arXiv:submit/3144908 [stat.OT] 24 Apr 2020

Learning as We Go – An Examination of the Statistical Accuracy of COVID19 Daily Death Count Predictions

Roman Marchant[1,2], Noelle I. Samia[4], Ori Rosen[5], Martin A. Tanner[4], and Sally Cripps*[1,2,3]

[1]ARC Centre for Data Analytics for Resources and Environments, Australia
[2]Centre for Translational Data Science, The University of Sydney, Australia
[3]School of Mathematics and Statistics, The University of Sydney, Australia
[4]Department of Statistics, Northwestern University, USA
[5]Department of Mathematical Sciences, University of Texas at El Paso, USA

April 24, 2020

Abstract

This paper provides a formal evaluation of the predictive performance of a model (and its various updates) developed by the Institute for Health Metrics and Evaluation (IHME) for predicting daily deaths attributed to COVID19 for each state in the United States. The IHME models have received extensive attention in social and mass media, and have influenced policy makers at the highest levels of the United States government. For effective policy making the accurate assessment of uncertainty, as well as accurate point predictions, are necessary because the risks inherent in a decision must be taken into account – especially in the present setting of a novel disease affecting millions of lives. To assess the accuracy of the IHME models, we examine both forecast accuracy using tools from the econometric literature, as well as the predictive performance of the 95% prediction intervals provided by the IHME models. We find that the initial IHME model underestimates the uncertainty surrounding the number of daily deaths substantially. Specifically, the true number of next day deaths fell outside the IHME prediction intervals as much as 70% of the time, in comparison to the expected value of 5%. In addition, we note that the performance of the initial model does not improve with shorter forecast horizons. Regarding the updated models, our analyses indicate that the later models do not show any improvement in the accuracy of the point estimate predictions. In fact, there is some evidence that this accuracy has actually decreased over the initial models. Moreover, when considering the updated models, while we observe a larger percentage of states having actual values lying inside the 95% prediction intervals (PI), our analysis suggests that this observation may be attributed to the widening of the PI's. The width of these intervals calls into question the usefulness of the predictions to drive policy making and resource allocation.

*We thank the authors of the IHME model for making their predictions and data publicly available. We agree with the statement on their website http://www.healthdata.org/covid/updates: *Having more timely, high-quality data is vital for all modeling endeavors, but its importance is dramatically higher when trying to quantify in real time how a new disease can affect lives.* Without access to the data and predictions this analysis would not have been possible. We also thank Noam B. Tanner for bringing the Brooks reference to our attention. Corresponding author: sally.cripps@sydney.edu.au

Key Summary Points

1. The discrepancy between the predicted death and the actual death in the USA has serious implications for the USA government's future planning and provision of ventilators, PPE, and the staffing of medical professionals equipped to respond to this pandemic.
2. Regarding the initial IHME model, between 49% and 73% of states in the USA have actual daily death counts which lie outside the 95% prediction interval.
3. The updated IHME models do not show any improvement in the accuracy of point estimate predictions.
4. The rather large level of predictive uncertainty implied by the updated models casts doubt on their usefulness to drive the development of health, social, and economic policies.

1 Introduction

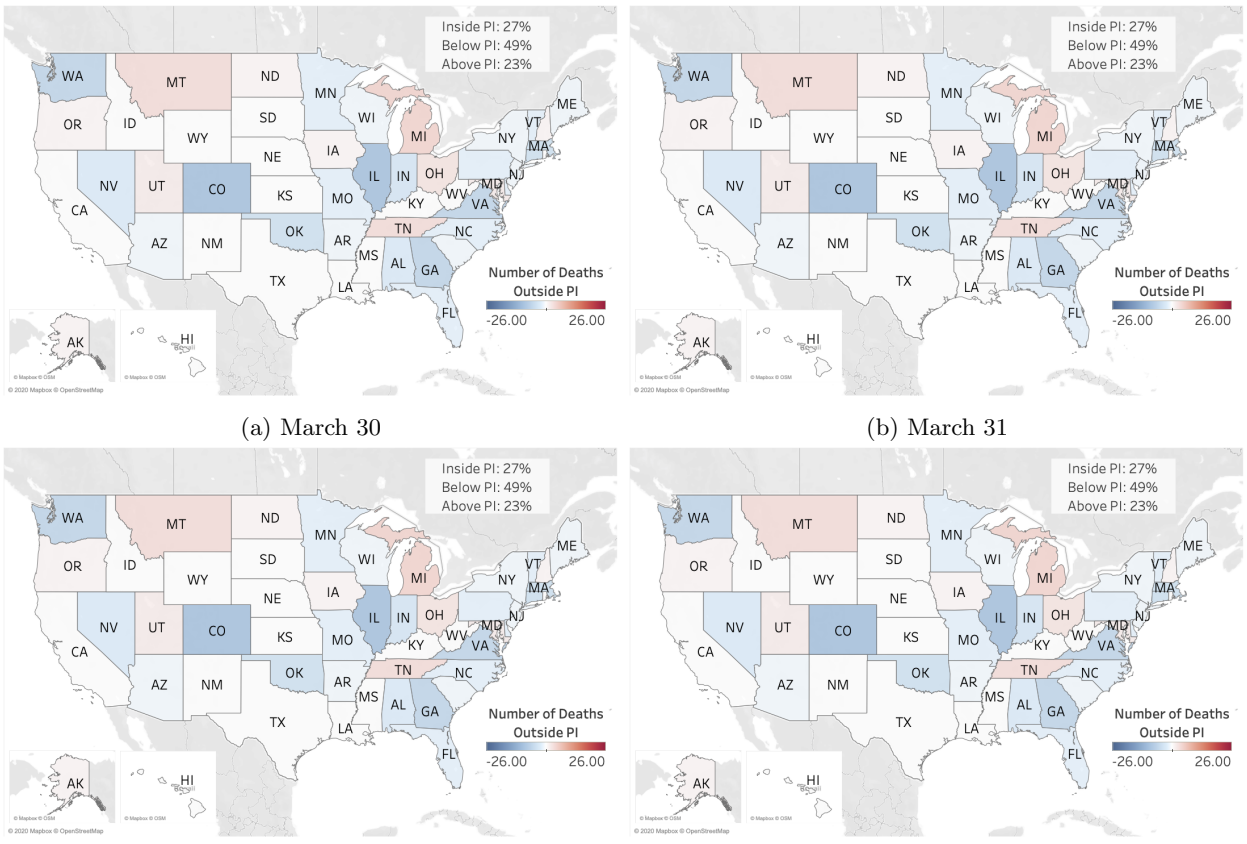
A recent model developed at the Institute for Health Metrics and Evaluation (IHME) provides forecasts for ventilator use and hospital beds required for the care of COVID19 patients on a state-by-state basis throughout the United States over the period March 2020 through August 2020 [7] (See also the related website https://covid19.healthdata.org/projections for interactive data visualizations). In addition, the manuscript and associated website provide projections of deaths per day and total deaths throughout this period for the entire US, as well as for the District of Columbia. This research has received extensive attention in social media, as well as in the mass media [2, 3]. Moreover, this work has influenced policy makers at the highest levels of the United States government, having been mentioned at White House Press conferences, including March 31, 2020 [2].

Our goal in this paper is to provide a framework for formally evaluating the predictive validity of the IHME forecasts for COVID19 outcomes, as data become sequentially available. We treat the IHME model (and its various updates) as a "black box" and examine the projected numbers of deaths per day in light of the ground truth to help quantify the predictive accuracy of the model. We do not provide a critique of the assumptions made by the IHME model, nor do we suggest any possible modifications to the IHME approach. Moreover, our analysis should not be misconstrued as an investigation of mitigation measures such as social distancing. We do, however, strongly believe that it is critical to formally document the operating characteristics of the IHME model – to meet the needs of social and health planners, as well as a baseline of comparison for future models.

2 Methods

Our report examines the quality of the IHME deaths per day predictions for the initial model over the period March 29–April 2, 2020, as well as for a series of updated IHME models over the period April 3, 2020–April 17, 2020. For these analyses we use the actual deaths attributed to COVID19 as our ground truth – our source being the number of deaths reported by Johns Hopkins University [1].

Each day the IHME model computes a daily prediction and a 95% posterior interval (PI) for COVID19 deaths, four months into the future for each state. For example, on March 29 there is a prediction and corresponding PI for March 30 and March 31, while on March 30 there is a prediction and corresponding PI for March 31. We call the prediction for a day made on the previous day a "1-step-ahead" prediction. Similarly, a prediction for a day made two days in advance is referred to as a "2-step-ahead" prediction, while a prediction for a day made k days in advance is called a "k-step-ahead" prediction.



(a) March 30

(b) March 31

(c) April 1

(d) April 2

Figure 1: Discrepancy between actual death counts and one-step-ahead PIs for specific dates (see sub-figures). The color shows whether the actual death counts were less than the lower limit of the 95% PI (blue), within the 95% PI (white), or above the upper limit of the 95% PI (red). The depth of the red/blue color denotes how many actual deaths were above/below the 95% PI.

3 Results

3.1 The Initial Model: 3/29-4/2

Figure 1 graphically represents the discrepancy between the actual number of deaths and the 95% PIs for deaths, by state for the dates March 30 through April 2. The color in these figures shows whether the actual death counts for a state were less than the lower limit of the 95% PI (blue), or within the 95% PI (white), or above the upper limit of the 95% PI (red). The depth of the red/blue color denotes the number of actual death counts above/below the PI. A deep red signifies that the number of deaths in that state was substantially above the upper limit of the 95% PI, while a light red indicates that the number of deaths was marginally above the 95% PI upper limit. Similarly, a deep blue signifies that the number of deaths was substantially below the lower limit of the 95% PI, while a light blue indicates that the number of deaths was marginally below 95% PI lower limit.

These figures show that for March 30 only 27% of states had an actual number of deaths lying in the 95% PI for the 1-step-ahead forecast. The corresponding percentages for March 31, April 1 and April 2, are 35%, 27% and 51%, respectively. Therefore the percentage of states with actual number of deaths lying outside this interval is 73%, 65%, 73% and 49% for March 30, March 31, April 1 and April 2, respectively. We note that we would expect only 5% of observed death counts to lie outside the 95% PI.

For a given day the initial model is also biased, although the direction of the bias is not constant across days. For the 1-step-ahead prediction for March 30, 49% of all locations were over-predicted, that is 49% of all locations had a death count which was below the 95% PI lower limit, while 23% were under-predicted. For

	1 - step	2 - step	3 - step	4 - step
March 30	29(47,24)			
March 31	33(16,51)	45(14,41)		
April 01	25(37,37)	37(29,33)	41(33,25)	
April 02	49(29,22)	49(25,25)	43(25,31)	45(25,29)
April 03		33(45,22)	33(45,22)	37(39,24)
April 04			37(45,18)	35(47,18)
April 05				35(51,14)

Table 1: Percentage of locations with actual death counts inside the 95% PI, as a function of the number of forecast periods. The values in parentheses indicate the percentage of locations that were (below,above) the limits of the 95% PI.

March 31 the reverse was true; only 16% of locations had actual death counts below the 95% PI lower limit while 49% had actual death counts above the 95% PI upper limit. This can be clearly seen from Figures 1a and 1b which are predominantly blue, and red, respectively. These figures are summarized in Table 1.

Table 1 also shows that the accuracy of predictions does not improve as the forecast horizon decreases, as one would expect. For March 31 and April 1 the forecast accuracy, as measured by the percentage of states whose actual death count lies within the 95% PI, decreases as the forecast horizon decreases. For March 31 the 2-step ahead prediction is better than the 1-step ahead prediction, while for April 1 the 3-step is better than the 2-step, which in turn is better than the 1-step. However, April 2 shows that accuracy slightly improves between the 3-step and the 2-step.

To investigate the relationship between the 2-step-ahead and the 1-step-ahead prediction errors by state, Figure 2 shows the March 31 1-step-ahead prediction errors, made on March 30, on the y-axis, versus the March 31 2-step-ahead prediction errors, made on March 29, on the x-axis. The colors in the graph correspond to different subsets of the data; red corresponds to those locations where the actual number of deaths was above the 1-step-ahead 95% PI upper limit, blue corresponds to those locations where the actual number of deaths was below the 1-step-ahead 95% PI lower limit, while grey corresponds to those locations where the actual number of deaths was within the 1-step-ahead 95% PI. This graph shows a very strong linear association between the predicted errors for the red locations ($R^2 = 96\%$, $n = 25$). This suggests that the additional information contained in the March 30 data did little to improve the prediction for those locations where the actual death count was much higher than the predicted number of deaths. The number of observations in the other two subsets of data was insufficient to draw any firm conclusions.

3.2 The Updated Models: 4/4–4/17

Per the IHME website, the IHME model underwent a series of updates beginning in early April and in this subsection we examine the performance of these later versions of the model. Our analysis focuses on two aspects of the IHME model predictions, first on the accuracy of the point estimates used for forecasting and second on the estimated uncertainties surrounding those forecasts.

We note that there are two ways in which the accuracy of the model, as measured by the percentage of states with death counts which fall within the 95% PI, can improve. Either the estimated uncertainty increases and therefore the prediction intervals become much wider, or the estimated expected value improves. The latter is preferable but much harder to achieve in practice. The former can potentially lead to prediction intervals that are too wide to be useful to drive the development of health, social, and economic policies.

3.2.1 Accuracy of Predictions of the Updated Models

Figure 3 is a heat map of the difference between the actual daily death count and the 1-step ahead predicted daily death count produced by the initial model for each state, expressed as a percentage of the actual daily death count for the days between March 30–April 2. This graph reproduces Figure 1 with two changes. First, instead of analyzing the discrepancy between actual daily deaths and the predicted daily deaths, we now analyze the discrepancy as a percent of the actual daily death count. This is done so that the discrepancy between observed and predicted counts is normalized across different states and on different days. If the actual value and the predicted value were both zero, we have set the percentage error to zero. If the actual value for a state

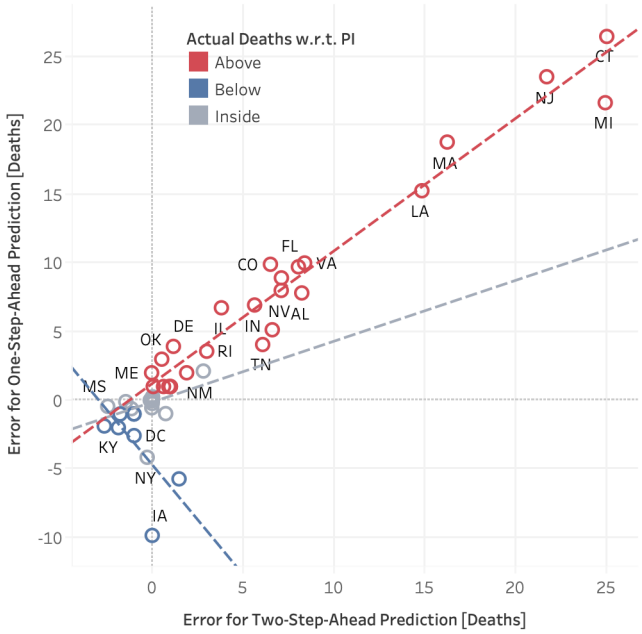


Figure 2: Actual minus predicted values of the 1-step ahead prediction for March 31, (y-axis) vs actual minus predicted value of the 2-step ahead prediction for March 31, (x-axis). The colors in the graph correspond to different subsets of the data; red corresponds to those locations where the actual number of deaths was above the 1-step-ahead 95% PI upper limit, blue corresponds to those locations where the actual number of deaths was below the 1-step-ahead 95% PI lower limit, while grey corresponds to those locations where the actual number of deaths was within the the 1-step-ahead 95% PI.

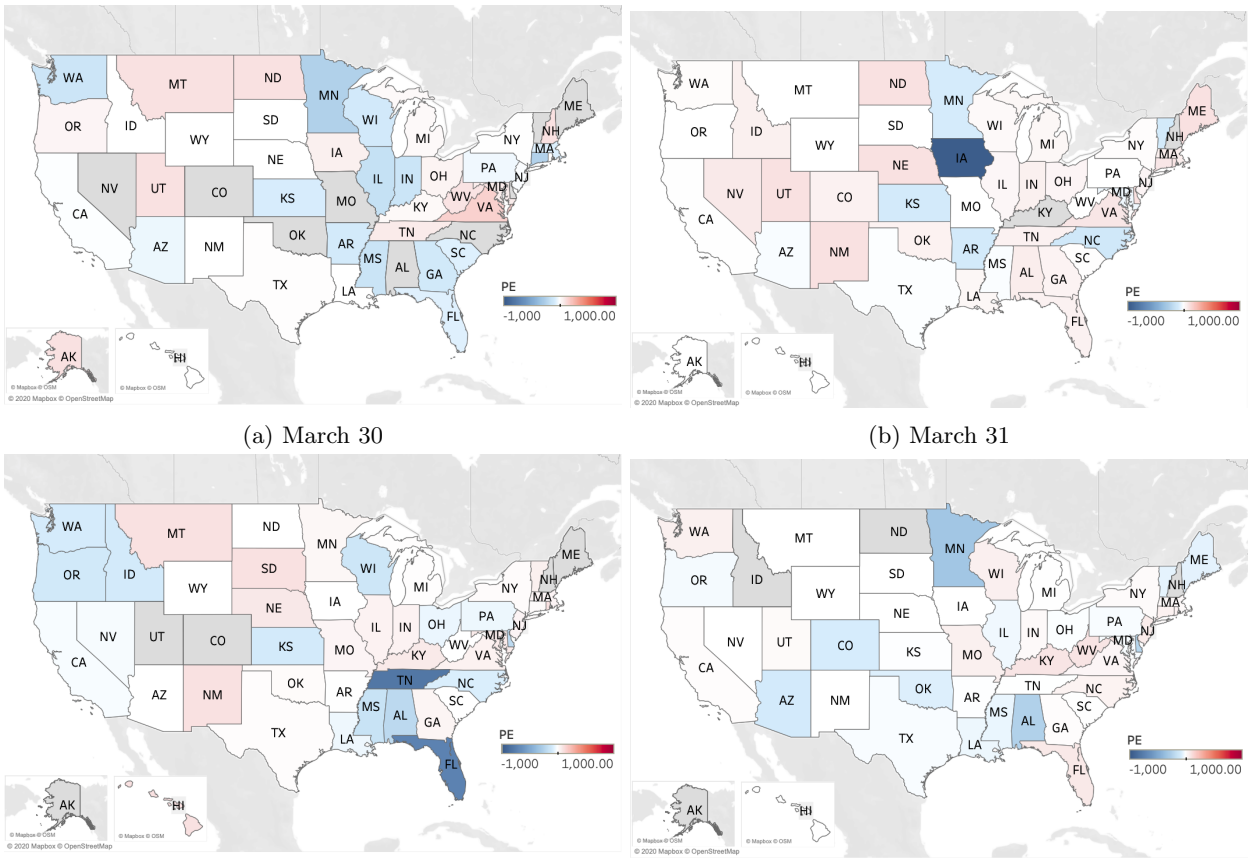
was zero but the predicted value value was not, we have labeled "NA" for that state and shaded it as grey. The second alteration is that the white color coding of states for which the actual death rate was within the 95% posterior interval is now omitted, so that Figure 3 is now a heat map of the percentage discrepancy.

Figure 4 is a similar map for the days April 4, April 8, April 13 and April 17 produced by the updated series of models. That is, Figure 4 is a heat map of the percentage difference between the actual daily death count and the 1-step ahead predicted daily death count produced by the updated models for each state, for the days April 4, April 8, April 13 and April 17. Note that these days are not consecutive due to the fact that these were the only days for which 1-step ahead predictions were made available by IHME.

A comparison between these two graphs highlights two features. First, the updated models are systematically biased across all days; that is the later models over-predict the number of daily deaths for April 4, April 8, April 13 and April 17. This can be seen by the predominance of blue across the maps in Figure 4. In contrast, the initial model produced predictions that were biased on any given day but the direction of the bias, that is whether the model over- or under-predicted the actual daily death counts, varied across days. The second feature is that the percentage discrepancy between actual deaths and predicted deaths, in addition to being biased, does not improve with the updated models. That is, the point estimates from the updated models are as inaccurate, sometimes more so, than those in the initial model.

To further investigate the accuracy of the point predictions, we computed the logit of the absolute value of the percentage error (APE) [4], denoted by LAPE. (The logit of $|x|$ is defined to be $\frac{1}{1+\exp(-|x|)}$.) Note that under this logit transformation, LAPE is a normed metric which approaches one for very large percentage discrepancies (especially when the observed count is close to or equal to zero), while equaling 0.5 for those instances with perfectly accurate predictions. Working on the logit scale avoids the need for ad-hoc rules for discarding outliers. When both observed and predicted death counts are equal to zero, the LAPE is equal to 0.5.

Figure 5 presents boxplots of these values for the dates March 30 to April 2, corresponding to predictions made with the initial model, as well as boxplots of the LAPE's for the dates April 4, April 8, April 13 and April 17, corresponding to predictions made with the updated models, where each row corresponds to 1-step through



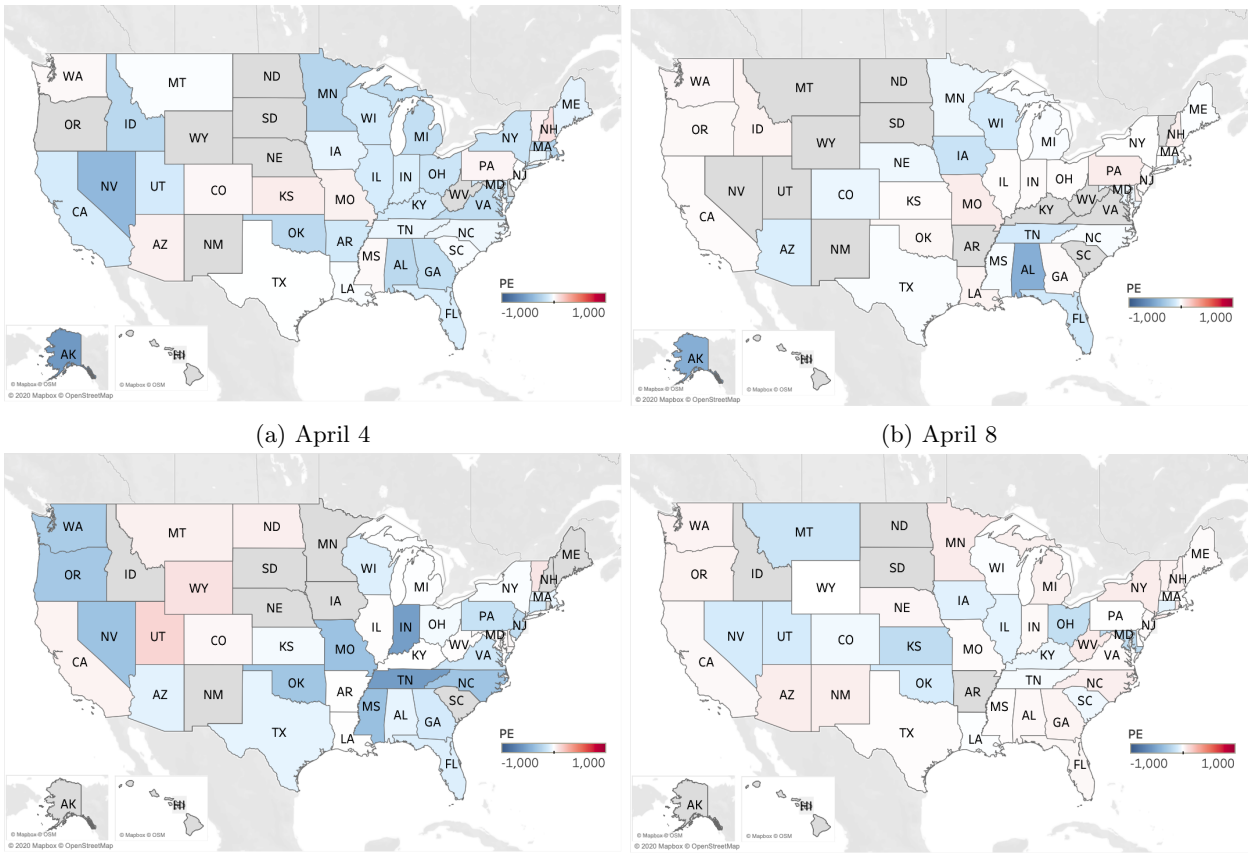
(a) March 30

(b) March 31

(c) April 1

(d) April 2

Figure 3: Heat map of the percentage error between the actual daily death count and the 1-step ahead predicted daily death count produced by the initial model for each state, expressed as a percentage of the actual daily death count for the days between March 30–April 2.



(a) April 4

(b) April 8

(c) April 13

(d) April 17

Figure 4: Heat map of the percentage error between the actual daily death count and the 1-step ahead predicted daily death count produced by the updated models for each state, expressed as a percentage of the actual daily death count for the days 4/4, 4/8, 4/13 and 4/17.

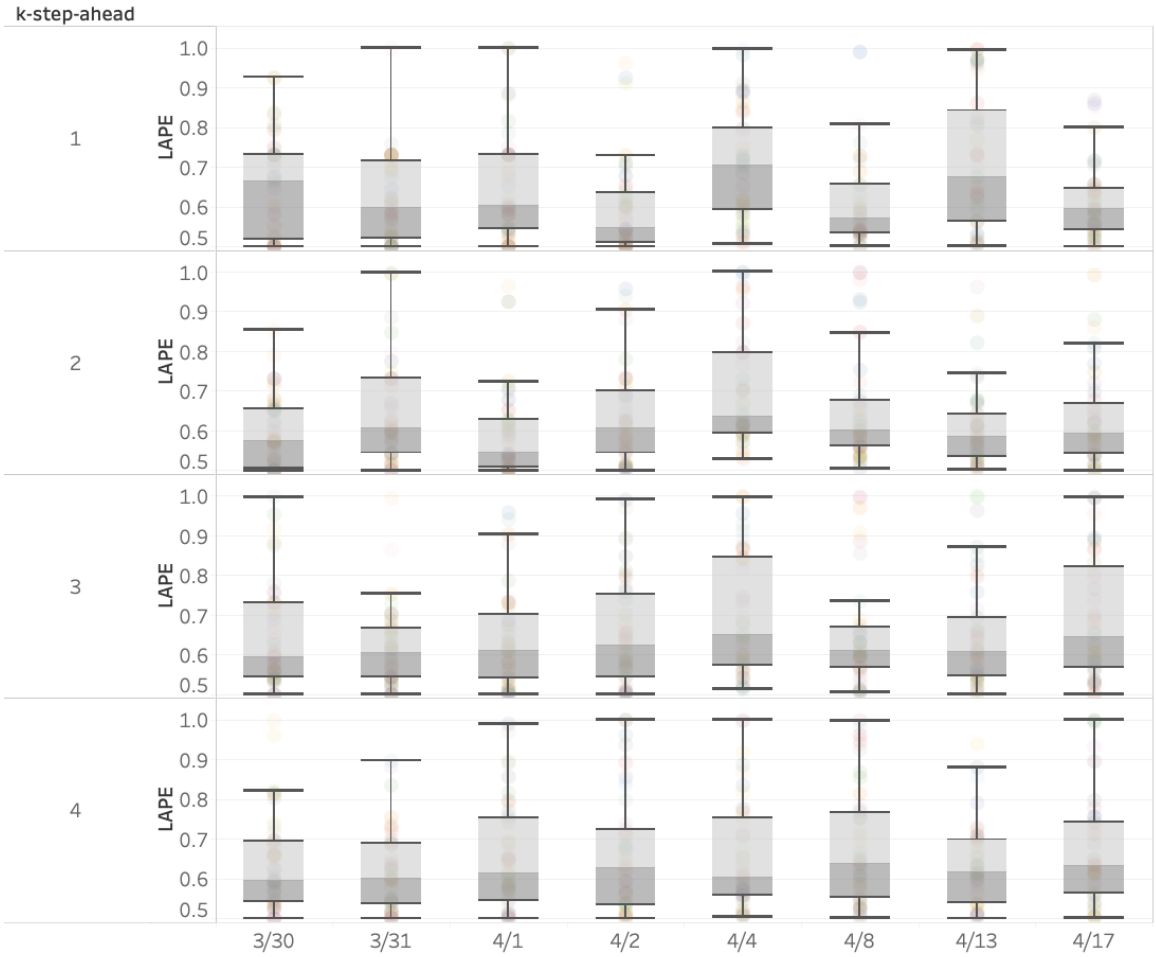


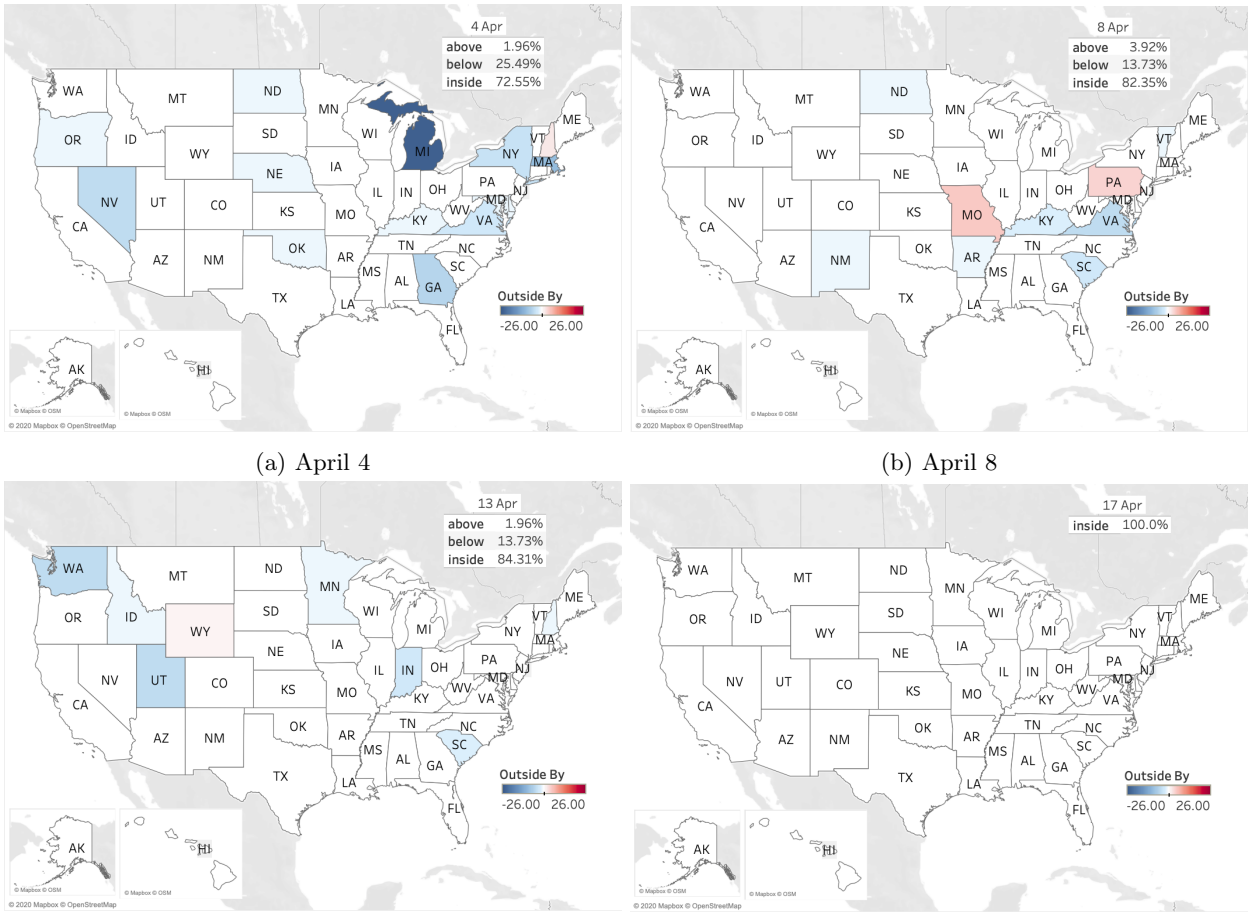
Figure 5: The logit of the absolute percentage error (LAPE) in multiple-step ahead predictions for the model's revision dates. The LAPE values for dates from April 4 onwards had k-step ahead predictions (corresponding to the particular row in the figure) made by the updated models, while those prior to this date had k-step ahead predictions made by the initial model.

4-step ahead predictions.

An examination of the first row of this figure (corresponding to 1-step ahead predictions) suggests that the predictive performance may have deteriorated somewhat with the updated models, as some boxplots on the right seem shifted toward 1. More formally, the Friedman nonparametric test [5], which accounts for possible correlation within states over time, revealed a difference across the eight time points ($p = 0.001$), with the corresponding post-hoc analysis indicating an elevation in the median LAPE on April 4 and April 13. By the 4-step-ahead prediction (i.e. the fourth row of the figure), the median LAPE are very similar ($p = 0.60$), with the LAPE values in general taking the full range of the logit from 0.5 to 1.0. Interestingly, as was noted earlier, the prediction accuracy of the initial model on March 30 seems to *deteriorate* as the number of steps ahead i.e. as k *decreases*.

3.2.2 Uncertainty Estimates of the Updated Models

We now turn to the evaluation of the uncertainty estimates produced by the models. Figure 6 is similar to Figure 4, except that those states for which actual daily deaths fell within the 95% PI are colored white, analogous to Figure 1. Figure 6 illustrates that many more states now have actual death counts which lie within the 1-step ahead 95% PI, as estimated by the updated models than as estimated by the initial model. In this way, we see that the percentage coverage improved substantially in early to mid April and Table 2 confirms this.



(a) April 4

(b) April 8

(c) April 13

(d) April 17

Figure 6: Discrepancy between actual death counts and one-step-ahead PIs for specific dates (see sub-figures) for the updated IHME models. The color shows whether the actual death counts were less than the lower limit of the 95% PI (blue), within the 95% PI (white), or above the upper limit of the 95% PI (red). The depth of the red/blue color denotes how many actual deaths were above/below the 95% PI.

However, in this regard, it is noted that all but two of the percentages in Table 2 are below the expected value of 95% and percentages below 0.88 are statistically significantly different from 0.95 at the 5% level, according to a one-tailed binomial test.

To explore this change in the uncertainty estimates of the predictions from the initial model to the updated models, we computed the range of the 95% PI at the date of the forecast peak of daily deaths for each state divided by the predicted value of the number of daily deaths at that peak (analogous to a coefficient of variation). In particular, division by the expected value of daily deaths at the peak takes into account the fact that those states with higher predicted peak daily deaths will have a larger 95% PI than those states with a lower expected peak daily deaths. Figure 7 presents boxplots of this quantity for all states for both the initial and updated models, As can be seen from this figure, the normalized range of the PI's expands dramatically with the updated models, with $p < 0.001$ according to the Friedman nonparametric test.

4 Discussion

Our results suggest that the initial IHME model substantially underestimated the uncertainty associated with COVID19 death count predictions. We would expect to see approximately 5% of the observed number of deaths to fall outside the 95% prediction intervals. In reality, we found that the observed percentage of death counts that lie outside the 95% PI to be in the range 49%–73%, which is more than an order of magnitude above the

	1 - step	2 - step	3 - step	4 - step
April 04	73(25,2)			
April 05		84(12,4)		
April 06			86(10,4)	
April 07				90(6,4)
April 08	82(14,4)			
April 09		82(12,6)		
April 10			84(12,4)	
April 11				94(6,0)
April 12				
April 13	84(14,2)			
April 14		96(4,0)		
April 15			98(2,0)	
April 16				92(2,6)

Table 2: Percentage of locations with actual death counts inside the 95% PI, as a function of the number of forecast periods. The values in parentheses indicate the percentage of locations that were (below,above) the limits of the 95% PI.

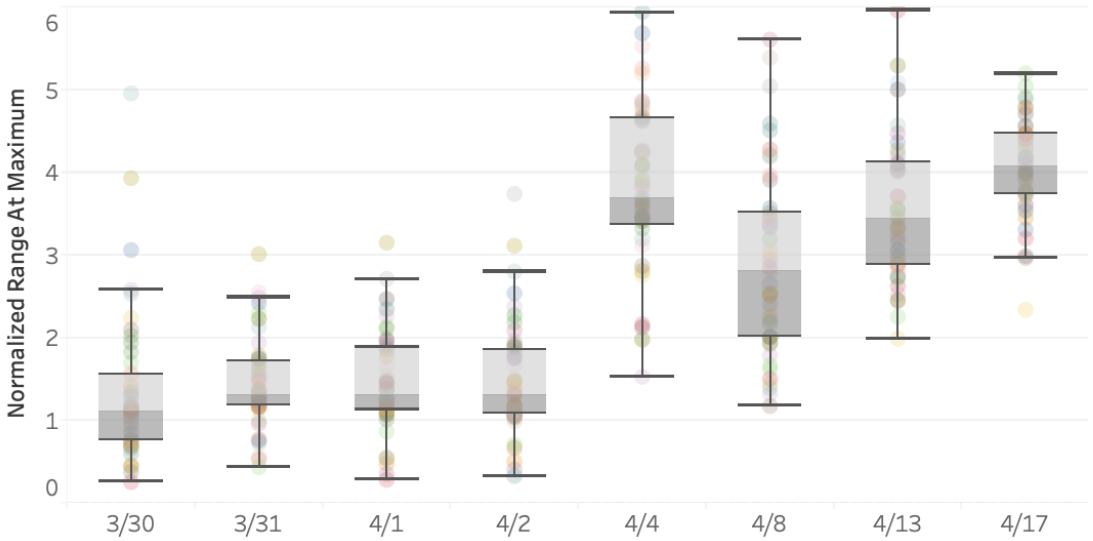


Figure 7: The range at the maximum predicted number of deaths, divided by the maximum predicted number of deaths across states. Each observation represents a state and boxplots are calculated across model release dates.

expected percentage. Moreover, we would expect to see 2.5% of the observed death counts fall above and below the PI. In practice, the observed percentages were asymmetric, with the direction of the bias fluctuating across days.

In addition, the performance accuracy of the initial model does not improve as the forecast horizon decreases. In fact, Table 1 indicates that the reverse is generally true. Interestingly, the model's prediction for the state of New York is consistently accurate, while the model's prediction of the neighboring state of New Jersey, which is part of the New York metropolitan area, is not consistently accurate.

Our comparison of forecasts made by the initial model versus forecasts to the updated models indicates that the later models do not show any improvement in the accuracy of point predictions. In fact, there is some evidence that this accuracy has actually decreased. Moreover, when considering the updated models, while we observe a larger percentage of states having actual values lying inside the 95% PI, Figure 7 suggests this observation may be attributed to the widening of the PI's. The width of these intervals does call into question the usefulness of the predictions to drive policy making and resource allocation. In this regard, it is noted that Jewell *et al.* [6] make general comments as to why the IHME model may suffer from the shortcomings formally documented in the present paper.

The accurate quantification of uncertainty in real time is critical for optimal decision making. It is perhaps the most pressing issue in policy making which is informed by mathematical modeling; decision makers need accurate assessments of the risks inherent in their decisions. All predictions that are used to inform policy should be accompanied by estimates of uncertainty, and *we strongly believe that these estimates should be formally validated against actual data as the data become available - especially in the case of a novel disease that has affected millions of lives around our entire planet.*

References

[1] Johns Hopkins University Coronavirus Resource Center, March 2020. retrieved from: https://coronavirus.jhu.edu, last accessed: April 8, 2020.

[2] White House press briefing on U.S. coronavirus response, March 2020. retrieved from: https://www.youtube.com/watch?v=H7CoCiH2F7U&feature=youtu.be, last accessed: April 8, 2020.

[3] A. Azad. Model cited by white house says 82,000 people could die from coronavirus by august, even with social distancing, 2020. https://www.cnn.com/2020/03/30/health/coronavirus-us-ihme-model-us/index.html, last accessed: April 8, 2020.

[4] C Brooks. *Introductory Econometrics for Finance, 4th edition*. Cambridge University Press, 2019.

[5] M Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32:675–701, 1937.

[6] NP Jewell, JA Lewnard, and BL. Jewell. Caution warranted: Using the institute for health metrics and evaluation model for predicting the course of the covid-19 pandemic. *Ann Intern Med.*, 2020. https://doi.org/10.7326/M20-1565, Epub ahead of print 14 April 2020.

[7] CJL Murray. Forecasting covid-19 impact on hospital bed-days, icu-days, ventilator-days and deaths by us state in the next 4 months. *medRxiv*, 2020.