THE CURIOSITY CUP 2022
A Global SAS® Student Competition

Accuracy of CDC Ensemble Model in Predicting COVID-19 Cases and Deaths

Deiriai Myers, Jenna Carcione, Octavio Alanis and Sarah Weinbrom, Drexel University – COVID Coders

ABSTRACT

Outbreaks due to emerging pathogens remain a concern and monitoring outbreaks in a timely manner is important for societal wellbeing. The objective of this study was to evaluate the accuracy of the Center for Disease Control and Prevention (CDC) Ensemble Forecast model at predicting COVID-19 cases and deaths. This study retrospectively evaluated the model's accuracy by comparing prior predictions to actual occurrence data. Secondarily, this study aimed to determine the relationship between predominant variants and the accuracy of the Ensemble predictions.

INTRODUCTION

As of the 15th of January 2022, the United States (US) has recorded over 65 million cases of COVID-19 with almost 850,000 deaths according to the John Hopkins COVID-19 dashboard (Johns Hopkins, 2022). Data modeling is important to assess risk of exposure and contributes information to understand the trajectory of the outbreak especially when rapidly evolving data impacts school and business closures. It is important to note, however, that data modeling is limited in that it provides a forecast of what is to be expected based on the current trajectory (Jewell et. al, 2020).

Ensemble models are the combination of different predictive forecasts into a single model (Cramer et. al, 2021). They tend to be more accurate and stable and are usually used to forecast events a maximum of four weeks in advance (Cramer et. al, 2021). Combining different models together into one reduces the biases that are introduced with any one of the ensemble components (Roberts, 2021). In conjunction with universities across the globe, the United States CDC developed and maintained the COVID-19 Forecast Hub which contains records of COVID-19 cases, deaths, and hospitalizations. Using the ensemble model to forecast seasonal influenza outbreaks previously (McGowan et. al, 2019) the model was used once again to predict COVID-19 cases, death, and hospitalizations for the subsequent 1-, 2-, 3-, and 4-week timeframes in the United States. We sought to evaluate the effectiveness of the CDC's Ensemble 1-week Forecast model to understand how reliable this model has been in predicting COVID-19 disease activity.

METHODS

DATA SOURCE

The national forecast data for COVID-19 Cases and Deaths were individually downloaded from the 72 weekly reports produced by the CDC for the prediction window of August 8th, 2020, to December 18th, 2021 (CDC, 2021, "COVID-19 Forecasting and Mathematical

Modeling"). Predictions of cases were accessed in the subsection "Previous Case Forecasts" under "Download national forecast data" for each week in the prediction window, and predictions of deaths were accessed in the subsection "Previous Death Forecasts" under "Download national forecast" for each week in the prediction window.
The dataset containing observed COVID-19 Cases and Deaths was downloaded from the COVID Data Tracker page produced by the CDC via the "Download Data" button (CDC, 2022, "COVID Data Tracker").
The dataset containing the information on the proportion of United States COVID-19 cases attributable to variants was downloaded from the "Download" button in the Our World in Data site, and the original source of this data is GISAID (Out World in Data, 2022; GISAID, 2022).

DATA CLEANING PROCESS

The ensemble model data for predicted deaths and predicted cases were imported and merged using a macro iterating for each date in the prediction window. The datasets were filtered to only import the data from the ensemble model, as specified by the tab of the excel file in the import statement. Data for the 1 week ahead case predictions were missing for December 7th, 2020, and December 14th, 2020, so these dates were excluded from analysis for the predicted case data.
Next, from one file, the incident case and incident death data were imported. While this file contained forecasted predictions of cases and deaths for prior weeks, it was missing a variable identifying the timeframe of the prediction (e.g., 1 week ahead) and there were duplicate observations for single dates. Therefore, it was necessary to import the predictions from the weekly files and use this file only to report observed cases. The imported file was merged separately with the datasets containing predicted values for cases and deaths resulting in one dataset for deaths and one dataset for cases.

ANALYSIS METHODS

University-licensed SAS version 9.4 software was provided by Drexel University and used to complete this analysis.
The difference and percent difference between predicted and reported cases and deaths were calculated using the formula: (Predicted – Reported) / Reported. Indicator variables were created report overestimates and underestimates of cases and deaths. Descriptive statistics were computed for the percentage difference in cases and deaths.
For the analysis and visualization of the observed versus predicted deaths and cases, overlayed line plots were created using the TEMPLATE procedure. Using the same method, another plot was produced for the percentage difference between prediction and actual values for deaths and cases.
To visualize the relationship between the predominant variants and the accuracy of the CDC predictions, the FREQ procedure was applied to the transposed variant dataset to identify the cumulative proportion of variants at each time point. The LAG function was used on the cumulative percent variable to define the lower bound of the band plot, that is, the upper bound of the prior variant corresponds with the lower bound of the subsequent variant. The variant data were merged separately with the death data and case data, and PROC TEMPLATE was used to overlay the variant band plot with the series plot showing the percent difference between predicted and reported cases and deaths.

RESULTS

COMPARISON OF FORECASED VS. OBSERVED EVENTS

The period in which the CDC Ensemble model was least accurate in predicting incident COVID-19 cases and deaths was July 2021 (-70.85% difference in predicted cases, -41.94% difference in predicted deaths). We hypothesized that the emergence of the Delta variant of the SARS-CoV-2 virus as a Variant of Concern (VOC) during this period may have played a role in the model's decreased accuracy. We compared the percent difference between forecasted and observed incident cases and deaths to the relative proportion of virus strains circulating in the US.

The accuracy of the forecast model drastically declined in the period when the Delta variant had become the dominant strain in the USA (Edwards, 2021). The case prediction model was less accurate in this period than the death prediction model, although both yielded substantial underestimates. This finding indicates a need for future study into adapting forecast models to better prepare for changing proportions of viral strains, as well as a need to account for global data trends when forecasting data for the US, as the Delta variant was declared a VOC by the WHO (World Health Organization) on May 15th, 2021, approximately 6 weeks before becoming the dominant strain in the US (Christensen, 2021).

An analysis of the model's accuracy throughout the more recent emergence of the Omicron variant of the SARS-CoV-2 virus is, unfortunately, outside the purview of this study, as this variant became dominant after December 18th, 2021. Future research should evaluate the predictive success of this model in relation to the Omicron variant.

COMPARING MODEL ACCURACY TO EMERGENCE OF VARIANTS OF CONCERN

The period in which the CDC Ensemble model was least accurate in predicting incident COVID-19 cases and deaths was July 2021 (-70.85% difference in predicted cases, -41.94% difference in predicted deaths). These percentages represent substantial underestimates. We hypothesized that the emergence of the Delta variant of the SARS-CoV-2 virus as a Variant of Concern (VOC) during this time period may have played a role in the model's decreased accuracy. The WHO declared the Delta a VOC on May 15th, 2021, and the CDC declared the same in June 2021 (Christensen, 2021). By early July 2021, the Delta variant had become the dominant strain in the USA (Edwards, 2021). To evaluate how this development affected the accuracy of the forecast model, we compared the percent difference between forecasted and observed incident cases and deaths to the relative proportion of virus strains circulating in the US.

The accuracy of the forecast model drastically declined in the time period when the Delta variant was rapidly spreading in the US. The case prediction model was less accurate in this time period than the death prediction model, although both yielded substantial underestimates. This finding indicates a need for future study into adapting forecast models to better prepare for changing proportions of viral strains. Although the Delta variant did not become dominant in the US until early July 2021, it was circulating globally and declared a VOC by the WHO two months earlier. This model may not have accounted for global data trends when forecasting data for the US.

An analysis of the model's accuracy throughout the more recent emergence of the Omicron variant of the SARS-CoV-2 virus is, unfortunately, outside the purview of this study, as this variant became dominant after December 18th, 2021. Future research should evaluate the predictive success of this model in relation to the Omicron variant.

IMPACTS AND CONCLUSION

Between August 2020 and December 2021, the CDC Ensemble Forecast Model generally succeeded in providing reliable predictions for incident COVID-19 cases and deaths in the US. The greatest variation between the predictions and the observed cases and deaths occurred in July and August of 2021, during the time that the Delta variant was rapidly spreading indicating a need for integration for global data trends into US-focused models. The model was more likely to underestimate both cases and deaths 1 week in the future than to overestimate, indicating that the model's predictions tend toward the assumption that the impact of COVID-19 would be less severe than it was. This is valuable information for Public Health entities who use this model's predictions to plan for resource allocation, as it would suggest a need for additional preparedness above and beyond the suggestions from the CDC.

REFERENCES

CDC. 2021. "COVID-19 Forecasting and Mathematical Modeling." Accessed December 26, 2021. https://www.cdc.gov/coronavirus/2019-ncov/science/forecasting/mathematical-modeling.html.

CDC. 2022. "COVID Data Tracker." Accessed January 6, 2022. https://covid.cdc.gov/covid-data-tracker/#forecasting_weeklydeaths.

Christensen, Jen. (2021, June 15). *CDC now calls coronavirus Delta variant a 'variant of concern'.* CNN Health. https://www.cnn.com/2021/06/15/health/delta-variant-of-concern-cdc-coronavirus/index.html

Cramer, E. Y.,Ray. E.L., Lopez, V.K.,Bracher, J.,Brennen, A.,Rivadeneir, A. J. C.,Gerding, A.,Gneiting, T., House, K. H.,Huang, Y.,Jayawardena, D.,Kanji, A. H., Khandelwal, A., Le, K.,Mühlemann, A., Niemi, J., Shah, A., Stark, A., Wang, Y… Reich, N. G.(2021) Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the US [Unpublished Manuscript].

Edwards, Erika. (2021, July 7). *The delta variant is now the dominant Covid strain in the US.* NBC News. https://www.nbcnews.com/health/health-news/delta-variant-now-dominant-covid-strain-u-s-n1273214

GISAID. 2022. "Tracking of Variants." Accessed January 6, 2022. https://www.gisaid.org/hcov19-variants/.

Howard, J. (2021, July 27). *CDC updates guidance, recommends vaccinated people wear masks indoors in certain areas.* CNN Health. https://www.cnn.com/2021/07/27/health/cdc-mask-guidance-vaccinated-people-bn/index.html

Jewell, N. P., Lewnard, J. A., Jewell, B.L. 2020. "Predictive Mathematical Models of the OCVID-19 pandemic." *JAMA,* 323:1893-1894.

Johns Hopkins University of Medicine Coronavirus Resource Center. 2022.Accessed January 15, 2022. https://coronavirus.jhu.edu/map.html.

McGowan, C.J., Biggerstaff, M., Johansson, M., Apfeldorf, K.M, Ben-Nun, M., Brooks, L., Convertino, M., Erraguntla, M., Farrow, D.C., Freeze, J., Ghosh, S., Hyun, S.,

Kandula, S., Lega, J. Liu, Y., Michaud, N., Morita, H., Miami, J., Ramakrishnan, N. … "The Influenza Forecasting Working Group. 2019. Collaborative Efforts to forecast seasonal influenza in the UNited States, 2015-2016." *Sci Rep*, 9.

Netburn, D. (2021, July 27). *A timeline of the CDC's advice on face masks.*LA Times. https://www.latimes.com/science/story/2021-07-27/timeline-cdc-mask-guidance-during-covid-19-pandemic)

Our World in Data. 2022. "SARS-CoV-2 variants in analyzed sequences, United States." Accessed January 6, 2022. https://ourworldindata.org/grapher/covid-variants-area?country=~USA

Roberts, S. (2021, May 28). *All together now: The most trustworthy covid-19 model is an ensemble.I*Technology review. https://www.technologyreview.com/2021/05/28/1025478/covid-ensemble-model-forecast-trustworthy/

ACKNOWLEDGMENTS

The COVID Coders team wish to extend their sincere gratitude and appreciation to Dr. Michael Levasseur for his contributions towards the completion of this project and to Drexel University for providing the SAS licensing used.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Deiriai Myers
Dtm89@drexel.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.

APPENDIX

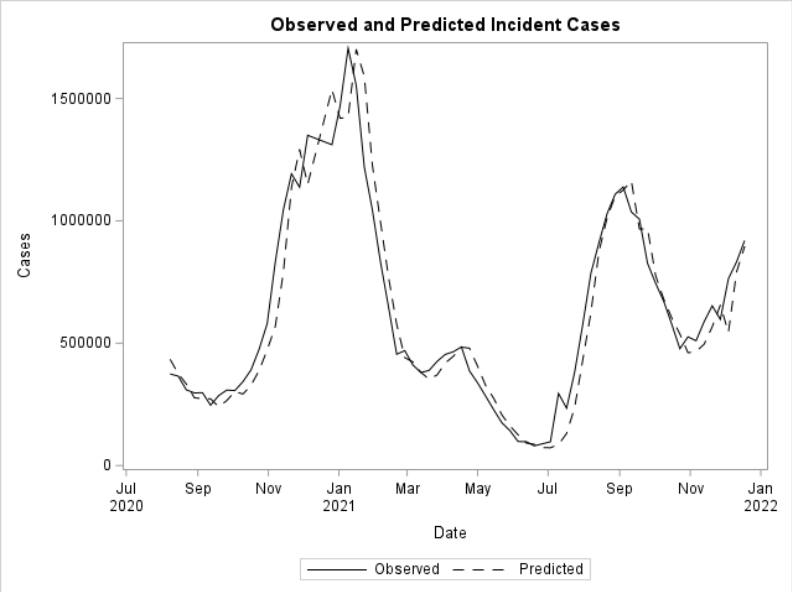


Figure 1. Observed and Predicted Incident COVID-19 Cases in US, 08/08/2020 to 12/18/2021

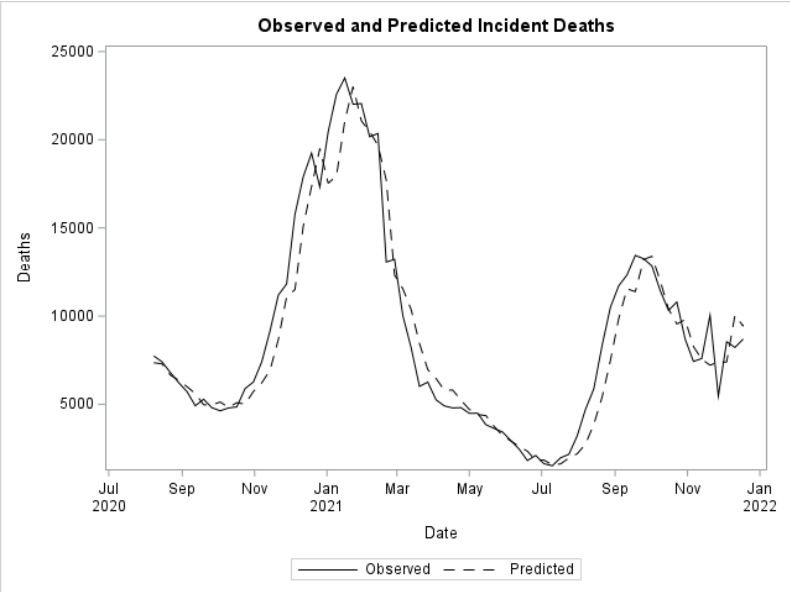


Figure 2. Observed and Predicted Incident COVID-19 Deaths in US, 08/08/2020 to 12/18/2021

Figure 3. Percent Difference Between Forecasted and Observed COVID-19 Cases and Deaths in US, 08/08/2020 to 12/18/2021

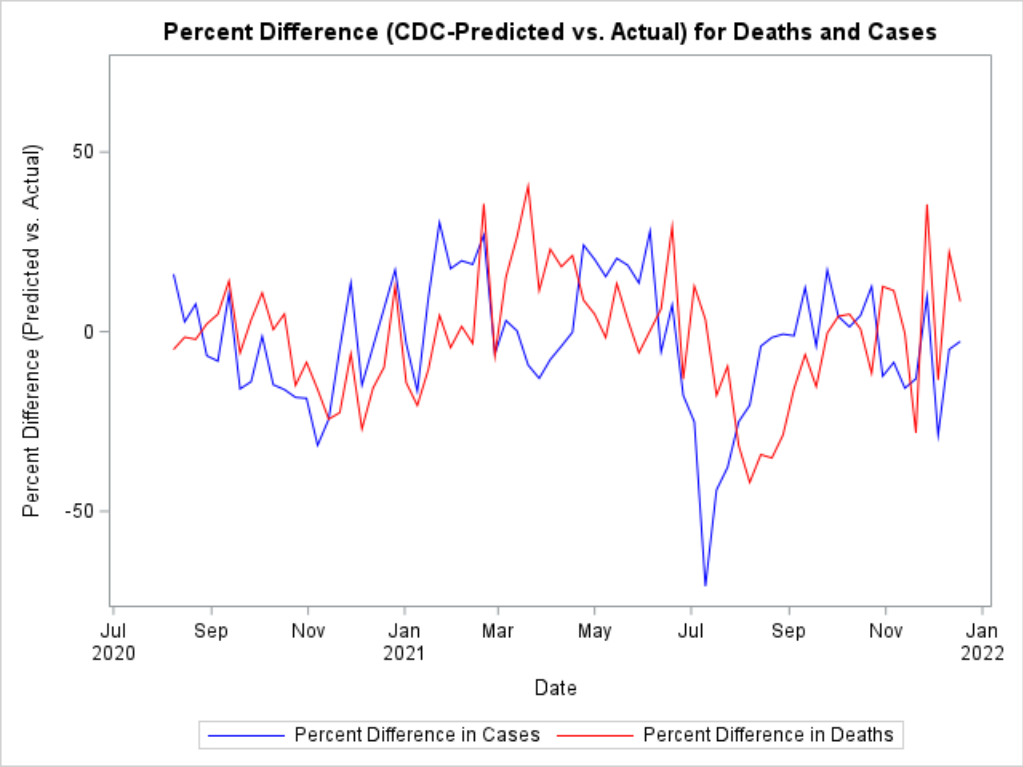
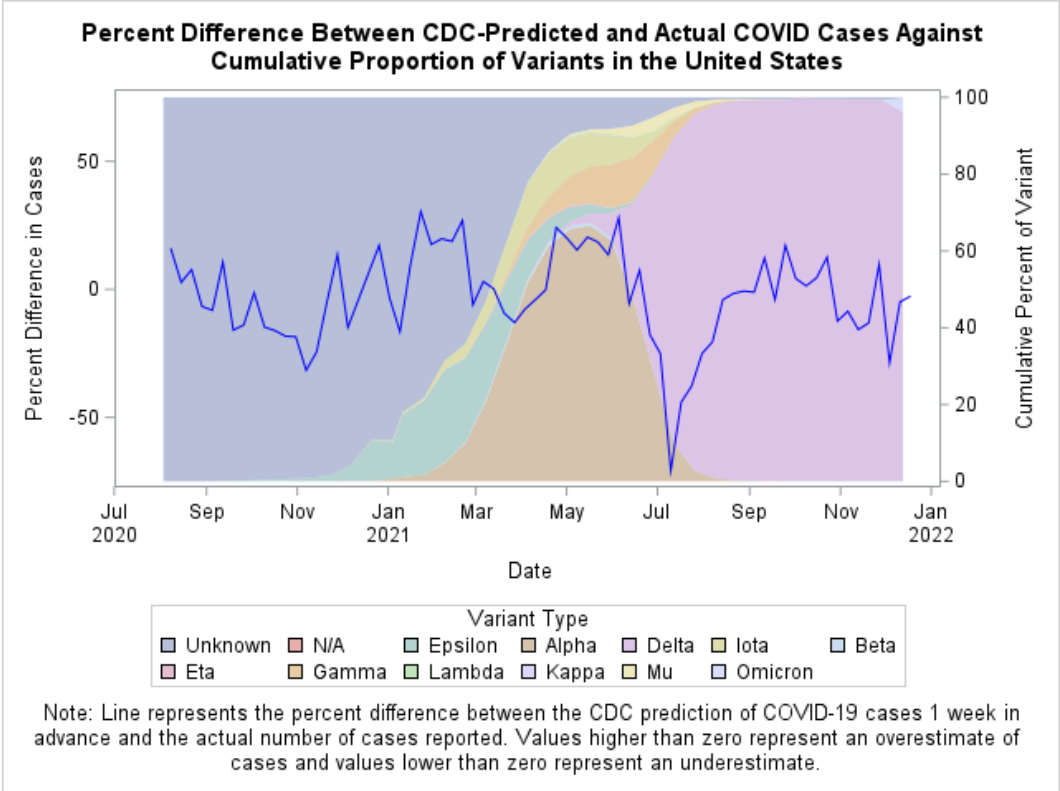
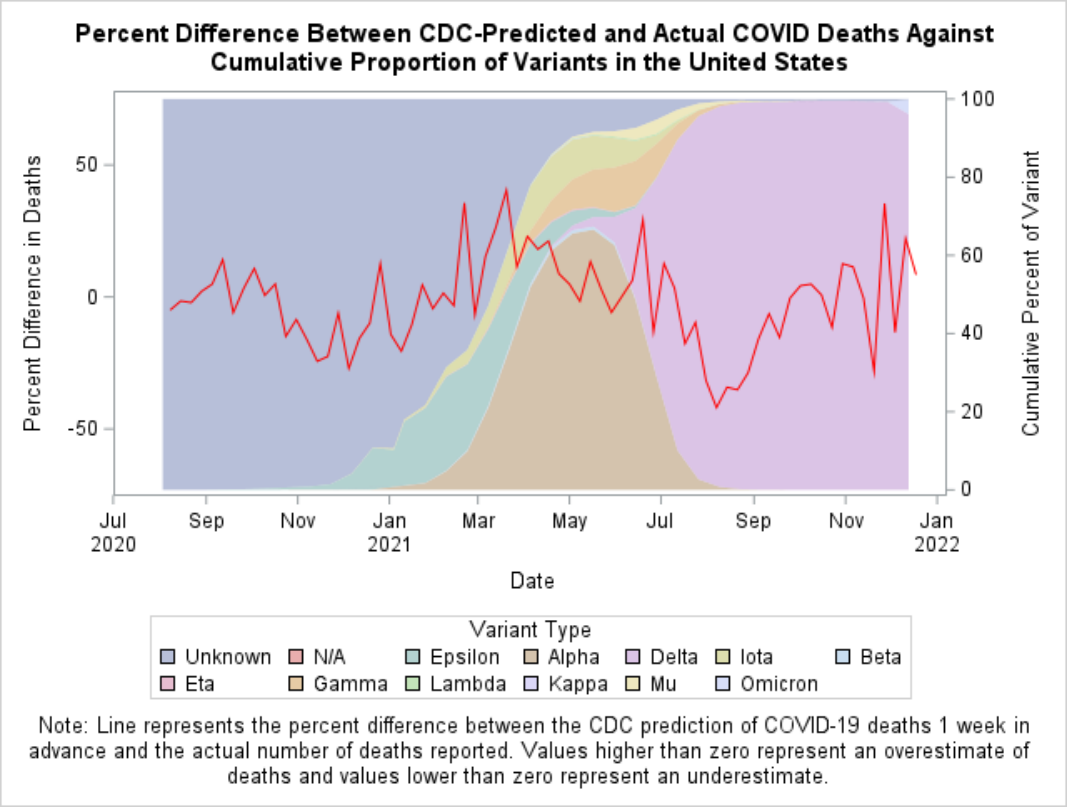


Figure 4. Percent difference in incident COVID-19 Cases compared to emerging variants



Note: Line represents the percent difference between the CDC prediction of COVID-19 cases 1 week in advance and the actual number of cases reported. Values higher than zero represent an overestimate of cases and values lower than zero represent an underestimate.

Figure 5. Percent Difference in incident COVID-19 Deaths compared to emerging variants



Note: Line represents the percent difference between the CDC prediction of COVID-19 deaths 1 week in advance and the actual number of deaths reported. Values higher than zero represent an overestimate of deaths and values lower than zero represent an underestimate.