Assessing socio-demographic biases and fairness of population-level COVID-19 forecasts in the US

Ariane Stark*, Dasuni Jayawardena*, Nicholas G. Reich*
* Department of Biostatistics and Epidemiology, University of Massachusetts Amherst

Abstract

Increasingly, policy-makers turn to data-driven computational models to inform their real-time understanding and decision-making regarding how to best serve populations at-risk for particular diseases. Forecasting models that attempt to predict future aggregate observations from public health surveillance systems have been used frequently in recent years to help anticipate health-care system burden in outbreaks of influenza, dengue fever, Ebola, chikungunya, Zika, and COVID-19. It is important to assess the "fairness" of such models by measuring whether the accuracy of these models varies by the socio-demographic makeup of different geographic regions. These efforts are critical to better understanding if, how, and to what extent disparities present in healthcare systems may translate to inaccuracy in model outputs. We evaluated the fairness of the COVID-19 Forecast Hub ensemble model predictions of weekly county-level incident cases from July 20, 2020, to March 6, 2021, using demographic data from the United States Census Bureau. Fairness was assessed by investigating whether average model error at the county level was associated with socio-demographic variables. There are observed associations between the proportion of underserved populations in a county and the ensemble forecast's mean absolute error (MAE). The association appears to be primarily driven by the overall number of cases the county experienced: more cases led to higher model error. Additionally, when using the scale-free metric of relative mean absolute error (RMAE) to evaluate model fairness, the initially observed relationships between predictive model error and demographic variables are no longer present. Health outcome disparities in the COVID-19 pandemic are a critical public health issue and these disparities are mirrored in overall forecast model error. However, using an objective relative model error metric we do not find evidence to suggest that ensemble forecasts performed have lower relative accuracy in counties with higher proportions of historically underserved populations.

Introduction

Healthcare systems and public health surveillance networks often have biases that prevent them from being able to effectively reach certain segments of the population. Data that come out of surveillance systems may not be representative of the underlying population due to the costs of seeking healthcare treatment, the differences in access to preventative care and clinical treatment, and/or differences in care-seeking behavior due to access or trust in the healthcare system. [CITATION]

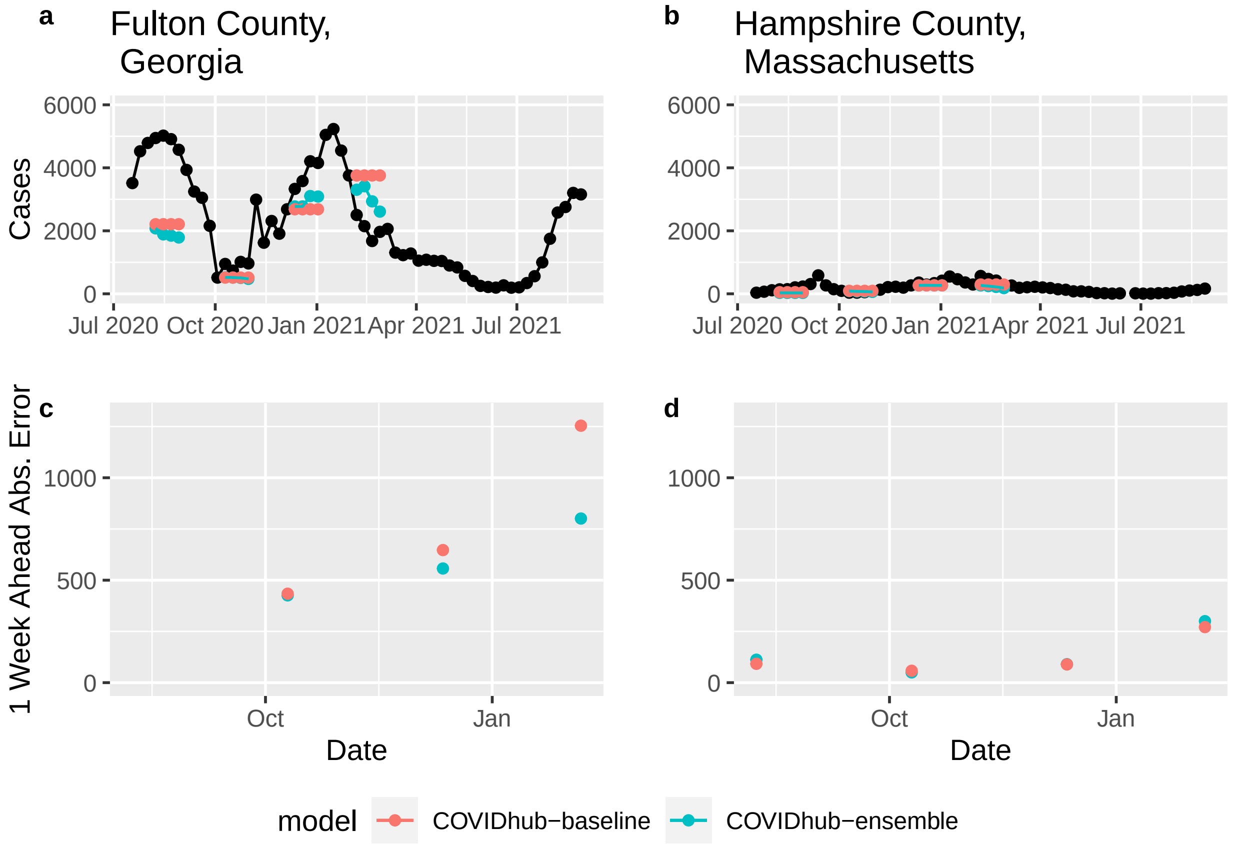
Increasingly, policy-makers turn to data-driven computational models to inform their understanding and decision-making regarding how to best serve populations at-risk for particular diseases [CITATION] . Forecasting models, defined here as models that attempt to predict future aggregate observations from a particular surveillance system (e.g. the total number of influenza-related hospitalizations in a future week), have been used frequently in recent years to help anticipate health-care system burden in outbreaks of influenza, dengue fever, Ebola, chikungunya, Zika, and COVID-19.[CITATION]

Recently, groups have started to assess the "fairness" of such models, measuring whether the accuracy of these models varies by the socio-demographic makeup of different geographic regions.(1,2) These efforts are critical to better understanding if, how, and to what extent biases present in the surveillance system may carry over to modeling efforts carried out with these data.

As it is important to distinguish between model fairness and the biases in the underlying data, we define a "fair model" as one that does not add any additional "unfairness" to what exists within the data, i.e. inaccuracy is not compounded. Different groups having different rates of a particular condition is commonly referred to as disparities in a health outcome. A system's ability to measure accurately the outcomes in different groups may vary, yielding different quality in available data: this could yield biases in the resulting data that measures the health outcome. Finally, a statistical model fit to data, while typically unable to see beyond the assumptions used to build the model, can introduce additional bias if appropriate modeling techniques are not followed to ensure, for example, that the analyzed sample is representative of a larger population.

In population-level models, such as outbreak forecasting models, often the same socio-demographic variables that are being considered as drivers of model (un)fairness are also associated with higher case counts of the predicted outcome [CITATION]. Higher case counts are typically associated with higher model errors on an absolute scale (e.g. mean absolute error or root mean squared error)(1,3) (Figure 1). Therefore, it is important to assess model fairness in a manner that takes into account and adjusts for the scale of the forecasted outcome data

We assess model fairness by comparing model error metrics for different locations to a naïve baseline model that objectively treats every location equally due to its simplicity.



Methods

Prior Work

Prior literature on this topic has assessed model fairness through adjusting for population size (Scarpino et al) and normalizing a measure of the predicted outcome over the window of prediction (Google). Scarpino et al. analyzed socioeconomic bias in influenza surveillance. They compared hospitalizations due to flu by "poverty quartile" by dividing their data by zip code into the quartiles based on the proportion of population living below the federally defined poverty line. To evaluate the models analyzed predictive performance they looked at the Out of Sample Root Mean Squared Error (ORMSE) per county where

$$\text{ORMSE}^{(l)} = \frac{\sqrt{\frac{1}{N}\sum_{t=1}^{N}(\hat{y}_{lt} - y_{lt})^2}}{\text{Pop}^{(l)}}$$

for location l , number of weeks N , and error $\bar{y}_{it} - y_{it}$ where is prediction and y_{it} truth at time t for location l and $Pop^{(l)}$ is the population of location l . They found that models had the highest ORMSE (i.e. the worst accuracy) for the most impoverished quartile of locations across all of the data sources they used.

Another white-paper analyzed US county-level forecasts of COVID-19 cases to look at the distribution of their model's prediction errors. They binned counties into four equally sized groups based on quartiles of a particular socio-demographic variable in a county, e.g. proportion of the county's residents who were Black, and then analyzed the distribution of the errors within those groups. The authors used a Normalized Mean Absolute Error (NMAE) across the quartiles where NMAE normalizes the sum of the absolute differences by the cumulative number of deaths in a county over the forecasting horizon:

$$\text{NMAE}^{(l)}(T, \tau) = \frac{1}{\tau}\frac{\sum_{t=T+1}^{T+\tau}|\hat{y}_{lt} - y_{lt}|}{(y_{l(T+\tau)} - y_{l(T+1)})_1}$$

[Paragraph explaining equation]

The above two approaches use slightly different normalization approaches. For computing the ORMSE, overall average model errors are normalized by a fixed population in a given location. For computing the NMAE, average model errors for a forecast made at a particular time (T) are normalized by dividing by a measure of future change in incidence. We note that these adjustments do not scale the forecast errors based on the absolute level of cases, but rather the trend. This means that if a forecast was made for a time-period for which cases increased by 100, the scaling factor is the same, regardless of whether the observed numbers of cases during this time went from 0 to 100 or 1000 to 1100.

Data

COVID-19 case data was retrieved from the Johns Hopkins University Center for Systems Science and Engineering (CSSE) COVID-19 dashboard.[CITATION] CSSE data were accessed via the COVIDcast platform.[CITATION] Daily cumulative case counts are available from CSSE, from which weekly incident counts may be inferred. For some analyses in this paper, counties were divided into four equally sized groups based on quartiles of the cumulative number of cases observed in each county from July 20th 2020 to March 6th 2021. There are 3,142 counties in the US. We analyzed data from the 3,117 counties with nonzero cumulative cases.

Population demographic data was taken from the American Community Survey which provides five-year demographic estimates by county from 2015-2019. [CITATION] We extracted the proportion of the population of each county belonging to a particular demographic group, e.g., White, Black and Hispanic, and the fraction of households above and below the poverty level.

The US COVID-19 Forecast Hub is a central repository for models making short-term weekly forecasts of COVID-19.[CITATION] The forecast data includes one to four week ahead predictions from the COVID-19 Forecast Hub ensemble model of new weekly cases due to COVID-19. The ensemble model produces quantile and point forecasts by taking the median of each prediction quantile for all eligible models for a given location.(4) Our data also include the baseline model which predicts for one to four week ahead incident cases the same value as the week prior.