Final Report

Ben Stan, Yongzhe Wang, Fred Yu
Winter 2022

Project: Applications of Nowcasting Methods to Notifiable Disease Surveillance
Sponsor: Washington State Department of Health
Contact: Ian Painter

Abstract

The Covid-19 pandemic has been the focus of public health entities since the Winter of 2020. Monitoring cases, hospitalizations, and vaccinations is of critical importance to inform health policies and mitigate the spread of the disease. The Washington State Department of Health (DOH) performs this function and reports their tracking of the disease on publicly available websites. Among other metrics, the organization is concerned with tracking patient hospitalizations by the date of infection to monitor disease spread and resource utilization. Due to various delay sources, this requires the DOH to estimate the total count of hospitalizations associated with a given day of infection using partial data. Our analysis implemented Bayesian frameworks to perform this prediction. We used several methods defined by Michael Höhle and Matthias an der Heiden and Sarah F. McGough et al. and compared performance to determine which model should be implemented by the DOH. Additionally, we defined a set of parameters to guide the use of the model and ensure proper outcomes. The goal is that this method can lead to better-informed decisions by the department and improved health outcomes in the state.

Introduction

Background

During the Covid-19 pandemic, the Washington State Department of Health has collaborated with local jurisdictions to collect data on all tests and hospitalizations associated with the disease. The DOH prioritizes the tracking of hospitalizations due to the impact on occupancy rate and downstream public health effects. To understand the spread of the disease, the department associates each hospitalization with the date of patient infection, which for simplicity is equated to the date of a positive Covid-19 test. This metric presents unique challenges due to the latency between contraction of the disease and the time that a patient is hospitalized. The Mayo Clinic estimates that the time from a viral load sufficient for a positive test to hospitalization is 10-12 days for most patients.[1] However, many patients are not tested until they display severe symptoms, causing data variability. Once reporting delays are included, the delay between test collection and hospitalization can range from 1-21 days. The DOH is interested in obtaining estimates for hospitalizations before the end of this 21 day period to inform proactive policy decisions. Estimating the full count of hospitalizations from partial data during this delay period, a process known as nowcasting, was our central problem.

[1] https://www.mayo.edu/research/remote-monitoring-covid19-symptoms/people-with-covid19#:~:text=While%20the%2010%20to%2012,they%20may%20predict%20hospitalization%20later.

Objective

The central objective of this analysis was to estimate hospitalization counts in the three week period in which only partial data is available. The secondary interest was to evaluate model performance with respect to several model candidates in order to identify the highest-performing model and the parameters for its use. The metric that was used to evaluate the model and compare it to alternative methods is relative root mean squared error. This metric is computed as such:

$$rRMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\frac{y_i - x_i}{y_i})^2|}$$

In this equation, y is the actual value, x is the predicted value and n is the total number of predictions. A relative measure of error was chosen due to the variability in values of hospitalization counts over the course of the pandemic. By using a relative value, it is possible to compare the performance of the model at different time points with different levels of hospitalization. We evaluated several different techniques to predict full hospitalization counts and to determine which technique is most accurate.

Study Population

The population of interest was subjects who were hospitalized with Covid-19 in the state of Washington. The data sources used by the WA State DOH provide testing results and hospitalizations separately. Records were joined by patients' first and last name along with date of birth. While it is possible to obtain information for all tests administered in the state, the priority was those that are associated with a hospitalization. The record of hospitalization occurs on or after the date of the testing, as patients are tested upon entry into the hospital.

Methodology

Through our research, we identified two nowcasting approaches in the context of disease surveillance. One approach is the most established and cited (implemented in the *surveillance* R package), and the other suggests improvements upon the first model. Both methods have published software packages in R. For our project, we applied the models to our data and evaluated their performance based on the primary error metric (relative root mean squared error), while examining the underlying mechanisms based on our context. In this section, we briefly explain the main ideas of each approach and their implementation.

Framework and Notation

The term nowcasting refers to approaches that predict the total counts of the occurrence of an event in the near future, but are unavailable at the current time due to reporting delays. The nature of nowcasting problems is different from the more familiar forecasting, in which the aim is to predict what will happen in the future and has not yet occurred.

Previous nowcasting work originated from actuarial literature to address the issue of reporting delays of insurance claims. The idea was first explored in the context of disease surveillance in 1994 by J.F. Lawless[2] to study the reporting delays in AIDS cases; this is due to the latency from HIV infection to the development of AIDS. The two papers of interest adapted the statistical framework of the occurred-but-not-reported-events (OBNR events) problem of Lawless (1994). The setting is illustrated as follows:

[2] https://www.jstor.org/stable/3315820

Variable name	Description
t	The time when the case actually occurred; from 0 to T (current time).
d	The delay time; from 0 (no delay) to D (maximum delay).
n(t,d)	The number of cases occurred at time *t* and reported after *d* days of delay.
N(t,T)	The number of cases occurred at time *t* and reported until the current time T; This information is assumed to be available to us. It is the cumulative sum of n(t,d).

Table 1. Abbreviations and Definitions

The delay times are assumed to occur only up to a maximum delay time *D*, which is reasonable because cases with larger delays provide information too far from the current time to be relevant. In particular, the number of cases is unknown to us when $d > T - t$, or equivalently, when $T < t + d$ (when the reporting time point is beyond the current time). Hence, the data that is available to us is assumed to be in the region illustrated as the right-angled trapezoidal region in the reporting triangle in Figure 1.

$$A_T^m = \{(t,d) : max(T-m,0) \leq t \leq T, 0 \leq d \leq min(D, T-t)\}$$

Denoted by *N(t,T)*, the number of cases occurred on time *t* and reported until the current time *T*, is a cumulative count of cases on time *t* and reported on each day up to the current time *T*, i.e.

$$N(t,T) = \sum_{d=0}^{min(T-t,D)} n_{t,d}$$

Given *N(t,T)*, the aim of nowcasting work is to predict the total number of cases that occurred at time *t*, with any delays up to the maximum *D,* denoted *N(t,t+D).*

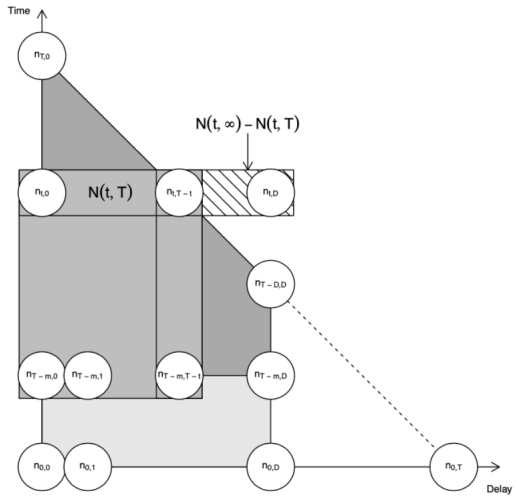

Figure 1. Reporting triangle from *Höhle, der Heiden (2014).*

An alternative visualization for nowcasting is shown below. Each of the dark purple squares corresponds to a known n(t,d) value, the number of cases occurring on day t with d days of delay. In this example, the maximum delay of interest (D) is five days. The information that is estimated using nowcasting is shown in the light purple squares. Note that the total number of hospitalizations up to the maximum delay of interest, N(t,t+D), is obtained by taking the sum of each row in the example.

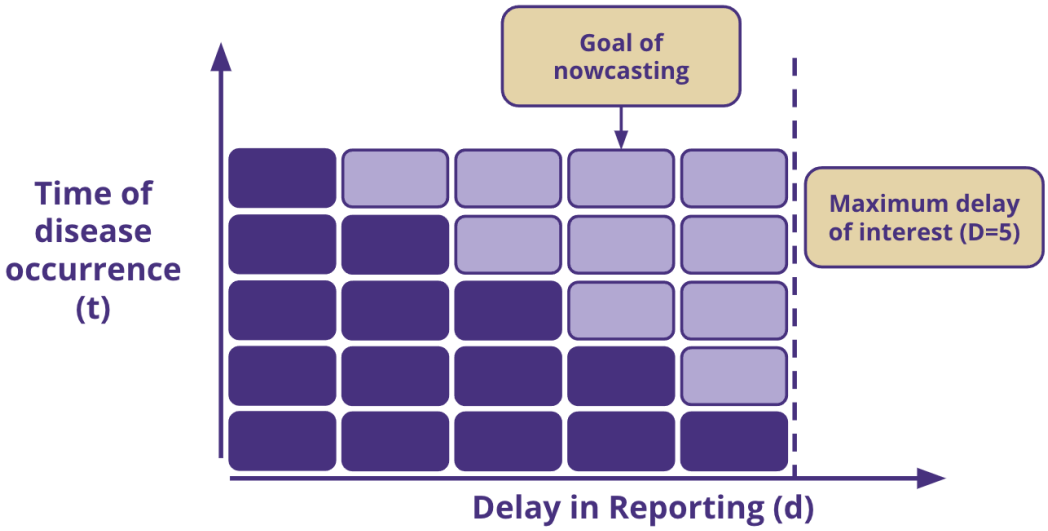

Figure 2. Alternative representation of the reporting triangle.

Next, we present the main ideas of the two delay-adjusting methods we used for this project. Both approaches model the full case count by assuming an underlying in-homogeneous Poisson process of the case count, where the Poisson mean at each time point and delay time is dependent on the epidemic curve that represents the natural disease occurrence rate and the probability of the case being delayed.

Model 1: Bayesian Nowcasting by *Michael Höhle and Matthias an der Heiden* (HH)[3]

The first approach we examined, proposed by Michael Höhle and Matthias an der Heiden (2014), is usually referred to as the HH model. We chose this approach since it is the most established approach with an existing implementation as the *nowcast* function in the R package *surveillance.*

Key Ideas

- The HH approach is a Bayesian hierarchical model with the following hierarchy:

$$\lambda_t \sim \Gamma(\alpha_\lambda, \beta_\lambda)(\text{prior}),$$
$$N(t,\infty)|\lambda_t \sim Pois(\lambda_t),$$
$$p_{t,d}|N(t,T) = p_d|N(t,T) \sim GD(\mathbf{a_{d,T}}, \mathbf{b_{d,T}})$$
$$N(t,T)|N(t,\infty), q_{T-t} \sim Bin(N(t,\infty), q_{T-t})$$
$$\text{where } q_{T-t} = \sum_{d=0}^{T-t} p_{t,d}$$

- The delay distribution is assumed to be a time-invariant probability mass function (a vector of cell probabilities, for each delay *d*) ***p(t,d) = p(d)*** for all time points within the time window.
- Assume ***n(t,d),*** the number of cases occurring at *t* but reported after *d* days of delay, follows multinomial distribution given the size ***N(t, t+D)*** and cell probabilities ***p(t,d)***.
- Using a Dirichlet prior on the delay distribution, if we take the column sums in the reporting triangle from the multinomial sampling, then we would have the posterior delay distribution to be again Dirichelet distributed (GD).
- By conjugate prior-posterior updating, we would continue updating the delay distribution based on

[3] https://pubmed.ncbi.nlm.nih.gov/24930473/

N(t, T) as we move along the time axis towards the current time ***T***.

- Given the $q_{T-t}(t)$, the proportion reported up to current time ***T,*** and the full count, the observed count ***N(t, T)*** follows binomial distribution. This gives us ***f(N(t, T) | N(t, infty), q).***
- Finally the output of the HH method is the posterior distribution of the full count given the partial count we observed, by the following formula:

$$f(N(t,\infty)|N(t,T)) = \int_0^1 f(N(t,\infty)|q_{T-t}, N(t,T)) f(q_{T-t}|N(t,T)) dq_{T-t}$$

where by application of the Bayes' theorem (denote the full count as N_t for simplicity)

$$\begin{aligned} f(N_t|q_{T-t}, N(t,T)) &= f(N(t,T)|q_{T-t}, N_t) f(N_t|\lambda_t) \\ &= \binom{N_t}{N(t,T)} q^{N(t,T)} (1-q)^{N_t - N(t,T)} * \binom{N_t + \alpha - 1}{N_t} p^{\alpha} (1-p)^{N_t} \end{aligned}$$

$$\text{where } p = 1/(1+\beta_\lambda)$$

Although we know the distribution of the delay PMF vector by GD-Multinomial conjugacy, ***f(q, N(t, T))*** does not have an analytical form. We therefore solve the integration by Monte Carlo sampling jointly for all q_{T-t} for all ***t = T-D, …, T.***

Model 2: Nowcasting by Bayesian Smoothing (NobBS)[4]

We refer to this approach, proposed by McGough et al. (2020), as NobBS, as it is referred to in the original paper. We chose this method as it is generally new, iterated upon the HH approach, and has been shown to outperform the HH because of its ability to capture the historical case count in modeling the epidemic curve.

Key Ideas

- Similar to the HH method, NobBS assumes the ***n(t,d)*** to occur in an underlying Poisson-process.
- The model is developed based on the following hierarchy:

$$n(t,d) \sim Pois(\mu_{t,d}) \qquad \textit{(Poisson case count)}$$
$$log(\mu_{t,d}) = \alpha_t + log(\beta_d)$$

$$\alpha_{t=1} \sim N(0, 0.001) \qquad \textit{(Epidemic signal)}$$
$$\alpha_{t\geq 2} \sim N(\alpha_{t-1}, \tau_\alpha^2)$$

$$\beta_d \sim Dir(\theta) \qquad \textit{(Time-homogeneous Delay distribution)}$$

- Assume a log-linear association of Poisson mean with the epidemiological signal and the delay distribution.
- The epidemiological signal is modeled as a first-order random walk to capture the temporal evolution of cases over time; this is in contrast to the gamma and spline approaches utilized in HH.
- The reporting probability with *d* units of delay (the delay distribution) is modeled similarly using the Dirichlet-Multinomial conjugacy scheme.

[4] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7162546/

Application of Related Work to WA State DOH Data

We used the methods outlined in the HH and NobBS publications to perform the nowcasting on the WA DOH data. As mentioned in the previous sections, we were mostly interested in predicting the *N(t,t+D),* which is the true count of hospitalization that occurred at a given time point *t*, assumed to be reported within the maximum length of delay *D.* Following discussion with the DOH, the maximum delay of interest was set to 21 days. This threshold was determined following consultation with the DOH around the delays commonly observed in the data. Our main predictor was the partial case count at a given time point and delay length. The above models (and their existing implementations in R) both required time-series data; each case must contain information on the day of the event (e.g. a positive test) and the day of reporting, a structure paralleled in our data.

Model Evaluation

The relative RMSE was used to evaluate the performance of the model. This relative error metric allowed for both comparisons across models as well as comparisons for a given model at different time points and levels of hospitalization. In the former case, the NobBS and HH models were compared to one another to determine which has the greatest performance (minimal error) for a given time period. In the latter case, any given model could be evaluated at periods of high and low hospitalization to determine how performance fluctuates given the stage of the pandemic. Another dimension on which the models were evaluated is lookback time. This is defined as the number of days between the date being estimated (referred to as “onset date”) and the date on which the nowcast is performed (the “now”). This value ranges from 0 to 21, and it is expected that as lookback time increases, error will decrease. This is because the difference between partial and full hospitalization count decreases with increasing lookback.The highest-performing model across all days represented in the data set was selected as the best candidate. Secondary analyses examined the performance of models as a function of the lookback time.

Analysis and Results

Descriptive Analysis

Data provided by the Washington State Department of Health included all hospitalizations occurring between January 12, 2020 and January 13, 2022 along with the date of the associated positive test and the date that the hospitalization was confirmed. The smoothed chart below shows the number of hospitalizations and positive Covid-19 testing results during the study period.

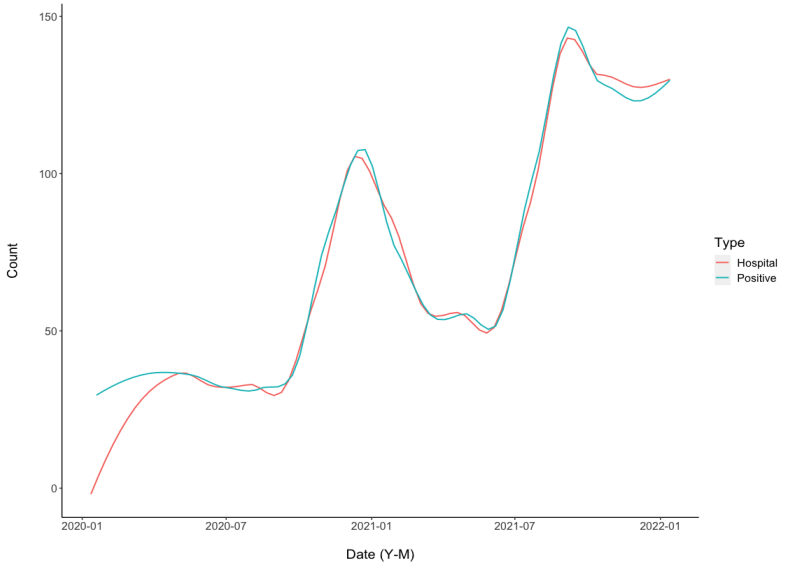

Figure 3. Epidemiological curves fitted by loess with span of 0.37.

The first confirmed Covid-19 case was reported by the CDC on January 20, 2020, and the cases that followed can be visualized by the first spike in counts of hospitalizations and positive tests. The subsequent large wave of cases occurred over the winter months of 2020-2021. The dominant variant during this time was Alpha, and the decline of this wave may be due to vaccinations. Then the Delta variant, which was first identified in India in late 2020, caused a new spike in confirmed cases. The last spike in the figure was caused by the Omicron variant beginning in November 2021. Notably, we can observe that the difference between the number of hospitalizations and the number of positive Covid-19 testing results in a given date was more apparent in spike periods than low volume periods, which indicates that the delay shows a time-varying pattern.

For the purpose of nowcasting, the delay of interest was between positive test collection and the confirmation of hospitalization. Due to a system change, the confirmation data for all hospitalizations before 6/11/2020 was 6/11/2020. This makes the data prior to this date invalid for the purpose of our modeling. As the maximum delay of interest (D) was 21 days, we also needed to limit the data under consideration to 21 days prior to the last date in the data set. Thus, we considered all data by positive test collection date from 6/11/2020 to 12/22/2021. The count of hospitalizations by test collection date is shown below.

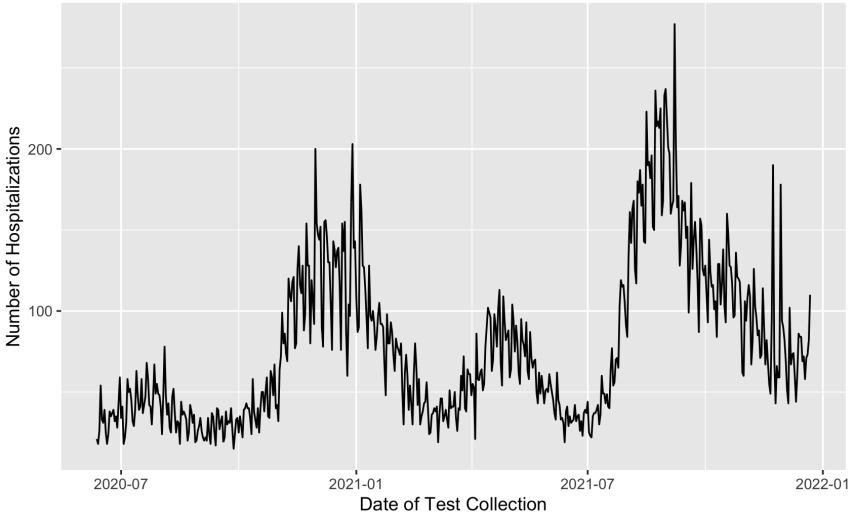

Figure 4. Counts of hospitalizations by date of positive test collection.

The delay distribution was also a critical element in the process of nowcasting. The delay distribution across the full data set is shown below.

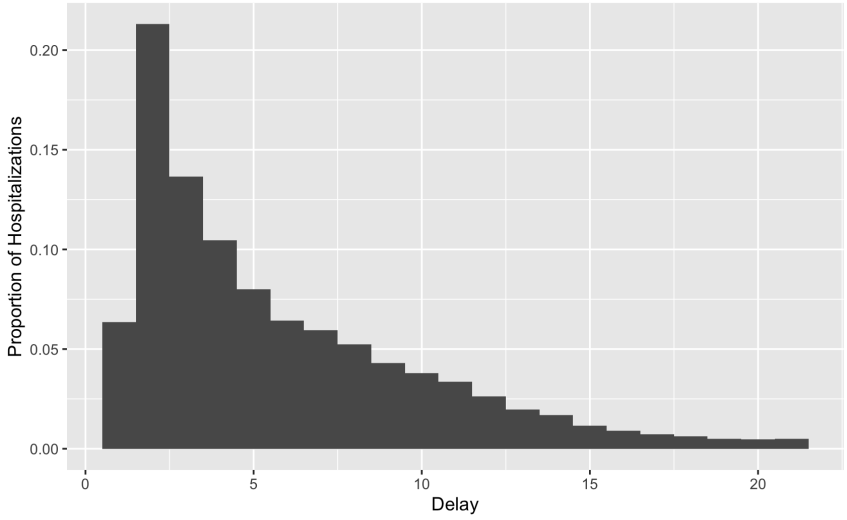

Figure 5. Delay distribution for all hospitalizations in date range of interest.

Note that no hospitalizations were reported on the same date as positive test collection. This is due to an inherent delay in the reporting system and a batch synchronization of data. Hospitalizations were reported with a minimum of one day of delay. The plurality, over 20% of hospitalizations, were reported with two days of delay, and this proportion decreased for all subsequent days up to the maximum delay of 21 days. The delay distribution also changed over the timeframe of interest, as shown by the plot below. The number of days of delay needed to record 50% of hospitalizations associated with a given test date fluctuated in line with the hospitalization volume. The observed time needed ranges from 3 to 5.5 days.

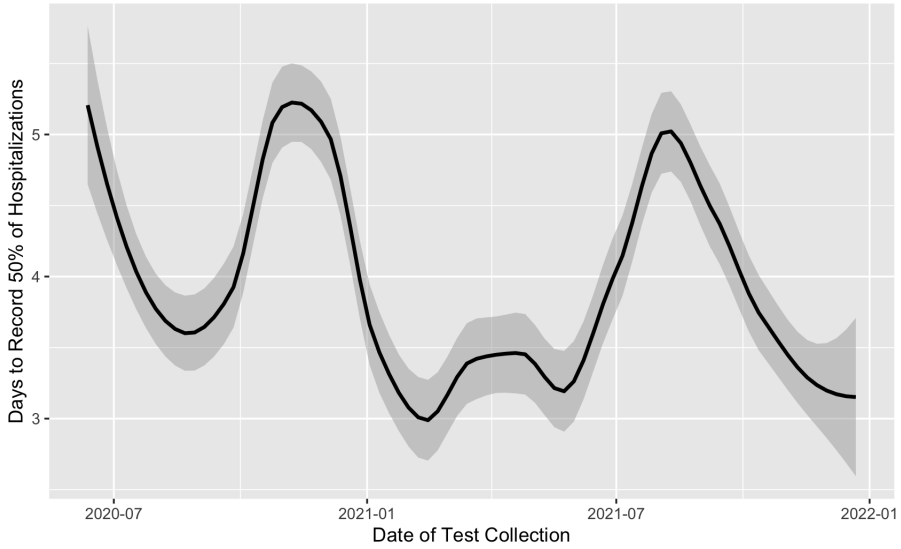

Figure 6. Days needed to record 50% of hospitalizations by date of positive test collection; a loess smoother was used with span of 0.37, and the gray area represents a 95% confidence interval for the curve

Modeling results

The nowcasting results for individual days can be observed and compared to the actual (full) counts. The figure below shows an example.

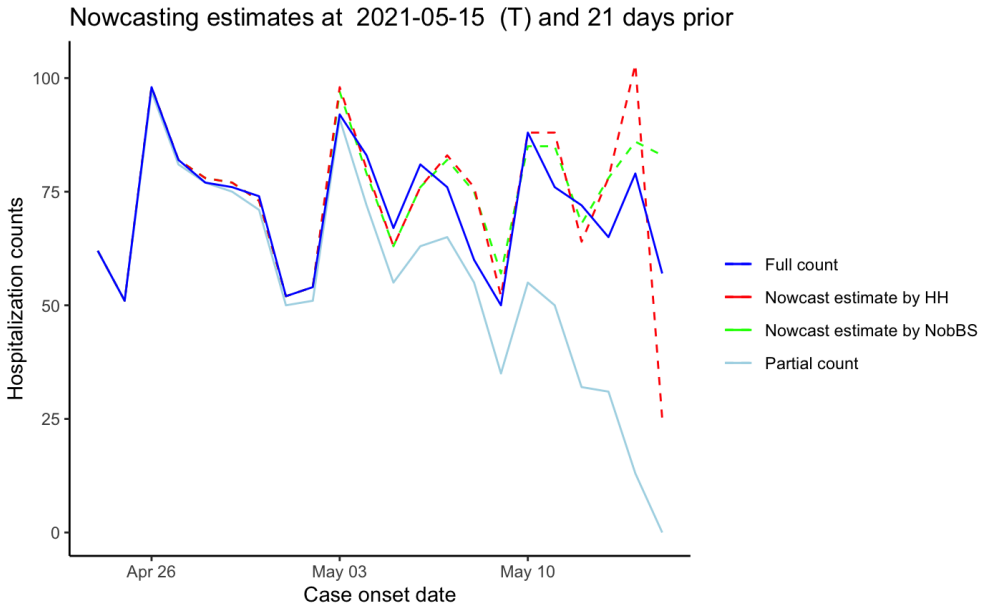

Figure 7. Nowcasting estimates and full and partial counts of hospitalizations for 5/15/2021.

The partial counts in the graph tend towards zero for the most recent days prior to nowcasting. This is expected given that these days have the fewest days of data collected. The performance of the models did not

appear to demonstrate a consistent pattern with respect to the actual counts. The same chart for the nowcast performed on August 15, 2021 is shown below.

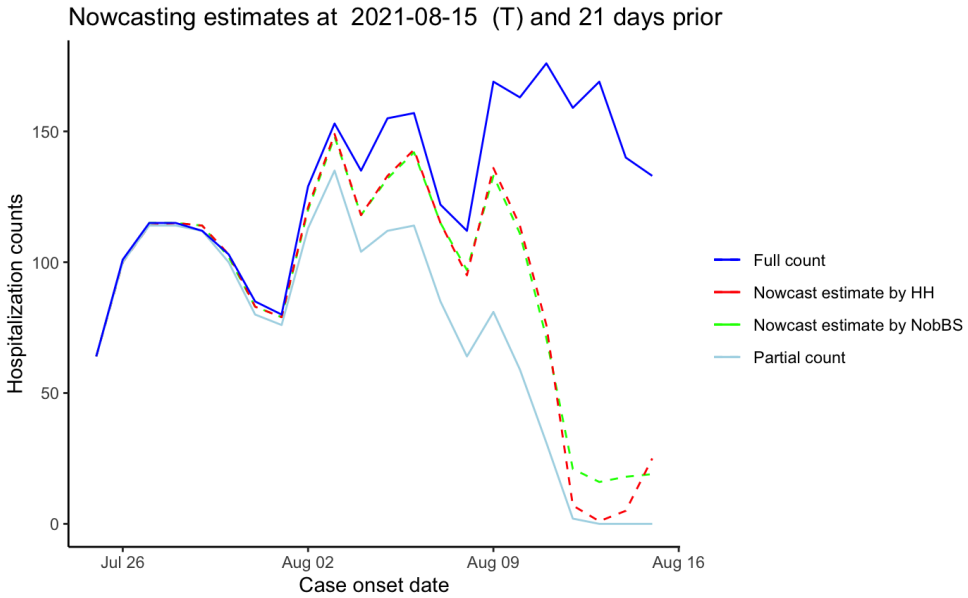

Figure 8. Nowcasting estimates and full and partial counts of hospitalizations for 8/15/2021.

The error of the two models fluctuated over the date range of the data. It can be seen that the error, particularly for the NobBS method, tended to spike in the periods of high volume in late 2020 and summer 2021.

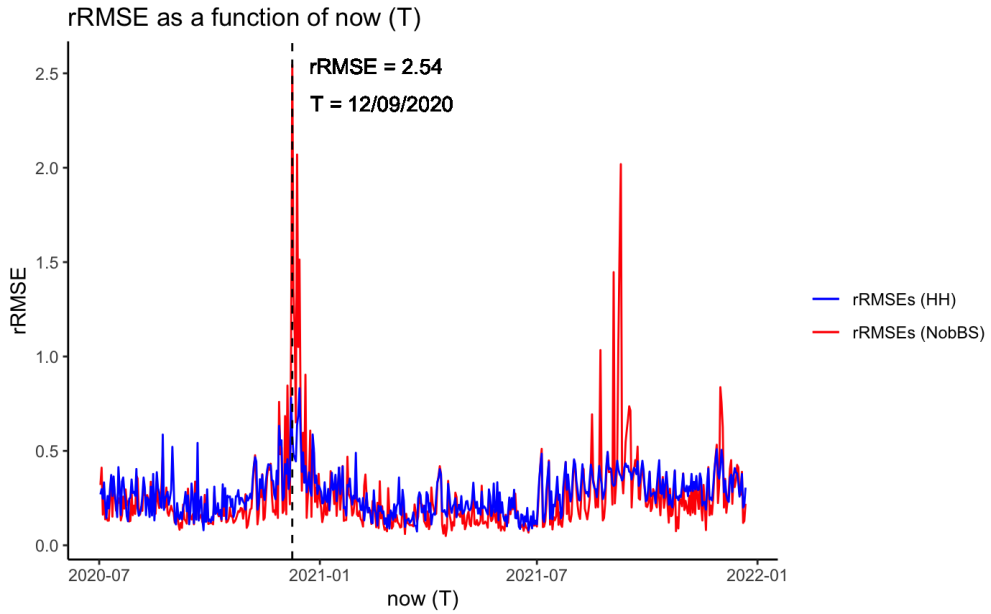

Figure 9. Model performance by date.

The overall performance of the models was determined by considering all estimates generated by the two approaches. The median rRMSE error for the HH method was 0.261, while the median rRMSE for NobBS was 0.197. Thus, the NobBS method appeared to have better performance across all estimates. One additional dimension to condition the results over was lookback time. We calculated that the median error for the HH method for lookback times of five days or greater was 0.089, while for the NobBS method, the same metric was 0.092. The two methods performed comparably in this metric, so we further considered the patterns in the error by calculating the metric by lookback time.

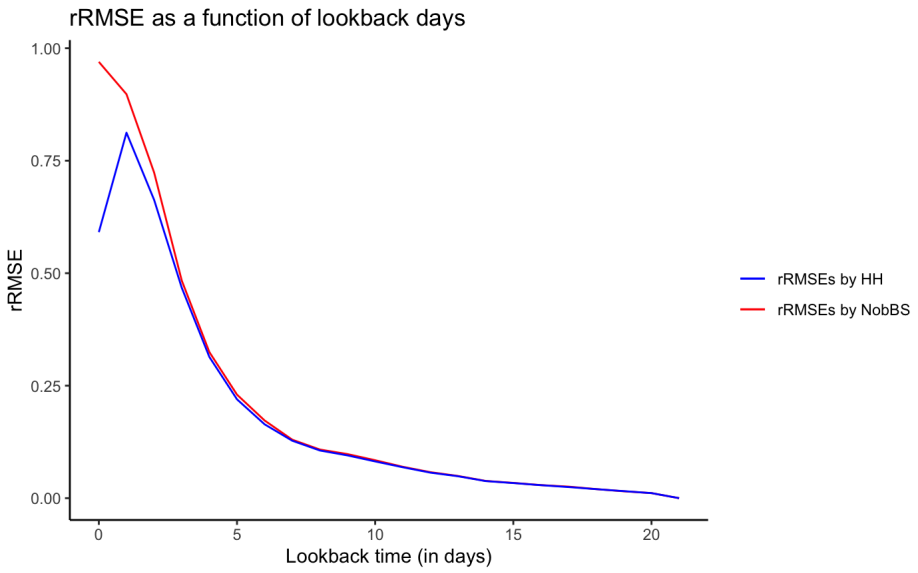

Figure 10. Model performance by lookback time.

As expected, the error tends to decrease with increasing lookback. The HH method appeared to have improved performance at lower lookback times, but we used the criteria of overall error to select the most performant model, which led to the selection of the NobBS model.

One of the goals of this analysis was to determine the parameters under which the model can be used by the Department of Health to expedite decision-making. We identified a threshold of 20% as one that would be acceptable given the natural variability in the hospital counts from day to day. We considered the proportion of estimates within this 20% error threshold and found that by a lookback time of eight, 90% of estimates were within it. This led to the conclusion that the DOH can view estimates as recently as eight days after a date to make decisions regarding hospitalization counts.

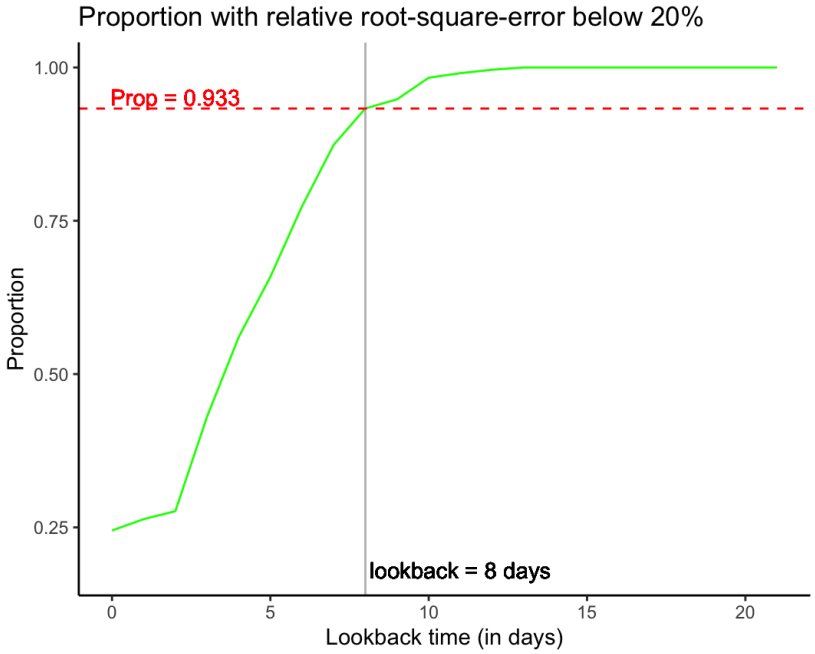

Figure 11. Proportion of estimates with acceptable threshold (20% error) by lookback time.

Impact

The current practice for the WA State Department of Health is to exclude the most recent 16 days of data when evaluating hospitalization counts. By using the NobBS model presented here, this timeline can be reduced by half. This improvement in decision-making time can be critical when combating a pandemic. Further patterns in the error of the model can be explored in order to establish alarms that signal potential greater uncertainty in the estimates. It is our hope that the output of this nowcasting model can be used by the DOH to monitor Covid-19 hospitalizations moving forward.

Transition Plan

The transition plan for the project is to provide the WA DOH with all related material, such as code used to perform the analysis and the primary results. The goal would be for a project owner in the DOH to operationalize the methods in such a way that they can be continuously monitored and utilized.

Next steps regarding the model would be to implement additional modeling methods and incorporate hospital information. Due to computational limitations, we were unable to fit a version of the HH model which included a time-varying delay distribution. This method also involves fitting the epidemiological curve using a spline function. The reporting hospital information could also be used as a covariate in the total hospitalization modeling process or as a variable on which the method could be stratified.

References

1. McGough SF, Johansson MA, Lipsitch M, Menzies NA (2020) Nowcasting by Bayesian Smoothing: A flexible, generalizable model for real- time epidemic tracking. PLoS Comput Biol 16(4): e1007735. https://doi.org/10.1371/journal. pcbi.1007735
2. Hohle M, an der Heiden M. Bayesian nowcasting during the STEC O104:H4 outbreak in Germany, 2011. Biometrics. 2014; 70: 993–1002. https://doi.org/10.1111/biom.12194 PMID: 24930473

Team Member Contributions

Ben Stan

- Primary writer for both slides and report documents
- Project management for coordinating with instructors, providing structure to meetings, etc
- Exploratory analysis and planning for results presentation

Fred Yu

- Investigator of previous methods and responsible for summarizing methods
- Implemented methods in R

Yongzhe Wang

- Researched additional applications of methods and summarized information
- Performed exploratory analysis