The Real Toll of the COVID-19 Pandemic:
Estimating County-Level U.S. Excess Mortality Rates

Katelyn Cranney, Khalel Corona, John Durant, John Larsen

April 2023

I. Introduction

As the world becomes increasingly removed from the most severe months of COVID-19, many interesting opportunities arise to analyze the pandemic's true impact. Our research applies various machine learning methods to measure the "excess mortality," or true death toll due to the pandemic. Inspired by a country-level model from *The Economist* (2020), we generate new US-county-level data on the hypothetical all-cause mortality rates if the pandemic had not occurred, thus helping us understand the true death toll, or "excess mortality" due to the COVID-19 pandemic. Similar to the model used by *The Economist*, we are defining excess mortality to be "the gap between the total number of deaths that occur for *any* reason and the amount that would be expected under normal circumstances" (The Economist*,* 2021). The following empirical analysis examines U.S. county-level data to create, test, and train a model that fits true mortality rates in a time period before the beginning of the pandemic. Our model is then applied to 2020 demographic variables and generates new predictions of what mortality rates would have been in the absence of the pandemic.

We compare our predictions to real county-level mortality data during 2020 to generate data on excess deaths for every county in the U.S. As described by Ackley et al, "Excess deaths not assigned to COVID-19 may reflect a variety of factors, including COVID-19 deaths that were ascribed to other causes of death due to limited testing, indirect deaths caused by interruptions in the provision of health care services, or indirect deaths caused by the broader social and economic consequences of the pandemic" (Ackley et al. 2022). By analyzing the difference between our generated data and true mortality rates, we can estimate county-level excess deaths and look for interesting discrepancies and trends in an attempt to estimate the pandemic's real human toll. This data will be useful to confirm other predictions of excess mortality from the CDC, WHO, and various academic journals to help policy-makers, economists, politicians, and health officials understand the impact of the pandemic and how to better prepare for any future health disasters.

II. Data

There are two main types of data we used for our project: county-level all-cause mortality data prior to as well as during the pandemic and predictors of mortality rates. To train our model that predicts county-level all-cause mortality rates (from before the COVID-19 pandemic), we used data from a variety of sources. We used county-level data from the CDC/ATSDR's Social Vulnerability Index, which contains relevant factors like poverty, lack of vehicle access, crowded housing, racial and ethnic minority status, disability, and unemployment (CDC/ATSDR Social Vulnerability Index, 2020). Because the Social Vulnerability Index is only collected on even years (2010, 2012, …, 2020), we built our model only on data from even years. We assume that the range of years accounts successfully trains our model and the lack of odd years in our training set does not bias our results. We argue that there are no statistically significant differences or changes in trends of odd years that are not accounted for in the even years in this time period.

We also used county-level health data from The University of Wisconsin Public Health Institute which contains data on various public health indicators, including sexually transmitted infections, violent crimes, and the number of single-parent households (University of Wisconsin Population Health Institute, 2020). In addition, we used income data from the Bureau of Economic Analysis which reports county-level income per capita (Bureau of Economic Analysis, 2020). We also created lagged historical mortality variables that indicated the amount of deaths and magnitude of crude rate in a given county two and four years prior.

We acquired general all-cause county-level mortality rates from before (2010, 2012, 2014, 2016, and 2018) and during the COVID-19 pandemic (2020) from the CDC Wonder mortality dataset, including by metropolitan status. Lastly, we use 2020 COVID-19 confirmed mortality rates from the New York Times to see how many of the excess deaths we estimate were attributed to COVID-19 (New York Times, 2020).

Combining all of these sources, we had a set of 469 demographic and health variables for each even year from 2010-2018 that could predict mortality. After dropping variables that had a significant amount of missing values, we were left with a set of 269 predictor variables. We also dropped any variables that we believed would be too correlated with either total deaths or the COVID-19 pandemic, including communicable disease raw value, premature deaths, and preventable hospital stays. Summary statistics for a selection of the remaining variables in this

dataset (including those that were not shrunk to zero in our Lasso model) can be found in Appendix A.

Additional data visualizations can be found in Appendix B. Figure B4 is a histogram of the predicted deaths in 2020 in the absence of COVID-19. Figures B5 and B6 highlight some of the predictors of a county's excess mortality. Figure B5 shows that there is a strong positive relationship between a county's total unemployment and excess mortality, while Figure B6 shows a weak positive linear relationship between a county's total rural population and excess mortality.

III. Method

There are two different measures of mortality that we used as outcomes: total all-cause mortality per county (raw value) and the number of all-cause mortality deaths per 100,000 people. Before training a model, we standardized our variables and split our data from 2010-2018 into a training set (75%) and a test set (25%). For the crude rate prediction, we fit a random forest and gradient-boosted regression model. Both models used cross-validation to find optimal parameters. As shown in Table 1, the gradient-boosted regression method had higher test set accuracy for crude rate.

Table 1: Machine Learning Estimation results for Crude Rate (total deaths/100,000)

Cross-Validated Model	Training Set Accuracy (MSE)	Test Set Accuracy (MSE)	Average Difference Between Predicted and Observed Crude Rates
Random Forest (CV)	0.904	0.787	-171.38722
Gradient Boosted Regression (CV)	0.890	0.795	-172.5212

The other outcome we predicted was total deaths. We used a cross-validated lasso, random forest, and gradient-boosted regression to train our model, and achieved very high test set accuracy scores ranging from 0.98 to 0.99 (see Table 2 below). We predict that our model fits total deaths better than the crude rate due to the likely linear relationship between the number of deaths and population, especially before the pandemic. It is expected that in the absence of confounding large shocks (like a pandemic), the rate of change between population and death rate is approximately linear. Because our MSE score accuracies were higher for our total deaths prediction, we used that outcome variable for further analysis. Since the Lasso and

Gradient-Boosted methods performed almost identically, we applied both of these models to our 2020 data and looked at how many counties had “negative” excess death results, meaning that our model predicted more deaths than actually happened (see figure 1). We found that the Lasso model had a lower “error rate” (or less counties with negative excess deaths) than the gradient-boosted model, and because our average for the Lasso model (-146.045) more closely resembled the prediction ofAckley et al.. (2022) we used this model for further analysis.

Table 2: Machine Learning Estimation Results for Total Deaths

Cross-Validated Model	Training Set Accuracy (MSE)	Test Set Accuracy (MSE)	Average Difference Between Predicted and Observed Deaths
Lasso (CV)	0.999	0.9992	-146.045
Random Forest (CV)	0.973	0.988	-136.675
Gradient Boosted Regression (CV)	1.000	0.999	-144.788

We chose our tuning parameters using k-fold cross validation with 5 folds. The two algorithms we used were LassoCV and GridSearchCV from scikit-learn. In both of these methods, the optimal tuning parameters are selected through an iterative process wherein a section of the training data is held-out as the “test set,” one possible set of values for the tuning parameters is used to train the model on the non-held-out data, the trained model’s out-of-sample error rate on the “test set” is recorded, and that rate is compared against other recorded out-of-sample error rates generated by other attempted tuning parameters upon the datasets. For GridSearchCV, we fed the algorithm a grid of values whereas LassoCV randomly chooses values to test from a distribution. The optimal tuning parameter for LassoCV is an alpha of 6.095.

Figure 1: County-Level Excess Mortality, Cross-Validated Lasso Model

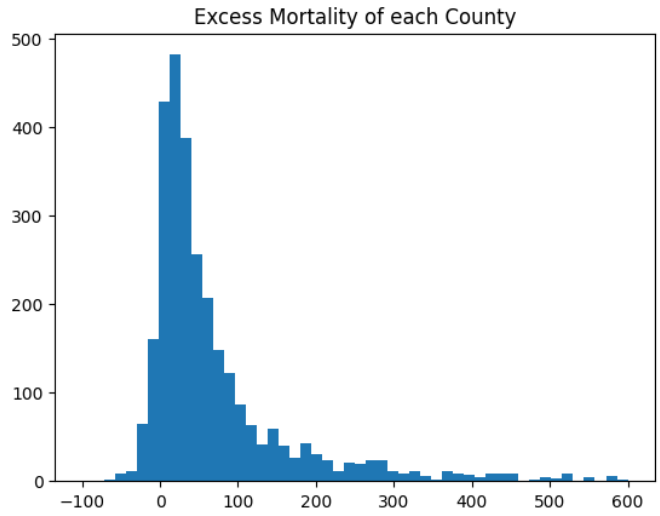


Figure 1, above, illustrates the county-level excess mortality. This histogram is particularly interesting as it shows the potential error of our model, meaning the counties in which we predicted higher mortality rates than what actually occurred. Because we are assuming that mortality should have increased during 2020 (or at least stayed the same), we choose our final model based on which model had the lowest prevalence of these negative excess deaths counties—the Lasso CV model. Because our estimates for these counties are likely incorrect or biased, for further analysis, we removed counties where our model predicted higher deaths than what actually occurred (as was done in Ackley et all 2022). A histogram of our county-level predicted deaths for 2020 using the Lasso CV model can be found in Appendix figure B4.

IV. Results

As described above, we chose to use a cross-validated Lasso model to estimate the number of deaths we would expect there to be in 2020 in the absence of the pandemic, based on the same demographic predictors used to train the model. The CDC reports 350,831 COVID-19 deaths in 2020 (Murphy et al, CDC, 2020). Based on our new data and the true mortality statistics for 2020 (as reported by the CDC), we predict that there were 453,747 excess deaths. In other words, there were potentially 102,916 additional deaths due to the COVID-19 pandemic that were not directly ascribed to the disease itself. As shown in Table 3 below, our model is quite close to other prominent organizations that estimated excess mortality, including the CDC, The Economist, the World Health Organization, and various other academic papers. These results not only validate that our model has similar levels of accuracy to other prominent excess death

models, but that using machine learning methods can be a valuable asset in predicting the effects of these kinds of shocks, and can be used alongside other methods, such as the ones described in the table below.

Table 3: Excess Mortality Results in 2020 Across Sources

Source	Estimated Total Excess Deaths	Method
Our model	**453,747**	Cross-Validated Lasso
CDC [1]	459,421	Farrington Surveillance Algorithms
The Economist [2]	490,934	Gradient Boosted Regression
National Center for Biotechnology Information [3]	438,386	Quasi-Poisson Generalized Linear Model
World Health Organization [4]	465,707	Negative Binomial Spline Model

[1] Centers for Disease Control and Prevention. (2022). Excess deaths associated with COVID-19. National Center for Health Statistics. Retrieved April 19, 2023, from https://www.cdc.gov/nchs/nvss/vsrr/covid19 /excess_deaths.htm
[2] The Economist. (2020, December 23). The pandemic's true death toll. *The Economist.* https://www.economist.com/graphic-detail/coronavirus-excess-deaths-estimates
[3] Ackley et al.. (2022). County-level estimates of excess mortality associated with COVID-19 in the United States. *SSM-Population Health*, 17, 101021.
[4] World Health Organization. (2021). Global excess deaths associated with COVID-19: Modeled estimates. Retrieved April 19, 2023, from https://www.who.int/data/sets/global-excess-deaths-associated-with-covid-19- modeled-estimates

As an additional robustness check of our results, we looked at a disaggregated subset of our population. Table B presents summary statistics for excess mortality and excess deaths not assigned to COVID-19 across rural/urban categories and the percentage of uninsured residents. As expected, because COVID-19 rates were higher in cities than in less-populated areas (largely because of the communicable nature of the disease), we see higher COVID-19 to excess ratios in urban counties (75.24%) than in rural counties (58.71%). It is also possible that these differences could be due to lower reported COVID-19 cases in more rural areas.

Regarding different rates of health insurance across counties, we see that counties with a higher level of residents without health insurance (more than 10%) have a higher excess death rate (15.98%) and lower COVID-19 to excess ratio (68.53%) compared to counties with a lower percentage of uninsured residents (with excess rate of 14.74% and COVID-19 ratio of 77.89%). While we do not have enough data to fully explain what causes these differences, these results could be explained by the theory that those who were uninsured were less likely to seek and receive medical assistance, leading to higher rates of deaths. As shown in Table 4 below, we also

observe that only 69.31% of our estimated excess deaths are attributed to confirmed deaths due to COVID-19 infection.

Table 4: 2020 Total Deaths Results Predicted by our Lasso Model Compared to Actual Observed Values

County type	Predicted Deaths	Total Deaths	Excess Deaths (Total Deaths - Predicted Deaths)	Excess Death Rate (Excess Deaths/Predicted Deaths	Total COVID-19 Deaths	COVID-19 To Excess Ratio %	N
All counties	2,849,115	3,302,862	453,747	15.93	314,509	69.31%	2,774
More Uninsured Counties (>10% of population is uninsured)	1,181,284	1,370,078	188,794	15.98	129,375	68.53%	1,593
Less Uninsured Counties (≤ 10% of population are uninsured)	1,612,128	1,849,823	237,695	14.74	185,134	77.89%	1,175
Urban Counties (population >250,000)	2011812	2,324,193	312,381	15.53	235,037	75.24%	769
Suburban Counties (20,000 ≤ population ≤ 250,000)	645,200	734,981	89,781	13.92	65,191	72.13%	1,135
Rural Counties (population < 20,000)	136,401	160,727	24,326	17.83	14,281	58.71%	865

To better understand and contextualize our results, we looked at the geographic distribution of our results. Figure 1, below, shows the county-level COVID-19 deaths as reported by the New York Times. Figure 2 shows our excess deaths by county. Clearly, the two figures correlate, and we can confidently say that confirmed COVID-19 deaths made up a significant portion of the additional deaths we saw in 2020 (in addition to our model's predicted levels). This correlation is also seen in Figure B1 in the appendix, where we report Covid-19 to excess ratio.

Figure 2: County-Level COVID-19 Deaths in 2020

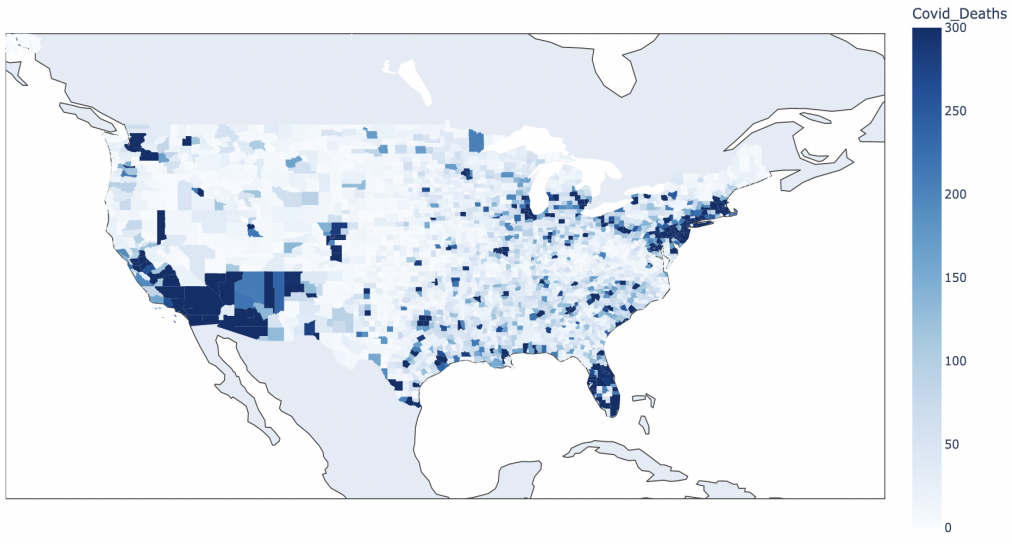
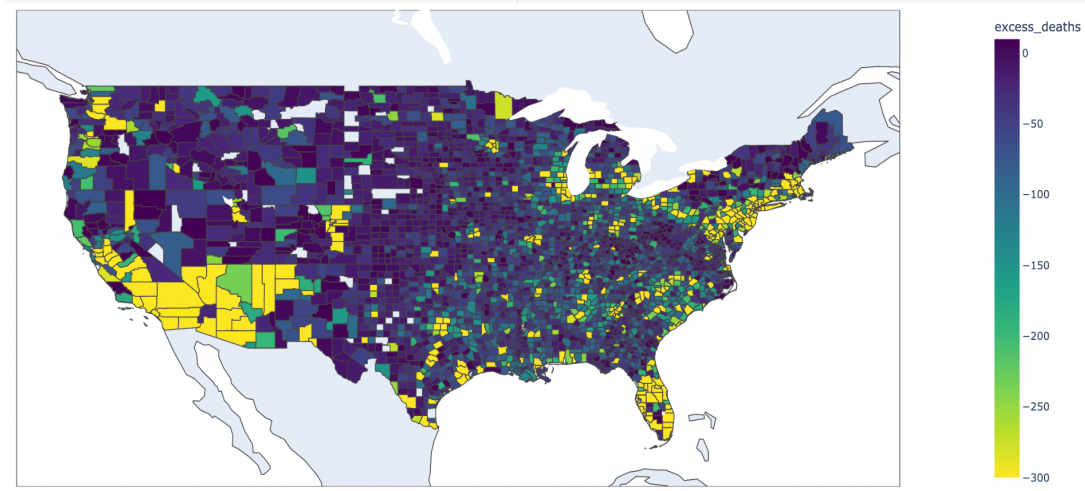


Figure 3: Excess Deaths (Predicted Deaths - True Deaths)



However, Figure 4, which shows the excess mortality rate (the number of excess deaths divided by predicted deaths) implies that confirmed COVID-19 deaths does not fully explain the significant increase in mortality. As seen in the figure below (particularly in Texas, Oklahoma, Nebraska, and the Dakotas), excess mortality is also highly concentrated in locations with lower reported COVID-19 mortality rates, implying that official COVID mortality dates do not capture the full death toll of the pandemic.

Figure 4: 2020 County-Level Excess Mortality Rate

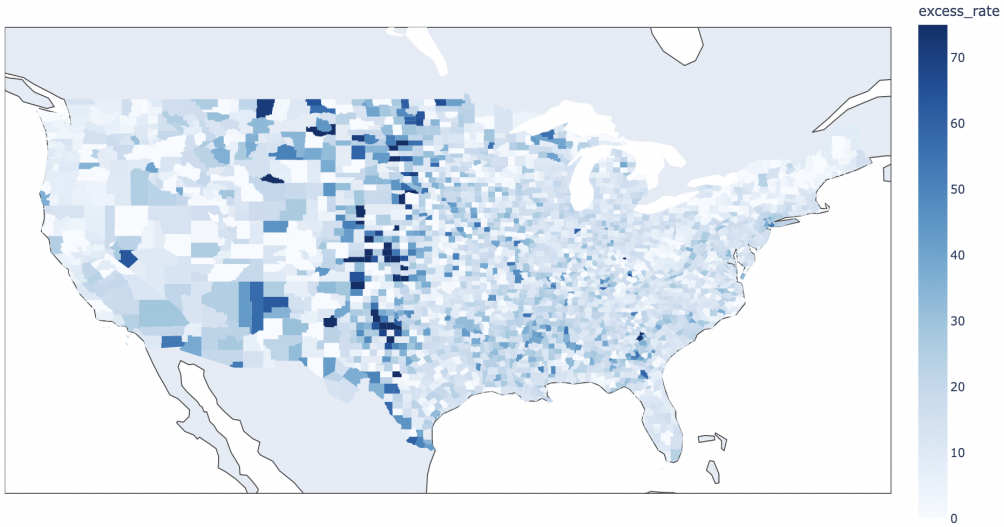
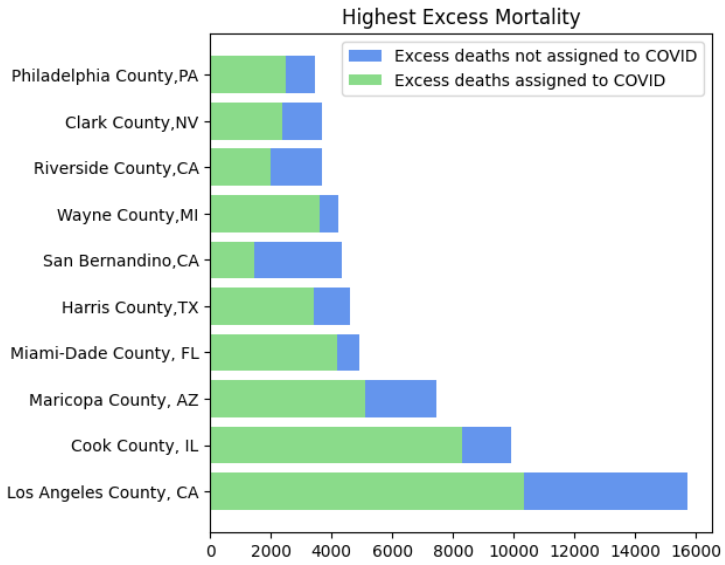


Figure 5 (below) shows the ten counties with the highest excess mortalities with a breakdown of how many of those deaths were assigned to COVID-19. It is important to note that the proportion of excess mortalities assigned to COVID-19 varies greatly. For example, Wayne County, Missouri and San Bernardino County, California have approximately equal excess mortalities, and yet Wayne County assigned a much higher proportion of those deaths to COVID-19 than San Bernandino did. It is beyond the scope of this project to determine the causes of this variance, but possible explanations are that these deaths are unreported COVID-19 deaths or that these deaths were caused by COVID-19 indirectly.

Figure 5: US Counties with Highest Excess Mortality Rates, with COVID-19 Prevalences



The appendix also contains other figures that further investigate these observational trends, including geographic maps of uninsurance rates (B3), the relationship between excess mortality and unemployment (B5), the relationship between excess mortality and rural counties (B6), and the predicted vs realized deaths in 2020 (B7).

V. Conclusion

Our research uses multiple machine learning methods to analyze county-level data both prior to, and during 2020 in order to estimate the true impact of the COVID-19 pandemic. Our findings suggest that the true death toll caused by the global pandemic was 453,747 deaths, which is much higher than the official amount of COVID-19 deaths measured by the CDC of just over 350,000 (Bialek et al., 2020). We further observe that only 69.31% of our estimated excess deaths are attributed to confirmed deaths due to COVID-19 infection. However, although we used the most complete data publicly available, employed various cross-validated models to ensure accuracy, and compared our results to those of similar research, our findings still have limitations that are important to acknowledge.

First, it is difficult to identify how these excess deaths occurred looking only at our data since they may have arisen from a variety of difficult causes. A pattern of under-reporting COVID-19 deaths in rural communities, lack of available testing, limited take-up of medical services, new challenges to mental health, health impacts of stay-at-home orders, and/or disruptions to health care for unrelated diseases due to the increased demand on hospitals and their staffs represent only a few of the possible factors that could explain these excess deaths. We

do not pretend to identify which of these are likely causing the discrepancy between excess deaths and reported COVID-19 figures but instead, believe that this is an important field of research that can be developed to prevent similar disruptions from occurring in the future.

A second limitation of our research is that we rely on the assumption that on a country-wide scale, the COVID-19 pandemic is the only major factor that could explain these differences in mortality rates. A presidential impeachment trial in the U.S. government or racial tension experienced across the United States following the death of George Floyd represent two large events during the same time period as the COVID-19 pandemic, a particularly unusual year impacting American society (Jacobson, 2020; Reny & Newman, 2021). However, upon further exploration of our descriptive summary statistics, the only variables we see varying severely from prior years to 2020 are ones that we would intuitively expect to change during a pandemic, such as communicable disease prevalence or preventable hospital stays. Given that the pandemic represents a relatively exogenous shock whose impact on daily life was impossible to predict, we believe that our previous assumption is a reasonable one. Although we believe it to hold, this assumption still provides opportunities for future research to determine its veracity and to examine the effects of how some of these other shocks may have impacted excess deaths in 2020.

Despite these and any other possible limitations, we believe our research provides a valuable starting point in assessing the real human toll of the pandemic in the United States. In addition, this research provides important implications in examining how a global event was experienced to differing degrees of severity among different subpopulations, and can inform healthcare professionals and policymakers in better protecting vulnerable populations in the event of a future catastrophe. Together with continuing official information about COVID-19 deaths, this project illustrates how estimating excess deaths during the pandemic using machine learning can play a valuable role in future research.

References

Ackley et al.. (2022). County-level estimates of excess mortality associated with COVID-19 in the United States. *SSM-Population Health*, 17, 101021.

Bialek, S., Bowen, V., Chow, N., Curns, A., ... & Wen, J. (2020). Geographic differences in COVID-19 cases, deaths, and incidence—United States, February 12–December 7, 2020. Morbidity and Mortality Weekly Report, 69(15), 465.

Bureau of Economic Analysis. (2020). Personal income by county, metro, and other areas. Retrieved April 19, 2023, from https://www.bea.gov/data/income-saving/personal-income-county-metro-and-other-areas

Centers for Disease Control and Prevention/ Agency for Toxic Substances and Disease Registry/ Geospatial Research, Analysis, and Services Program. CDC/ATSDR Social Vulnerability Index 2020, 2018, 2016, 2014, and 2012 Database: United States. https://www.atsdr.cdc.gov/placeandhealth/svi/data_documentation_download.html. Accessed on February 15, 2023.

Centers for Disease Control and Prevention. (2022). Excess deaths associated with COVID-19. National Center for Health Statistics. Retrieved April 19, 2023, from https://www.cdc.gov/nchs/nvss/vsrr/covid19/excess_deaths.htm

Jacobson, G. C. (2020). Donald Trump and the parties: Impeachment, pandemic, protest, and electoral politics in 2020. *Presidential Studies Quarterly*, 50(4), 762-795.

Reny, T. T., & Newman, B. J. (2021). The opinion-mobilizing effect of social protest against police violence: Evidence from the 2020 George Floyd protests. *American political science review*,

115(4), 1499-1507.

Murphy SL, Kochanek KD, Xu JQ, Arias E. Mortality in the United States, 2020. NCHS Data Brief, no 427. Hyattsville, MD: National Center for Health Statistics. 2021. DOI: https://dx.doi.org/10.15620/cdc:112079

New York Times, (2021, June 23). Coronavirus Case Data for Every U.S. County. *The New York Times*. https://www.nytimes.com/article/coronavirus-county-data-us.html

The Economist. (2020, December 23). The pandemic's true death toll. *The Economist.* https://www.economist.com/graphic-detail/coronavirus-excess-deaths-estimates

The Economist. (2021). How we estimated the true death toll of the pandemic. *The Economist.* https://www.economist.com/graphic-detail/2021/05/13/how-we-estimated-the-true-death-toll-of-the-pandemic

United States Department of Health and Human Services (US DHHS), Centers for Disease Control and Prevention (CDC), National Center for Health Statistics (NCHS), Compressed Mortality File (CMF) on CDC WONDER Online Database.

University of Wisconsin Population Health Institute. (2020). 2020, 2018, 2016, 2014, and 2012 County Health Rankings National Data. https://www.countyhealthrankings.org/explore-health-rankings/county-health-rankings-model

World Health Organization. (2021). Global excess deaths associated with COVID-19: Modeled estimates. Retrieved April 19, 2023, from https://www.who.int/data/sets/global-excess-deaths-associated-with-covid-19-modeled-estimates

Appendix A: Descriptive statistics of the variables

	Pre-2020 (2012, 2014, 2016, 2018)			2020		
County-level variables	N	Mean	St. Dev.	N	Mean	St. Dev.
General county and demographic variables						
Population	12,238	7,870.650	4,757.102	3,076	7,827.664	4,822.795
Square Mileage	12,238	975.675	3,139.565	3,076	1,089.698	3,556.819
% Rural	12,238	15.2	3.59	3,076	20.6	4.05
% Age ≥ 65	12,238	10.1	5.24	3,076	21.7	1.359
% Racial or Ethnic Minority	12,238	20.6	1.75	3,076	23.7	2
% Age ≤ 17	12,238	22.1	2.9	3,076	21.6	3.55
Employment variables						
Socioeconomic status measure	12,238	1.999	0.634	3,076	1.983	0.736
Median household income	12,238	50,812.215	11,149.401	3,076	57,687.414	14,382.892
Number of unemployed	12,238	2,641.782	10,330.402	3,076	2,046.076	7,037.425
% No high school diploma	12,238	13.364	5.693	3,076	12.004	5.867
Limited English Proficiency	12,238	1.016	2.492	3,076	1.233	2.600
Public health variables						
Violent crimes	12,238	237.321	168.447	3,076	250.601	187.039
Sexually transmitted infections	12,238	336.149	216.170	3,076	398.486	279.272
Premature deaths	12,238	7,970.479	2,106.988	3,076	8,518.874	2,732.492
Preventable hospital stays	12,238	64.775	23.424	3,076	4,875.235	1,826.978
Drinking water violations	12,238	0.237	0.425	3,076	0.370	0.483
Air pollution - particulate matter	12,238	10.210	1.710	3,076	8.634	1.941

Teen births	12,238	34.426	14.530	3,076	24.594	12.389
Food environment index	12,238	6.866	1.065	3,076	7.043	1.121
% Uninsured	12,238	11.570	4.758	3,076	9.023	5.001
Prevalence of housing that is crowded	12,238	1.709	1.993	3,076	1.898	2.171
Prevalence of multi-unit housing	12,238	3.684	4.967	3,076	4.414	5.850
Citizens without access to a vehicle	12,238	208.128	243.764	3,076	264.800	342.918
Number of single-parent households	12,238	253.434	261.012	3,076	267.974	309.724
Prevalence of mobile phone usage	12,238	2,467.113	4,091.423	3,076	2,719.590	4,725.327

Appendix B: Additional Figures

Figure B1: Covid-19 Deaths over Excess Deaths, 2020

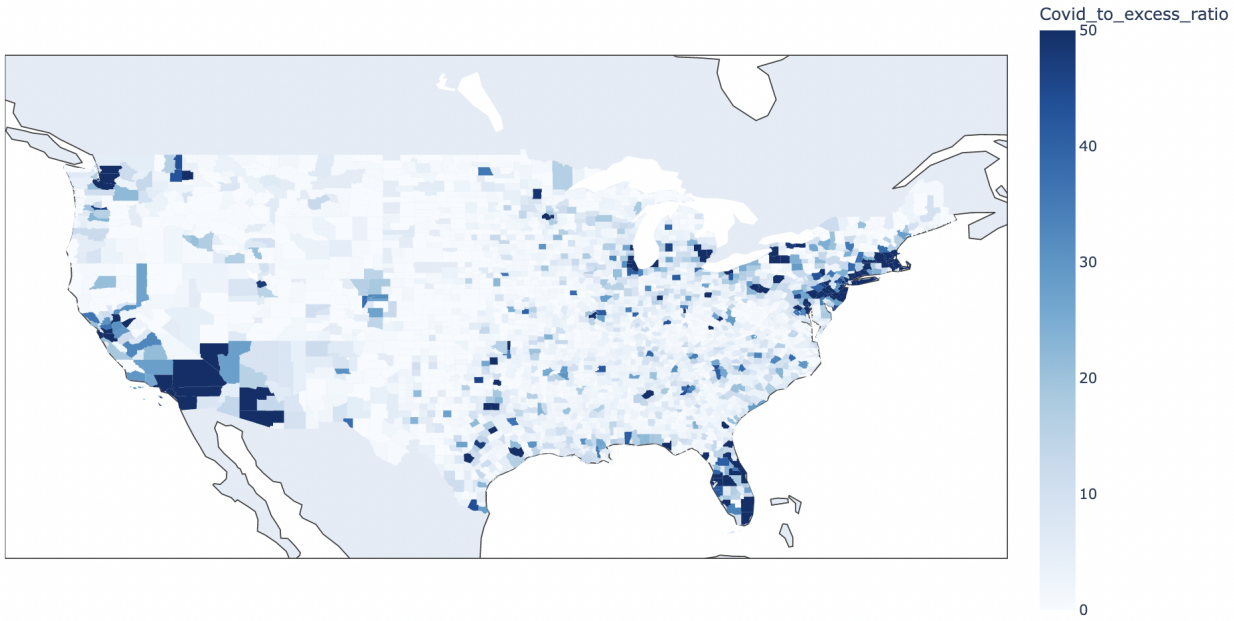


Figure B2: Excess Death Rate Percentile, 2020

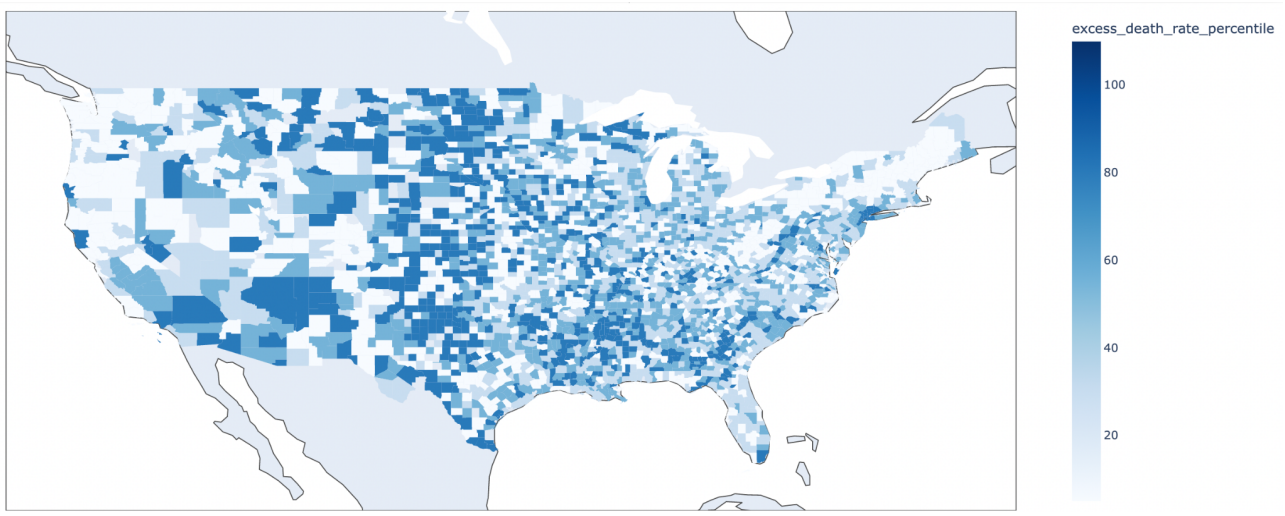


Figure B3: Percent of Population that is Uninsured, 2020

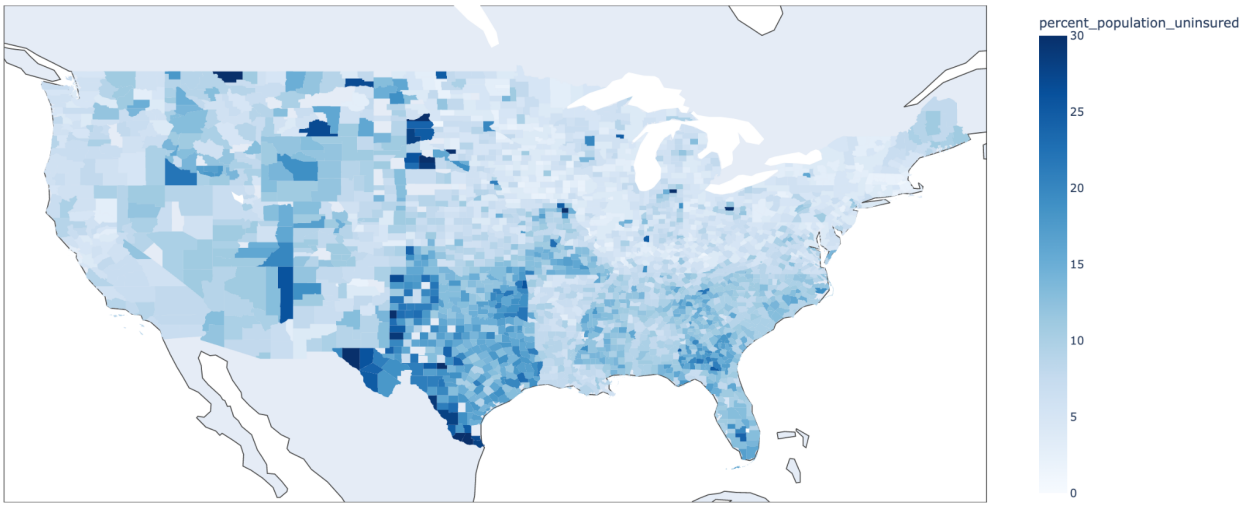


Figure B4: County-Level Predicted Deaths for Cross-Validated Lasso Method, 2020

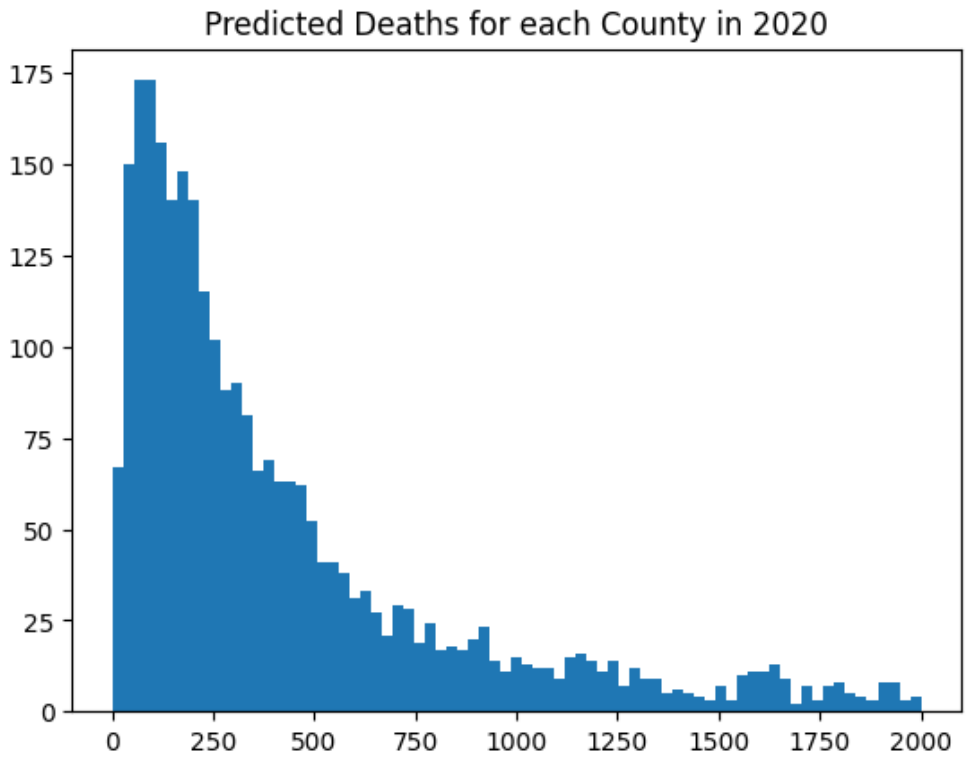


Figure B5: Relationship Between Total Unemployment and Excess Mortality, 2020

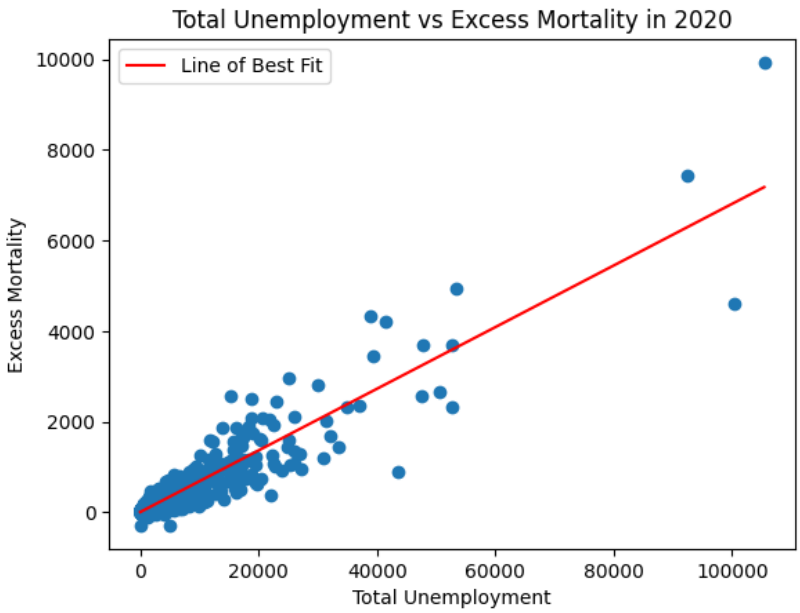
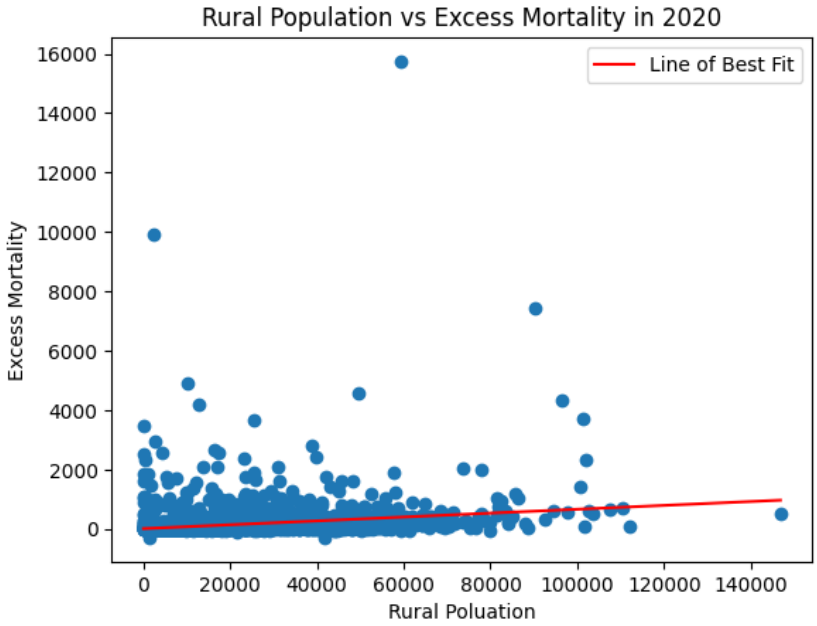
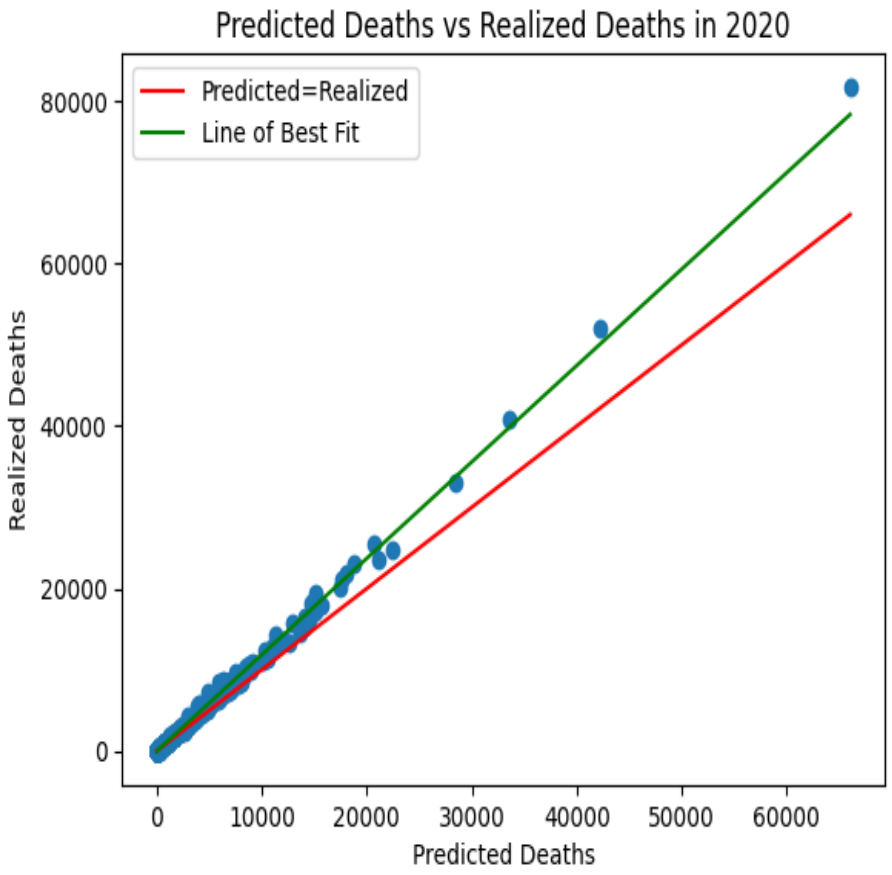


Figure B6: Relationship Between Rural Population and Excess Mortality, 2020*



** We believe the reason why this is slightly upward sloping is because higher rural population correlates with higher overall population, leading to higher excess deaths. However, we believe this figure is still interesting in that you can see that there are quite a few outliers with higher levels of rural population and excess mortality, indicating that COVID-19 did not only affect urban areas. Future research should further explore this relationship, specifically looking at the relationship between the rates of excess mortality and rural population.*

Figure B7: Predicted Deaths vs Realized Deaths, 2020


Predicted Deaths vs Realized Deaths in 2020
Predicted=Realized
Line of Best Fit
Realized Deaths
Predicted Deaths
0
20000
40000
60000
80000
0
10000
20000
30000
40000
50000
60000