COVID-19 Public Forecasts: Fairness Analysis

Introduction

Google is committed to a core set of AI principles. These principles lay out how we think about using technology to solve important problems, but they also identify important challenges that we need to address clearly, thoughtfully, and affirmatively to responsibly develop technology. In developing the COVID-19 Public Forecasts, we paid close attention to the disproportionate impact the disease has had and how that would impact our adherence to these principles, particularly principle #2: "Avoid creating or reinforcing unfair bias."

CDC research has shown that communities of color in the United States have been the hardest hit by COVID-19 with disproportionately high rates of cases and deaths. This is likely due to issues related to structural racism, various systemic inequities in access to healthcare, inherent systemic bias, and underlying negatively impacting social determinants of health. In developing our forecasting model, our team was highly conscious of these inequities, and undertook a thorough analysis of how they might surface in our forecasting model.

Predicting county-by-county COVID-19 cases is harder than predicting state-by-state, because the number of cases in each county is usually much smaller than the statewide counts. So, accurate predictions are difficult. The Google team developed increasingly accurate prediction algorithms over a period of weeks and then let the machine learning algorithm run without intervention. We then compared the predictions with the actual counts and also other available county level forecasts. We observed our model made errors typically around 10-20% in predicting case counts, which is meaningfully lower than other benchmark county level models.

For simplicity in calculations below, we'll assume a 10% average error for our model. For example, if the model predicted 100 cases over the next two weeks, the actual count would likely end up between 90 and 110, and even if the true number falls outside those limits, it is likely that it will be between 80 and 120. Note that these "plus or minus 10%" prediction accuracies remain constant for counties with large numbers of cases and counties with small numbers of cases. If we predict 50, the actual number is likely between 40 to 60. If we predict 500, the actual number is likely between 400 and 600. And if we predict 5000 then the actual number is likely between 4000 and 6000.

This phenomenon has a significant issue: if instead of considering percentage error (which is roughly constant around 10-20%) you look at the absolute error (the difference between actual and predicted counts) then you see that for counties with larger caseloads, even a small percentage error can translate to a significant under- or over-counting of total cases. The sections below explain why this is serious when we consider the relationship between the absolute errors and the demographics of counties.

We identified this important point after evaluating through our AI Principles governance process, which led to this fairness analysis supported by health and AI fairness experts from across the company. Though we remain troubled by the underlying disproportionate impact of COVID-19 and higher absolute error it causes on communities of color, we concluded after considerable deliberation, consultation with external experts, and statistical analysis that the COVID-19 Public Forecasts would serve as a valuable input for decision making, so long as the outputs were appropriately contextualized and understood.

We hope that this approach and focus on fairness will contribute to a broader discussion around how AI can not only support the response to COVID-19 but also support efforts to identify and address disproportionate impacts. Beyond COVID-19, we also hope this approach will spur greater consideration of fairness implications in AI research and deployments overall. We will continue working to improve the COVID-19 Public Forecasts and invite further feedback and discussion from others on this important topic.

Absolute Errors Vary with Different Demographic Makeups

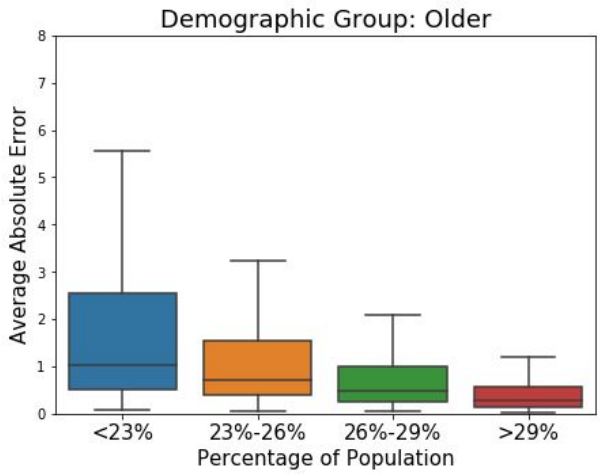
Different counties in the United States have different demographic makeups. Some counties tend to have a higher proportion of elderly residents, some have a higher proportion of African American residents, and so on. Wherever there is demographic variability, it is important to audit a prediction method for algorithmic bias, using a methodology such as SMACTR, proposed in https://arxiv.org/pdf/2001.00973.pdf. In this case, when we examine the relationship of demographic makeup to absolute errors in prediction, we observe a critically important point: although the relative errors are fairly constant, the absolute errors are not.

We observe this important point of higher absolute errors for many different demographic makeups — here we describe it for three makeups: age, racial and ethnic demographics, and income. It's important to say that our model only makes predictions at the county or state level, not for any individuals. So, we looked at demographic information from the US Census. The Census tells us things like "What is the average age in a county?" or "What percentage of a county's residents self-identify as Black?"

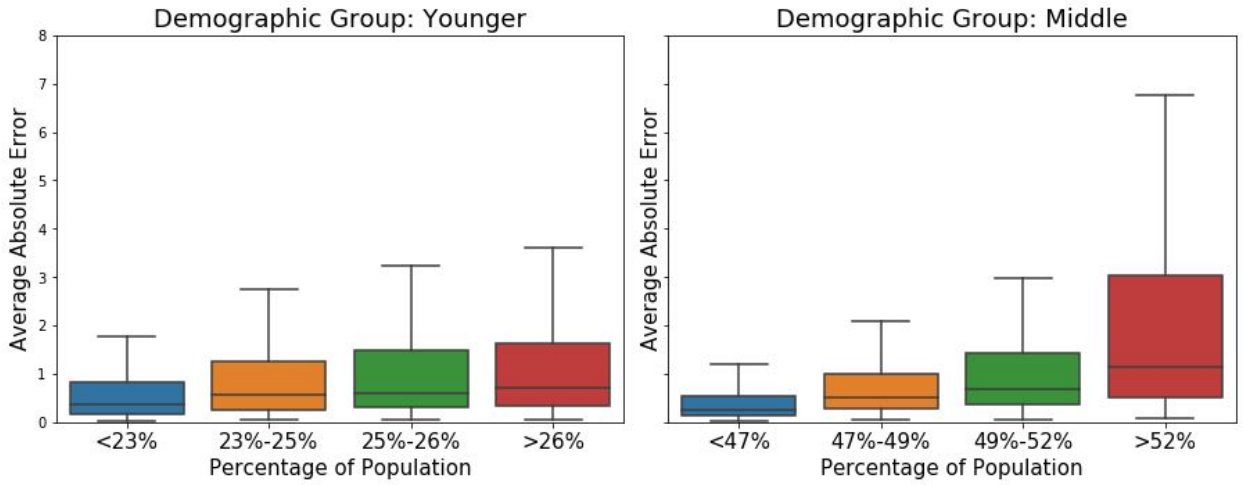
Age

In the graph below, we look at how our forecasting error (y-axis) changes as counties have an increasing percentage of a certain demographic (x-axis moving left to right). For this analysis, to better be able to observe trends across demographic groups, we bin the counties into four groups of an equal number of counties according to the percentage prevalence of that demographic group. For each of the four bins, we show a box plot which shows the maximum, minimum, median and lower/upper quartiles for the absolute error in that bin.

First, we focus on the older age demographic group (>60 yo). The first bar from the left (shown in blue) shows the average absolute error for counties in the first bin (<23% percentage prevalence) for older people, and the other bars show that same error for the other percentage prevalence groups. We observe that the absolute error gets progressively lower for counties with a higher proportion of older people.



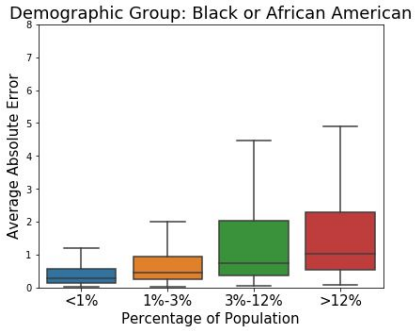
Below we provide the same results for two other age groups: younger (<20 yo) and middle (20 - 60 yo) aged, one graph each.

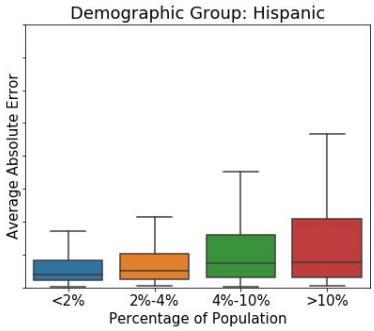


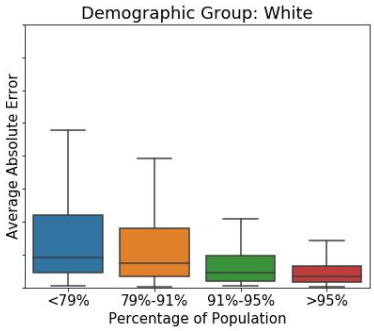
As can be observed, the absolute errors vary significantly. In particular, the error is significantly higher for counties with a higher proportion of younger and middle-aged people. This increased error for younger and middle-aged groups is proportional to the increased COVID-19 case counts for these groups.

Race and Ethnicity

The following graphs show a similar pattern for race and ethnicity. The population is again divided into three demographic groups, this time according to race and ethnicity (we pick the three most common groups in the United States: African American, Hispanic, White) using US Census data. Similar to age, we observe differences between groups in absolute errors. This time, we observe higher absolute errors for counties with higher prevalence of African American and Hispanic individuals and lower absolute error for majority white counties.

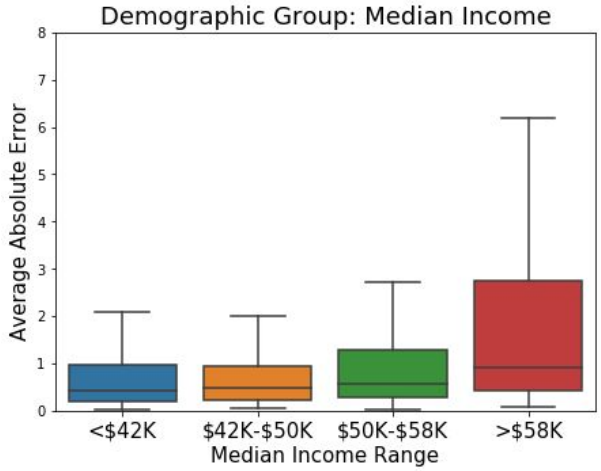






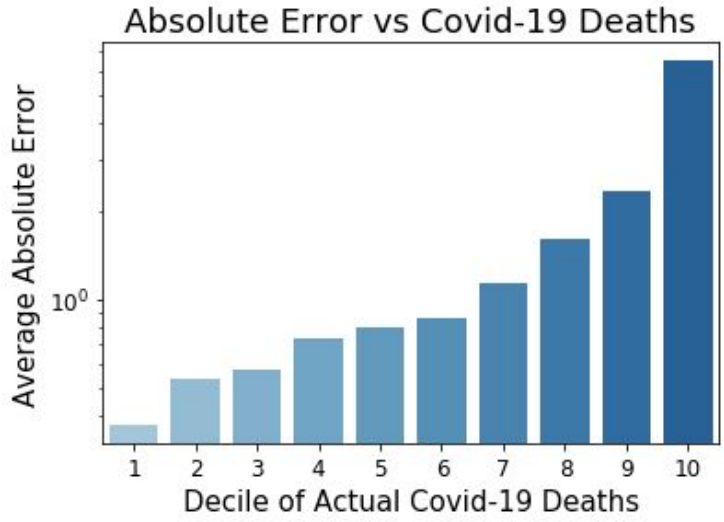
Income

For income we do not need to divide the population into separate plots but instead directly bin county populations according to their income. Similar to the above pattern we observe differences, here higher absolute errors for higher income counties.



Investigating the Causes of Higher Absolute Error

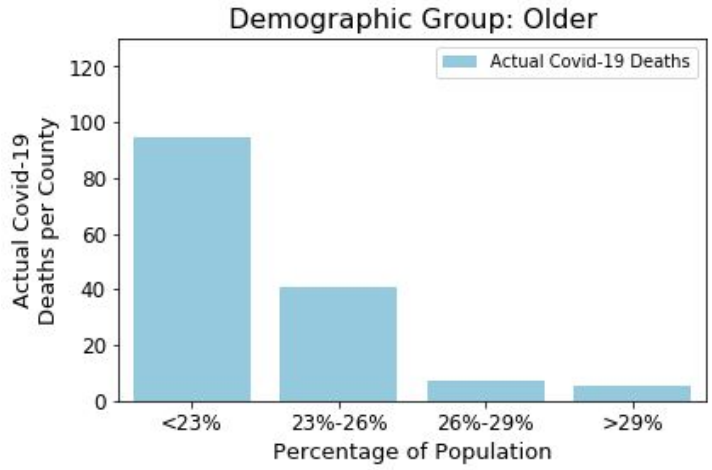
We worked with experts across Google to analyze the COVID-19 Public Forecasts' accuracy and in particular this issue of higher absolute errors for certain demographic groups. During our investigation, we identified a hypothesis which looks like it may explain these observations: higher absolute error correlates with higher case counts. This can be seen in the plot below showing absolute error (y-axis) plotted by actual death counts (x-axis) in counties.



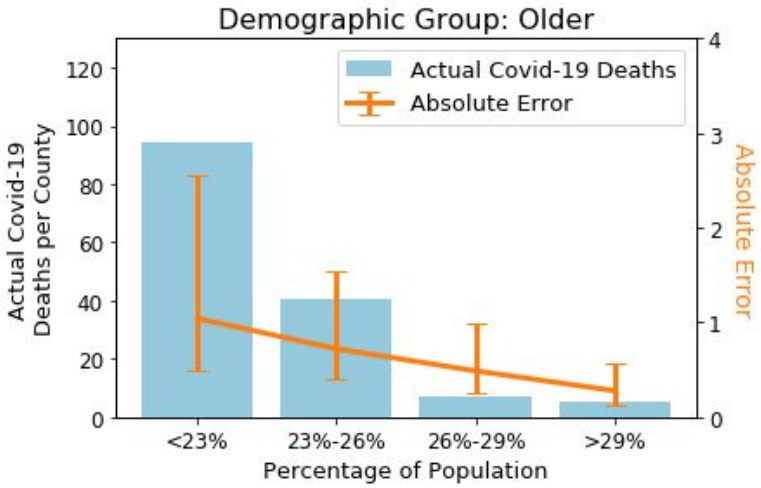
In particular, we observe that all of the absolute error differences between demographic groups are directly correlated with higher case counts.

Age

Similarly to above, we will look at the older age demographic. First, we show the actual COVID-19 deaths for the older age group across the same percentage prevalence bins as in the previous section. Interestingly, we observe that the actual death counts decrease for counties with lower older populations. This may be due to these counties being less dense, more rural, and less populous.



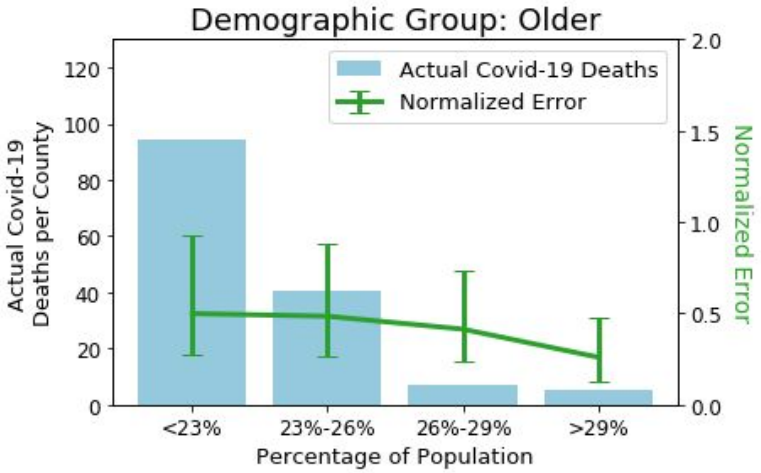
The graphs below show the actual death counts overlaid with the absolute errors from the previous section (errors shown with orange lines, with a line connecting the medians and the range showing the 25th and 75th percentiles). As can be observed there is a direct correlation between absolute error and actual deaths: as the actual deaths decrease, so does the absolute error, and vice versa.



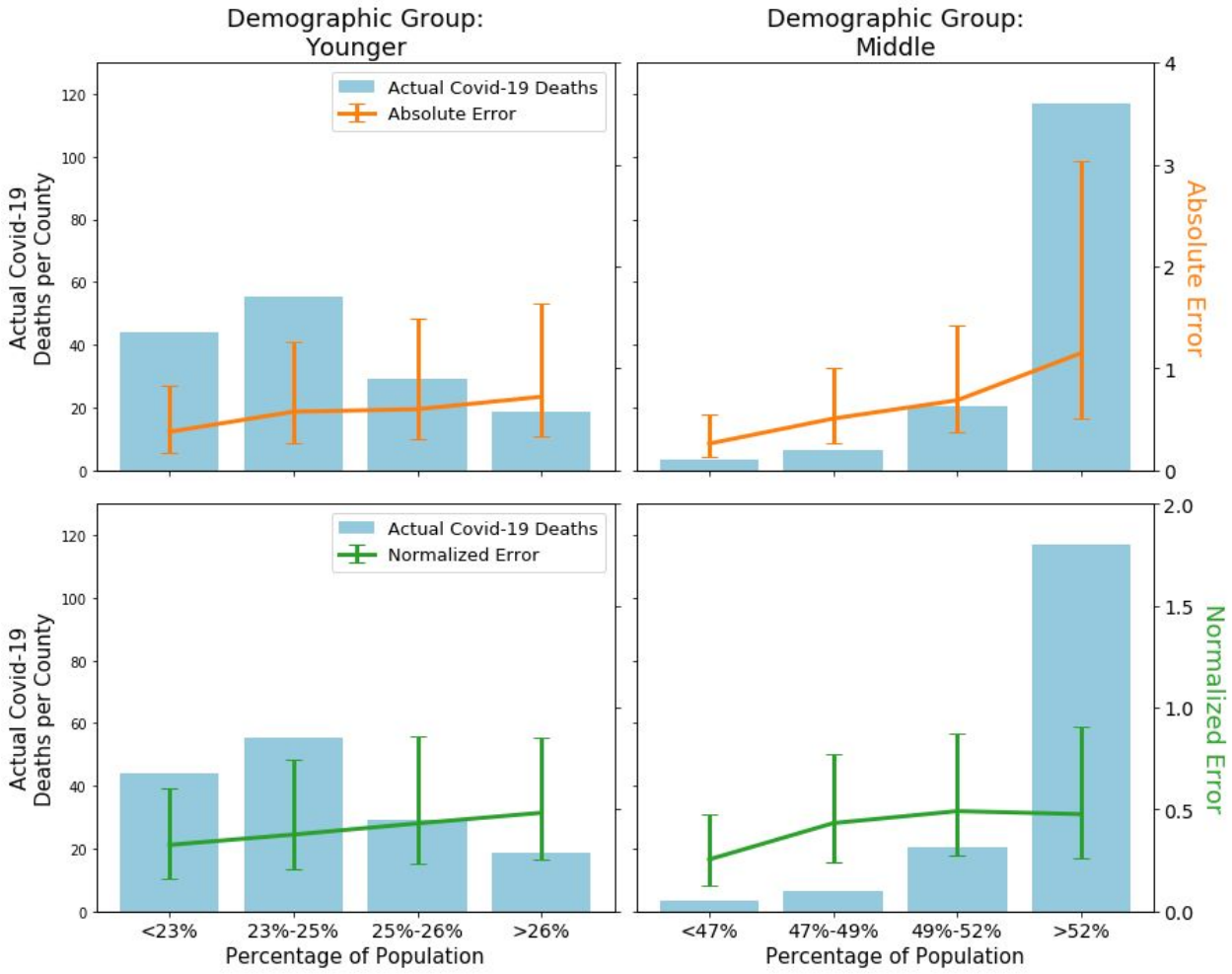
Normalizing Absolute Error Helps Reduce Difference Across Demographic Makeups

To further explore the correlation of the absolute errors with the COVID-19 deaths, we investigate whether the difference in deaths may in fact explain the difference in error. One struggle our team ran into during our fairness analysis is the lack of consensus on how to define and measure normalized error, which in this case is important to evaluate how the model performs due to the bias in actual case counts. After consulting with various experts, we moved forward with a metric that normalizes based on the number of deaths in a county. However, we recognize this is an imperfect definition and encourage further exploration of better ways to normalize error metrics for time-series forecasts.

In the graph below we show that if this correlation is taken into account by normalizing the error with the death count (dividing the error by death count), the difference across demographic groups is significantly reduced and the confidence intervals are overlapping. This indicates that the difference in absolute error across demographic groups can be largely explained by the difference in COVID-19 deaths/case counts.

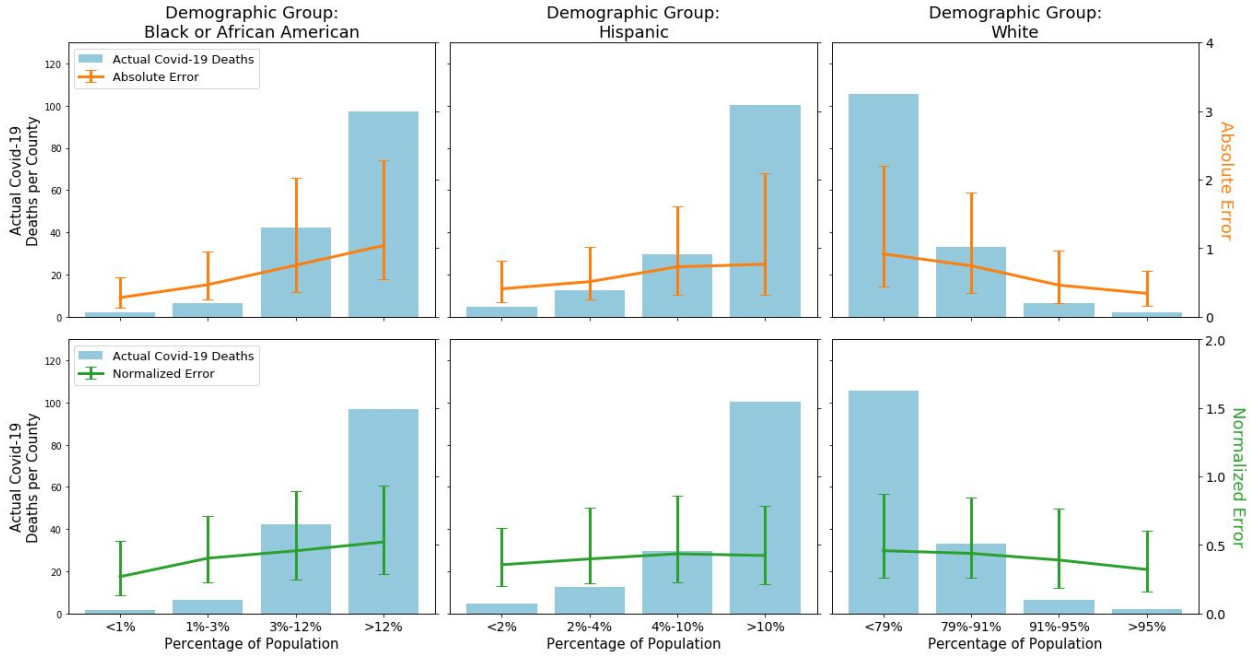


We next perform this same analysis for the other demographic groups (younger and middle age groups) and observe the same pattern: after the absolute errors are normalized by actual death counts, there is less difference between the confidence intervals across the demographic groups.



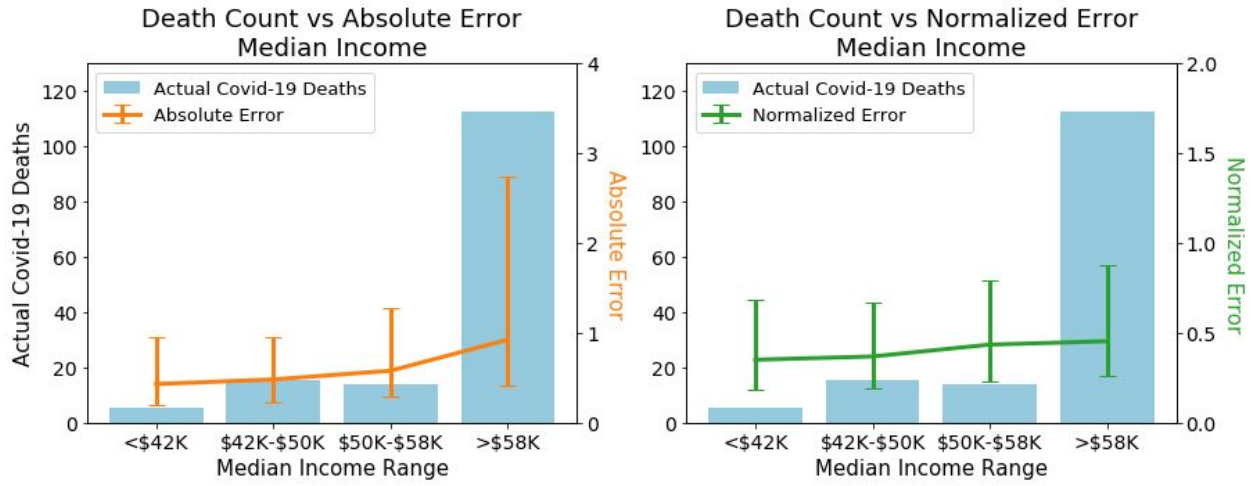
Race and Ethnicity

Similar to above, there is a direct correlation between the absolute errors and death counts, and this is meaningfully reduced when the error is normalized by the death count.



Income

We see the same pattern for the other breakouts: the difference in absolute error is directly correlated with death counts.



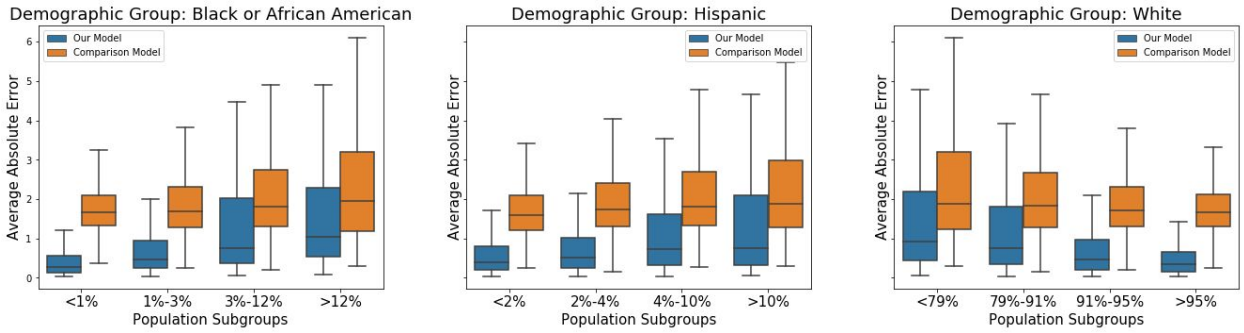
Benchmarking with Other Models

In addition to performing the above analysis, it is important to consider how our model's behavior compares to other available public forecasts to evaluate how our model and the bias in absolute error may impact decision-makers. As most demographic variability will occur across counties, we benchmark with another county level forecast in the United States. Unfortunately, there are very few models that are

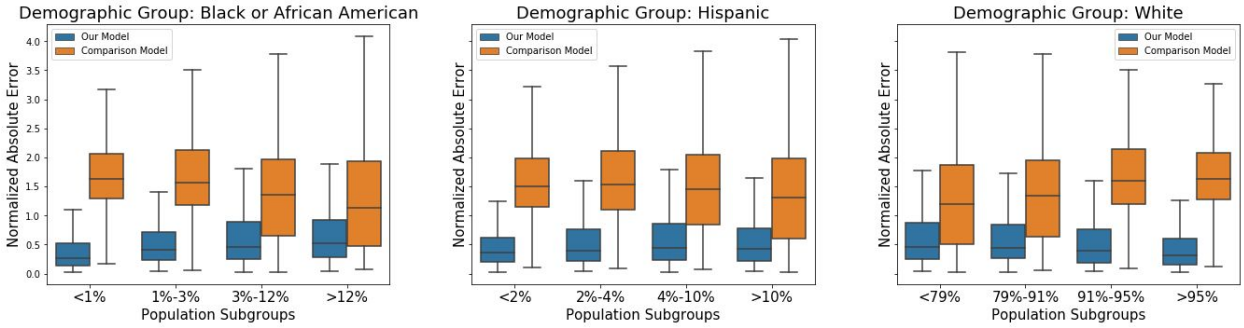
able to achieve county level forecasts. For benchmarking, we use a top publicly available comparison model and specifically look at racial and ethnic minority groups.

As shown below, we observe that our model produces meaningfully lower absolute error and normalized (relative) error as compared to the comparison model across predominantly African American, Hispanic, and white counties.

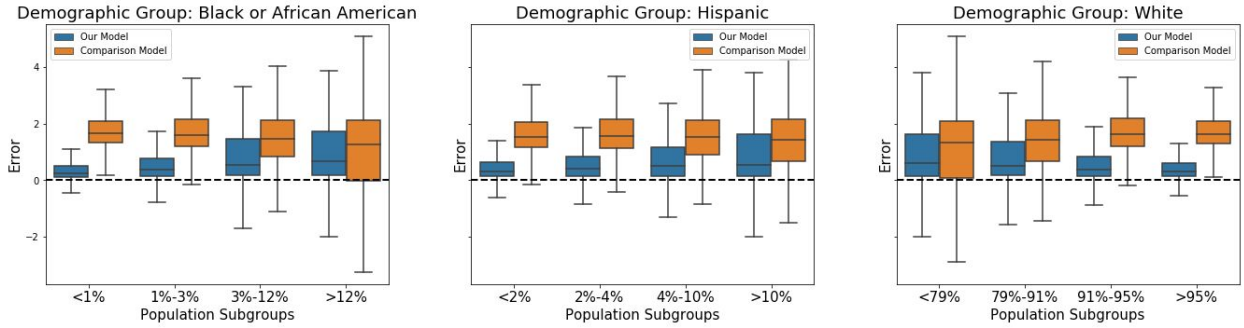
The graphs below show the same data as in the previous sections, but this time with two box-plots for each group, one for our model (blue) and another for the comparison model (orange). The comparison shows that our model has lower absolute error across racial and ethnic minority groups.



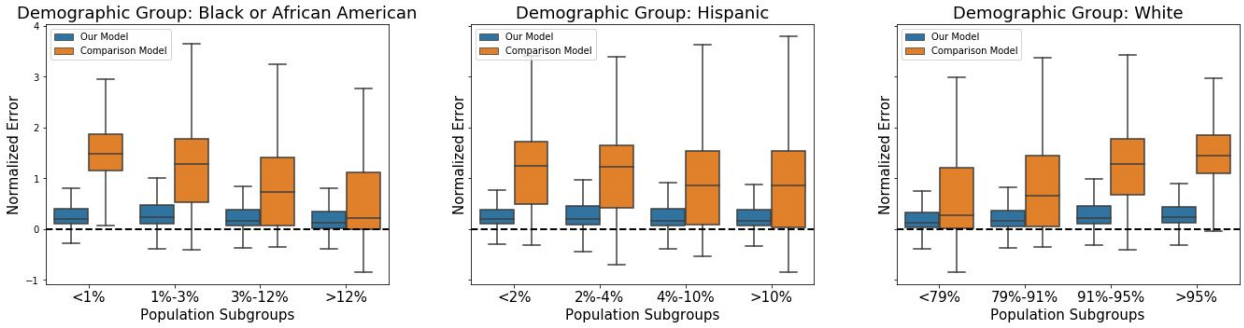
Normalized Error also shows meaningfully lower errors for the our model across all demographic groups:



Finally, Mean Error (another metric used to compare models) also shows lower error for our model. In addition, it shows our model has a very slight propensity to over-forecast rather than under-forecast. Based on guidance from public health experts, this slight propensity to over-forecast is actually preferred.



Normalized Mean Error shows the same trend as above as well as our model's equal likelihood to slightly over-forecast across subgroups:



As can be seen in all of the above plots, our model reduces forecasting errors across all of the demographic groups that we considered.

Conclusions

Our model optimizes for high accuracy across all US counties to provide the best overall forecast for most communities. Yet COVID-19 has also highlighted long-standing, yet unacceptable, health disparities across the country in its disparate impact on communities of color, and our model's errors reflect these disparities. Mainly, locations with high case counts and deaths in the historical data also yield higher absolute errors, although the relative (or "normalized") errors are comparable across locations.

When we compare performance of our model in counties grouped by racial and ethnic demographics, we observe that our model yields higher absolute errors for locations with higher ratios of racial and ethnic minority groups; importantly, the relative errors are comparable. It's important to note that the model serves as another way in which these pre-existing persistent national disparities and inequities are highlighted, showing how some communities are impacted by COVID-19 more than others— as we have seen, locations with the highest case counts tend to have higher ratios of racial and ethnic minority groups. While it is reassuring that our model's relative error remains similar across groups, the absolute error trends are a reminder of the challenges we face and of how the nation must do better.

We also demonstrate that our model yields lower absolute and relative errors compared to alternatives

both overall and for the demographic makeups we analyzed. However, we strongly encourage users to consider the COVID-19 Public Forecasts as one of many inputs in their decision making, and specifically point to the shortcomings highlighted in this analysis to caution against reliance on the absolute values of the forecast. For example, in determining resource allocation needs (e.g. planning for personal protective equipment, hospital beds, etc.), users should be aware that the disparities in the ground truth data that are used to train these forecasts lead to higher absolute errors for some demographic groups, including counties with more African American, Hispanic, and younger residents.

We remain committed to ensuring that our AI models are both valuable to the broader public and fair to vulnerable populations. Accordingly, we recognize that there is more work that needs to be done to help address the societal inequities that are brought to the foreground in this analysis. We encourage more dialog amongst all concerned parties including public health officials and the AI community in how to address these inequities and measure as well as improve how they may appear in various AI models.