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

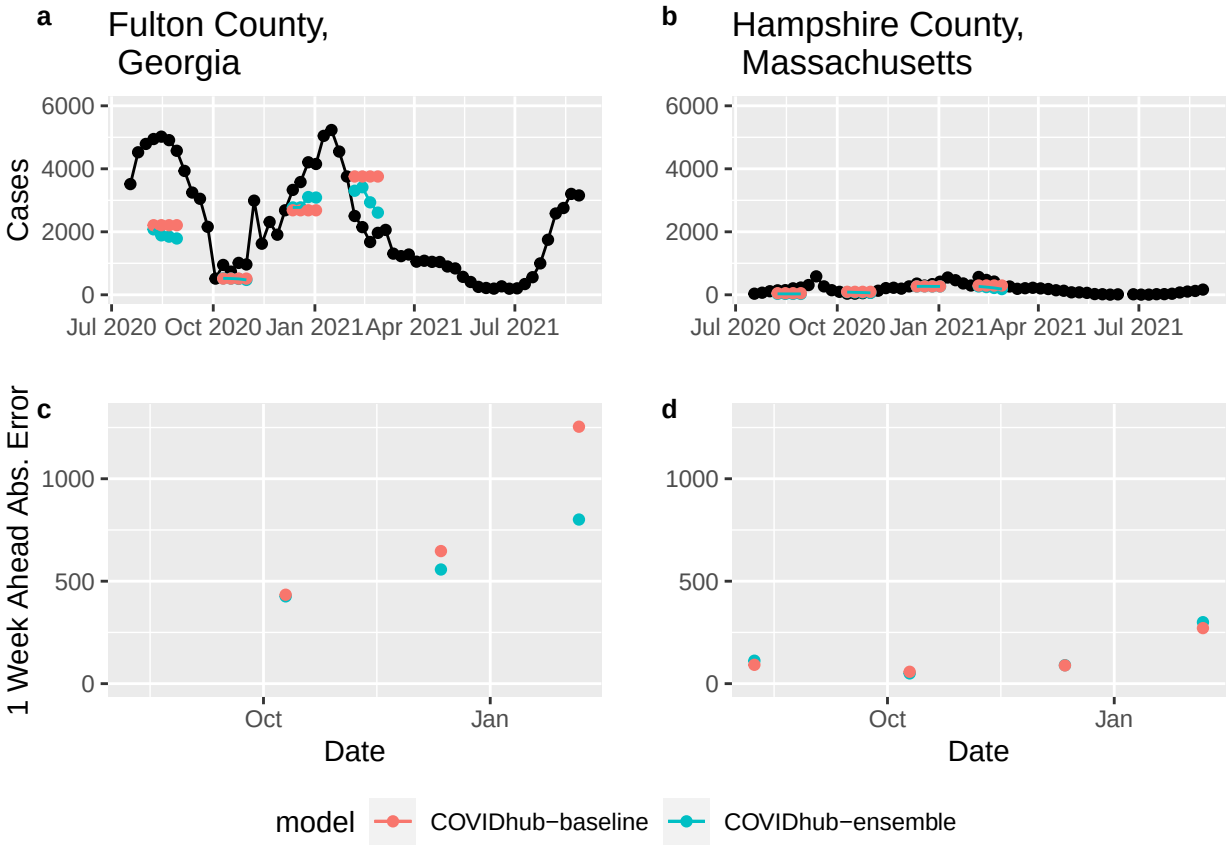
critical to better understanding if, how, and to what extent biases present in the surveillance system may carry over to modeling efforts carried out with these data.

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.

Figure 1



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.

Metrics

We define the mean absolute error (MAE) as

$$\text{MAE}_{lh} = \frac{1}{N} \sum_{w} |\hat{y}_{lwh} - y_{lwh}|$$

for a specific location l over a horizon h , by averaging over the N weeks (w) starting from forecasts made July 20, 2020, to forecasts made March 6, 2021. Higher values of MAE indicate less accuracy of the point forecasts from the model. We used regression methods to evaluate associations between the ensemble model's MAE and other independent variables including county-level incident cases and sociodemographic measurements. We performed one set of linear regressions for 1 week ahead MAE and another for 4 week ahead MAE. Within each set we predicted the MAE 4 times as a function of proportion in poverty and proportion black, total population, total cases, and cases with minority proportions respectively.(table 1) For 1 week ahead MAE as a function of total cases, the output of the linear regression model was:

$$\text{One Week Ahead MAE} = \beta_0 + \beta_1 * \text{Total Cases}$$

We further analyzed the data by looking at the relative mean absolute error (RMAE) . The RMAE was calculated for a specific location over a horizon , by calculating the MAE

$$RMAE_{l \cdot h} = \frac{MAE_{l \cdot h}(E)}{MAE_{l \cdot h}(B)} = \frac{\frac{1}{N} \sum_{w} |\hat{y}_{lwh}^{\text{Ensemble}} - y_{lwh}|}{\frac{1}{N} \sum_{w} |\hat{y}_{lwh}^{\text{Baseline}} - y_{lwh}|}$$

at location l and horizon h for the Covid-19 Forecast Hub ensemble model and dividing that value by the MAE at location l and horizon h for the Covid-19 Forecast Hub Baseline Model. The Baseline Model which is a naive model and predicts for horizons one to four weeks out the same number of incident cases as the prior week at location . RMAE is nonnegative with focus being placed as to whether it is above or below one. When the RMAE is below one it means that the ensemble model was a more accurate predictor at location l and horizon h than the baseline model was. Conversely, when the RMAE is above one it means that the ensemble model was a less accurate predictor at location l and horizon h than the baseline model was.

We used nonparametric Wilcoxon rank-sum tests to assess differences in 1 week ahead median MAE and RMAE between groups of counties. Specifically, we assigned every county to a "case quartile" based on how many total COVID-19 cases the county reported by XX date, and to a "race quartile" based on the proportion of Black constituents in the county. Within each particular case quartile (that is, among counties with similar numbers of total cases) and among all counties overall, we performed 3 Wilcoxon tests. Within each strata, each test compared counties with higher proportions of black constituents to the counties that had the lowest number of black constituents. Since three tests were conducted for each of five groups and two Since three tests were conducted for each of five groups and two methods were used (MAE and RMAE), to adjust for the large number of comparisons, we applied a Bonferroni correction by indicating a significance level of 0.05/30.

Table 1

	R Squared	Adj. R Squared
1 wk Ahead MAE as a funct. of Prop. Black and Prop. Pov	0.0472	0.0463
1 wk Ahead MAE as a funct. of Total Population	0.8711	0.8711
1 wk Ahead MAE as a funct. of Total Cases	0.9382	0.9382
1 wk Ahead MAE as a funct. of Cases and Minority Prop.	0.9389	0.9389
4 wk Ahead MAE as a funct. of Prop. Black and Prop. Pov	0.0405	0.0396
4 wk Ahead MAE as a funct. of Total Population	0.9437	0.9437
4 wk Ahead MAE as a funct. of Total Cases	0.9767	0.9767
4 wk Ahead MAE as a funct. of Cases and Minority Prop.	0.9768	0.9768

Results

Untangling the relationship between demographics, case counts and model error

The magnitude of the absolute error in the forecasting of 1 week-ahead county-level COVID-19 reported cases is associated with changes in sociodemographic variables such as poverty and race. We looked at aggregate errors across four groups of counties, where counties were grouped into quartiles based on the percentage of the population belonging to specific socio-demographic groups. The model made less accurate predictions for counties with higher proportions of black constituents (Figure 2a, Table 2). The median county-level MAE of the ensemble forecast in the 25% of counties with the lowest percentages of black constituents, was 10.3 (IQR: 5.917- 17.890) compared with a median MAE of 29.5 (IQR: 14.721- 86.246) in the 25% of counties with highest percentages of black constituents. Each of the 2nd, 3rd, and 4th quartiles had distributions of MAE that were significantly higher than those in the 1st quartile (non-parametric Wilcoxon Rank-Sum p-values all less than 1x10^-15, Table 2). When the relationship between proportion of minority population and accuracy was assessed with the proportion as a continuous variable and not grouped into quartiles, the error increased rapidly and then decreased (Figure 3a).

However, there was a strong positive association between case counts and demographic variables at the county level. Higher proportions of black constituents were associated with larger cumulative case counts at the county level (Spearman correlation coefficient of 0.43, p-value < 0.0001). Meanwhile, higher proportions of households in poverty showed a small but significant correlation with smaller cumulative case counts at the county level (Spearman correlation coefficient of -0.08, p-value < 0.0001).

An investigation of the relationship between both total COVID-19 cases and socio-demographic variables with 1 week-ahead ensemble forecast MAE reveals that case counts capture substantially more variation in model accuracy than socio-demographic variables alone. Highlighting the impact of total cases on MAE, within every quartile of counties grouped by the proportion of black constituents, the median county-level MAE monotonically increased as case counts increased (Figure 2b). This is exemplified by the example shown in Figure 1, comparing Fulton County, Georgia to Hampshire County, Massachusetts, where higher case counts led to higher errors.

When stratifying by the level of cases a county experienced, the proportion of black residents was no longer consistently associated with higher model error. We used the Wilcoxon rank-sum test to evaluate the significance of the differences in errors across the racial demographics within counties with similar levels of cases (Table 2). We observed that some differences in errors were statistically significant even among counties that were in the same quartile of overall case counts. For the case quartiles representing the lowest numbers of cases, the practical significance of the differences in model error was small [add a specific number here?]. The quartile of counties with the highest case counts spanned a range between xxx to yyy cumulative cases per county. Within this quartile of the counties with highest cumulative incidence, an increasing proportion of black residents was associated with increasing case counts, indicating that the quartile stratification approach (as implemented in [add citation number]) may not fully adjust for the confounding between case counts and demographics (Supplemental figure).

Additionally, a regression analysis on all county-level observations showed that cumulative case counts in a county explained substantially more variation in model error than sociodemographic components. Looking at the regression model for one week ahead MAE as a function of total cases we see that its adjusted R2 value (0.89) is identical to that of a model that also includes socio-demographics (R2=0.89). When adjusting for total cases, county-level socio-demographic variables did not show a statistically significant association with 1 or 4 week-ahead MAE.

figure 2

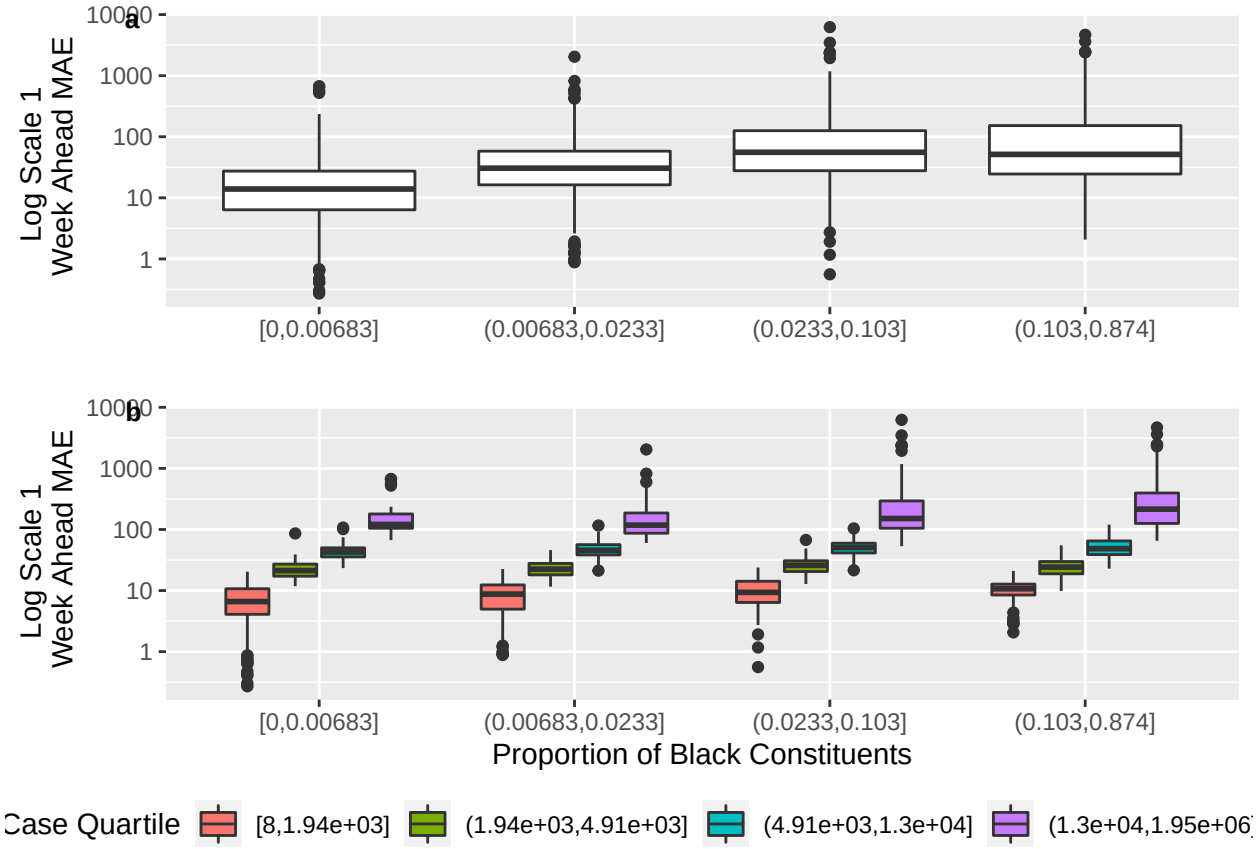


Figure 3

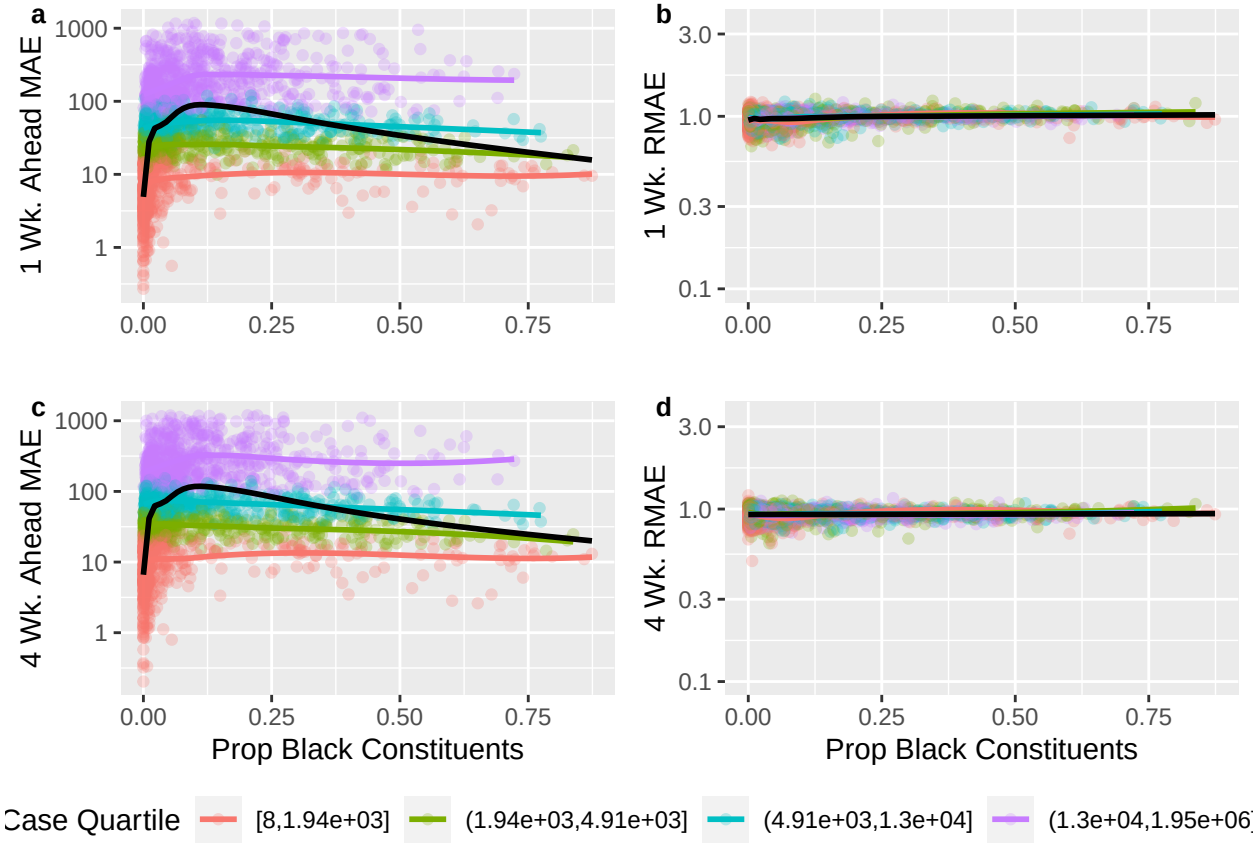


Table 2

		Prop. Black Quartile				
Case Quartile	Range	lowest 25%	Q2	Q3	highest 25%	Overall
lowest 25%	[2, 1.09e3]	6.059	6.559	7.794	7.632	6.647
Q2	(1.09e3, 2.64e3]	14.103	14.882	17.706	15.147	14.912
Q3	(2.64e3, 6.82e3]	25.853	29.088	29.603	28.912	28.515
highest 25%	(6.82e3, 1.25e6]	79.059	73.368	91.324	111.779	93.353
Overall		10.279	19.588	35.618	29.5	20.735

Table 2: Pink for statistically significant at $\alpha = 0.05$ and red for statistically significant at $\alpha = 0.00333$ (from the Bonferroni Correction over 15 tests) when comparing to Quartile 1 of Prop. Black. across the case quartiles and overall

Using a relative error metric to adjust for outcome level

The relative mean absolute error (RMAE), computed as the error of the ensemble forecast divided by the error of the baseline forecast, showed less variation overall than the unscaled MAE values (Figures 3 and 4). The naive baseline model makes the same objective prediction in every county (predicting the most recent observation into the future), and therefore serves to adjust for the scale of cases in each county. The overall median 1-week ahead RMAE across all 3,117 US counties for which forecasts were made was 0.90, meaning

that the ensemble forecast in 10% less error on average than the baseline when predicting one week ahead. For the middle 80% of counties sorted by relative performance of the ensemble, RMAE varied from 0.79 to 1.00. The ensemble made more accurate forecasts than the baseline forecast in over 90% of all counties.

The stability in the values of the relative error metric (RMAE) contrasts with the variation in the values of the mean absolute error (MAE), where differences between counties with different socio-demographics are measurable and in some instances significant (Figures 3 & 4, Table 2). The MAE was significantly larger in counties with larger proportions of black constituents (even when stratifying by the total number of cases in each county, Table 2). However, the RMAE values showed little variation as a function of the proportion of black constituents (Table 3). Within the counties with the highest number of cases, we observed a relatively constant pattern of RMAE across the race quartiles. Overall, these counties had an RMAE of 0.89 (11% less error than the baseline on average) and values for specific race-quartiles ranged from 0.89 to 0.92 (Table 3). This contrasts with the high variance in MAE across the same high-case counties, which showed an overall MAE value of 93.4 and ranged from 73.4 to 111.8 across race quartiles (Table 2). Tests for significant differences between groups of counties with different percentages of black constituents showed that in some comparisons ensemble showed more error (higher MAE) in groups of counties with larger black populations but the ensemble was significantly more accurate relative to the baseline (lower RMAE) in those same counties.

We performed the Wilcoxon Rank- Sum test on the RMAE values similar to our analysis with MAE, in order to see if there are statistically significant differences in RMAE amongst the differing case and race quartiles. The results from comparing per case quartile and overall race quartiles 2, 3, and 4 to race quartile 1 resulted in 3 statistically significant lower RMAE values all within race quartile 4 with 2 remaining statistically significant after the bonferroni correction. These results show less variation in RMAE amongst the race quartiles compared to MAE and the significance values showing slight improvement in ensemble model accuracy. Of the 15 comparisons performed 13 (86.7%) were insignificant after bonferroni correction for RMAE and 8 of the 15 comparisons (53.3%) for MAE.

Figure 4

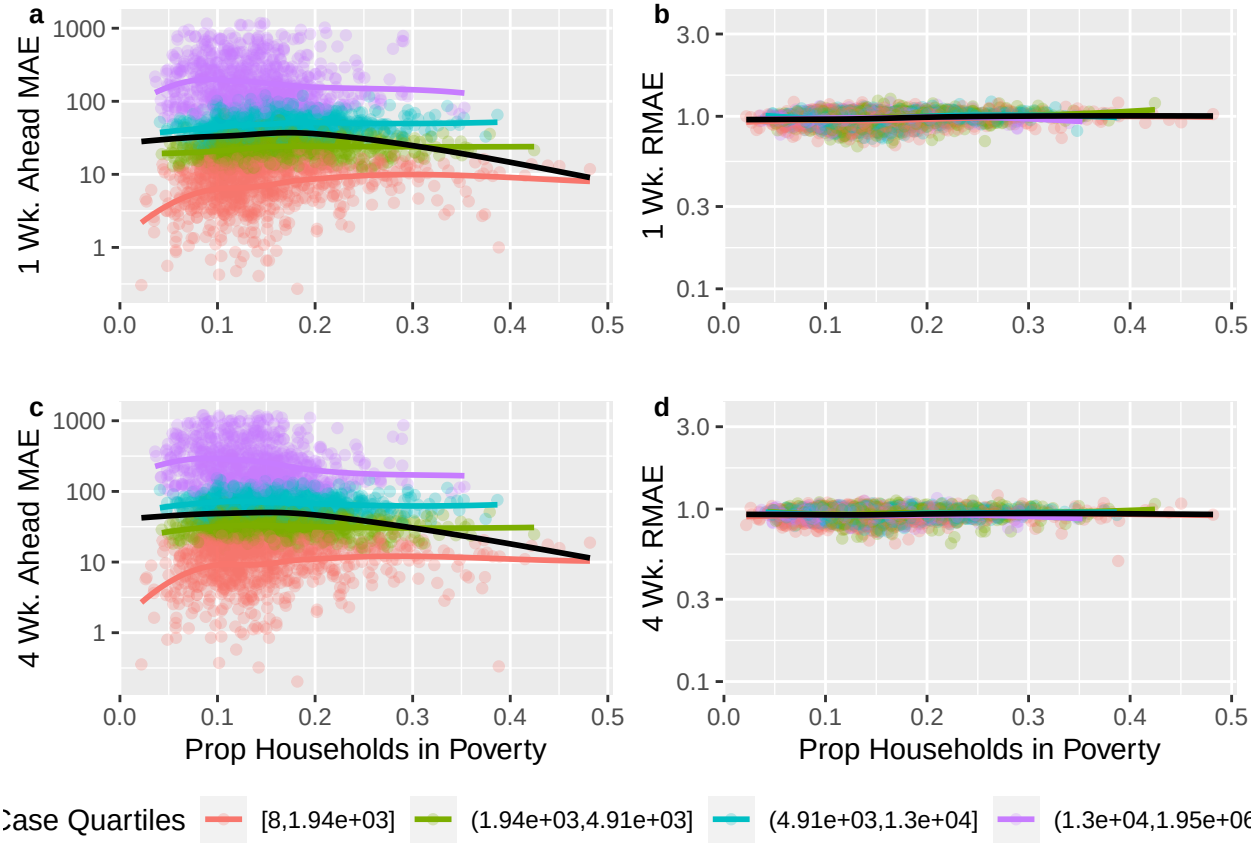


Table 3.

		Prop. Black Quartile				
Case Quartile	Range	lowest 25%	Q2	Q3	highest 25%	Overall
lowest 25%	[2, 1.09e3]	0.898	0.901	0.910	0.885	0.898
Q2	(1.09e3, 2.64e3]	0.906	0.887	0.902	0.887	0.897
Q3	(2.64e3, 6.82e3]	0.911	0.899	0.905	0.879	0.898
highest 25%	(6.82e3, 1.25e6]	0.923	0.890	0.889	0.894	0.891
Overall		0.903	0.896	0.897	0.887	0.896

Table 3: Pink for statistically significant at $\alpha = 0.05$ and red for statistically significant at $\alpha = 0.00333$ (from the Bonferroni Correction over 15 tests) when comparing to Quartile 1 of Prop. Black. across the case quartiles and overall

Discussion