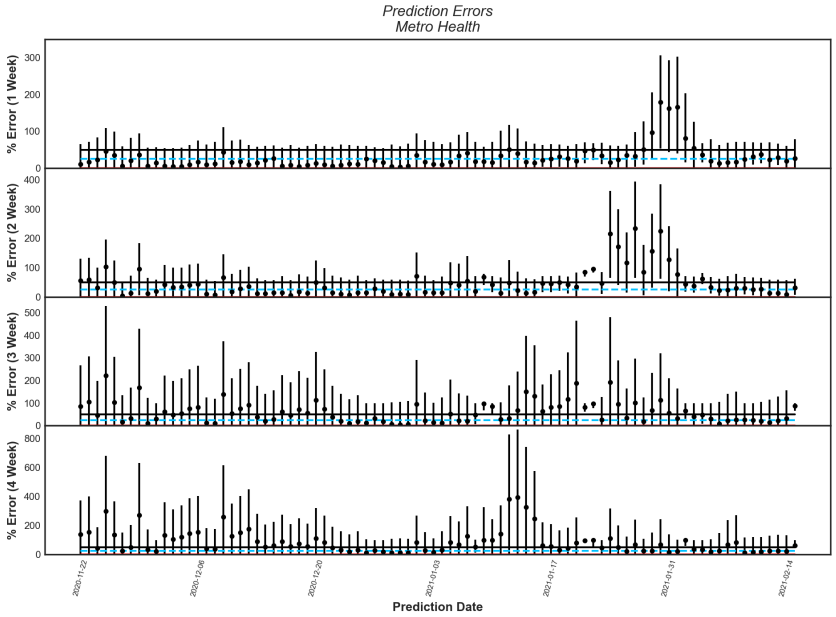

Figure 4.7: Daily one through four week ahead COVID-19 census prediction errors for the Metro Health system, by prediction date, with custom prediction intervals. *Black solid line:* 50% error. *Blue dashed line:* 25% error.

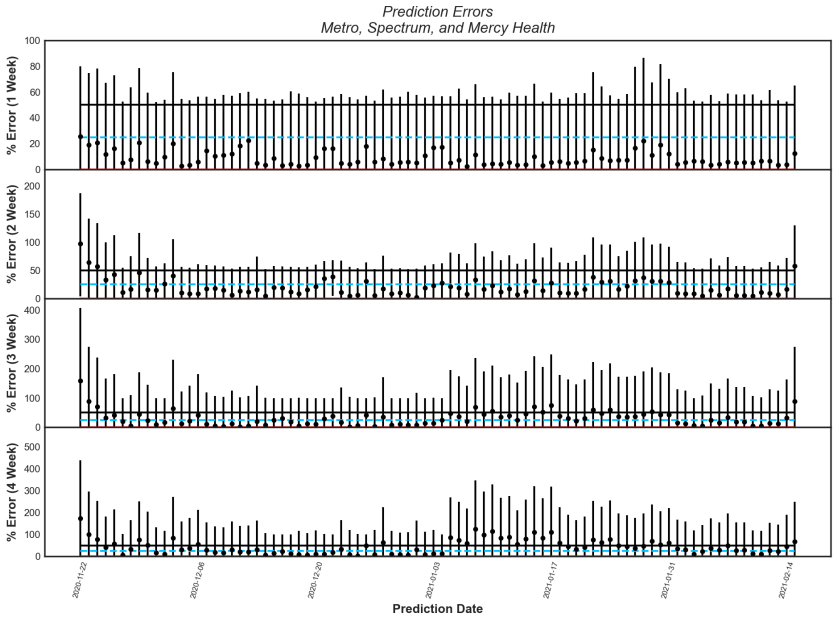

Figure 4.8: Daily one through four week ahead COVID-19 census prediction errors for the Metro, Spectrum, and Mercy Health systems, by prediction date, with custom prediction intervals. *Black solid line:* 50% error. *Blue dashed line:* 25% error.

In general, we see much stronger performance from predictions made for the combination of Metro, Spectrum, and Mercy Health. Specifically, we see that the majority of predictions up to four weeks ahead fall within the 50% MPE margin. In addition, we see that virtually every MPE interval has a minimum of 0%, meaning that the model's prediction intervals capture the health systems' truth data nearly 100% of the time. In contrast, while most Metro-specific predictions fall within the 50% error margin up to two weeks ahead, we see significantly higher error rates between three and four weeks ahead, and for specific days along the prediction curve.

This discrepancy has two main sources, the first of which stems from the nature of MPE as a metric. In Fig. 3.2, it is clear that Metro Health's COVID-19 census counts are typically small, on the order of 10%, compared to the combined health system census. Thus, when MPE is calculated relative to the Metro Health truth data, a marginal difference between predicted and truth data may produce an extremely large percent error. For example, if the model predicts a two-patient COVID-19 census for a day in which Metro Health has only one COVID-19 patient, this will register as a 200% error. While MPE generally scales well based on a dataset's order of magnitude, extreme cases like this can contribute to high MPEs for relatively strong predictions.

The second discrepancy source is slightly more complicated, as it arises from the relationship between Metro Health and the other two hospital systems. Based on an individual system's resources and capacity for COVID-19 patient care, its proportion of COVID-19 patients relative to the other two systems may increase or decrease dramatically. This phenomenon is apparent in Fig. 3.2, in which the Metro Health COVID-19 census increases intermittently following February 8, 2021, whereas the combined census for the three systems consistently decreases. In instances like these, the U-M COVID-19 model may recognize decreases in both Kent County COVID-19 cases and the Metro Health COVID-19 census and subsequently predict a decrease in the Metro census. However, since the model has no way to predict human-determined changes in patient proportions, it may fail to predict an increase.

Together, these explanations of the performance discrepancy between Metro-specific and combined health system predictions underscores the benefits of producing both sets of predictions. If the health professionals at Metro Health anticipate an increase or decrease in their proportional COVID-19 patient share, they can use the combined health system forecast to determine their anticipated patient counts. Together, the observations in Figs. 4.7 and 4.8 can be summarized as:

Table 4.5: Average Health System Prediction Errors

Prediction Type	1-Week [%]	2-Week [%]	3-Week [%]	4-Week [%]
Metro Health	27.85	44.75	58.22	83.14
Combined Systems	8.75	19.68	30.20	43.53

Table 4.6: Health System Prediction Interval Coverage Rates

Prediction Type	1-Week [%]	2-Week [%]	3-Week [%]	4-Week [%]
Metro Health	70.93	62.79	89.53	90.70
Combined Systems	100	95.35	100	98.84

The data in the tables clearly reflects the visual trends in Fig.s 4.7 and 4.8. Specifically, predictions made for the combination of Metro, Spectrum, and Mercy Health are significantly more accurate than Metro-specific predictions. While average Metro Health COVID-19 census prediction errors range from 27.85% - 83.14%, average combined health system prediction errors range from only 8.75% to 43.53%.

In addition, while Metro Health prediction interval coverage rates are relatively high, reaching as high as 90.70%, predictions for the combined health systems are noticeably more precise, with coverage rates falling no lower than 95.35%. In total, these observations point to not only the benefits of making both sets of predictions, but also of using customized prediction interval definitions in cases where high precision is pertinent for resource allocation.

CHAPTER 5

Conclusions and Ongoing Investigations

5.1 Performance Conclusions

Based on the results presented in Chapter 4, we can determine that the U-M COVID-19 model is a reliable epidemiological tool in multiple contexts, from the national down to the regional levels. Specifically, we find that at the national level, the model produces high-accuracy case and death predictions with error rates $< 46\%$ up to four weeks ahead. Similarly, in a high-noise state-level context, it produces case and death predictions with heightened precision and error rates $< 117\%$ up to four weeks ahead. Finally, at the regional health system level, the model produces individual hospital system census predictions with average error rates $< 84\%$ and multi-hospital system census predictions with average error rates $< 44\%$.

While we observe relatively low prediction interval coverage rates for national-level case and death predictions, with average rates as low as 33.72%, we find that the model tends to produce higher-coverage prediction intervals when trained using high-noise datasets, as in the case of United States versus Michigan case and death predictions. Based on this accuracy-precision tradeoff between national and state predictions, the model has proven adaptable to varying noise conditions, with interval coverage increasing in tandem with decreasing prediction accuracy.

In addition, we find that the use of relatively large training datasets and the introduction of custom prediction interval definitions can help the model achieve up to 60%-100% coverage rates, as in the case of our system-level hospitalization predictions. While the custom bounds for the combined system-level census predictions are broader than necessary, in that they produce $> 95\%$ coverage, they are appropriate for high-risk situations in which underallocation of resources like hospital beds can lead to life-threatening consequences.

From our observations of the model's accuracy and precision, we conclude that the model is a reliable tool for predicting pandemic trajectories in terms of cases, hospitaliza-

tions, and deaths for virtually any region within the United States. The model's adaptability is a great asset, in that its accuracy-precision relationship varies based on the data used to train it, and simple modifications can be implemented to create enhanced, target-specific uncertainty computations.

5.2 Building Upon the Model

While the U-M COVID-19 model has proven to be a reliable stand-alone resource for disease forecasting, it has also provided a foundation for the further involvement of machine learning in epidemiological forecasting. With its flexibility with regard to inputs, the model can be adjusted to make predictions for virtually any illness in any global region with minimal tweaking of equation (2.9). For example, a potential future use for the model is the creation of annual influenza epidemic predictions within a variety of countries.

In addition, the script used to generate the U-M COVID-19 model's predictions can support the use of any machine learning algorithm to make its predictions - not just ridge regression. Thus, as of March 2021 we are actively investigating the effects of using neural networks, rather than ridge regression, to generate predictions. A comparable investigation of any other algorithm simply requires that a user import their desired model package in Python and replace any Ridge() or RidgeCV() instances with their preferred method. Through comparisons of algorithms such as decision tree and lasso regressors, an optimal trade-off between time efficiency and prediction accuracy can be determined for diverse prediction contexts.

As it stands, the script can also be used to address fundamentally different questions than we have previously explored. For example, continuing with COVID-19 forecasts, can we train a single model to make predictions for any global region, rather than training independent regressors for each location of interest? Or, can we create category-specific regressors geared towards adjacent regions or locations with similar population characteristics? By exploring categorical prediction schemes based on different shared characteristics, underlying regional relationships can be exploited and used towards the creation of even stronger predictions.

Entirely beyond COVID-19 forecasting, the model and its statistical methods can be repurposed for studies in numerous other fields in which ridge regression has shown predictive promise. For example, as explored in [24], large-scale quantitative genetics analyses can involve large enough sample sizes that ridge regression provides substantially higher prediction accuracies than the classical approach. In an entirely different direction, the American Psychological Association Dictionary of Psychology describes ridge regression

as a method to determine whether certain independent variables can be removed from an analysis [25]. Thus, by changing the inputs of the COVID-19 model to be independent variables in a psychological study, one can study the linear combination coefficients assigned to each variable and eliminate those with near-0 coefficients.

By expanding upon the model's foundations, future researchers in epidemiology, genetics, psychology, and countless other fields will have the opportunity to create robust prediction and analysis mechanisms tailored to the demands of their fields. Through this repurposing and adaptation process, the U-M COVID-19 model and others like it have the potential to contribute to the amplification of machine learning as a tool for mapping and understanding even the most subtle trends in the systems around us.

Bibliography

[1] K. Albertsson et al., *Machine learning in high energy physics community white paper*, 2019.

[2] E. Y. Cramer et al., "Evaluation of individual and ensemble probabilistic forecasts of covid-19 mortality in the us", medRxiv, 10.1101/2021.02.03.21250974 (2021).

[3] E. Cramer et al., *Covid-19 forecast hub: 4 december 2020 snapshot*, version v1.2 (Zenodo, Dec. 2020).

[4] S. Corsetti et al., *Covid-19 collaboration repository*, (Accessed 15 March 2021), https://gitlab.com/sabcorse/covid-19-collaboration.

[5] D. Smith and L. Moore, "The sir model for spread of disease - the differential equation model", Journal of Online Mathematics and its Applications **4**, Accessed: 2021-03-12.

[6] *Seir and seirs models*, (Accessed 17 March 2021), https://docs.idmod.org/projects/emod-hiv/en/latest/model-seir.html.

[7] Y.-C. Chen, P.-E. Lu, C.-S. Chang, and T.-H. Liu, "A time-dependent sir model for covid-19 with undetectable infected persons", IEEE Transactions on Network Science and Engineering **7**, 3279–3294 (2020).

[8] V. B. Ramirez, *What is r0? gauging contagious infections*, (Accessed 25 March 2021), https://www.healthline.com/health/r-nought-reproduction-number#meaning.

[9] I. V. Tetko, D. J. Livingstone, and A. I. Luik, "Neural network studies. 1. comparison of overfitting and overtraining", Journal of Chemical Information and Computer Sciences **35**, 826–833 (1995).

[10] A. E. Hoerl and R. W. Kennard, "Ridge regression: biased estimation for nonorthogonal problems", Technometrics **12**, 55–67 (1970).

[11] *Charts showing spike in michigan's coronavirus cases are misleading*, (Accessed 14 March 2021), https://www.freep.com/story/news/local/michigan/2020/06/11/coronavirus-covid-19-cases-spike-data/5334492002/.

[12]"Leave-one-out cross-validation", in *Encyclopedia of machine learning*, edited by C. Sammut and G. I. Webb (Springer US, Boston, MA, 2010), pp. 600–601.

[13]C. C. Testa, N. Krieger, J. T. Chen, and W. P. Hanage, "Visualizing the lagged connection between covid-19 cases and deaths in the united states: an animation using per capita state-level data (january 22, 2020 – july 8, 2020)", The Harvard Center for Population and Development Studies (HCPDS) Working Paper **19** (2020).

[14]D. E, D. H, and G. L, "An interactive web-based dashboard to track covid-19 in real time.", Lancet Infect Dis. **20**, 533–534 (2020).

[15]*Google covid-19 community mobility reports*, (Accessed 15 March 2021), https://www.google.com/covid19/mobility/.

[16]F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: machine learning in Python", Journal of Machine Learning Research **12**, 2825–2830 (2011).

[17]J. Goeman, "L1 and l2 penalized regression models", 18–19 (2009).

[18]S. Kumar and A. N. Srivistava, "Bootstrap prediction intervals in non-parametric regression with applications to anomaly detection", in (2012).

[19]D. S. Nielsen, *Bootstrapping prediction intervals*, (Accessed 16 March 2021), https://saattrupdan.github.io/2020-03-01-bootstrap-prediction/.

[20]T. Hastie, R. Tibshirani, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction* (New York: Springer, 2001).

[21]*Forecast evaluation dashboard*, (Accessed 26 March 2021), https://delphi.cmu.edu/forecast-eval/.

[22]S. McConnell, *Covid complete data center*, (Accessed 17 March 2021), https://stevemcconnell.com/covidcomplete/.

[23]A. Bergman, Y. Sella, P. Agre, and A. Casadevall, "Oscillations in u.s. covid-19 incidence and mortality data reflect diagnostic and reporting factors.", mSystems **5** (2020).

[24]J. Wang, R. de Vlaming, and P. J. F. Groenen, "The current and future use of ridge regression for prediction in quantitative genetics", BioMed Research International **2015**, 10.1155/2015/143712 (2015).

[25]*Ridge regression*, (Accessed 14 March 2021), https://dictionary.apa.org/ridge-regression.