A Quantitative Evaluation of COVID-19 Epidemiological Models

Osman N Yogurtcu[1], Marisabel Rodriguez Messan[1], Richard C Gerkin[2], Artur A Belov[1], Hong Yang[1], Richard A Forshee[1], Carson C Chow[3]
(1) Office of Biostatistics and Epidemiology, Center for Biologics Review and Research, Food and Drug Administration, (2) Arizona State University, (3) LBM/NIDDK National Institutes of Health

Abstract

Purpose. COVID-19 pandemic has taken a significant human toll. Quantifying how accurate epidemiological models of COVID-19 forecast the number of future cases and deaths can help frame how to incorporate mathematical models to inform public health decisions.

Methods. Here we analyze and score the predictive ability of publicly available COVID-19 epidemiological models on the COVID-19 Forecast Hub. Our score uses the posted forecast cumulative distributions to compute the log-likelihood for held-out COVID-19 positive cases and deaths. Scores are updated continuously as new data become available, and model performance is tracked over time.

Results. We use model scores to construct ensemble models based on past performance. We found that different modeling strategies score comparably despite varying assumptions. When we look at the distribution of scores, we see that scores decrease for long-term forecasts substantially and models are in general much better (with higher score) at forecasting cumulative death counts than in forecasting weekly incidental case counts. We also noticed that the models failed to capture abrupt changes in the observed case counts, specifically first week of July 2020 and mid November 2020. We calculated the scores of the unweighted and score-weighted ensemble models at each target end date.

Conclusions. We found that the score-weighted ensemble performed better than the unweighted ensemble model. Our publicly available quantitative framework may aid in improving modeling frameworks and assist policy makers in selecting modeling paradigms to balance the delicate trade-offs between the economy and public health.

Introduction

Epidemiological models of COVID-19 have proliferated quickly but it is unclear how well they forecast true values of key underlying variables: the number of infected and dead individuals. Establishing the quality of inference and predictive power is essential to make data-driven public health decisions, including those that manage the delicate trade-offs between the economy and public health. Model evaluation is therefore of critical importance.

There have been efforts to assess the accuracy of published and unpublished COVID-19 models. Some have focused on evaluating model performance on their weekly cumulative predictions. For instance, two different studies (Friedman et al., Gola et al.) assessed the performance of different models with the median absolute percent error (MAPE) of cumulative deaths. Friedman et al. observed that the calculated MAPE increased for longer forecasts, and the best performance model varied by region. Others have focused on evaluating model performance based on weekly incident case forecasts by ranking them according to the mean weekly percentile, while Leaderboard uses the root mean squared error of weekly new deaths and reporting recent and running average performance of eight models. Another group (Brooks et al.) evaluated ensemble models strictly containing probabilistic forecasts by computing a weighted interval score. They found that combining forecasts in various ways consistently leads to improved performance over single model forecasts.

- Friedman et al. IHME COVID, and Model Comparison Team. Predictive performance of international covid-19 mortality forecasting models. medRxiv, 2020.
- Gola et al. Review of forecasting models for coronavirus (COVID-19) pandemic in India during country-wise lockdown. medRxiv, 2020.
- Brooks et al. Comparing ensemble approaches for short-term probabilistic COVID-19 forecasts in the US. forecasters.org.
- Leaderboard: https://scc-usc.github.io/ReCOVER-COVID-19/#/leaderboard

Materials and Methods

We have created a uniform objective scoring system for COVID-19 models that assess their predictive performance. A scoring system has previously been deployed to assess Flu prediction models in CDC's FLUSIGHT contest two years ago. For COVID-19, investigators have been collating COVID-19 model epidemiological forecasts from multiple research groups in the COVID-19 Forecast Hub (covid19forecasthub.org). These results are concurrently presented on the CDC coronavirus forecasting website. While this effort provides insight into various model forecasts and their various assumptions, it stops short of providing an actual score for cumulative predictive performance.

We score individual models using a Leave-Forward-Out-Cross-Validation scheme. The score is computed by taking the log of the likelihood of the model forecasts on current data using the model predictions from the past. Each weekly projected quantity is scored separately, making it possible to assess model accuracy by the forecasted time increment. Thus, we have a matrix of scores where each entry is the computed log-likelihood for a model fitted to observed data up to a certain date of a quantity for each week forward from that date. This matrix continually expands in size as new weekly forecasts are made and new forecasting teams join the COVID-19 Forecast Hub collaborative. A global score is computed from the matrix by averaging over the desired elements. Some models may do well for short time predictions but not for longer ones and vice versa. Our score keeps track of how models improve (or degrade) over time and how far into the future they can reliably forecast. Our score dashboard is available at www.covidforeca.st.

To calculate scores of epidemiological forecasts on the COVID-19 Forecast Hub, we first converted the reported cumulative distribution function (CDF) quantiles into approximate probability density functions (PDFs) and then computed the log-likelihood of the held-out future observation as depicted in **Figure 1**. We constructed a scoring function that rewarded models for assigning high probabilities to the true values. Thus, models with broad predictive distributions are penalized compared to models with narrower distributions even if the true values are at the mode in both distributions. Conversely, models with narrow distributions are severely penalized if the true value falls outside of their predicted distribution.

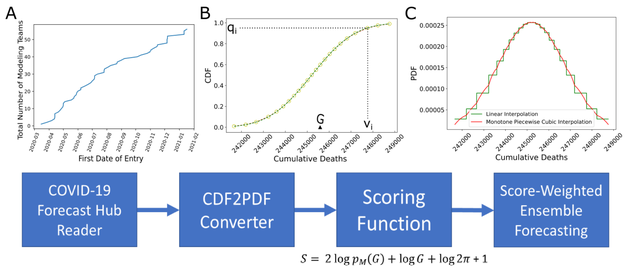

Figure 1. Scoring framework and analysis. **A.** Total number of teams which deployed US country-level epidemiological forecasts on COVID-19 Forecast Hub as of January 24, 2021. **B.** Scoring starts with reading forecast data available at COVID-19 Forecast Hub. An example forecast is shown for the model BPagano:RtDriven forecast made on 2020-11-9 targeting cumulative number of deaths on target end date 2020-11-14 (as denoted by G). Each forecast has a set of quantiles q and a set of corresponding values v. **C.** We calculate probability density functions using forecast data {q,v}. We apply our scoring function on every forecast available. The past performances of models are used to form score-weighted ensemble model forecasts.

Results and Discussion

The number of models on the COVID-19 Forecast Hub increased substantially since March 2020 (**Figure 1A**). We focused on weekly incidental case forecasts and cumulative death counts (**Figure 2**). We identified 39 models for the weekly case counts and 54 models for cumulative death counts. Most of the forecasts target less than 4-week ahead COVID-19 epidemiology. There are 7402 unique entries for cumulative death count forecasts and 3759 for the weekly incidental case counts as of January 24, 2021 and the cumulative death counts forecasts span a much larger time frame (up to 21 weeks ahead).

When we look at the distribution of scores, we see that scores decrease for long-term forecasts substantially and models are in general much better (higher score) at forecasting cumulative death counts than in forecasting weekly incidental case counts (**Figure 3**). We also noticed that the models failed to capture abrupt changes in the observed case counts, specifically first week of July 2020 and mid November 2020.

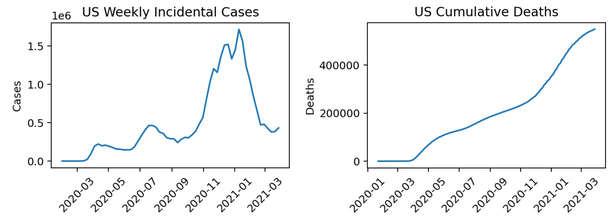

Figure 2. COVID-19 epidemiological curves in the US.

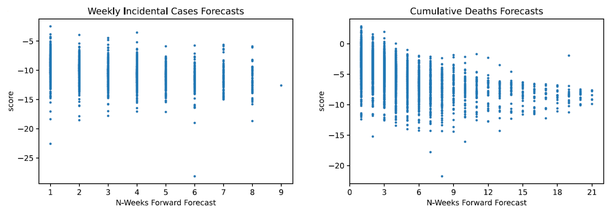

Figure 3. Score scatter plots for weekly incidental case count forecasts and cumulative death forecasts in COVID-19 Forecast Hub.

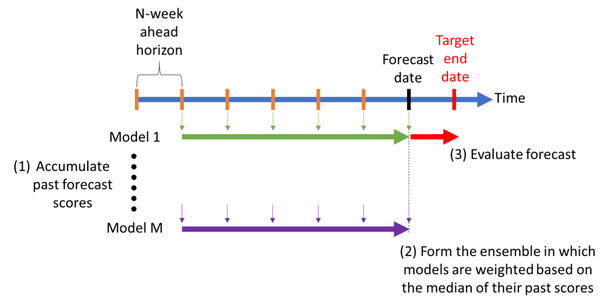

Figure 4. Ensemble forecast formation.

We formed unweighted and score-weighted ensemble models which forecast US weekly incidental case counts, and US cumulative death counts over the horizon of 1-week to 6-weeks (**Figure 4**). We calculated the scores of the unweighted and weighted ensembles at each target end date. In general, the score-weighted ensemble performed better for 1-week ahead forecasts. As a quantitative measure of forecast performance of our ensemble model, we calculated the median of the difference between ensemble model scores and the median of all available scores across the pandemic over different forecast horizons (large positive difference means better ensemble model performance). Based on calculations, score-weighted ensemble model performed consistently better and had higher score than average model for 1 through 6-weeks-ahead horizon (**Table 1**). We also observed that Sweight ensemble model tends to make more stable forecasts based on the median absolute deviation (MAD) calculations.

Table 1. Performance evaluation of the score-weighted ensemble model (FDANIHASU:Sweight) forecasts over time for US cumulative death and weekly incidental cases. We compared median difference in scores (Ensemble - Average Model) for different forecast horizons (i.e., 1- through 6-weeks-ahead). Across all horizons and forecast target types the score-weighted ensemble model performed better based on this measure (i.e., positive values).

Median Difference in Scores (Ensemble - Average Model)		
Weeks-Ahead	**Case Forecasts**	**Death Forecasts**
1	0.77	2.08
2	1.14	1.17
3	1.61	1.03
4	1.32	1.18
5	2.28	2.07
6	2.03	2.74

Conclusion

We developed a scoring framework for the forecasts of COVID-19 epidemiological models using forecast and observed data available at COVID-19 Forecast Hub. Scores can be used to evaluate past performances of all models. Additionally, quantitatively evaluating model forecasts enable score-weighted ensemble model forecasting. We provide a systematic review of available model types which show what type of modeling efforts are more successful in forecasting the COVID-19 epidemiology in the US. Our results suggest that models are unable to capture abrupt changes in COVID-19 epidemiology (e.g., the first two weeks of July and the first two weeks of November in US). Death count forecasts so far have been more accurate than weekly case counts. That could be because incidental case count curves are much higher than death counts and non-monotonic in behavior and may have varying reporting practices across health institutions, counties, and states.

Our work has some notable limitations. We only consider models on the COVID-19 forecast hub and this may inadvertently lend to selection bias of groups willing to format their model output according to required metrics and upload to the hub. Currently, our scoring framework considers only the US national data and scores on the individual US state model forecasts can be made available. We have also only considered weekly incidental case numbers and number of cumulative deaths, which does not consider modeling efforts that predict hospital capacity and utilization.

You may read more about our work from https://www.medrxiv.org/content/10.1101/2021.02.06.21251276v1.