2022 National Inter-University Big Data Challenge

Stepwise Regression Analysis: Healthcare Shortage Impacts on U.S. States COVID-19 Development

Nicholas William Susanto, Joyce Nguyen, Anisha Huq, and Justin Alianto
University of Toronto, Ontario, Canada

July 3th, 2022

ABSTRACT

The pandemic had tremendously affected all aspects of our lives, especially our healthcare system. Severe shortages were seen throughout the pandemic ranging from hospital capacity to healthcare workers shortages. In the light of the pandemic, our study aims to determine potential shortcomings of our healthcare systems in overcoming future pandemics. Focusing on COVID-19 developments in U.S states, as measured by number of new cases, we utilized a stepwise regression algorithm to identify correlations between the pandemic development and resource shortages. We identified three significant predictors towards the success of states in managing the current COVID-19 pandemic, namely: number of hospital beds, healthcare providers, and state's healthcare responsiveness ratings. As a result, we identified that while hospital capacity plays a critical role in managing demand, future mitigation will also need to focus on both providing efficient services and managing sufficient healthcare human resources.

KEYWORDS: Stepwise regression, COVID-19 cases, hospital beds, healthcare providers, hospital quality

I. INTRODUCTION

Hospitals serve as our frontline to any anticipated or unanticipated public health crisis. In other words, hospital resources are indispensable and essential. Unfortunately, the ongoing global pandemic has uncovered shortcomings and deficiencies in the global hospital sector and there is a pressing need for the hospitals to evolve to cater to the ever changing demands of public health (Ducarme, 2022). A vicious cycle of rising demands for specialized acute care and declining availability of hospital resources and manpower is engendered (Blumenthal et al., 2020). In fact, employment in the health sector dipped by 1 million by May 2020 in the United States together with the then forecasted value of $323.1 billion to lose (Blumenthal et al., 2020). Going forward, what we can work towards is a community that is equipped and ready for possible unforeseen future health crises. Acknowledging the numerous aspects that build a prepared community, our study will be taking a closer look within the healthcare sector amongst all.

During this extensive and rapid spread of the pandemic, cracks were observed in relation to hospital resource allocation. More specifically, at the start of this year, 19% of U.S. hospitals were experiencing shortages in staff members due to the ripples of the pandemic (Plescia and Gooch, 2022). Considering it has been two years since the birth of the virus phenomenon and the expected changes and growth in hospital resources by now, this deepens the extent of this specific problem. Beyond the quantity of the healthcare workforce, the quality should also be given attention to. In such urgent scenarios, it is critical that the responsiveness and efficiency of the hospital staff are up to standard. Apart from the manpower of the health sector, the non-living resources needed in hospitals demand our attention too. In particular, we have seen that in the U.S., there were apparent insufficiencies in the number of hospital beds across the nation's health systems (Abelson, 2020). Hospitals were turning away patients not by choice and placing more patients in a room than they should (Abelson, 2020). With these in mind, our study wants to explore the number of beds, the number of health providers, and the responsiveness of the hospital staff in a large and developed country like the U.S. when it handles such a colossal event.

Utilizing stepwise regression, we can possibly recognize patterns of the COVID-19 development in U.S. states in relation to some predictor factors like number of beds, number of health providers, and the responsiveness of hospital staff. Correspondingly, we may identify loopholes in hospitals that need to be filled for future improvements.

II. METHODOLOGY

A. Sourcing Our Data

The data for US COVID-19 development by state was sourced from Centers for Disease Control and Prevention through the website healthdata.gov (Centers for Disease Control and Prevention, 2022). The data for the number of hospital beds in each U.S. state and the population of each state were sourced from KHN and arcgis combined through a Kaggle database (Kiulian, 2020). The data for the number of healthcare providers in each U.S. state was sourced from Health Resources & Services Administration website (HRSA Data Warehouse, 2022). Lastly, the data for the responsiveness of the hospital staff in each U.S. state was sourced from The Commonwealth Fund website (CMS Hospital Compare, 2022).

B. Data Preprocessing

No one can forecast a global crisis like this and it becomes clear that globally, we have to be ready and able to handle anything that walks in our path. Since the pandemic hit at the end of 2019, we believe it would be reasonable to analyze the data from 2019 for the U.S. states. This would show the level of readiness of these states when it comes to managing an unpredictable large-scaled public health downturn. The world is driven by tail events, so it is our responsibility as a community to be ready for whenever such low-probability, but high consequential events occur. This is our justification as to why we chose to take data from 2019 instead of the years before or after.

We used Python inside the Jupyter Notebook to complete our data preprocessing. We first extracted the various datasets into pandas.DataFrame (dataframe) objects from csv files. We wanted to split our datasets into two types: predictor and target.

For our target variable data, we used the dataset on COVID-19 development. We picked the 'submission_date', 'state', 'tot_cases', 'new_case', 'tot_death', 'new_death' columns and dropped the remaining columns from our dataframe object. Before exporting it as a new csv file for the data analysis using R, we first arranged the data by the various states and by the submission date using pandas.

For our predictor variables data, we used the datasets on the number of beds, the number of providers, and the responsiveness of hospital staff in each U.S. state. We picked the 'state', 'county', 'type', 'beds', 'population', 'year' columns for the dataframe object containing data on hospital beds. We picked the 'MUA/P Update Date', 'State Name', 'State Abbreviation', 'Complete County Name', and 'Providers per 1000 Population' columns for the dataframe object containing data on the number of providers. We picked the 'time_period', 'state', 'point_estimate', and 'map_group' columns for the dataframe object containing data on the responsiveness of hospital staff. We dropped the remaining columns from the respective dataframe objects. Using features and functions from pandas and numpy, we constructed a new dataframe object that possesses data on the year, state, point estimate for the responsiveness of the hospital staff of each state, population of each state, number of beds per 1000 people in each state, and the number of providers per 1000 people in each state. Using simple divisions of some column values and simple formulas, we also created 3 additional variables for the number of beds per provider in each state, the total number of beds in each state, and the total number of providers in each state. Similarly, we exported this dataframe as a new csv file for the data analysis using R. The two tables for the predictor variables and the target variables will be placed under Appendix A and Appendix B respectively after the References section for better numerical context of our analysis. Appendix A represents the full data table for the predictor variables and Appendix B represents the data table in brief for the target variables.

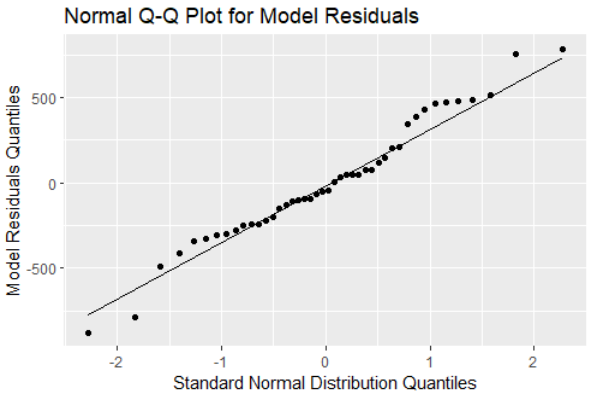
Out of the 50 U.S. states, we removed data from 6 of the states due to the lack of data values in some of the predictor variables, namely: Alaska, Connecticut, Maryland, Mississippi, North Dakota, and Rhode Island.

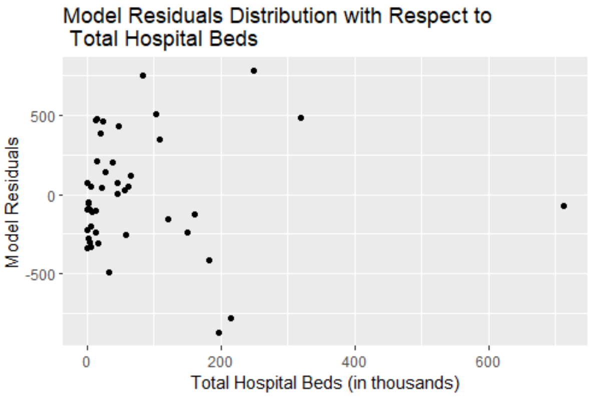
C. Selecting the Model

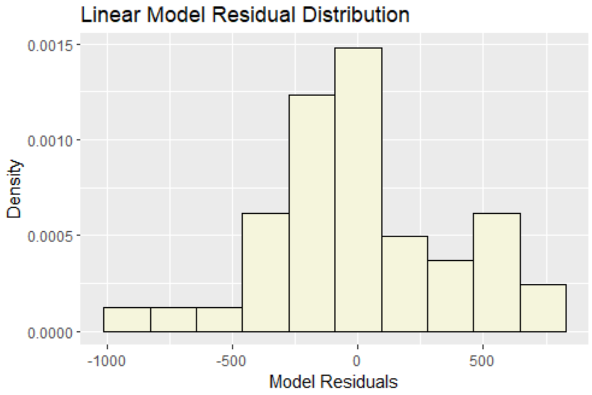
We measured COVID-19 development in each state using the number of new cases, taking the median value of each state. We then investigated their relationship with healthcare quality and shortage measures such as number of hospital beds, healthcare providers, hospital responsiveness score as measured by the Commonwealth Fund, and more. Using R's stepwise regression algorithm, we then fit two separate models to study the development of cases and deaths in a particular state. The algorithm modeled different combinations of independent variables to find the best model, relative to different statistical measures (p-value, R^2, etc.), for both proxies. Based on its results, we modeled new COVID-19 cases with the predictors: total number of hospital beds, healthcare providers, and the hospital responsiveness score. The models results and coefficients are as follows:

Predictors	**Coefficients**	**p-value**
Total Beds	0.01122	< 0.1%
Total Providers	-0.006858	< 0.1%
Responsiveness Rating	-53.65	< 5%

We also conducted some preliminary analyses to verify the assumption of our model, namely: linearity, homoscedasticity, normality, and independence. The analyses can be seen as follows:



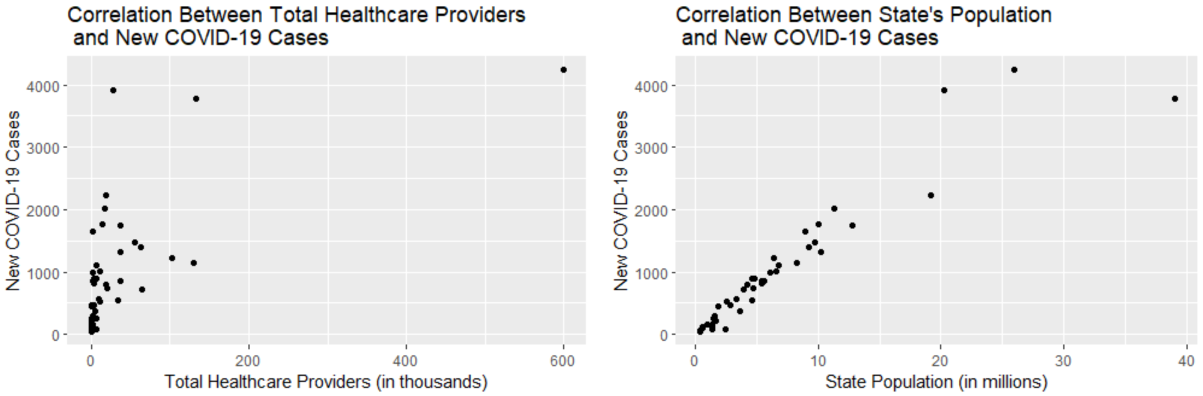


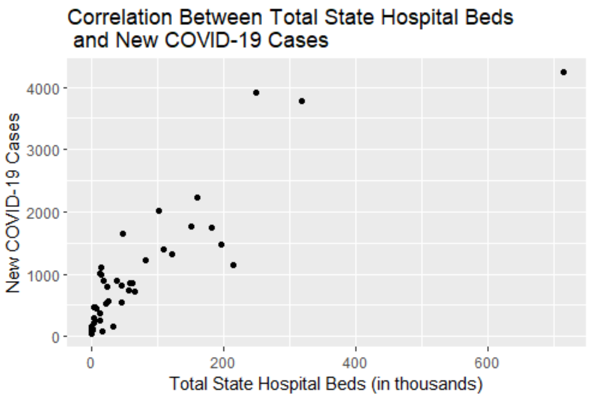


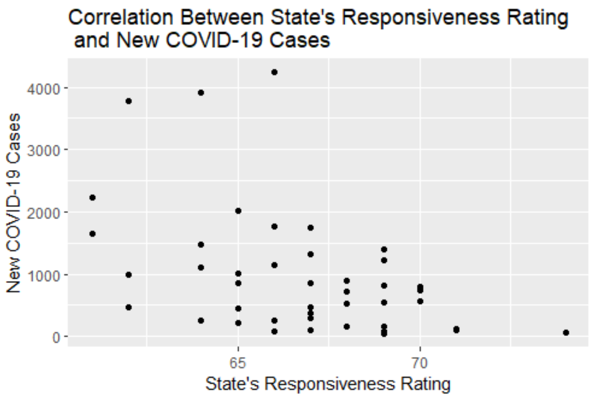
Utilizing a residual analysis and graphical approaches, we observed that our model appears to satisfy the assumptions. The model appears to have a normal distribution as seen in both the normal quantile-quantile plot and the histogram. However, we do see some slight deviations in both tails of the distribution. Nevertheless, these variations are normal especially with a relatively small sample size. On the other hand, we could also observe that the residuals seem to have equal variances and independently distributed with respect to its predictors, one of which is the total number of hospital beds in a state. Therefore, it appears that the model suffices its assumptions.

III. RESULTS AND ANALYSES

A. Data Exploration







Relating to our study, we could observe some apparent trends pertaining various potential predictors of COVID-19 case development. We could observe that new COVID-19 cases strongly correlates with the number of state's population, with more populous states facing a faster spread of the virus. Similarly, we could also observe an interesting trend regarding the total hospital beds available in each state. Contrary to one's belief, it appears that states which have higher hospital capacity faced a larger increase in the number of cases.

We could also observe how a state's responsiveness rating appears to correlate with the number of new COVID-19 cases. It appears that states with higher ratings are able to better handle the pandemic as demonstrated by achieving a low number of new cases. In addition, while appearing weak, we could also notice that the quantity of healthcare providers negatively correlates with new COVID-19 cases.

B. Interpretation of results

Based on our model, we discovered that both hospital responsiveness score and total healthcare providers are negatively correlated with our proxy. On the other hand, our model suggested a positively correlated relation between new cases and the total number of hospital beds available. The positive correlation observed appears particularly odd especially when studying about the severe capacity shortages observed throughout the pandemic. Hence, the result suggested further study to confirm its relation. Nevertheless, we found that all three predictors appear to be statistically significant on the 5% level. It is, however, important to note that each predictor, except for responsiveness score, only carries a minor effect on the number of new cases.

Based on these results, we could observe that both capacity and efficiency are significant predictors towards a state's success in mitigating future pandemic. As one might expect, greater human resources will tend to result in a slower spread of the pandemic. Similar observation is also seen towards hospital efficiency with higher rated states performing better in COVID-19 development. While it appears contradictory, the correlation observed between the number of new cases and hospital bed capacity suggest that huge expansion in hospital capacity might not be a solution in controlling future pandemics. A plausible argument for this is the delay or avoidance of seeking medical care due to the fear of the pandemic pervasiveness. By the end of June 2020, 4 in 10 adults in the United States reported to have postponed or dodged medical treatment due to the virus spread and the fear of catching it or worsening their symptoms (Czeisler et al., 2020). By staying untreated for longer times, this may have induced higher chances of exposure to the rest of the public, generating the rising number of new cases. Beyond the physical improvements of the hospitals, we should focus on the sector's communication and care delivery efforts to evoke a peace of mind for the public to trust and seek medical care promptly. The model suggested an importance to strike a balance between a hospital's capacity and its capability to handle them, highlighting the urgency in handling healthcare worker shortages seen throughout the pandemic.

C. Limitations

Our study limitations include two categories: data limitations and model limitations.

With regards to our data, we were not able to acquire more significant predictor variables for our model. For the quantity of hospital resources, data on variables such as the providers' working hours in each state and the availability of hospital equipment were not able to be sourced (Ren et al., 2022). For the quality of hospital resources, data on variables such as the tendencies of the providers working overtime, the education and skill level of the providers, the providers' job satisfaction were not able to be sourced (Ren et al., 2022). The allocation and rostering of nurses could also be investigated to measure the efficiency of the hospitals in utilizing their manpower (Ren et al., 2022). Moreover, we are also assuming that our data on the number of beds and the number of providers are all allocated for COVID-19 development due to a lack of information. Consequently, the total numbers of beds and providers computed are both an overestimation relative to the actual proportion of beds and providers catered to COVID-19 cases only. With this assumption, the extent of the hospital resource shortages is actually greater than computed.

With respect to our fitted model, our data limitations generate an effect on the accuracy of our model. We were unable to obtain a model that considers more accurate and suitable predictors to predict the development of the pandemic. Furthermore, due to the erratic nature of the virus spread, identifying a pattern in COVID-19 development was not possible, which led us to pick instead the median value for the daily number of new cases in each state. This helps to preserve consistency for our model. In order to perform our regression algorithm, we ran a simple residual analysis to check for violations in assumptions instead of more rigorous testing due to our current level of statistical knowledge.

D. Further Exploration

Some noteworthy areas to explore include amending on our study limitations. We could explore other predictor variables to add to cover and increase the practicality of our model. As mentioned in our limitations, these include the providers' working hours, education and skill level, and workers' wellness. We could also explore other methods of statistical analyses and machine learning models to verify our model performance. One interesting aspect that may be worth looking further into is our computation for the correlation of number of beds and the number of COVID-19 cases. A positive correlation value was not something we have expected and although the value is small, this was a unique insight to the relationship. Should our data be actually sufficient and this correlation be accurate, perhaps having more beds within the same vicinity may increase chances of exposure to the virus from longer waiting times and more sources of transmissions and that there is a need to strike a balance on the capacity limit (Zhuang et al. 2021).

Additionally, we could look at other aspects outside of the hospital but have an impact on the hospital. For example, we could source data for the proportion of budget expenditure of each state spent on healthcare.

IV. CONCLUSION

In conclusion, we discovered that both quantity and quality of hospital resources are valuable to a state or a country's ability to manage a pandemic and its spread. As seen from our study, there were correlations between our predictor variables, total number of providers in a state and the responsiveness of hospital staff in a state, and our target variable, the number of new COVID-19 cases per day. It is, hence, apparent that governments should improve the number of hospital resources, but also augment the quality and efficiency of hospital resources. The wellness and satisfaction of the hospital manpower should also be considered to maintain healthy and efficient work ethics from the staff when dealing with future unpredictable global health crises (Ren et al., 2022). Moreover, telehealth services should be given more weight in this technological era to boost the public reach that the hospital sector can obtain at a time as well as the communication quality given to the public (Blumenthal et al., 2020). This would benefit people psychologically and generate a better ease of mind even without being physically at the hospital. Using the context of the current pandemic, this is critical to areas where there are shortages in hospital resources or to individuals who choose not to seek treatment right away.

Looking into the future, it would be ideal for the global community to be better equipped with the right amount and the right standard of resources to avoid shortages from appearing again. As our frontline soldiers to any health-related needs, we need to acknowledge the important and irreplaceable responsibility that the hospitals and their people have to take up.

V. REFERENCES

Abelson, Reed. "Covid Overload: U.S. Hospitals Are Running Out of Beds for Patients." *The New York Times*, 27 November 2020, https://www.nytimes.com/2020/11/27/health/covid-hospitals-overload.html.

Blumenthal, David, et al. "Covid-19 — Implications for the Health Care System." *The New England Journal of Medicine*, vol. 383, no. 15, 2020, pp. 1483-1488. *The New England Journal of Medicine*, https://www.nejm.org/doi/full/10.1056/nejmsb2021088.

Centers for Disease Control and Prevention. "United States COVID-19 Cases and Deaths by State over Time." *HealthData.gov*, June 2022, https://healthdata.gov/dataset/United-States-COVID-19-Cases-and-Deaths-by-State-o/hiyb-zgc2.

CMS Hospital Compare. "Responsiveness of hospital staff when called by patients (out of 100 points)." *The Commonwealth Fund*, 2022, https://www.commonwealthfund.org/datacenter/responsiveness-hospital-staff-out-100-points.

Czeisler, Mark E., et al. "Delay or Avoidance of Medical Care Because of COVID-19–Related Concerns — United States, June 2020." *Morbidity and Mortality Weekly Report*, vol. 36, no. 69, 2020, pp. 1250-1257, https://www.cdc.gov/mmwr/volumes/69/wr/mm6936a4.htm.

Ducarme, Thibault. “What will be the impact of the Covid-19 pandemic on healthcare systems?” *Deloitte*, 2022, https://www2.deloitte.com/fr/fr/pages/covid-insights/articles/impact-covid19-healthcare-systems.html.

HRSA Data Warehouse. “MUA/P - CSV.” *data.HRSA.gov*, 2022, https://data.hrsa.gov/data/download#SHORT.

Kiulian, Igor. “Global Hospital Beds Capacity (for covid-19).” *Kaggle*, April 2020, https://www.kaggle.com/datasets/ikiulian/global-hospital-beds-capacity-for-covid19?select=hospital_beds_global_v1.csv.

Plescia, Marissa, and Kelly Gooch. “19% of US hospitals critically understaffed, 21% anticipate shortages: Numbers by state.” *Becker's Hospital Review*, 10 January 2022, https://www.beckershospitalreview.com/workforce/19-of-us-hospitals-critically-understaffed-21-anticipate-shortages-numbers-by-state.html#:~:text=in%20America%20%7C%202021-,19%25%20of%20US%20hospitals%20critically%20understaffed%2C%2021%25,anticipate%20shor.

Ren, Hong-Fei, et al. “Nursing allocation in isolation wards of COVID-19 designated hospitals: a nationwide study in China.” *BMC Nursing*, vol. 21, no. 1, 2022, p. 23. *BMC*, https://doi.org/10.1186/s12912-021-00795-w.

Zhuang, Zian, et al. “The shortage of hospital beds for COVID-19 and non-COVID-19 patients during the lockdown of Wuhan, China.” *Annals of Translational Medicine*, vol. 9, no. 3, 2021, pp. 1-9. *NIH*, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7940947/.

VI. APPENDIX A

	time_period	state	point_estimate	state_abbreviation	population	beds_per_1000	providers_per_1000	beds_per_provider
1	2019	Alabama	70	AL	4731663	11.96759	4.15	2.883757
2	2019	*Alaska	65	AK	636213	4.597883	0	N/A
3	2019	Arizona	65	AZ	6630442	2.050268	1.671	1.226971
4	2019	Arkansas	68	AR	2583665	8.403947	4.349	1.932386
5	2019	California	62	CA	38982847	8.182707	3.426	2.388414
6	2019	Colorado	68	CO	4640885	4.159082	0.58	7.170831
7	2019	*Connecticut	65	CT	3594478	1.174298	0	N/A
8	2019	Delaware	69	DE	943732	0.558183	0.41	1.361422
9	2019	District of Columbia	53	DC	672391	0.46699	0.2	2.33495
10	2019	Florida	64	FL	20278447	12.31543	1.344	9.163263
11	2019	Georgia	64	GA	9736487	20.15766	5.703	3.534572
12	2019	Hawaii	66	HI	1421658	0.520208	0.3	1.734027
13	2019	Idaho	67	ID	1597117	2.516173	0.881	2.856042
14	2019	Illinois	65	IL	11312822	9.060001	1.44	6.291667
15	2019	Indiana	67	IN	5553596	11.23409	6.761	1.661602
16	2019	Iowa	65	IA	1886498	4.386061	0.25	17.54424
17	2019	Kansas	69	KS	2439525	6.782206	2.44	2.779593

18	2019	Kentucky	68	KY	3966829	16.5192	16.329	1.011648
19	2019	Louisiana	69	LA	4663461	9.872175	7.397	1.334619
20	2019	Maine	68	ME	1330158	24.51081	1.826	13.42323
21	2019	*Maryland	N/A	MD	5996079	3.50896	1.622	2.163354
22	2019	Massachusetts	64	MA	6789319	2.144799	1.02	2.102744
23	2019	Michigan	69	MI	9256726	11.80835	6.734	1.753542
24	2019	Minnesota	70	MN	4237839	5.658129	4.37	1.294766
25	2019	*Mississippi	69	MS	2564095	11.36365	0	N/A
26	2019	Missouri	65	MO	5376886	10.78106	0.393	27.43273
27	2019	Montana	71	MT	609018	3.571003	0.67	5.329855
28	2019	Nebraska	64	NE	1458310	4.385605	0.25	17.54242
29	2019	Nevada	62	NV	2873396	1.748019	1.32	1.324257
30	2019	New Hampshire	67	NH	1331848	1.937659	0.28	6.920211
31	2019	New Jersey	61	NJ	8960161	5.34243	0.12	44.52025
32	2019	New Mexico	65	NM	1620289	2.960478	0.53	5.585808
33	2019	New York	61	NY	19164551	8.382213	0.943	8.888879
34	2019	North Carolina	66	NC	10038356	15.00682	1.461	10.27161
35	2019	*North Dakota	61	ND	431894	1.800766	0	N/A
36	2019	Ohio	67	OH	10208692	11.94855	3.602	3.3172
37	2019	Oklahoma	70	OK	3383874	7.896221	2.938	2.687618
38	2019	Oregon	67	OR	3658297	3.407778	1.108	3.075612
39	2019	Pennsylvania	67	PA	12790505	14.26061	2.854	4.99671
40	2019	*Rhode Island	68	RI	1056138	0.720353	0	N/A
41	2019	South Carolina	68	SC	4843956	7.950473	1.443	5.509683

42	2019	South Dakota	71	SD	547780	2.378135	2.272	1.046714
43	2019	Tennessee	69	TN	6418018	12.88652	15.918	0.809556
44	2019	Texas	66	TX	25918960	27.50246	23.095	1.19084
45	2019	Utah	67	UT	2881247	2.0039	0.01	200.39
46	2019	Vermont	69	VT	377375	1.019483	0.74	1.37768
47	2019	Virginia	66	VA	8260121	25.94782	15.727	1.64989
48	2019	Washington	62	WA	6083428	2.495058	0.33	7.560782
49	2019	West Virginia	66	WV	1436549	8.690672	4.786	1.815853
50	2019	Wisconsin	69	WI	5361872	8.46172	0.52	16.27254
51	2019	Wyoming	74	WY	432677	2.163612	3.126	0.692134

Table 1: Data for the predictor variables

*These states are excluded from the analysis due to the lack of data values in at least one of the parameters

VII. APPENDIX B

submission_date	state	tot_cases	new_case	tot_death	new_death
3/12/2020	AK	0	0	0	0
3/13/2020	AK	1	1	0	0
3/14/2020	AK	1	0	0	0
3/15/2020	AK	1	0	0	0
3/16/2020	AK	1	0	0	0
3/17/2020	AK	3	2	1	1
3/18/2020	AK	8	5	1	0
3/19/2020	AK	11	3	1	0
3/20/2020	AK	14	3	1	0

3/21/2020	AK	15	1	1	0
…	…	…	…	…	…
5/29/2022	AK	251425	0	1252	0
5/30/2022	AK	251425	0	1252	0
5/31/2022	AK	251425	0	1252	0
6/1/2022	AK	253184	1759	1252	0
3/10/2020	AL	0	0	1	0
3/11/2020	AL	3	3	1	0
3/12/2020	AL	4	1	1	0
3/13/2020	AL	8	4	1	0
3/14/2020	AL	15	7	1	0
3/15/2020	AL	28	13	2	1
…	…	…	…	…	…
5/30/2022	AL	1315018	573	19664	0
5/31/2022	AL	1316044	1026	19664	0
6/1/2022	AL	1317029	985	19664	0
…	…	…	…	…	…
3/29/2020	MP	0	0	0	0
3/30/2020	MP	2	2	0	0
3/31/2020	MP	2	0	0	0
4/1/2020	MP	2	0	0	0
4/2/2020	MP	6	4	0	0
…	…	…	…	…	…
5/28/2022	MP	11333	0	34	0

5/29/2022	MP	11333	0	34	0
5/30/2022	MP	11333	0	34	0
5/31/2022	MP	11366	33	34	0
6/1/2022	MP	11366	0	34	0
…	…	…	…	…	…
3/16/2020	WV	0	0	0	0
3/17/2020	WV	1	1	0	0
3/18/2020	WV	2	1	0	0
3/19/2020	WV	5	3	0	0
3/20/2020	WV	8	3	0	0
…	…	…	…	…	…
5/28/2022	WV	513953	0	6945	0
5/29/2022	WV	513953	0	6945	0
5/30/2022	WV	513953	0	6945	0
5/31/2022	WV	515925	1972	6948	3
6/1/2022	WV	516553	628	6962	14
3/10/2020	WY	0	0	0	0
3/11/2020	WY	0	0	0	0
3/12/2020	WY	1	1	0	0
3/13/2020	WY	1	0	0	0
3/14/2020	WY	2	1	0	0
…	…	…	…	…	…
5/28/2022	WY	157861	0	1820	0
5/29/2022	WY	157861	0	1820	0

5/30/2022	WY	157861	0	1820	0
5/31/2022	WY	158472	611	1820	0
6/1/2022	WY	158472	0	1820	0

Table 2: Brief version of the data for the target variables