Determinants of COVID-19 Mortality at the U.S. County-Level

Stanford Data Science for Social Good Summer 2022

Jiacheng Ge, Charles Hendrickson, Scout Leonard

August 12. 2022

Technical Mentor: Min Woo Sun (Stanford University)

Faculty Mentors: Dr. Robert Tibshirani (Stanford University), Dr. Balasubramanian Narashiman (Stanford University)

Summary

The goal of this project is to build a model which generates reliable estimates of excess mortality at the U.S. county level for 2020 and 2021 using publicly available data. Using these excess mortality estimates, the fellows build models to analyze the effect that various socioeconomic, health, vaccination, and policy factors have on excess death at the U.S. county level.

Introduction

Since the start of the COVID-19 pandemic, more than one million Americans have died from COVID-19 (CDC 2020). These deaths have been tracked and reported using death certificate cause of death claims. This method of tracking COVID-19 deaths can misrepresent the true mortality burden of the pandemic for a variety of reasons. Reported COVID-19 deaths may not reflect true mortality burdens because COVID-19 testing has not been accessible for the duration of the pandemic. There are also indirect COVID-19 deaths that contribute to the mortality burden of a pandemic, including strained healthcare systems (Ackley et al. 2022).

Estimating excess mortality has been utilized as a more achievable way to capture the full mortality burden of the pandemic (Ackley et al. 2022) (Knutson et al., n.d.). Excess death describes the difference between all cause mortality, the reported death from all causes for a

given period of time, and expected mortality, the expected amount of death based on historic death rates.

In this project, we model the COVID-19 excess mortality rate in the United States at the county level and analyze the effect of socioeconomic, health, vaccination, and policy factors on estimated excess mortality for 2020-2021 using statistical learning models. This study will help elucidate the factors contributing to the U.S. county-level excess mortality rate and guide policy-makers in U.S. counties in improving their response to pandemics, with the goal of ultimately reducing the number of American deaths by guiding effective and equitable policy interventions. Our study utilizes open, reproducible, and rigorous science and can be used as a template for further research on the determinants of COVID-19 mortality.

It is noteworthy that previous studies have investigated the effect of determinants of COVID-19 mortality, but have focused solely on confirmed COVID-19 deaths (Fielding-Miller, Sundaram, and Brouwer 2020) (Squalli, n.d.). Instead, our study models excess deaths, which is widely considered a more objective indicator of the COVID-19 death toll. Furthermore, there has not been a holistic analysis including variables describing policies, such as stay at home orders, mask mandates, and COVID-19 vaccination rate.

Our Contribution:

This report describes our methods for understanding COVID-19 mortality at the U.S. county level in two phases: first for predicting county-level expected deaths in order to generate excess death estimates using all cause mortality data, and second, for comparing models which include intrinsic, policy, and public health measure feature importance to understand important factors in determining county-level COVID-19 mortality. Our results quantify excess death estimates for the 500 most populous U.S. counties for 2020 and 2021. Furthermore, this work demonstrates that income ratio, which represents income inequality of a county, and diabetes prevalence and obesity prevalence rank high in determining excess death outcomes in U.S. counties for the years 2020 and 2021.
Our code is version-controlled in a publicly available GitHub repository so that others may reuse and build on these analyses. Final datasets used in modeling are also stored in this GitHub repository.

Methodology

Data Sources

All of the data used in this project are provided by open-access repositories as outlined below. For more detailed descriptions of how each dataset was downloaded, see the Appendix.

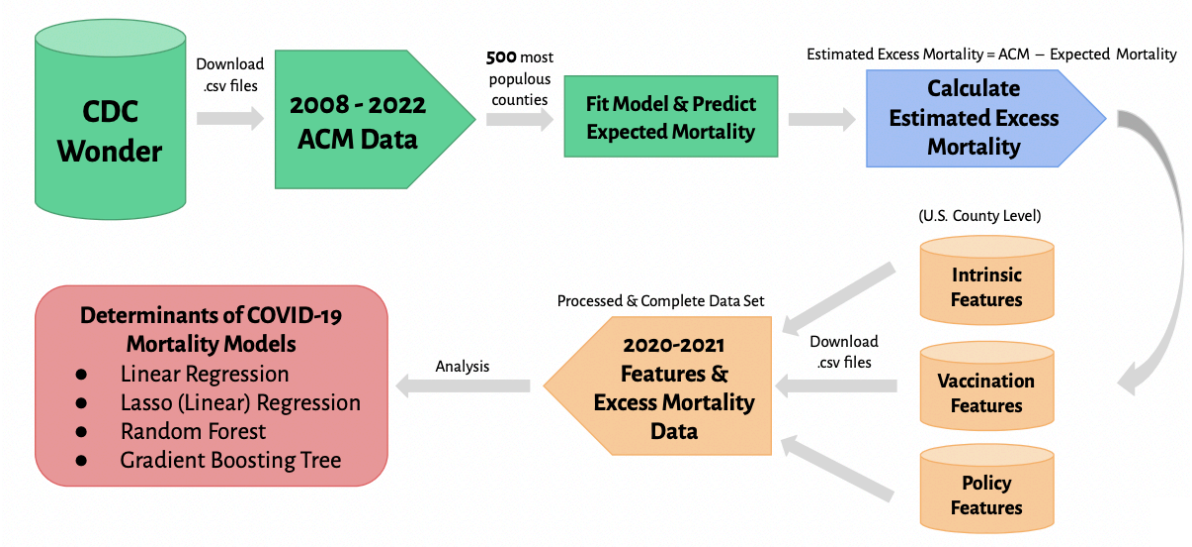


Figure 1: Phase I and Phase II data aquisition and modeling overview.

Phase I Modeling Data

Models for estimating county-level excess death utilize historic all cause mortality and population data at the U.S. county level. This data is provided by the CDC Wonder online database as cause of death data based on the death certificates of U.S. residents. 2012 - 2020 all cause mortality data is final, whereas 2021 data is provisional. Note that the CDC Wonder database gives 2020 and 2021 population values which are the same at the county-level.

Phase II Modeling Data

Intrinsic county features, such as health metrics and socioeconomic factors are included in phase II models to understand county-level intrinsic features importance in excess death outcomes for 2020 and 2021. These features were accessed via the University of Wisconsin's Population Health Institute County Health Rankings and Roadmaps database. These features are available as county level crude rates for any given year.

Vaccination features are also included in phase II models as county-level features. Vaccination data is sourced from the CDC for 496 U.S. counties of the 500 in the models. This data shows daily county-level vaccination percentages for different age groups and variations of dose completion.

The remaining four counties which did not have daily vaccination data available at the U.S county-level were Honolulu County, Hawaii County, and Maui County in Hawaii, and Barnstable County in Massachusetts. The next best available daily vaccination data for Honolulu

County, Hawaii County, and Maui County was available at the state level. This data was available on the State of Hawaii Department of Health website. Honolulu County, Hawaii County, and Maui County's vaccination percentages were imputed using state-level vaccination rates, and AUC values were calculated from these. The final missing vaccination data for Barnstable County, Massachusetts was imputed with the mean vaccination AUC value of Berkshire County, Hampden County, and Hampshire County Massachusetts, which had the closest populations to Barnstable County.

Finally, the data for policy features is drawn from healthdata.gov. This data shows state and county-level COVID-19 policy orders, as well as their major groupings. The phase II models include six policy order groupings or "types," including shelter in place policies, food and drink policies, non-essential business policies, outdoor recreation policies, entertainment policies, and mask requirements.

Modeling

Phase I Data Processing

Data for modeling county level expected mortality has yearly coverage from 2007 to 2022. Raw data from CDC Wonder contains the following variables: state, county name, county code, year, population, deaths, and death crude rate.

To prepare a dataset for phase I modeling and predict expected deaths for counties in 2020 and 2021, an exposure term is calculated. The exposure term is equal to the county population divided by 100,000. Additionally, the following 2-year-lag terms are added to the dataset: population, deaths, and death rate. The raw values of these variables are available as numeric yearly data.

The processed data is filtered to only include the top 500 most populous counties, as determined my mean population for years of coverage in CDC Wonder raw data.

The Phase I model output is county-level expected mortality estimates for 2020 and 2021. These estimates are used to calculate excess mortality for 2020 and 2021. With these excess mortality counts we calculate excess mortality rates by dividing the product of excess mortality and 100,000 by the population of that county. Our final data product from Phase I are county level estimates of excess mortality and excess mortality rate for 2019 (for model validation), 2020, and 2021.

Phase I Modeling

To predict county-level expected mortality for 2020 and 2021, we used a regularized poisson generalized linear model. Prior work estimating county-level excess death utilizes a quasi-poisson model to estimate county level excess mortality in 2020, and we modify this approach based

on the RMSE model performance metric of a quasi-poisson model and regularized poisson alternative. The model's form appears below:

$ln(\frac{\lambda}{pop_t}) = \beta_0 + \beta_1 \frac{death_{t-2}}{pop_{t-2}} + \beta_{2i} + \beta_{3i} t + \theta(\frac{1-\alpha}{2} L_2 + \alpha L_1)$

$L_1 = |\beta|$

$L_2 = |\beta|^2$

t = current year - base year (2011 in this case)

death_t = ACM at time t

$\text{pop}_t = \frac{Population}{100{,}000}$ at time t

β_{2i} for i_{th} county

β_{3i} for the interaction between i_{th} county and time

Here, $death_{t-2}$ is all cause mortality lagged two years from a given year relative to the base year t. The base year used for the data input to this model is 2011. β_{2i} is measuring the differences across counties besides population. Additionally, the time trend is allowed to vary across counties as determined by β_{3i}. Furthermore, there is a Lasso penalization L_1 and Ridge penalization L_2 in our model, which are added for regularization to avoid over-fitting our model.

The model was trained on our processed county-level all cause mortality data from 2011 to 2018 and was validated on 2019 all cause mortality data. With our final fitted model we generated expected mortality estimates for 2020 and 2021 and visualized comparisons for expected all cause mortality versus actual all cause mortality.

Phase II Data Processing

For Phase II data, the response variable, excess death rate, is the average excess death rate from the two values generated in Phase I modeling for each year.

Intrinsic features of counties examined include percent of adults with obesity, percent of adults with diabetes, primary care physicians rate, percent uninsured, income ratio, overcrowding and severe housing cost burden. The raw values of these variables are available as numeric yearly data for the county, and the average yearly value is used in Phase II model data.

Policy features for 2020 and 2021 are the same. More specifically, they are processed as follows: We collect the start date of the following policy order types: shelter in place, food and drink restrictions, nonessential business restrictions, entertainment restrictions, and mask requirements, as well as the length of mask requirements order. We subtract 2020-01-20 from the collected policy start dates to get the number of days before county policy-makers enacted the policy. We generate ranks based on the government response time: the sooner the implementation, the lower the rank , and the longer the shelter in place order, the lower

the rank. For counties that did not impose a policy, the state-level policy is considered as its policy order. If there are no state-level orders, we assign the county with the largest rank. All policy orders which we use took effect in 2020, except for two policy records. Since many policy orders are implemented at the same time, we perform a Principal Component Analysis (PCA), shown in Figure 3a, for the policy features and only keep the first and second PCs for modeling.

Vaccination data include the daily cumulative percentage of a county's population which received their first dose and the cumulative percentage of a county's population which received their complete series of doses. We calculate the Area Under the Curve (AUC) for the cumulative percentages for fair comparison between counties having a significant change in the vaccination rate during a year. Few counties had administered vaccines until 2021. We sum up the AUC across 2020 and 2021.

Finally, correlation between features is evaluated to understand multicollinearity in phase II models. We find that percent of adults with obesity and percent of adults with diabetes are highly correlated, and that all of the vaccination features are highly correlated. Correlation plots are shown in Appendix Figure 4

Phase II Modeling

In the first round model selection, we compare four models: linear regression, lasso linear regression, Random Forest, and Extreme Gradient Boosting tree. These models are run with all features on a validation of 100 U.S. counties, randomly generated in the data processing step which generates a random fold, 1-5 for each county. For each model, a RMSE for the validation set is calculated to compare performance between the four models.

By comparing the RMSE, we select the two best models: one regression model, lasso regression, and one tree-based model, Extreme Gradient Boosting, for further comparison.

To mitigate the issue of multicollinearity, we remove percent of adults with obesity (correlated with percent of adults with diabetes), percent of adults who are uninsured (correlated with vaccination features), severe housing cost burden, and percent of population with complete series of doses AUC (correlated with percent of population with their first dose completed) for both models. With this, four intrinsic variables remain: percent of adults with diabetes, overcrowding, income ratio, primary care physicians rate, two policy order variables: policy principal component 1 and policy principal component 2, as well as one vaccination variable: Administered Dose 1 Population Percent AUC. Two selected models are run again after feature selection. The regularization constant used for lasso regression is 0.01. The hyperparameters used for XGBoosting are number of trees 420, max depth 3, learning rate 0.06. The validation RMSE of Lasso decreases from 50.9 to 48.2 while the one of XGBoosting increases from 42.7 to 46. Removing collinear variables does not precipitate the prediction but it helps us understand the contribution of features better.

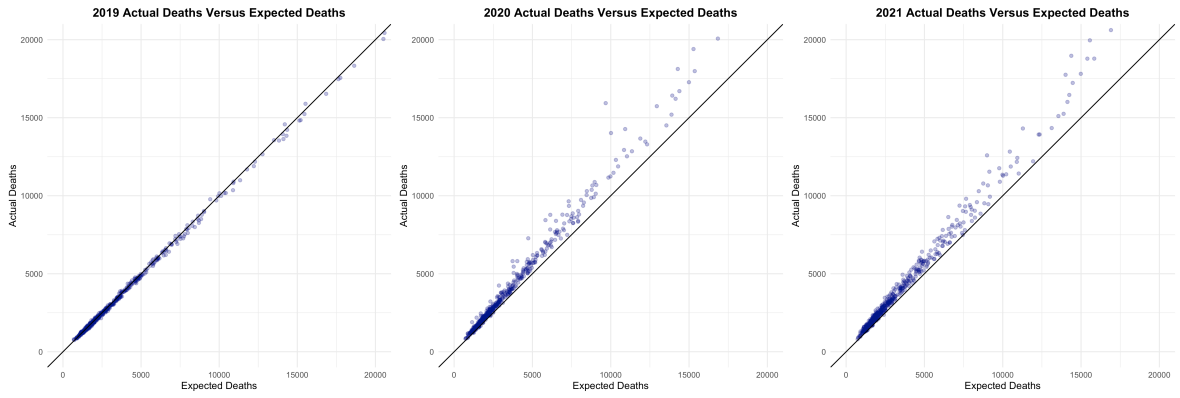
Results

Phase I Model Results

Using 2012-2018 all cause mortality data to predict estimated all cause mortality and 2019 to validate the model, we generate expected mortality and expected mortality crude rates for the years 2020 and 2021 for the 500 most populous U.S. counties. This value is then used to generate excess death and excess death crude rate.

We use Root Mean Squared Error (RMSE) values to measure the error of predictions for 2019, 2020, and 2021, with the expectation that 2019 has a small RMSE because it is a "normal" year in which the model can predict all cause mortality with higher accuracy. RMSE for 2019 predictions is 128.6, whereas RMSE for 2020 and 2021 are 1479.58 and 1042.35, respectively.

Predictions for 2019, 2020, and 2021 are also compared to all cause mortality measured outcomes, shown in Figure 2. As expected, and as indicated in reported RMSE values, measured all cause mortality outcomes are larger than predicted mortality for 2020 and 2021, whereas measured all cause mortality and predicted mortality are much closer for 2019.



(a) (b) (c)

Figure 2: Predicted deaths for the 500 most populous counties versus measured all cause mortality for 2019, 2020, and 2021.

Excess death rates were shown visually on U.S. maps, which are available in Appendix Figure 5.

Phase II Model Results

We measure the contribution of county-level features by comparing permutation feature importance score. Permutation importance score is calculated by first computing negative RMSE on validation set D. For each feature j, data in column j is randomly permuted to generate

a shuffled version of the data, Dk,j. Then, the score Sk,j of the model on shuffled data is computed. Finally, we compute importance i for feature j defined as the decrement in the negative RMSE. We run the procedure 10 times for each variable. Feature importance plots are included in Appendix Figure 7

Both models, lasso regression and Extreme Gradient Boosting, suggest the percentage of diabetic adults and the income ratio are the top two important features. More specifically, the models show that counties with higher percentages of diabetic adults showed higher crude rates of excess deaths, and counties with higher income ratios also showed higher rates of excess deaths.

For some counties, for example in San Francisco County, California and Taylor County, Texas, when policy and vaccination features are added to the model, excess death rate predictions tend to be closer to the phase 1 excess death rate.

Not every county, however, follows this pattern due to remaining uncertainty in the predictions.

Conclusion

Phase II modeling showed that intrinsic features play critical roles in determining county-level excess death. Furthermore, diabetes prevalence and income ratio are closely related to the outcomes..

For policy measures and preventive actions, strict COVID-19 restrictions and higher rates of vaccination are likely to reduce county-level excess death crude rates. Understanding the intrinsic and policy intervention determinants of excess death rates has the potential to help counties take preventative action in future pandemics and prevent death.

While policy features do not compare to the importance of intrinsic features, It is possible that the addition of compliance features to these models could account for differential compliance to county-level COVID-19 policies.

In future work, we see accessing policy compliance data possibly improving model predictions and understanding policy order feature importance. Meanwhile, policy and vaccination features are time-dependent. Their influence can be weak in the short term, but it is worthwhile to see if there are long term effects. There is still much work ahead in quantifying the magnitude of variable importance.

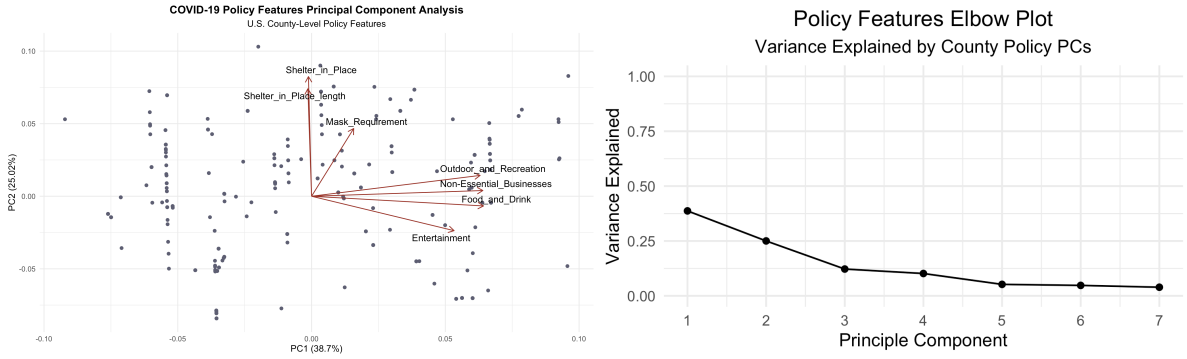
Finally, this project only studies the correlation between county-level features and excess deaths, but causality requires further exploration.While more granular data for the features is less available at the U.S. county-level, a state-level or country-level analysis could further elucidate the importance of these features.

Appendix

Figures

Additional project figures are included below.

Policy Features PCA Figures



(a) (b)

Figure 3: PCA autoplot and elbowplot for policy features.

Phase II Features Correlation

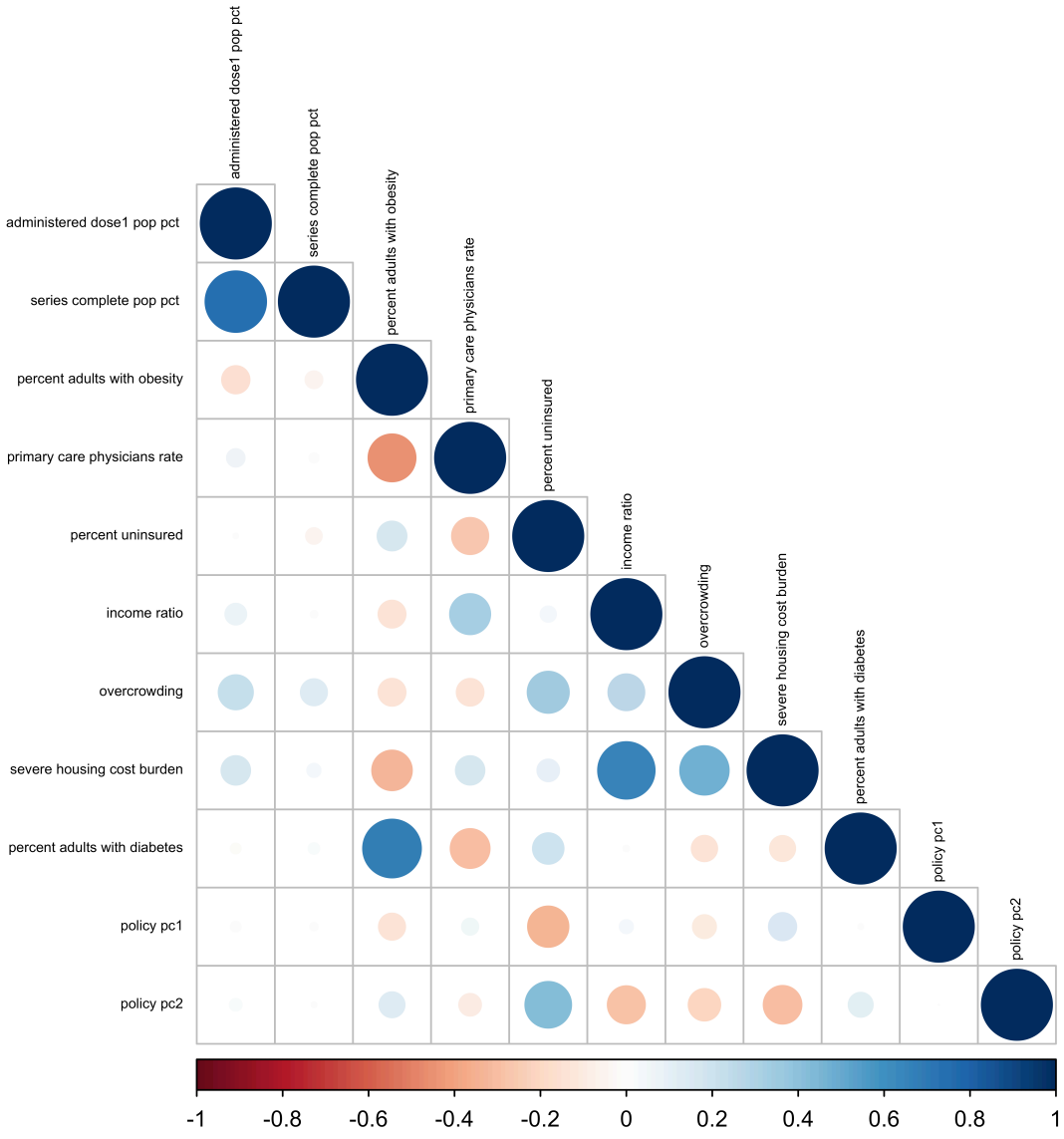


Figure 4: Correlation plot for 2020 policy, vaccination, and intrinsic features.

Predicted Excess Death Rate Maps

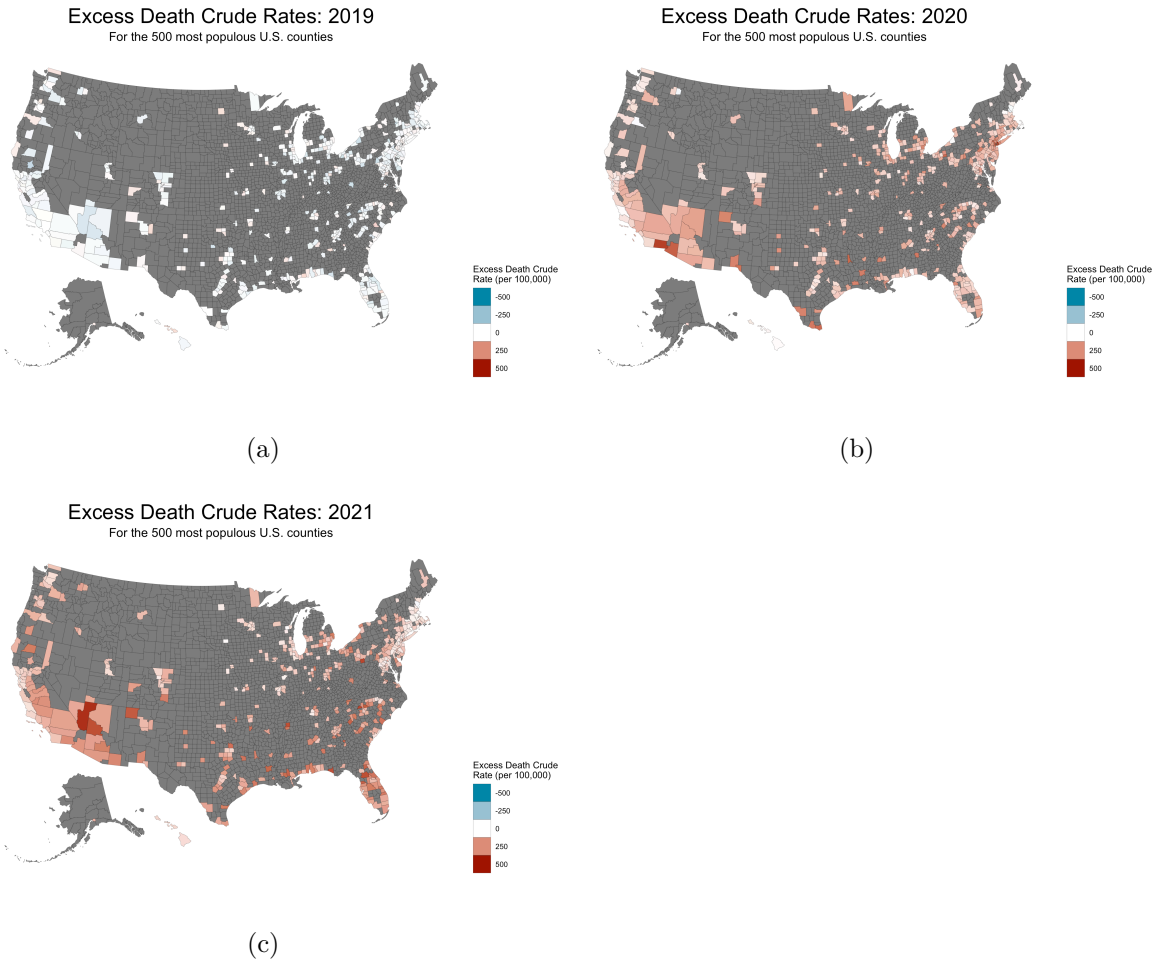
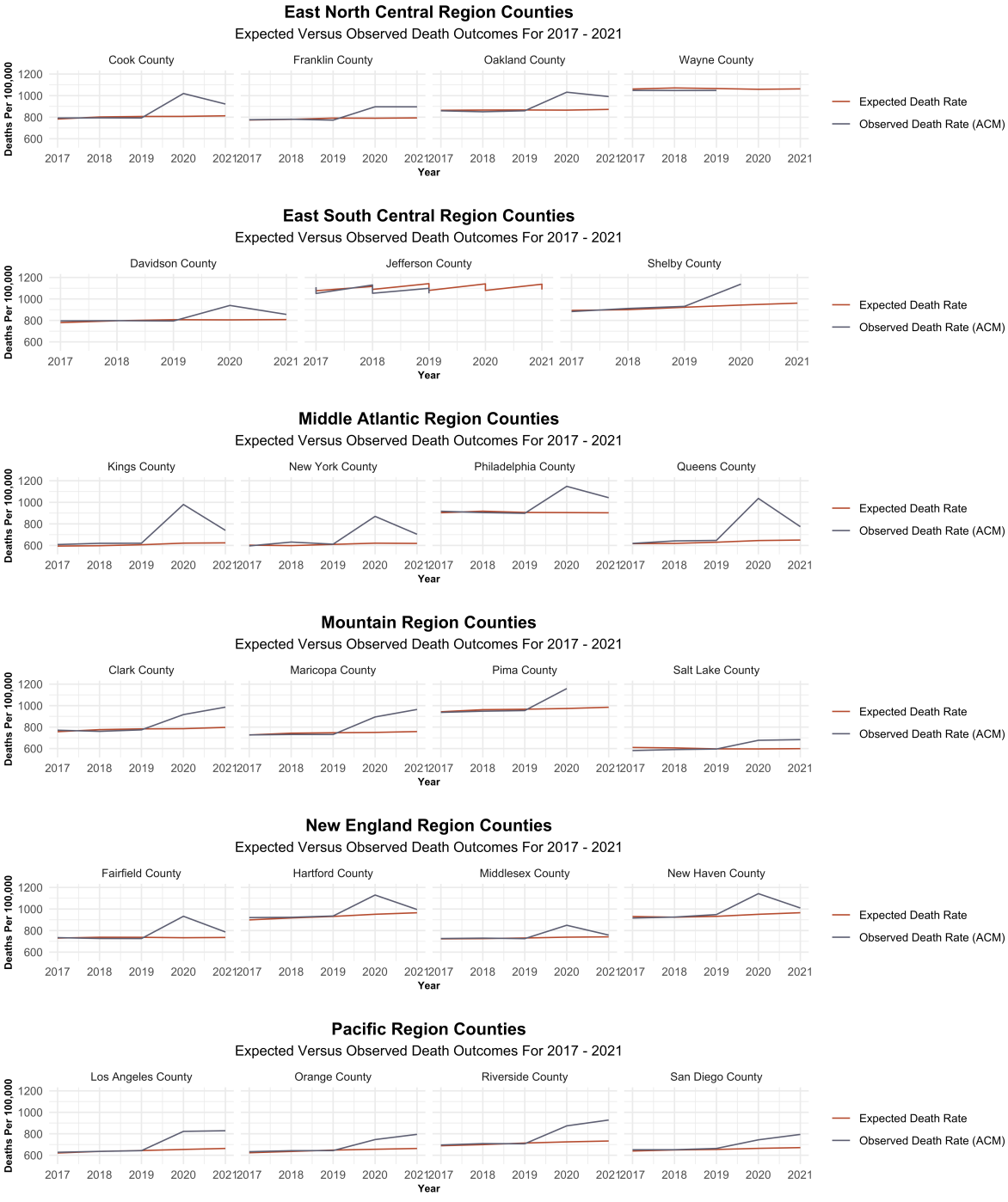
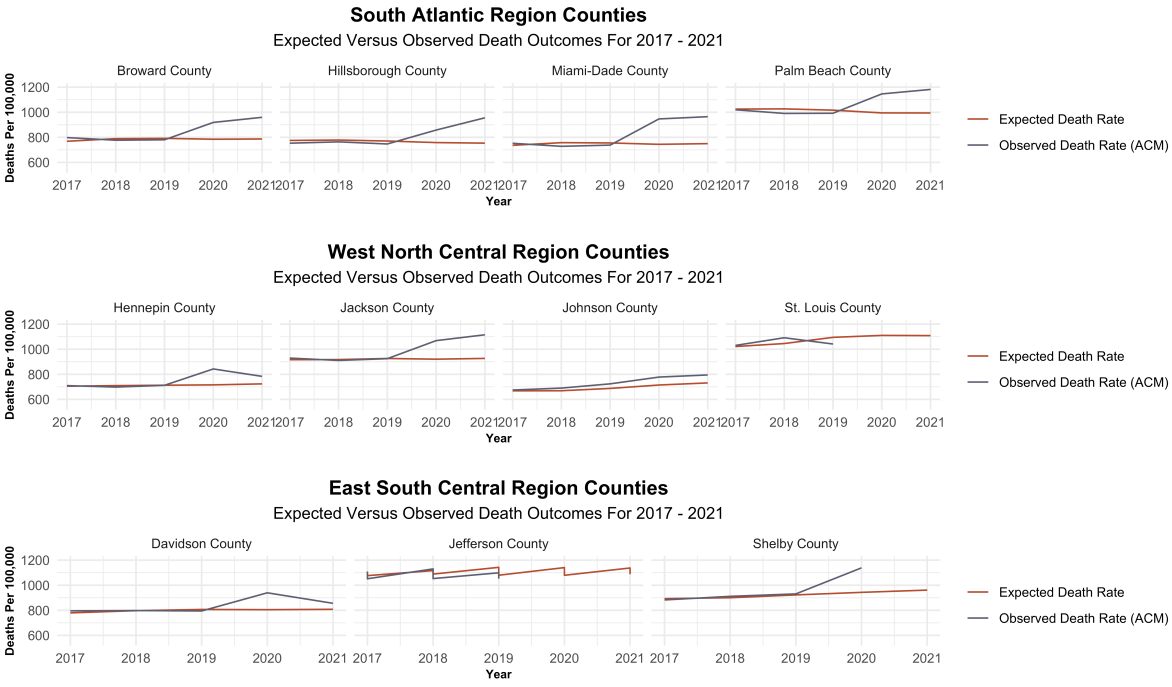


Figure 5: Excess death rate predictions for the 500 most populous U.S. counties.

Example County-Level ACM and Expected Death Rates

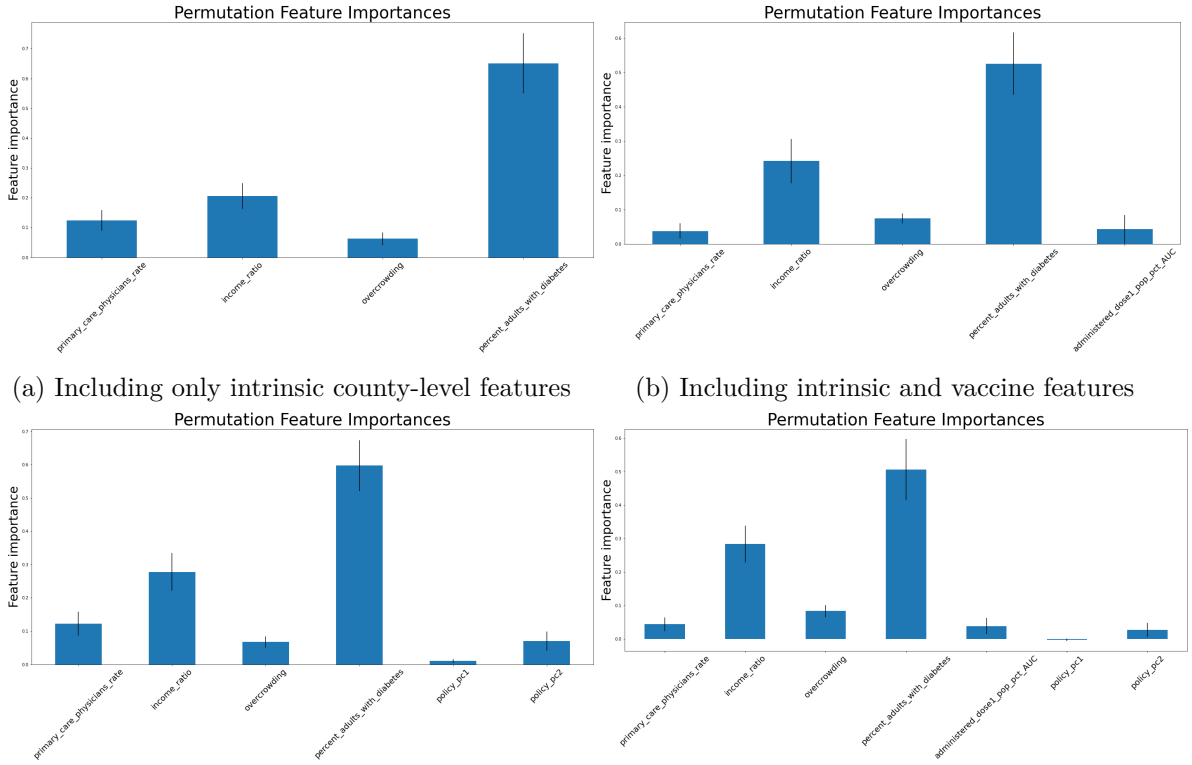



South Atlantic Region Counties
Expected Versus Observed Death Outcomes For 2017 - 2021
Broward County
Hillsborough County
Miami-Dade County
Palm Beach County
Deaths Per 100,000
1200
1000
800
600
2017
2018
2019
2020
2021
Year
Expected Death Rate
Observed Death Rate (ACM)

West North Central Region Counties
Expected Versus Observed Death Outcomes For 2017 - 2021
Hennepin County
Jackson County
Johnson County
St. Louis County
Deaths Per 100,000
1200
1000
800
600
2017
2018
2019
2020
2021
Year
Expected Death Rate
Observed Death Rate (ACM)

East South Central Region Counties
Expected Versus Observed Death Outcomes For 2017 - 2021
Davidson County
Jefferson County
Shelby County
Deaths Per 100,000
1200
1000
800
600
2017
2018
2019
2020
2021
Year
Expected Death Rate
Observed Death Rate (ACM)

Lasso Model Feature Importance Plots



(a) Including only intrinsic county-level features

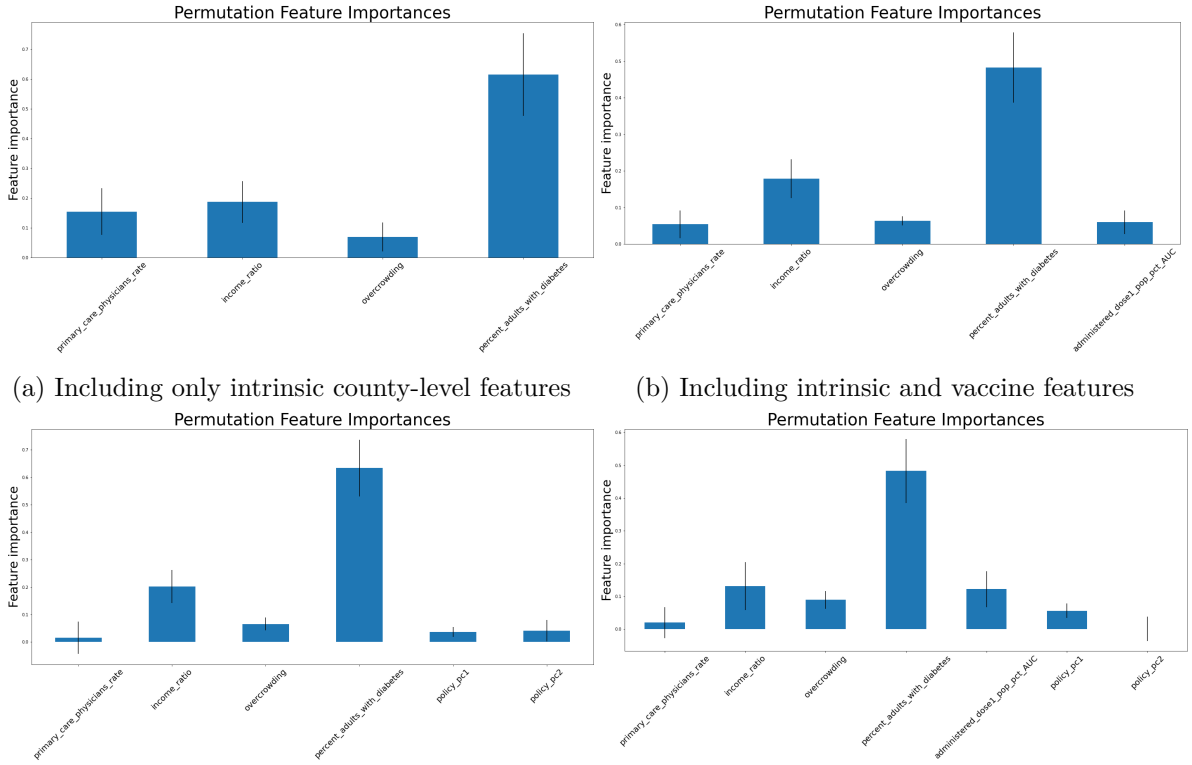
(b) Including intrinsic and vaccine features

(c) Including intrinsic and policy features

(d) Including intrinsic, policy, and vaccine features

Figure 6: Lasso feature importance

Extreme Gradient Boosting Feature Importance Plots



(a) Including only intrinsic county-level features

(b) Including intrinsic and vaccine features

(c) Including intrinsic and policy features

(d) Including intrinsic, policy, and vaccine features

Figure 7: Extreme gradient boosting feature importance

Manual Data Download Documentation

Certain raw data files used in this analysis were downloaded maually with file sizes too big for this GitHub repository.

Data download information was recorded in a .txt file in the raw data directory.

For any other questions about data used in this project, please contact the DSSG Covid Mortality team using a GitHub Issue in the project's public repository.

Student Fellows

Jiacheng Ge is a graduate student in the Statistics Department at Stanford. In Summer 2022, he served as a fellow of the Stanford Data Science for Social Good program. His research topic

is on the excess mortality of COVID-19.

Charles Hendrickson (University of California, Santa Barbara) recently graduated with a Master of Environmental Data Science from the Bren School of Environmental Science & Management at UC Santa Barbara. He is passionate about advancing scientific knowledge by leveraging data science to monitor, model, and manage the Earth's marine ecosystems. Charles has worked on projects ranging from assessing Nassau grouper spawning aggregation demographics to visualizing spatial and temporal patterns of coral reef stressors surrounding Moorea, French Polynesia.

Scout Leonard (University of California, Santa Barbara) holds a masters degree in environmental data science from the Bren School of Environmental Science & Management at UC Santa Barbara. Scout is passionate about data visualization and open, reproducible science, and her goal is to contribute to sustainable and equitable food systems.

References

Ackley, Calvin A., Dielle J. Lundberg, Lei Ma, Irma T. Elo, Samuel H. Preston, and Andrew C. Stokes. 2022. "County-Level Estimates of Excess Mortality Associated with COVID-19 in the United States." *SSM - Population Health* 17 (March): 101021. https://doi.org/10.1016/j.ssmph.2021.101021.

CDC. 2020. "COVID Data Tracker." https://covid.cdc.gov/covid-data-tracker.

Fielding-Miller, Rebecca K., Maria E. Sundaram, and Kimberly Brouwer. 2020. "Social Determinants of COVID-19 Mortality at the County Level." *PLOS ONE* 15 (10): e0240151. https://doi.org/10.1371/journal.pone.0240151.

Knutson, Victoria, Serge Aleshin-Guendel, Ariel Karlinsky, William Msemburi, and Jon Wakefield. n.d. "ESTIMATING GLOBAL AND COUNTRY-SPECIFIC EXCESS MORTALITY DURING THE COVID-19 PANDEMIC," 24.

Squalli, Jay. n.d. "Evaluating the Determinants of COVID-19 Mortality: A Cross-Country Study," 14.