Risk Factors Associated with New COVID-19 Deaths per State

Ravi Sharma, Benjamin Pham

Scientific background:

SARS-CoV-2 (also called COVID-19) is a pathogenic, highly infectious virus[1]. It is classified as a coronavirus, a certain RNA virus type that includes the viruses responsible for SARS and MERS syndromes. A COVID-19 infection can produce upper respiratory symptoms, muscle aches, fevers, and other symptoms[1,2]. The virus often invades the bronchial tract and lungs. In addition to causing respiratory damage itself, the virus may also trigger a large inflammatory response, known as a cytokine storm. This cytokine storm can lead to significant autoimmune damage of the lungs and fluid build-up in the lungs. These effects are referred to as ARDS - Acute Respiratory Distress Syndrome. In severe cases of ARDS, the lung damage is too extensive, and it renders the lungs unable to provide enough oxygen to sustain the patient. In some of these cases, patients can be placed on a mechanical ventilator to assist in respiration until their lungs heal and they overcome the infection. However, in other cases, the lungs are beyond repair, and death can ensue. This trajectory represents the most common cause of death due to COVID-19[3].

Currently, there is a global pandemic of COVID-19 infections, and groups worldwide are searching for methods to curb the numbers of cases and deaths. It is puzzling that the daily death counts for COVID-19 are so variable among states in the US, even after normalizing for population. Below, we look through current models to attempt to explain this problem.

There currently exist several models from different institutions, which aim to predict numbers of new cases and deaths per state. On the whole, these models excel in their goal of ***prediction of new cases and deaths***. Expectations about shelter-in-place orders are typically taken into account, probabilistic models are often used to model cases and death individually, and sophisticated mathematics are employed to capture trends in the time series data. For example, the Los Alamos National Lab model models cases by combining a dynamic growth function with the size of the remaining susceptible population in an area, and it models deaths separately. In addition to prediction, the models also (to a lesser extent) function reasonably well with ***interpretation of new cases.*** Expectations of phenomena such as social distancing compliance can be tuned and observed to see how the case estimates change. However, these models generally do not focus on easy-to-understand ***interpretation of new deaths***. Non-technical policy-makers interested in reducing death counts through means other than reducing case counts will be left confused[4,5].

We seek to fill this gap. We aim to associate various variables of a state (such as a state's age distribution information, its healthcare quality, its number of new cases, etc) with new deaths - specifically, the average number of new deaths per 100,000 capita per day across a 4-month time-period. The results of this study can pave the way for causal analyses of death counts. The results, when combined with future causal analyses, can suggest to policy-makers where to allocate more funding. For example, if high quality of healthcare shows a strong association with lower death count, then more funding to hospitals might be merited.

Lastly, It is worth noting that, to our knowledge, the other major models do not incorporate state-level metrics to the extent that we do. We decided to do so because such information allows us to easily understand state-by-state differences, and because such metrics can often be directly affected by policy actions.

Objectives:

Our main objective is to find and fit an accurate, interpretable linear regression model that predicts new deaths per 100,000 people per day from a combination of state-level metrics. We have two secondary objectives. First, we want to determine the quantitative association between each state-level metric and expected new deaths per 100,000 capita per day. Secondly, using our findings from our model, we want to suggest actionable recommendations to policymakers that will reduce COVID-19 death counts specific to their state.

Data:

We utilized various sources to generate our final dataset. We obtained case and death data from the publicly available COVID tracking project[6] dataset, which is updated nearly daily. The data is gathered from state health department dashboards and press conferences manually via volunteers. This dataset reports, on a per-state per-day basis, new cases and new deaths due to COVID-19, as well as other metrics.

We obtained hospital capacity metrics from the CDC[7]. This dataset is an aggregation of data feeds from hospitals nationwide. The data is publicly available and regularly updated.

To confirm that the number of hospitals reporting capacity metrics in the above data set was sufficiently close to the total number of hospitals per state, we compared the data against a different CDC dataset[8]. This new dataset reports hospitals in a specific state registered with Medicare. We used this data as a proxy for the total number of hospitals per state, because nearly all hospitals in the US are required to accept Medicare patients. Our comparisons showed that the numbers were very similar: the differences were ~5-10% per state.

The real GDP per state was gathered from the US Bureau of Economic Analysis (BEA)[9]. Q2 GDP was used, as it was the most recent data available. We assumed that the Q2 GDP per day was sufficiently similar to the GDP per day during our studied time period of August 1 to November 23. Furthermore, population information was gathered from the US Census[10].The most recent report was in 2019. We assumed 2019 data was sufficiently representative of 2020 data. Next, data describing percentages of people insured with health insurance was gathered from the Current Population Survey (CPS)[11] administered by the US Census. The most recent release of this is in 2017. We assumed the values did not change significantly since 2017. Land area data was gathered from the 2010 Census of Population and Housing[12]. We assume that the area of land in each state remains the same in 2020.

Data processing:

We started off with data with a state x date granularity (~6000 rows). The metrics included new cases (per state per day) and new deaths (per state per day). All 50 states were included, and the date range was August 1 to November 23. From here, we merged in our hospital metrics time series, which was also on a per-state per-day basis.

Next, for each state, we averaged the daily new cases across the entire time range, to produce an average new cases per day value for each state. We repeated for the death metric and relevant hospital metrics. At this stage, our data structure contained 50 rows, where each row corresponded to a certain state.

We then normalized the cases and death metrics by population, specifically 100,000 capita. From here, other metrics (such as Population, GDP, Area, etc) were merged into the data structure and normalized by population when appropriate.

One of the metrics merged in was the number of hospitals per state. This information was calculated by combining the hospital metrics dataset with the Medicare dataset, via taking the max value of the two data sources for each state.

In summary, our final data structure we used for model building was at a state-level granularity and had 50 rows. Each metric corresponded to a time-averaged value or a value assumed to be constant over the relevant time period.

Initial Data and Model:

Initially, we constructed a different dataset and different response variable. We tried predicting cumulative mortality rate by dividing cumulative deaths on a certain day with cumulative cases on that day. Regarding our dataset, each row corresponded to a certain state x date combination (e.g. Colorado October 15th). Hospital occupancy metrics varied by date, but most variables were static (such as GDPperCapita).

There were several issues with this approach. First, our calculation of mortality rate was likely incorrect because the cumulative cases likely included a significant number of unresolved “open” cases. In other words, we were not using closed cases for our denominator. Also, we significantly violated the iid assumption for regression for two reasons: cumulative numbers depend on previous time points, and we also only have states in the US we are sampling.

Therefore, we tried solving some of these issues. We aggregated all time points together to get rid of the time-dependent component and solve the issue that only hospital metrics varied with time. Furthermore, we made the assumption that cases that were 30 or more days prior to the max date could be considered closed cases, because the vast majority of people who are hospitalized for COVID-19 are likely either deceased or recovered by day 30[13–17]. Similarly, we then assumed that people who die, do so 10 days after they are diagnosed (admittedly a larger time-to-death would be more appropriate). More specifically, we averaged daily case count from August 1 to Oct 31, then we averaged daily death count from Aug 11 to Nov 9. Then we divided to get case fatality rate. However, when we ran this model, we did not receive good results. The model was not significant, the R^2 value was not the best (0.5), and the calculated numbers were too high. Thus, our method of associating the number of cases with their corresponding number of deaths was likely not accurate.

Due to the failure in our above method, we chose to switch to a more promising approach: predicting new death count (per capita per day) instead of case fatality rate. This approach has significant value in interpretation of death count, as previously described. We aggregated cases and deaths from Aug 1 to Oct 31, normalized each by population. The time component is still removed, like before. We ultimately received good results using this approach.

Data Analysis:

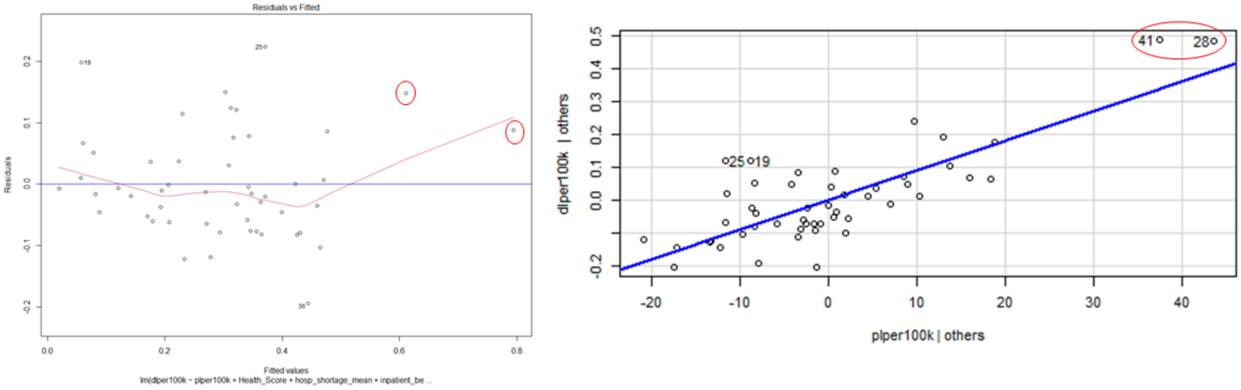
Our new approach consists of three parts. We first formulate different potential models out of the available predictors. We would then select the most optimal model by the AIC metric through bidirectional stepwise regression. Lastly, we would validate the optimal model through 10-fold cross validation.

Using our prior information on COVID19 deaths, we constructed a list of candidate predictors that potentially contributes to COVID19 deaths. Our initial list contained 13 predictors (Table 1) that can affect average death increases per day in a state.

Variable Name	Description (units)
dIper100k	Average Increase in Deaths per 100000 capita (deaths/day per 100000 capita)
pIper100k	Average Increase in New Cases per 100000 (new cases /day per 100000 capita)
inpatient_beds_utilization_mean	Average proportion of inpatient beds used
adult_icu_bed_utilization_mean	Average proportion of adult icu beds used
hosp_shortage_mean	Average proportion of hospitals reporting critical staff shortages
hosp_shortage_in1week_mean	Average proportion of hospitals reporting potential critical staff shortages 1 week from reporting
prop65	Proportion of people aged 65 years or older in a state
uninsured_percent_65	Proportion of people aged 65 years or older who are uninsured
GDPperCapita	State GDP per population (GDP per capita)
Health_Score	{Ravi write here please}
critical_care_doctors_and_nurses_per_100k	Average number of critical care doctors and nurses per 100000 capita (critical care personnel per 100000 capita)
first_and_second_line_workers_per_100k	Average number of first and second line workers per 100000 (first and second line personnel per 100000 capita)
population_density	Population in an area (Population/mi^2)

Table 1: *Candidate Predictors.* These candidate predictors were chosen from our working knowledge of risk factors that could affect COVID19 mortality. The name of the predictor in our dataset is in the first column, and its description is shown in the second column. The units of the metric are noted in the parentheses otherwise the metric is unitless.

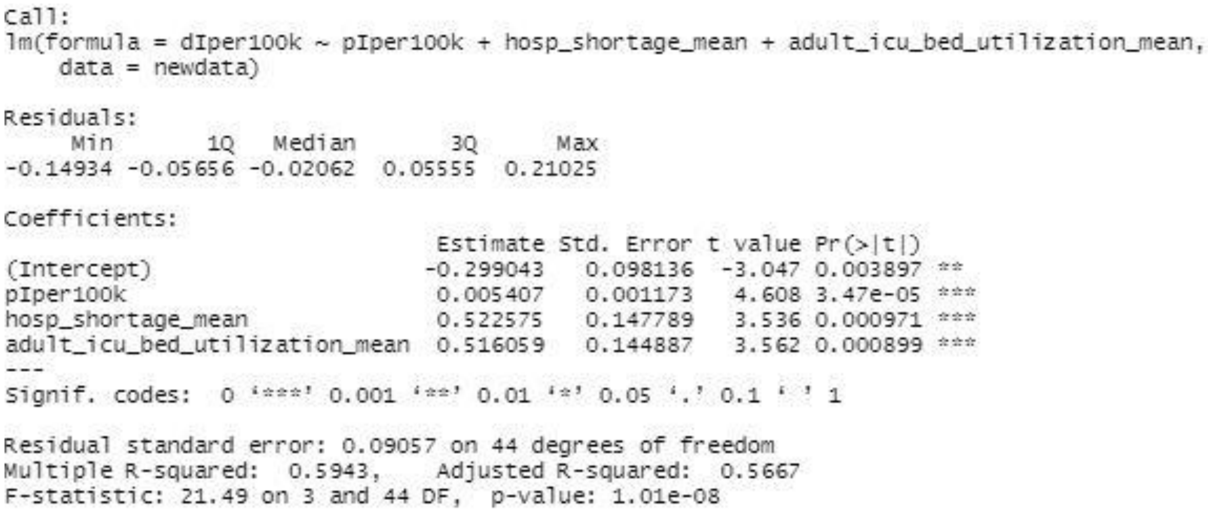
We constructed an initial model with the best possible combination of predictors out of our initial list through bidirectional elimination stepwise regression. This model consisted of 6 predictors: plper100k, Health_Score, hosp_shortage_mean, inpatient_beds_utilization_means, uninsured_percent_65, and prop65.



<u>Figure 1: </u> *Residuals vs Fitted plot and dlper100k (left) vs plper100k Added Variable plot of initial model (right).* Although not explicitly shown in the Residuals vs Fitted plot, the residual points that are unevenly distributed are circled, and they are North Dakota (28) and South Dakota (41). These points are circled in red and are also unequally distributed in the dlper100k vs plper100k added variable plot in the initial model.

This initial model was not a suitable ordinary least-squared regression model since the error residuals violated heteroscedasticity and normality. We identified two residual error points that had high influence on the Residuals vs Fitted plot through examination of its added variable plot (Figure 1). These states were North Dakota and South Dakota. Removing these states from the model resulted in eliminating the heteroscedasticity present in the model.

After removing these states, uninsured_percent_65 and prop65 predictors lost significance. We decided to remove these from the model. Because most cases of COVID-19 deaths are from extremely serious hospitalizations, we had also replaced the inpatient_bed_utilization_mean with adult_icu_bed_utilization_mean. We also removed Health_Score from the model because this metric was extremely difficult for non-technical end users to find.



```
Call:
lm(formula = dIper100k ~ pIper100k + hosp_shortage_mean + adult_icu_bed_utilization_mean,
    data = newdata)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.14934 -0.05656 -0.02062  0.05555  0.21025 

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                    -0.299043   0.098136  -3.047 0.003897 ** 
pIper100k                       0.005407   0.001173   4.608 3.47e-05 ***
hosp_shortage_mean              0.522575   0.147789   3.536 0.000971 ***
adult_icu_bed_utilization_mean  0.516059   0.144887   3.562 0.000899 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09057 on 44 degrees of freedom
Multiple R-squared:  0.5943,	Adjusted R-squared:  0.5667 
F-statistic: 21.49 on 3 and 44 DF,  p-value: 1.01e-08
```

Figure 2: *The most optimal model built from bidirectional stepwise regression.*
The most optimal model consists of average new cases per day (pIper100k), number of hospitals reporting staff shortages (hosp_shortage_mean) and average proportion of adult ICU beds utilized (adult_icu_bed_utilization_mean).

The most optimal model consists of three predictors: pIper100k, adult_icu_bed_utilization_mean, and hosp_shorage_mean. These predictors have extremely low p-values indicating high significance. To check if this is truly the most optimal model possible under the new dataset, we performed bidirectional stepwise regression with the same initial list of thirteen candidate predictors over the new dataset. The output from this is a four parameter model that consists of our elucidated three parameter optimal model and the Health_Score predictor that we had previously chosen to remove.

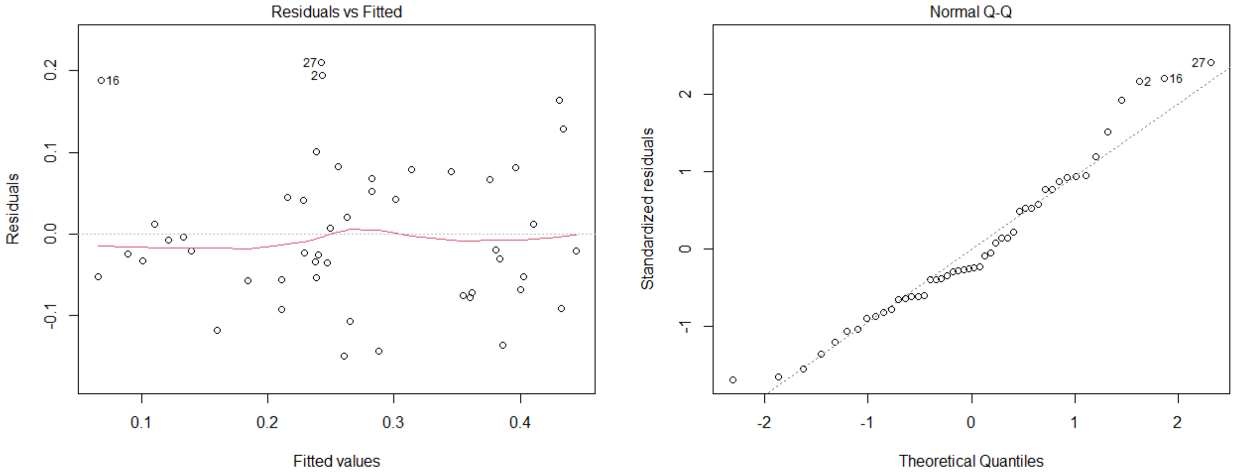


Figure 3: *Residuals vs Fitted plot and Q-Q plot of error residuals in optimal model.*
Unlike with the initial model (Figure 1, left), the residuals seem evenly distributed. In addition, the Q-Q plot (right) seems mostly normal indicating that the residuals are somewhat normally distributed.

The Residuals vs Fitted plot of the most optimal model satisfies the homoscedastic assumption that underlies OLS models. In addition, the error residuals appear to be distributed mostly normally (Figure 3). This shows that the interpretations of this model are most certainly valid since the estimations from this model are accurate.

Analysis of Overfitting:

K-fold Cross Validation (R caret package)[18] was then used to estimate the ability of our model to generalize on future data and make accurate assessments. We set k=10. The reported average RMSE value was 0.0885 deaths/100k capita/day. Given our response variable ranges from 0.01 to 0.6 deaths/100k capita/day, this RMSE value is okay but not great. As a gut check, to check whether the fact that our model only had 3 predictors was causing the high RMSE, we tested other fuller models as well. Unfortunately, we received similar RMSE values. Log transforming the response variable and adding predictors resulted in a much worse RMSE (0.47).

In order to improve the model's accuracy, county-level data instead of state-level data could be acquired, so that the number of rows become greatly expanded. Acquiring and processing this data would be very time-consuming, because there does not look to be any centralized source for all counties' data. Having more rows could result in less overfitting, thus potentially better accuracy and more accurate interpretations of our coefficients. Furthermore, in our data exploration, we observed that there existed much data that was not publicly available, but available to academic researchers. In future studies, this data should be taken into account to potentially improve accuracy on existing covariates and generate new candidate covariates. Furthermore, we cannot be certain about the accuracy of our CDC-provided data. Given the scramble to output data and given potentially different definitions of different metrics per state, there is the risk of significant error in the data. Lastly, having a larger time range available to us would also assist us in generating more accurate models because data noise would be reduced.

Interpretation:

Because the RMSE of our model was not excellent, our interpretations should be taken with a grain of salt. Given that we find that the estimates of each covariate in the model match our expectations. An increase of 10 average cases per 100k capita in a day would be associated with an expected 0.05 increase of new deaths per 100k capita in a day. This is because wide exposure of COVID-19 would increase the frequency of COVID-19 deaths. A 10% increase in occupancy of ICU bed space would be associated with an expected 0.0523 increase in new deaths per 100k capita in a day. This is not a very large increase, but still is worrisome (for CA, this would be ~2 more deaths each day). Intuitively, the shortage of staff increases mortality because afflicted patients would not be able to receive the care they need. Furthermore, from our model, a 10% increase of hospitals experiencing staff shortages in a given day would be associated with an expected 0.052 increase in new deaths per 100k capita in a day. This shows that the limitation of hospital working capacity is not only limited to hospital equipment, but also in specially-trained staff needed to operate equipment where an increase in understaffed hospitals contribute to an increase of daily mortality.

Action Items:

Our linear model shows aspects of the healthcare system that could potentially merit policy change. First, we suggest further optimization of the current model for accuracy as well as future casual analyses. In particular, causal analyses determining to what degree hospital shortages cause high death rates should be examined, as well as to what degree ICU bed occupancy rates do. Increasing funding to hospitals to hire more critical care doctors can be used to alleviate staff shortages. In addition, funding to hospitals to potentially expand their ICU capacity could help, though such actions would likely take time to implement. In addition, it is completely possible that one of the origins of these staff shortages is from the specialized staff themselves getting sick due to their frequent exposure to COVID-19. This could potentially merit a need for training additional staff to take on these responsibilities as well as a need for proper personal protective equipment (PPE) to protect themselves.

Reducing the number of average new cases per day can be done with policies that prevents the active spread of COVID-19. This can be done with social distancing mandates as well as lockdown procedures such as shelter in place or limiting social gathering capacity. The use of PPE such as face masks should also be encouraged.

Conclusion:

We were successful in formulating a model that is valid across 48 states. Although the model has a mediocre RMSE, we believe the model is still moderately useful for interpretation due to its significant predictor components and good adjusted R^2. The risk factors inside of the model matches expectations of previous COVID19 observations since many current policies in place target them. More work is needed to formulate a model that can be generalized for all 50 states. This could be done through investigating more sophisticated models such as logistic regression, logistic regression, or even survival analysis.

Contributions:

Ravi Sharma did initial data processing, helped with later staged data processing, and conducted iterations of model formation. He also implemented the final cross validation. Benjamin Pham conducted initial data processing and worked on most of later staged data processing. He also conducted iterations of initial model formation and helped with the final model formation. Both authors wrote the report and agreed that the report and project reflect their best effort. Both consulted one another in weekly meetings throughout the project where they contributed approximately equally to the total workload of the whole project.

References:

1. Wang, L., Wang, Y., Ye, D. & Liu, Q. Review of the 2019 novel coronavirus (SARS-CoV-2) based on current evidence. *Int J Antimicrob Agents* **55**, 105948 (2020).
2. Kaur, N. *et al.* Genetic comparison among various coronavirus strains for the identification of potential vaccine targets of SARS-CoV2. *Infect Genet Evol* 104490 (2020) doi:10.1016/j.meegid.2020.104490.
3. Acute Respiratory Distress Syndrome (ARDS). *Yale Medicine* https://www.yalemedicine.org/conditions/ards.
4. Boice, R. B., Jay. Where The Latest COVID-19 Models Think We're Headed — And Why They Disagree. *FiveThirtyEight* https://projects.fivethirtyeight.com/covid-forecasts/ (2020).
5. LANL COVID-19 Cases and Deaths Forecasts. https://covid-19.bsvgateway.org/#uncertainty.
6. The COVID Tracking Project. *The COVID Tracking Project* https://covidtracking.com/.
7. COVID-19 Reported Patient Impact and Hospital Capacity by State Timeseries | HealthData.gov. https://healthdata.gov/dataset/covid-19-reported-patient-impact-and-hospital-capacity-state-timeseries.
8. Hospital General Information | HealthData.gov. https://healthdata.gov/dataset/hospital-general-information.
9. GDP by State | U.S. Bureau of Economic Analysis (BEA). https://www.bea.gov/data/gdp/gdp-state.
10. Bureau, U. C. State Population by Characteristics: 2010-2019. *The United States Census Bureau* https://www.census.gov/data/tables/time-series/demo/popest/2010s-state-detail.html.
11. US Census Bureau, D. I. S. Current Population Survey (CPS), CPS Table Creator. https://www.census.gov/cps/data/cpstablecreator.html.

12. Census Report. https://www.census.gov/prod/cen2010/cph-2-1.pdf
13. COVID-19 patients die 11 days after diagnosis on average. *www.donga.com* https://www.donga.com/en/article/all/20200409/2033350/1/COVID-19-patients-die-11-days-after-diagnosis-on-average.
14. Boston, 677 Huntington Avenue & Ma 02115 +1495-1000. Data animation shows time lag between COVID-19 cases and deaths. *News* https://www.hsph.harvard.edu/news/hsph-in-the-news/data-animation-shows-time-lag-between-covid-19-cases-and-deaths/ (2020).
15. Verity, R. *et al.* Estimates of the severity of coronavirus disease 2019: a model-based analysis. *The Lancet Infectious Diseases* **20**, 669–677 (2020).
16. Mar 10, M. V. B. | N. W. | C. N. | & 2020. Old age, sepsis tied to poor COVID-19 outcomes, death. *CIDRAP* https://www.cidrap.umn.edu/news-perspective/2020/03/old-age-sepsis-tied-poor-covid-19-outcomes-death.
17. Wang, L. *et al.* Real-time estimation and prediction of mortality caused by COVID-19 with patient information based algorithm. *Science of The Total Environment* **727**, 138394 (2020).
18. Kuhn, M. Building Predictive Models in *R* Using the **caret** Package. *J. Stat. Soft.* **28**, (2008).