Covid-19 was first characterized as pandemic by the World Health Organization on 11 March 2020 (WHO, 2020). By this date there were approximately 1181 confirmed COVID-19 cases in the United States with only 10 confirmed deaths (The New York Times, 2021). As more cases were confirmed, public health officials tried to determine what factors increased spread and what measures might prevent the spread of COVID-19. Initially masks were not recommended. This guidance changed over time as more was learned about the virus and mask supply increased. However, mask use was not widely adopted in all regions of the United States. In this report we aim to determine if the wearing a face mask differing amounts of time impacted the number of cases or deaths in a county. In addition, through the PCA analysis, we try to find the components of the data that explains the most variability. We also determine the number of days deaths lag behind a positive case and if the expected increase in death, as we have one more person infected, changed through the course of the pandemic.

This data, and therefore the conclusions, are limited on the data available to us. We are only evaluating the total number of cases and deaths, which will be impacted by population density. We believe there are other factors that should be considered such as number of individuals over the age of 65, number of families living below the poverty level, and number of individuals who identify as a minority. These factors have been shown in other reports to have an impact of COVID-19 cases in a community. As this data was not provided to us, we did not investigate its impact but care should be taken when generalizing the following results.

Methods

Data analyzed and presented in this report came from The New York Times (2021). Mask use was classified as either “Always”, “Frequently”, “Sometimes”, “Rarely”, or “Never.” We summed the proportions of “Always” and “Frequently” to classify mask use as either compliant, intermediate compliance, or not compliant for each county. A county was categorized as “Compliant” if the proportion of respondents that said either “Always” or “Frequently” were greater than or equal to 75% and “Not compliant” if less than or equal to 50% of the respondents said “Always” or “Frequently”; intermediate was used if between 50% to 75% responded “Always” or “Frequently.”

Once the data was classified, an analysis of variance (ANOVA) was performed to determine if there were differences in the three mask usage categories. Log-transformed total cases and log-transformed total deaths were used as the response variables. When mask usage category was significant ($p < 0.05$), pairwise comparisons were performed for the three categories using Tukey’s adjustment.

The PCA analysis uses the data organized by counties with the total cases and total death so far with the survey results. The analysis considers four variables: log of total cases, log of total deaths, death to cases ratio, and the proportion of residents responded that they wear a mask always or frequently. Since each variable has different measurements and variabilities, the correlation matrix was used to obtain the results for the PCA analysis.

To determine the number of days that deaths lag behind cases, we first calculated the rolling seven-day average of cases and deaths. Then, we found that each variable had three peaks throughout the pandemic, and the average of the differences in dates of the peaks for each variable was used as the estimated lag. With the estimated lag, we defined linear regression models using 60 days per model over the period of the pandemic. The dependent variable here is the seven-day

average of death after taking account of the expected lag and the explanatory variable is the seven-day average of the total cases. The slope coefficient is set at 0, since 0 cases obviously indicates 0 death.

Results

For the total cases, there were significant differences between the three mask usage categories ($p < 0.0001$). The black bars indicate that all three categories were significantly different from each other with respect to the total cases (Figure 1). In locations that were classified as compliant, there were the most total cases, followed by intermediate and not compliant (Figure 1). There were also significant differences in the total deaths ($p<0.0001$). The red bars indicate that the trend was the same for the total deaths (Figure 1). Compliant locations had the most deaths, followed by intermediate locations, and lastly not compliant locations.

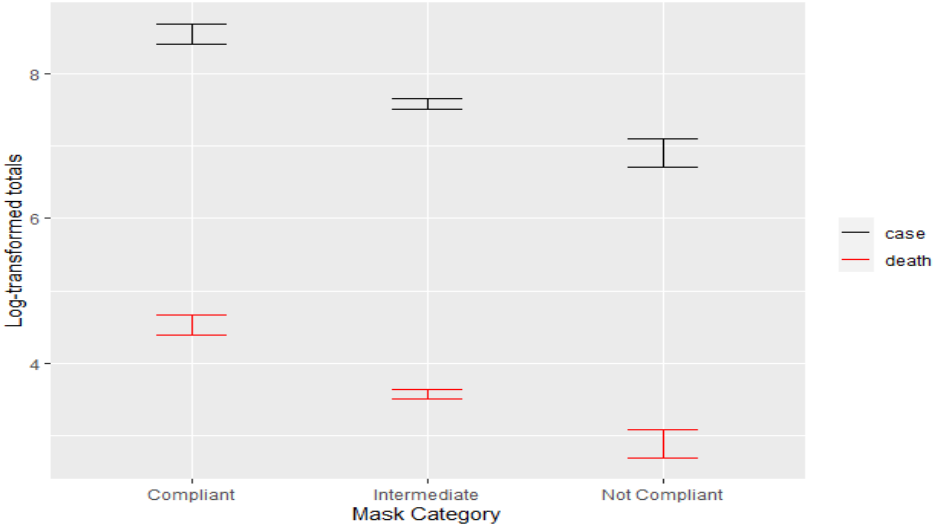

Figure 1. Log-transformed total cases (black; capital letters) and deaths (red; lower-case letters) +/- Bonferroni simultaneous interval for the three mask usage categories: Compliant, Intermediate, Not Compliant. Within cases or deaths, different letter indicates significant differences.

Total cases and total deaths had a high correlation ($r= 0.971$), so it is unsurprising that the trends between the two response variables are similar. The fact that compliant locations had the highest cases and deaths is a key finding that should be examined further. There could be several reasons this relationship was found. First, if there are many cases in a location, the chances that any individual knows someone who has tested positive for COVID-19 increases. Once this individual experiences the disease first- or second-hand they may be more inclined to wear a mask. Second, there could be demographic differences in locations with higher mask compliance that also make them more likely to come in contact with COVID-19, thus increasing the case number despite mask usage. These possibilities are not discussed here due to lack of available data for analysis but should be explored further when such data are available.

The PCA analysis gives log of total cases, log of total deaths, death to cases ratio, and the proportion of responses of "Always" and "Frequently" from the survey on mask uses

$\lambda = (2.16, 1.05, .76, .015)$				
Scores	1: (.64, .64, .04, .4)	2: (.20, -.12, -.97, -.00)	3: (-.28, -.29, -.02, .91)	4: (.68, -.69, .22, -.00)

Table 1 : Eigenvalues and Eigenvectors of the correlation matrix of log of total cases, log of total deaths, death to cases ratio, and the proportion of responses of "Always" and "Frequently" from the survey on mask uses

The result indicates that the first score alone explains more than the half of the total variance [2.16/(2.16+1.05+.76+.015)] and the first two scores together explain more than 80% of the total variance [(2.16+1.05)/(2.16+1.05+.76+.015)]. Let X_1 be standardized log of total cases, X_2 be standardized log of total deaths, and

X_3 and X_4 be the standardized third and fourth variables above. Then $Y_1 = .64\,X_1 + .64X_2 + .04\,X_3 + .4\,X_4$ explains more than 50% of the total variance and this tells us that as each of the standardized variable increases, we obtain greater value for the first score outcomes. This makes sense as the more cases we have, the more deaths we will experience and the people, as discussed from the ANOVA test, are more likely to comply with the mask policies. However, the second score, $Y_2 = .20\,X_1 - .12X_2 - .97\,X_3 - .00\,X_4$, gives a different aspect of the data. This tells us that about 25% of the total variance is explained by the score that will increase as the total cases increase while the total death and death to cases ratios fall. One possibility is that the more we test people, the less deaths from positive cases to be observed. That might be because as we increase the tests, we then detect more positive cases before the symptoms develops. However, to confirm this explanation, we need more variables such as the positivity rates of tests. Score 3 and Score 4 only explain 19% and 0.4% of the total variance, respectively.

Since a death occurs a number of days after testing positive for COVID-19, finding the regression between the cases and deaths seemed to generate incorrect inference about the two variables. Therefore, the lag between the seven-day average of deaths and seven-day average of cases was the data used for this analysis. Figure 2 shows that there is a lag between the seven-day average deaths (dashed-red) and the seven-day average of cases (blue).

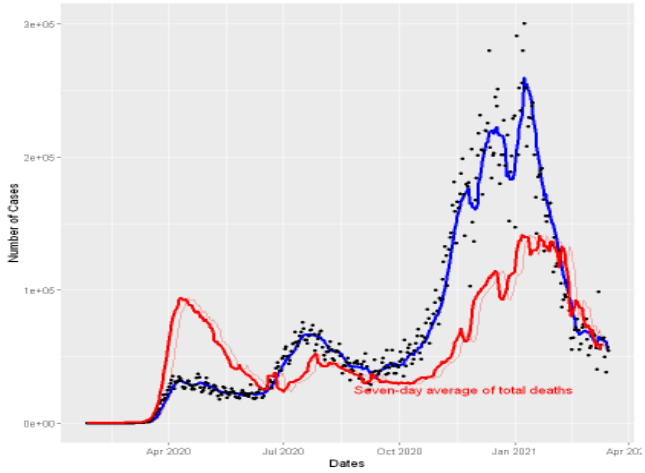

Figure 2. Seven-day average of total deaths (bold, red line), seven-day average of total cases (blue line), and seven-day average of total deaths shifted backwards seven days (light red line).

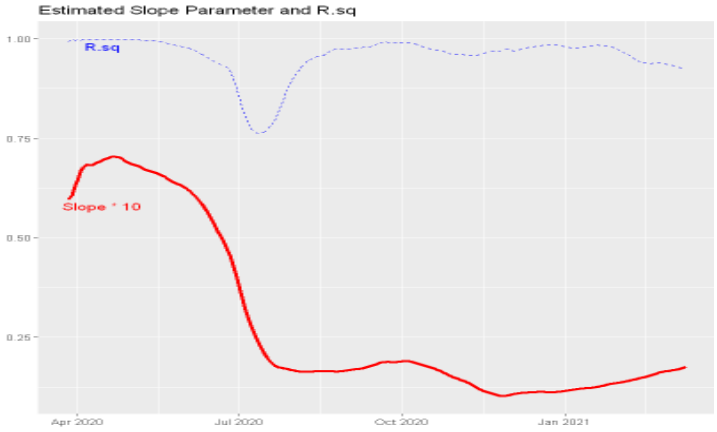

Figure 3. R-squared values (blue) and the estimated slope coefficients (red line; multiplied by ten for ease of visualization) from linear regression models using 60 day periods during the pandemic.

Three peaks are observed for both the seven-day average of deaths and the seven-day average of cases, and the mean difference between the peaks is 7 (Figure 2). The regression model, therefore, tells the expected changes in the seven-day average of death given the prior seven-day average cases. The bold red line in the figure 2 shows how the relationship changes after taking account of the lag between the two variables.

Every model has a R squared greater than .75 and this indicates the two variables are linearly associated and the strength of the association is strong (Figure 3). Strikingly, the slope has dropped precipitously as the US has faced the first few months without the proper preparation against the COVID-19 (Figure 3). The estimates of the slope parameter have been increasing lately (Figure 3). However, this may be because the average cases have been more volatile than the average deaths, and when the US began to experience the downturn in both variables the slope estimates tended to rise.

References

The New York Times. 2021. Coronavirus (Covid-19) Data in the United States. Available at https://github.com/nytimes/covid-19-data. Accessed 17 March 2021.

WHO (World Health Organization). 2020. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020. Available at https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020. Accessed 26 March 2020.