Ensemble Machine Learning of Factors Influencing COVID-19 Across US Counties

David McCoy
Division of Environmental Health Sciences, UC Berkeley https://orcid.org/0000-0002-5515-6307
Whitney Mgbara
Dept. Env. Science, Policy and Management, UC Berkeley
Nir Horvitzc
School of Mathematical Sciences, University of Kwazulu-Natal
Wayne M. Getz
Dept. Env. Science, Policy and Management; School of Mathematical Sciences, University of Kwazulu-Natal
Alan Hubbard (hubbard@berkeley.edu)
Division Biostatistics, UC Berkeley

Research Article

Keywords: COVID-19, Machine Learning, Super Learner, Variable Importance

Posted Date: October 14th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-90547/v1

Ensemble Machine Learning of Factors Influencing COVID-19 Across US Counties

David McCoy[a,1], Whitney Mgbara[b,1], Nir Horvitz[c], Wayne M. Getz[b,c], Alan Hubbard[a]

Corresponding author: Alan Hubbard

[1] *Co-first authors listed alphabetically*

All code for collecting data, collected data, statistical scripts, up-to-date outcome data, visualizations, and statistical results are available on GitHub: `https://github.com/blind-contours/Getz_Hubbard_Covid_Ensemble_ML_Public.git`

email addresses: david_mccoy@berkeley.edu, wmgbara@berkeley.edu, nirh@berkeley.edu, wgetz@berkeley.edu, hubbard@berkeley.edu

[a] *Division Environmental Health Sciences, UC Berkeley, CA 94720, USA*
[b] *Dept. Env. Science, Policy and Management, UC Berkeley, CA 94720, USA*
[c] *School of Mathematical Sciences, University of Kwazulu-Natal, Durban 4000, South Africa*
[d] *Division Biostatistics, UC Berkeley, CA 94720-3114, USA*

Abstract

Background: Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) the causal agent for COVID-19, is a communicable disease spread through close contact. It is known to disproportionately infect certain communities due to both biological susceptibility and inequitable exposure. In this study, we investigate the most important health, social, and environmental factors impacting both the early and later phases of COVID-19 transmission and mortality in US counties.

Methods: We aggregate county-level physical and mental health, environmental pollution, access to health care, demographic characteristics, vulnerable population scores, and other epidemiological data to create a large feature set to analyze COVID-19 outcomes. Because of the high-dimensionality and multicollinearity of the data, we use ensemble machine learning and marginal prediction methods to identify the most salient factors associated with several COVID-19 outbreak measures.

Findings: Our variable importance results show that measures of ethnicity,

public transportation and preventable diseases are the strongest drivers for both incidence and mortality. Specifically, the CDC measures for minority populations, CDC measures for limited English, and proportion of Black/African-American individuals in a county were the most important features for COVID-19 cases at day 25 and to date. For mortality at day 100 and total to date, we find that public transportation use and proportion of Black/African-American individuals in a county are the strongest predictors. The methods predict that, keeping all other factors fixed, a 10% increase in public transportation use increases mortality at day 100 by 2012 (95% CI [1972, 2356]) and likewise a 10% increase in the proportion of Black/African-American individuals in a county increases total deaths to date by 2067 (95% CI [1189, 2654]). In terms of cases to date, ethnicity turns out to almost twice as important as the next most important factors, which are location, disease prevalence, and transit factors.

Interpretations: Our findings indicate that a more focused approach should be taken when managing COVID-19, by considering features of the economy most responsible for transmission and sectors of society most vulnerable to infection and mortality. In particular, our results strongly reinforce others pointing to the disproportionate impact of COVID-19 on minority populations. They also suggests that mitigation measures, including rolling out vaccinations as they become available, will be most efficacious for the US population as a whole when, beyond healthcare workers and first responders, are focused first on the highest-risk communities.

Funding: UC Berkeley, Biomedical Big Data Training Fellowship; NSF Grant 2032264 to WMG and AH.

Keywords: `COVID-19`, Machine Learning, Super Learner, Variable Importance

---

1. Introduction

COVID-19 Background. Coronavirus disease 2019 (COVID-19), caused by the novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has spread rapidly around the world resulting in a pandemic. [1, 2, 3, 4, 5, 6]. Within a month of discovering the first cluster of cases in Wuhan, China [1, 6, 3], 18 additional countries had reported a case of COVID-19 [7]. The World Health Organization declared the resulting outbreaks a Public Health Emergency of International Concern by January 30, 2020 and a pandemic by March 11, 2020 [7, 8].

In the United States, to the best of our knowledge, the COVID-19 pandemic began to spread on January 21, 2020 when an infected individual traveled from Wuhan, China to Snohomish County, Washington [9]. Community circulation, within US states and territories, followed shortly after [10, 11]. By May 20, 2020, the U.S. had the most confirmed active cases and the most deaths of any country, with a reported total of 1,559,750 cases and 92,333 deaths across 55 US jurisdictions including 50 states, District of Columbia, Guam, Puerto Rico, the Northern Mariana Islands, and the U.S. Virgin Islands [8].

There are a few consistent observations regarding the epidemiology of the COVID-19 pandemic in the US. Most prominent is the relatively high infection and death rates of minority populations, particularly Black/African Americans [12] [13], a disparity researchers have noted occurred in previous pandemics, such as HIV [14]. This disparity has been observed in both adults and children [15]. There is much previous work on the causes of health disparities among Black/African Americans, and others have speculated on which of these causes are related to the differential impact of COVID-19 [16] [17]. Thus, to tease out the impact on vulnerable groups, one needs data on other baseline health factors, such as obesity [18] [19], co-morbidities, age [20] [21] environmental exposures, transportation use and employment factors, including types of occupations [13] [22].

US Health Response and Forecasting. To control the spread of COVID-19, the US has implemented complex and regionally uneven community-level, non-pharmaceutical interventions, including travel restrictions, social distancing, and stay-at-home orders. These interventions aim to reduce the opportunities for person-to-person and fomite-to-person contact routes for transmission of the virus. Although these interventions have shown to mitigate the community spread in certain communities, the trend did not hold for all communities. Many counties experienced an uptick in cases after an initial decline. There are several reasons why certain communities continue to see a growing number of cases, including: 1) lifting shelter-in-place or other social distancing restrictions earlier than advised; 2) lax controls on gatherings that resulted in super-spreader events [23], and; 3) unknown affects of plausible seasonality that impacts viral transmission [24]. As such, complex epidemiological contexts have emerged in US communities. The complexity is a result of dynamic environmental factors constituting social and physical environments for US populations that impact an individual's risk for contracting COVID-19. Thus, to adequately control the spread of COVID-19, it is important to identify the most salient social and physical environmental factors within US communities, driving transmission and and effecting susceptibility.

Despite awareness that disease transmission is related to social and physical environmental factors, few studies have rigorously analyzed the underlying drivers of the dynamics of disease emergence for COVID-19 in the US. Though models accounting for the specific vulnerabilities of local populations have been proposed, only a few models exist that assess the importance county-level variation of such variables in fueling COVID-19 outbreaks [25, 26]. Altieri et al. [25], for example, use county-level data from similar sources to this paper, to create an ensemble forecasting model, dubbed "Combined Linear and Exponential Predictors" to predict death counts from COVID-19. Their goal is to curate a data repository that can be used to forecast exponential and sub-exponential cases weeks in advance in order to help nonprofit organization disseminate much needed personal protective devices and respirators to areas projected to have

higher mortality rates.

In our study, we use similar data sources to model cross-sectional COVID-19 outcomes. Our goal, rather than prediction, is to explore the relative importance of different types of social, physical and environmental factors on COVID-19 transmission and mortality. Hospital case and mortality data, and seropositive surveillance studies have shown there are subgroups of the population that are more susceptible to higher cases of morbidity and mortality. These include people older than 65 and and communities affected by racial disparities. We attempt in this paper to expand on previous studies of county level variation in COVID-19 (e.g., see [27]) by evaluating additional socio-environmental data to understand if these disparities have direct effects on COVID-19 outcomes or are indirect through additional risk factors, such as diabetes, food security, air pollution or access to health care. Likewise, no study has been carried out to determine which of these factors are most associated with COVID-19 outcomes while still controlling nonparametrically for all other factors.

Added Value of this Study. We aggregate 5 types of COVID-19 outcomes, (i) day of the first case in a county relative to the first case reported in the U.S. (Snohomish County, Washington), (ii) number of cases 25 days after the initial case in a county, (iii) number of all-cause deaths 100 days after the first case in a county, (iv) total number of cases in a county to-date after initial case and (v) total number of all-cause deaths in a county to-date after first case in a county. From many sources including the CDC, U.S. Census Bureau, USA facts, google mobility data, and others, we collect a large number of pertinent variables for COVID-19. Using ensemble machine learning (ML) we create models that make no assumptions on the distributions of the data, these models are thereby non-parametric and allow all degrees of interactions and distributions in order to make the best fit.

In each of these models we identify the most important variables in the model by removing the variable and measuring the difference in model risk (model error). We then make marginal predictions for the number of cases and mortalities

from COVID-19 when increasing or decreasing these top variables while controlling for all other factors. Confidence intervals (CIs), significance and robustness of findings are measured via bootstrapping the model. Additionally, we investigate the predicted number of cases and mortalities from our model when controlling for only variables outside the target variable category (i.e ethnicity, public transportation subcategories) and univariate predictions (not controlling for other variables). Given this approach, our contributions are, rather than predictive forecasting the number of cases, an approach to measure the relative importance of risk factors for COVID-19 from an environmental perspective.

There are many known risk factors measured from case hospitalization data, including diabetes and heart disease. As such, we hypothesize these factors will show high variable importance, specifically for case mortality. Additionally, we hypothesize that environmental dynamics, which increase exposure time to the virus (i.e., number of occupations in a county, public transit use), will also be strongly associated with COVID-19 cases. When such factors are identified, this information could also be used to update or improve the public health response to specifically target factors related to high case counts in order to further mitigate new cases or prevent a resurgence of cases. Finally, we estimate variable importance of the high dimensional assembled county features without constraining (and bias-inducing), using a combination of machine learning and intuitive substitution estimators. Lastly, we make both the data and methods used in this paper accessible to others, thereby providing open source access and enhancing the utility of our results. All code, data, and county variables used are available and outcome data are updated daily on GitHub.

2. Methods

Data Sources. For all outcomes and predictors we compiled publicly available data for US jurisdictions reported at the state-level (e.g. Google Mobility Data) and county-level, excluding Alaska, Hawaii, and Puerto Rico. Our final dataset includes county-level case counts, death counts and a wide variety of county-level

demographic, epidemiological, health, and environmental data used as predictors in our analysis. Our analysis was based on the cumulative confirmed cases and deaths of COVID-19 in US counties, starting on January 22 2020 (referred to as Day 1) until July 14 2020 (USA Facts). The sources for our county-level features were: USA Facts, Bureau of Economic Analysis, American Community Survey, Tiger/Line GeoDatabase, CDC Interactive Atlas of Heart Disease and Stroke, County Health Rankings, Centers of Medicare and Medicaid Services, National Centers for Environmental Information, CDC Vulnerability Index, Bureau of Labor Statistics, MIT Election Lab, Google Community Mobility Reports; a total of 12 different sources which were joined on county FIPS codes.

Outcomes. We use five COVID-19 outcome scenarios: 1) we transformed the case data into day of the first case after the first confirmed case on January 21 2020; 2) we determined the number of cumulative cases on the 25 days after the day of the first case for each county (i.e., day 25 of the outbreak in each county); 3) we used the number of all-cause deaths on day 100 after the day of the first case of the outbreak for each respective county up to July 14 2020; 4) we determine the total number of cases to date after the day of the first case for each county and likewise scenario five is the total number of all-cause mortalities to-date after the day of the first case for each county.

For each outcome variable excluding number of cases at day 25, we divide the counts by population size for each county to create per-capita COVID-19 case and all-cause mortality outcome. This was done to remove population size bias. Likewise, all predictor variables measured in counts were also standardized by population size. All-cause deaths were used rather than reported fatalities due to COVID-19 for several reasons related to unreliable case data, differences in testing, and co-morbidities between COVID-19 and other fatal acute diseases. By using all-cause deaths measured since day of first case reported in a county, we hope to get a better estimate on the impact of COVID-19 on mortality. As discussed, our predictors cover a wide scope, **Table 1** gives a review of the data sources and variables collected from each source. **Table S1** in the

supplementary section gives more details of this process.

Predictors. Data on our predictor variables include demographics, health resource availability, health risk factors, social vulnerability, and other COVID-19-related information. The predictor variables used are collected from different sources including Altieri et al. [28] and Killeen et al. [29]. Because the aims of this paper are not purely predictive, but focus on understanding the relative impact each variable has on COVID-19 outcomes, our data curation process is different when compared to these two papers. We aggregate data from different stratified variables to create an overall public transportation use feature. Likewise for social vulnerability scores, we attempt to include core variable that represent a specific type of risk feature. For example, given our interest are variable importance measures and marginal predictions, if we included both the aggregated CDC vulnerability index with many features collected from other sources that are proxy measures for this index (percent non-English speaking, poverty levels etc.) then findings for aggregate vulnerability index would be conservative given these other variables are also included in the model. That being said, given the large overlap and interactions of all variables collected, it it likely that all variable importance estimates are conservative. In our curation process, only unstratified variables are used (not stratified by age or sex), we also create sub-categories for variables (ethnicity, geography, disease prevention, etc.) These sub-categories we use in later analysis to explore possible over correction of the model for a respective target variable.

Briefly, some predictor variables include proportions of individuals by poverty level, gender, age distribution, race distribution, household income, healthcare access, occupation type, and so on, these were collected from USA Facts and the Census Bureau. Airport data, including, distance from county polygon centroid to airports were calculated from the Federal Aviation Administration. The 2020 county health rankings and Center for Medicaid and Medicare services were used to gather information on a range of health data including smoking, diabetes, obesity, air pollution and many other physical and mental health metrics. Pre-

cipitation by month was gathered from the National Oceanic and Atmospheric Association. Vulnerability index scores from each theme (described in **Table S1**) were aggregated from the Center of Disease Control. In total, over 150 predictor variables were gathered before curation. This curated data along with all relevant code, documentation and results are provided on our GitHub page.

Data Cleaning and Curation. The data curation process is described in more detail in the supplementary information alongside the data dictionary. All resulting data was numerical (no factor variables). In addition, we screen out any variables with more than 70% missing values. Similarly, we removed variables with close to zero variance. Variables that were nearly perfectly correlating were removed (Pearson correlation = 0.95). Missing data in this cleaned dataset were imputed with the mean. For Google mobility data, we scraped data from the published mobility trend reports from Feb 16 2020 to March 29 2020. These data represent the general increase or decrease in movement to the respective destination (grocery stores, parks etc.) compared to baseline (pre-pandemic period). To create an aggregate score representing the mobility trend for each movement category for each county, we use the slope from linear regression to measure this trend over time. The slope for each movement category was included in our Super Learner models.

Exploratory Analysis. To graphically represent how our feature data are related to one another, and likewise how counties are related to one another through these variables, we use unsupervised hierarchical agglomerative clustering of both county features and counties. We present the results of this clustering as a heatmap. Our first goal of this method is to understand if there are counties that have a trend for early first case reported, high COVID-19 case rates at day 25, and COVID-19 case rates to-date, as well as high all-cause mortalities at day 100 and to-date. If this trend was seen, we next wanted to investigate what variables were 'highly expressed' in these counties. As such, all feature data was z-score standardized. We then took the quantiles of each outcome to create factor dummy variables that can be plotted alongside clustering of counties. Clustering

was done for both counties and county features and reordered accordingly using Euclidean distances. As this is an unsupervised approach, outcome data are not included in this machine learning method but are simply plotted alongside the clustering results to visually identify correspondences. Groups of county features that were found to be associated with groups of COVID-19 outcomes are presented.

Machine Learning Pipeline. Although our ultimate goal is not to use our final models for forecasting and prediction, one still needs to estimate a regression model in order to determine our measures of variable importance. By estimating this model as accurately as possible, one can better estimate the variable importance measures that rely on the prediction model. As such, instead of choosing one machine learning (ML) algorithm to model county features for each outcome, we used an ensemble approach (Super Learner) to fit a prediction function for each of our outcomes. The Super Learner combines the predictive probabilities of COVID-19 outcomes across many ML algorithms. The Super Learner finds the optimal combination of a collection of algorithms by minimizing the cross-validated risk [30, 31]. This method is an improvement over methods using only one ML algorithm because no one algorithm is universally optimal. The Super Learner has been shown in theory to be at least as good as the best performing algorithm in the ensemble and often times performs considerably better [32]. Given the high-dimensionality and complex relationships of the county data and the fact that there are no known distributions for these relationships, we chose a wide range of algorithms for the ensemble in order to optimize performance. For COVID-19 cases at day 25 and to-date per-capita rates and all-cause mortalities at day 100 and total to-date per capita, we use a large number of linear Gaussian based algorithms including conditional mean (control algorithm) simple generalized linear model, a series of penalized regressions setting alpha at levels to create ridge regression, lasso regression, and elastic net regression. Similarly, we use a number of gradient boosted decision trees that differed in depth and shrinkage. Because these algorithms require

hyper-parameter tuning for optimal performance, we create a grid of all possible hyper-parameters and choose algorithms across this grid for inclusion in the ensemble. The same procedure was applied for day-of-first-case in a county relative to-day-of-first case in the U.S. Instead of using Gaussian algorithms, however, custom learners were made for the Super Learner environment that model Poisson outcome data. The same parameters were chosen for this set of learners to create the Poisson ensemble. To address possible over-fitting and to get cross-validated risks for each algorithm in the ensemble sets, five-fold cross-validation was used for internal SL cross-validation both to build optimal models with each classifier and to determine optimal weighting across classifiers in the ensemble.

Unpacking the black box. The algorithm used to create our predictor given covariates has desirable optimality properties, being asymptotically guaranteed to have a fit as good as any of the candidate members (the "oracle property"), with no risk of overfitting. If a library of both smooth (e.g., parametric models) and flexible, nonparametric learners, then one can find it hard to outperform [33, 31]. However, the result is a black box that creates predictions as a complex ensemble of different learners, some having their own internal variable selection process and model selection framework. Thus, the resulting black box needs to be intelligently queried to estimate the independent impact of the various predictors used in the model. We do so in two ways. One, is using a straightforward leave-one-variable out method and re-examining the change in prediction accuracy. However, this provides no information about the direction of the impact, which is why we follow with a query inspired by causal inference methods. In that case, we use the model to forward model situations where we change the distribution of predictors across the counties in sequential fashion and then calculate the marginal predicted counts (so called substitution estimators, or G-compuation - [34, 35]). The combination of these two versions of non-parametric variable importance measures provide both the importance of the variable (or sub-category of variable) to the resulting predictor as well as an intuitive mea-

sure of the adjusted association of single variables.

Variable Importance. We built Super Learners from the same county-level data for each of the COVID-19 outcomes. To measure variable importance in each model, we take the fitted model and make predictions using all county features and measure the model risk (average in-squared differences in model prediction versus truth, or mean-squared error). We then scramble (sets) of variables and re-do the prediction and derive the new MSE. The plots (**Figures 2 and 3**) show the resulting ranked list of variables (most to least change in the MSE by scrambling). We use a risk-ratio (MSE-ratio) for each variable to measure its relative importance in the model for each outcome. The risk-ratio is the risk in the model without the respective variable (numerator) over the risk when the variable is included in the model (denominator). As such, a risk-ratio of 1.5 indicates that the model MSE rises by 50% when the variable is scrambled while controlling for all other variable affects. We use a similar approach to measure the variable sub-category importance on each outcome. Each variable was given a sub-category (described in supplementary material) resulting in a total of 15 categories. Blocks of variables in each category were scrambled and the model risk-ratio measured to attain information on category importance.

Marginal Predictions. Given that we fit a black-box to derive our prediction models, we have to unpack the black-box to understand what it implies about the adjusted relationship of the important variables to the outcomes. We thus use substitution methods to evaluate the predicted change in the mean if county characteristics are changed. We examine how the mean outcomes would be predicted to change if the inputs of the specific variable of interest are modified, such as reducing a variable in some equivalent way across counties. Other modifications of the inputs could be used to examine these variable importance plots, but we looked at % changes in the variable across counties. Using these models we then make marginal predictions on the predicted number of COVID-19 cases and all-cause mortalities when increasing (or decreasing) the top variables found by the variable importance procedure.

Suppose a particular observation $O_i = (W_i, A_i, Y_i)$ in county i, $i = 1, \ldots, n$, depends on explanatory variable A_i, other adjustment covariates W_i, and outcome Y_i, all in county i. If we wish to generate an estimate that characterizes the association of Y across all counties with A adjusting for W, but does not rely on a linear approximation, then we do this through plotting the estimate

$$\phi(\pi) = E_{A,W}\{E(Y|A = \max(\tilde{A}, (1-\pi) * A), W)\} \tag{1}$$

as a function of π, where $\tilde{A}$ is the minimum observed value of A across all counties in the data and π is interpreted as the proportional reduction in the county-specific value of the variable. To avoid extrapolation, we truncate A at the minimum observed value for the variable among counties. In essence, we exam the resulting predicted mean outcome across all counties as though the particular variable were reduced by $\pi\%$ in all counties. This plot then provides a relevant function of the importance of the variable to the outcome.

Under several strong assumptions, including that the other covariates (the W, being either all other predictors or all but the ones in same sub-category) contain all the confounding information, sufficient experimentation (no positivity violations [[36, 35**?**]), and independence of outcomes across counties, one could interpret (1) as identifying the marginal mean had, contrary to fact, all counties been set at the stochastic value, $\max(\tilde{A}, (1-\pi) * A)$. Of course, we have time-ordering among the covariates that we can assert, so we treat these plots as a nonparametric form of association measure.

To derive inference, we use the nonparametric bootstrap (randomly sampling counties with replacement) for each π reduced mean. The procedure is as follows: (i) fit each outcome using the aforementioned set of learners, (ii) iteratively remove variables and determine risk-ratios, (iii) set the variable with the highest risk-ratio as the target variable for marginal predictions, (iv) for each percentage from 0-1.0 at 0.10 intervals (a) resample the county data with replacement, (b) refit the Super Learner with this resampled data, (c) reduce or increase the target variable for the respective percentage, and (d) predict the expected number of cases or mortalities with this new fit. Here, in step

(iv) we bootstrap this procedure 1000 times by resampling, refitting, and making marginal predictions, in order to create confidence intervals (CI) at each percent change to the target variable.

Note that we use three different models to estimate the regression $E(Y|A,W)$ depicted in (1): fully adjusted, adjusted only variables not in the sub-category of the variable of interest, and unadjusted. To evaluate the performance of our model, we compare each marginal prediction to that of a univariate model of the target variable (found by variable importance) for each outcome. Here, a general additive model was retrained at each iteration of the bootstrap for each reduction in the target variable and cases/mortalities were predicted through this univariate model. Likewise, to investigate possible over-corrections of our Super Learners, we also remove similar variables that may have strong multicollinearity with the target variable. This was done by removing variables in the target variable sub-category and training a Super Learner on this set of covariates through the bootstrap. Results are presented as line plots and the actual observed average or sum for each outcome are plotted as horizontal lines for comparing actual outcomes to model predictions.

All data aggregation, curation, cleaning, exploratory analysis, ML pipleine, and marginal predictions were performed in R [37]. All coding scripts are available on our GitHub page for open-access and use: **(https://github.com/blind-contours/Getz_Hubbard_Covid_Ensemble_ML_Public.git)**.

3. Results

COVID-19 Outcomes and County Feature Distributions. There are 3,142 counties in the U.S. After our data cleaning and curation process, there were 2,620 counties included in the analysis and 101 county-level features. As such, our analysis covers 83% of the U.S. population as represented by counties. In these counties, as of July 15, 2020 there were 243,065 cases at day 25, 2,531,134 cases to-date, 53,018 all-cause deaths at day 100, 111,991 all-cause deaths to-date, and the average number of days to the first case in a county relative to the first

case in Snohomish County, Washington was 68 days. A description of the variables used and their sources are given in supplementary materials (**Table S1**). A breakdown of these features with the respective mean, standard deviation, and range of values are also given in the supplementary materials (**Table S2**).

Exploratory. The heatmap for exploring the patterns in these data are given in **Figure 1**. The marked section of the heatmap show an outcome trend for 1: First quantile of day of first case (Q1 = earlier days of first case): 2. highest total deaths to-date (Q4): 3. highest deaths at day 100 (Q3): 4. highest total COVID-19 cases to-date (Q4): and 4. highest COVID-19 cases at day 25 (Q4). The distribution of the outcome quantiles across the county dendrogram groups are provided in the supplementary materials (**Table S8**). The cluster of counties with most severe outcomes is marked on the left in **Figure 1**. These patterns indicate there are counties that cluster together based on similar characteristics and these counties correspond with an earlier first case in the county and higher COVID-19 case and mortality rates. For a breakdown of the number of counties in each state in this cluster see the supplementary material; briefly, however, the states with the highest number of counties in this cluster with highest outcomes are 1. Virginia (36), 2. Florida (33), and 3. Texas (30). The highest column values (red and orange pixels in the heat map) in this highest county row cluster occur occur in branches 3-16 and branches 46-53 of the column-wise dendrogram. A full list for each cluster is given in the plot but the highest county features for this subset were: 1. obesity, 2. sexually transmitted diseases, 3. income inequality, 4. food environment index, 5. CDC limited English scores, 6. latitude, 7. poverty income ratio, 8. GDP, 9. preventable hospital stays, 10. arthritis, 11. asthma, and 12. ischemic heart disease.

Super Learner. As discussed, we use cross-validation to generate a coefficient that defines the weight for a respective learner in the ensemble. This procedure is done for each outcome and the same learners are used for each outcome (outside of day of first case where Poisson outcomes were defined). **Tables**

S3-S7 in supplementary give a detailed breakdown of how each algorithm was used in the Super Learner, the risk of the respective algorithm and the overall risk of the Super Learner. Overall, our Super Learners were able to achieve good fits by utilizing multiple algorithms. Results show that for each outcome the Super Learners were largely and consistently built from multiple elastic net models, multiple xgboost models, and random forest. **Table 2** shows resulting risk for each Super Learner for each outcome. Because the learners were fit to the per-capita standardized outcome data, we multiply each risk by the total population in the dataset to get absolute error based on total numbers of cases or mortalities. We also calculate the r-squared for each Super Learner to show variance explained by each model.

Variable Importance. The top variable categories for each outcome were: 1) Day of first case in a county: demography; 2) COVID-19 cases at day 25: ethnicity, transit and preventable disease; 3) total COVID-19 cases to-date: ethnicity and preventable disease; 4) mortalities at day 100: transit and ethnicity; 5) total mortalities to-date: ethnicity and transit. The top individual variables across all the models were: overall population of a county, CDC vulnerability scores for minority and limited English, public transportation use, and proportion of Black/African American individuals in a county. To visualize results, we present the risk-ratio (RR) results for each COVID-19 outcome collectively in **Figure 2** and **Figure 3** as a series of dot-plots (RR threshold at 1.01). Based on these figures, it can be seen that for day of the first case in a county the total population (RR: 1.38) was the most important variable. For per-capita COVID-19 cases at day 25, the top variable is the CDC minority score (RR: 1.04). For per-capita COVID-19 cases to-date, the CDC's score for limited English speaking (1.40) the CDC's score for minority populations (1.17) and proportion Black/African-American individuals in a county (RR: 1.12) were the top variables. For per-capita all-cause mortalities at day 100, proportion taking public transportation (RR: 1.14) and proportion Black/African-American individuals (RR: 1.07) were the top variables. For per-capita all-cause mortalities to-date,

proportion Black/African-American individuals (RR: 1.08) and proportion taking public transportation (RR: 1.03) were the top county features.

Marginal Prediction Results. For the day of first case, population size was clearly most important predictor. For examining the association of cases and deaths of COVID-19 we choose two outcomes (total deaths and cases by July 14, 2020) and three of the most consistently important variables: two related to demographic features of the population (CDC Minority Score and Proportion of Black/African-American individuals) and one related to transportation (metric of public transportation use). We estimate the relationship of proportional reductions in each of the predictor variables on the marginal outcome (using the substitution estimator of (1)) based upon a machine learning fit when controlling for: 1) all other variables (including possibly strongly collinear variables within the same sub-category), 2) only variables outside the target variable sub-category, and 3) nothing (unadjusted). The latter estimator of $E(Y|A)$ is based upon a smooth regression of the outcome versus the continuous covariates, specifically using general additive model with identify link (GAM;[38]).

Figure 4 shows the predicted average day of first case across all counties for each proportion of population size reduced across the counties. Generally, smaller county populations are predicted to have a delay relative to counties with higher population sizes, *all other factors in the model staying constant* (the current average being 68 days after the initial U.S. case in Washington) for all three estimators, with somewhat larger effects in the adjusted models. For all models, a 0.50 proportional reduction in population size across the counties suggest a delay of around 2 to 3 days.

Figure 5 shows plotted results for two outcomes (total counts and deaths), both versus proportion reductions in the three predictor variables. Proportional decreases in CDC Minority Score is associated with a decrease in COVID-19 cases and death. The large attenuation of the relationship in the adjusted models suggest strong confounding by the other covariates, where in the fully (perhaps over) curve approaches the null line. Using the curves not adjusting for

other variables in the sub-category, the association suggest a 0.5 proportional reduction in the score would predict a reduction of 750,000 cases (out of around 2,400,000 total number by July 14) and approximately 10,000 deaths (out of around 115,000). Note that, particularly with deaths, the unadjusted curve is quite different from the actual number (blue line is substantially below the horizontal black line at the point of no intervention), this should be equal to that value if the model fits the data well. This is due to a few counties with extreme large counts (of both death and cases), which are poorly predicted by the bivariate smooths resulting in large positive residuals. Thus, when one estimates counts based upon reductions in the variable of interest (CDC Minority Score in this case), you get a prediction of the count that underestimates the true count. Note, by substantially reducing residual variation, the adjusted curves tend to be much closer to the observed count at 0 reduction.

Reducing public transportation suggests significant reduction in deaths, but little impact on case counts. For cases, we again have a poor fit of the bivariate smooth (unadjusted), along with the suggestion of significant confounding by other covariates. For deaths, a reduction of 50% suggest a reduction in deaths of 10,000, but the fact that the intercept for both adjusted curves (and unadjusted) is less than the observed count suggest again the influence of outliers. Bootstrapped linear regression of the marginal predictions showed that for a 10% reduction in public transit use, total deaths reduce by 2012 (95% CI [1972, 2356]).

Finally, for reductions in the proportion of Black/African-American individuals, there appears to be quite different estimates between the adjusted curves of COVID-19 counts, suggesting that variables in the sub-category of this variable create the possibility of over-adjustment in the full model. For deaths, the curves are nearly identical and imply that a reduction of the disparity between Blacks and Whites in 50% of the population (one way to interpret an actual reduction in this variable) suggest a reduction of about 9,000 total deaths. Likewise bootstrapped linear regression of these predictions show a 10% increase in the proportion of Black/African-American individuals in a county increases to-

tal deaths to date by 2067 (95% CI [1189, 2654]).

4. Discussion

In this paper, we took a semi-parametric (machine learning) approach to evaluating a wide range of county-level features which may impact the spread, number of cases, and deaths of COVID-19 in the U.S. Our contributions are the following: (i) curating an open-source data repository that includes variables from many sources, categorized into sub-groups and filtered such that strongly collinear variables are removed for statistical analysis; (ii) demonstrating the use of these variables in ensemble machine learning to build 5 Super Learners for each COVID-19 outcome measure; (iii) evaluating the features to identify the top variables that influence each outcome; (iv) adjusting the top variables in our model to make marginal predictions for each COVID-19 outcome, while controlling for all other factors to establish the strength and directionality of the relationships; (v) constructing confidence intervals around all effects via bootstrapping to evaluate significant trends from baseline and between modeling approaches. Overall, using 101 county-level features our models show very good fits to the outcomes (all observed outcomes were within model confidence intervals apart from mortality which were slightly outside our CIs at baseline).

These fits establish that our models are able to accurately predict each outcome given the county-level feature variables. Our variable importance measures for each model fit generally show a trend that the total population drives day of first case in a county and the proportion of Black/African-American individuals in a county and CDC minority scores are most important independent contributions of COVID-19 cases and deaths as of mid July.

Causal inference pertaining to the individual relationships of these variables on each outcome is speculative at best given that these study variables are ecological and also are a static snap-shot of county variables collected before the pandemic hit the U.S. However, the general trend in these results seem to represent what is anecdotally or univariately reported as the U.S. faces this

continuing pandemic. That is, for day of first case in a county, the total population as the most important variable makes sense given the larger the population the higher the probability of someone being infected traveling to the respective county. Likewise, CDC minority scores and Black/African-American individuals are correlated with reports that suggest that minority populations and People of Color are disproportionately impacted by COVID-19 [13] [15]. In addition, we also show a significant potential impact of baseline public transportation use and mortality [39][40] [27]. This could indicate that there is higher probability of exposure in counties where travel on public transportation leads longer and closer duration contains. This may also lead to higher infectious doses that may possibly increase severity of infection and consequent mortality, a phenomenon thought to be the case for influenza [41].

Our finding that counties with larger CDC minority population measures have higher COVID-19 outcomes, even when controlling for one-hundred other county-level variables (Table 1 and S1), show the value of such measures when trying to determine the impact of risk level based on social factors for those disproportionately impacted by COVID-19. Additional measures, however, are necessary to understand the reasons for this. For instance, although our models adjust for income, access to health, and occupation types, our data are limited to reported factors that may not account for systematic or institutional levels of cultural/societal factors placing individuals at risk. Such factors may confound or modify others for which have been adjusted and may place certain individuals at greater risk for SARS-CoV-2 infection. The main difference between our findings and those reported to date is that our analysis controls for many other possible mediating factors (e,g., access to health-care, smoking, diabetes, heart-disease, and food security).

5. Conclusion

The goal of our study is to identify the most salient factors that put populations at risk for COVID-19, thereby providing some guidance to individu-

als making difficult policy decisions at this critical time to quell the evolving pandemic. Specifically, racial composition of counties and intensity of public transportation use therein seem to be the most important risks factors for both the initial rapid growth and subsequent high incidence, and also help explain variations in mortality rates across counties. More work, however, is needed to establish causal rather than purely statistical relationships. Future work with detailed individual data will be important for getting more robust estimates of the individual impact of the factors examined. Whether causal or statistical, these results should be taken into account when developing policies for lifting restrictions. Additionally, as efforts continue to disseminate services and funding, and to roll out vaccination programs, once effective vaccines have been developed, consideration of these factors will facilitate the efficacious allocation of resources to the benefit of the US population as a whole.

Tables

Source	N Var.	Var. Examples
USAFacts	6	COVID-19 Outcome Data, Population
Bureau of Economic Analysis (BEA)	1	GDP
5-Year American Community Survey (ACS), 2014 - 2018	14	County percentages by Sex and Ethnicity, Employment, Household Income, use of Public Transportation
TIGER/Line Geodatabases	7	Latitude, Longtitude, Land Area
TIGER/Line Geodatabases; Federal Aviation Administration (FAA)		Distance to Airports
Interactive Atlas of Heart Disease and Stroke (2014-2016)	4	Number of Hospitals, Stroke, Access to Parks
County Health Rankings & Roadmaps	21	Life Expectancy, Smoking, Obesity,, Food Access, Mental Health, Physicians, Houshold Overcrowding etc.
Centers for Medicare & Medicaid Services (CMS)	15	Druge Abuse, Hypertension, Hyperlipidemia, Osteoporosis, etc.
National Centers for Environmental Information	1	Precipitation
CDC's Social Vulnerability Index (SVI)	11	Percentile over 65 or under 17, Minority Scores, Limited English, Low Income Housing Estimates, Number Institutionalized
Quarterly Census of Employment and Wages	14	Labor force types, farming/mining, private industry, education/healthcare etc.
MIT election lab	1	Calculated Proportion Voted Republican 2016
Google	6	Google mobility to location type, Residence, Grocery etc.

Table 1: Number of variables used from respective sources with some examples given, complete list with distributions given in supplementary material

COVID-19 Outcome	Model Risk (per capita)	Model Risk (counts)	R-squared
Day of First Case	NA	159.58	0.75
COVID-19 Cases at Day 25	4.22 e-05	10539.64	0.59
Total COVID-19 Cases to-date	5.22 e-05	13053.75	0.87
All-Cause Death at Day 100	2.80 e-08	7.00	0.57
All-Cause Death at to-date	1.42 e-07	35.52	0.59

Table 2: Cross Validated Super Learner Risk Across COVID-19 Outcomes

6. Figures

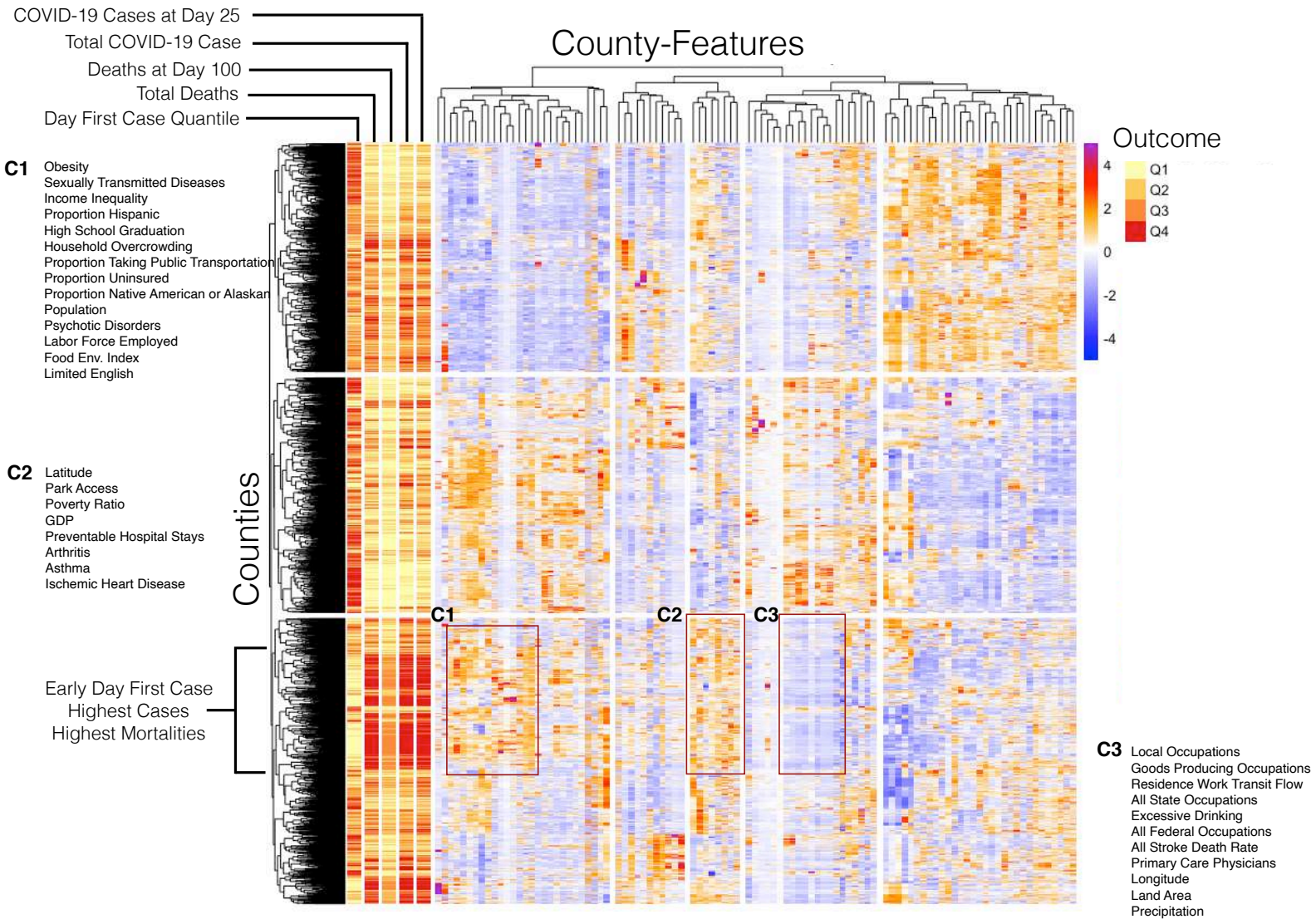

Figure 1: COVID-19 heatmap visualization of the distribution of county-level data

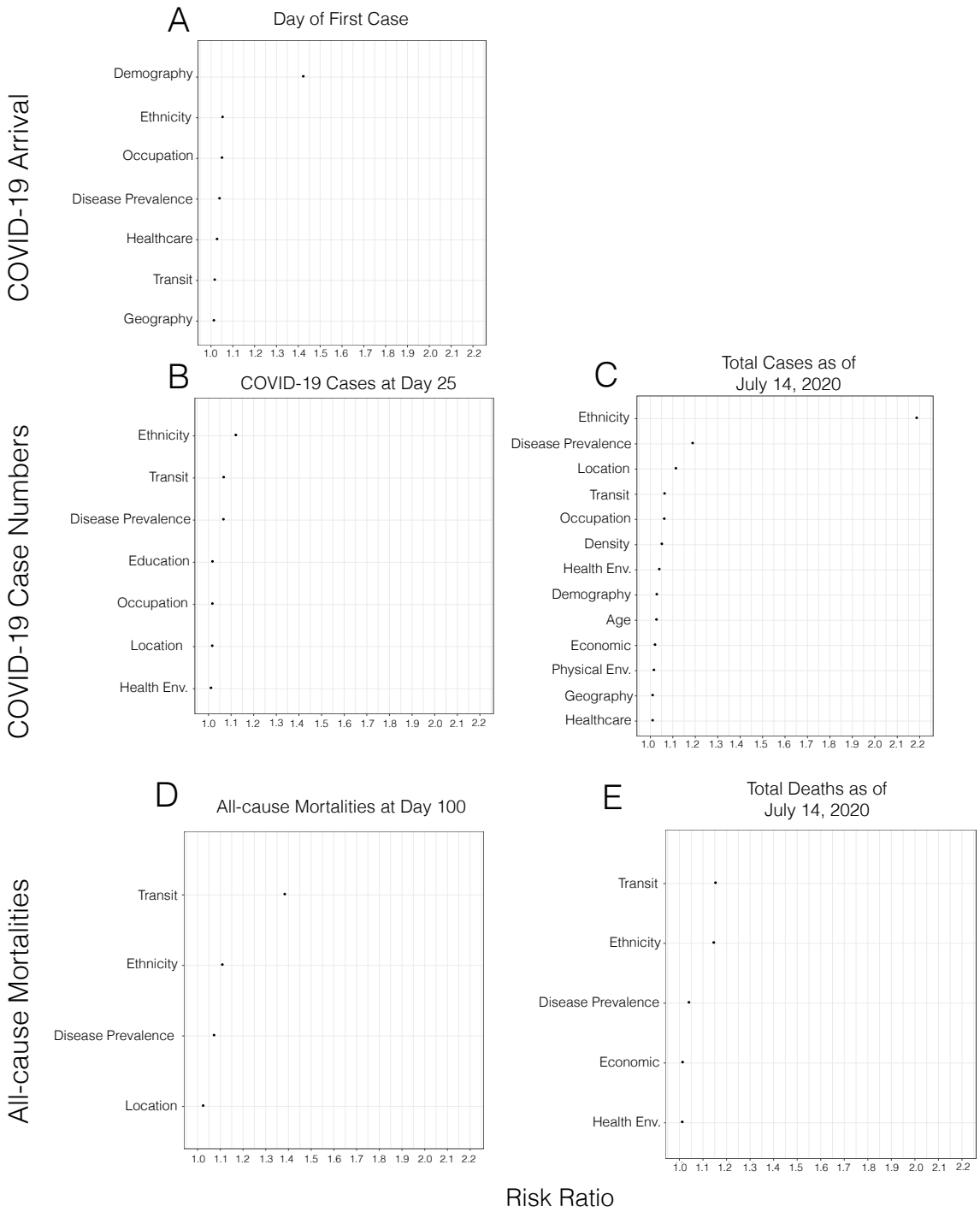

Figure 2: Variable importance as indicated by the relative increase of mean-squared error when the block of variables is removed

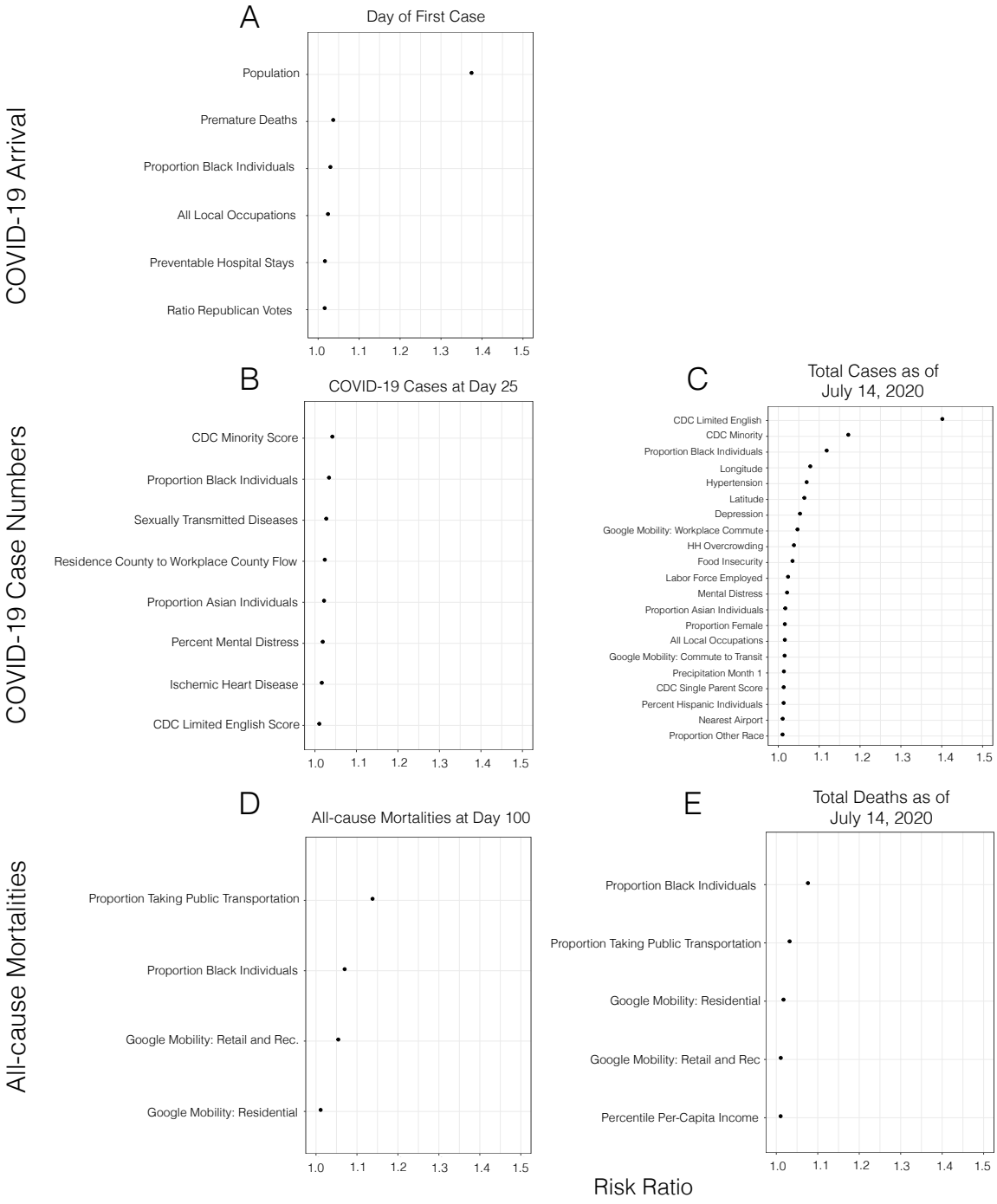

Figure 3: Variable importance as indicated by the relative increase of mean-squared error when a single variable is removed

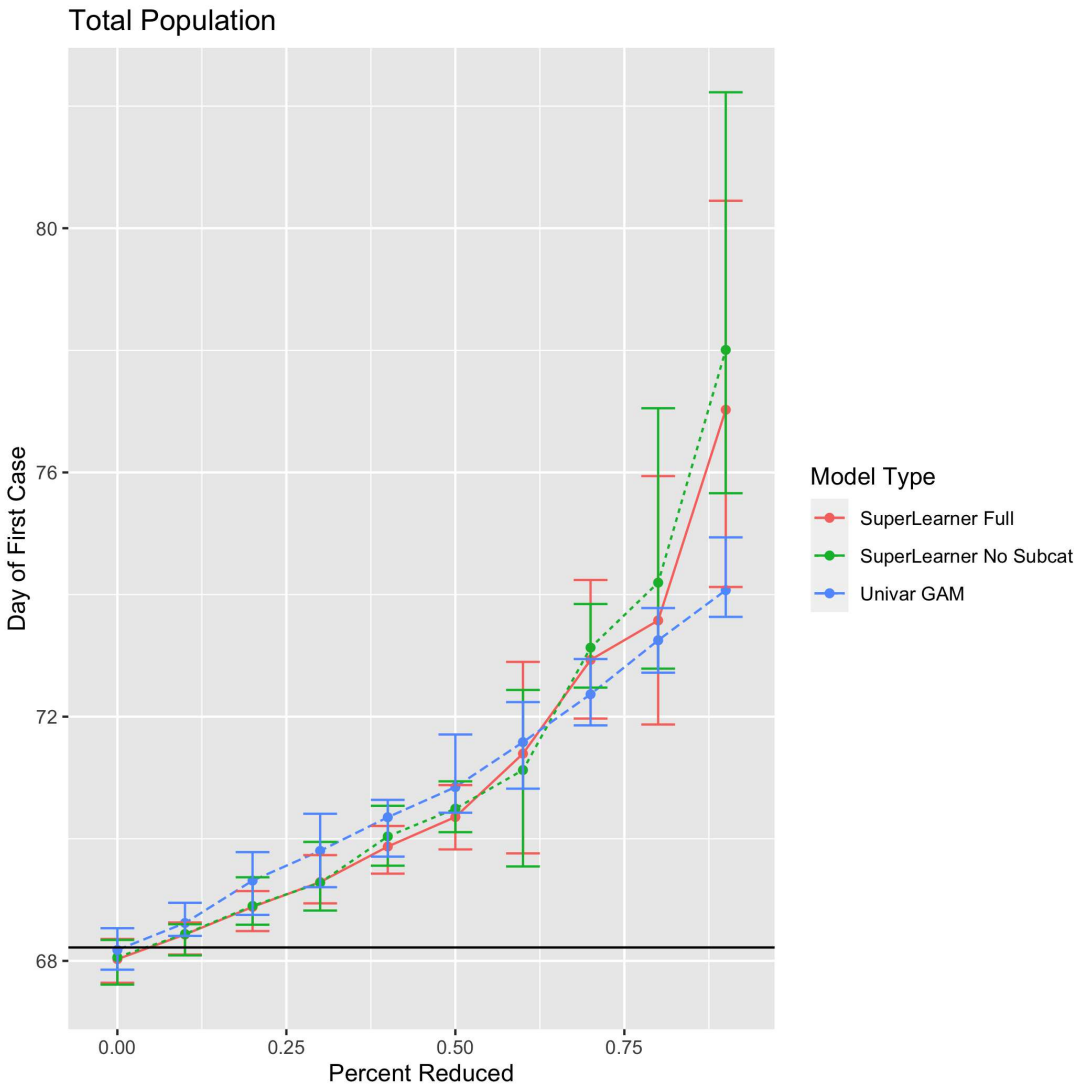

Figure 4: Marginal predictions of day of first case (relative to index time) for different proportional reductions of total population size for models adjusting for all other covariates, only covariates not in sub-category (see supplement table 1) and unadjusted.

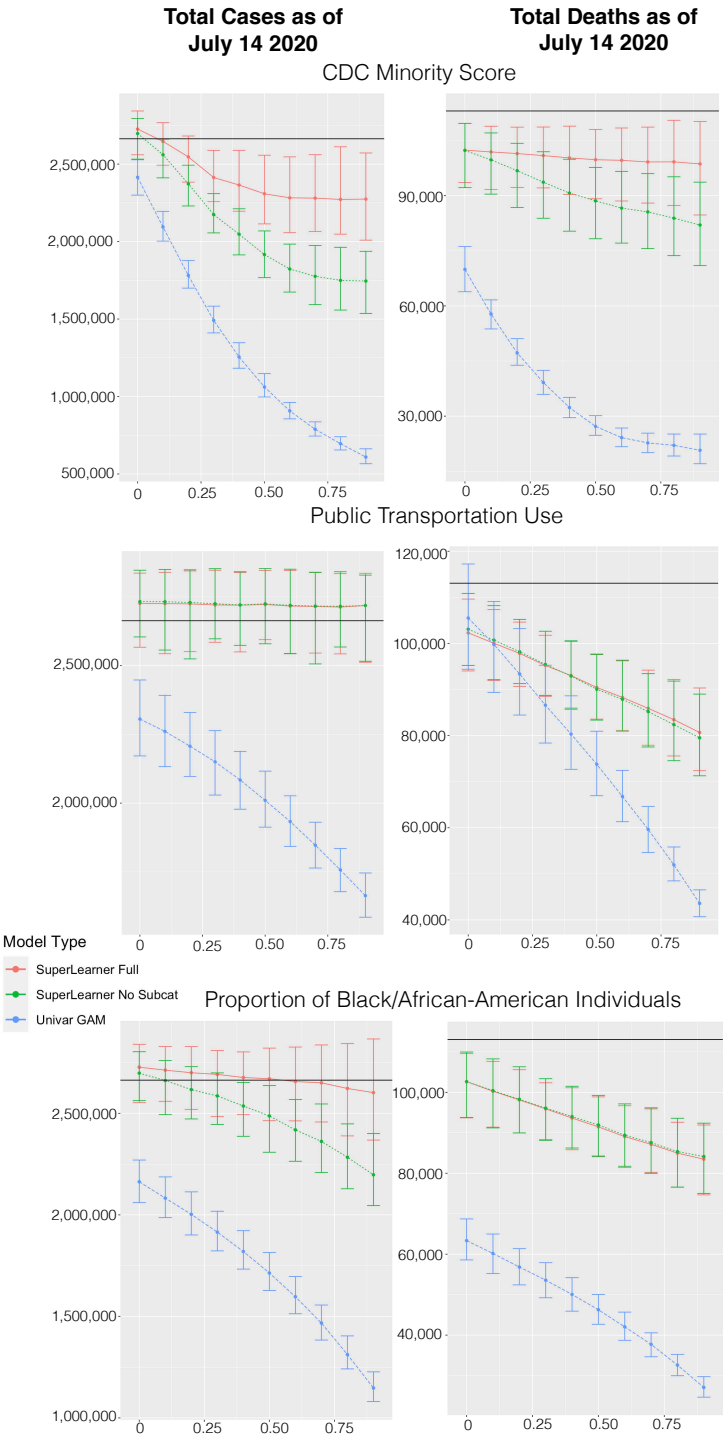

Figure 5: Marginal predictions of total cases and deaths by July 14, 2020) for three of the most consistently important variables in predicting the count outcomes: CDC minority score, proportion of Black individuals and a metric of public transportation use. X-axis is different proportional reductions of each of the three predictors, the Y-axis is the marginal predicted total counts for models adjusting for all other covariates, only covariates not in sub-category (see supplement table 1) and unadjusted.

References

[1] N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu, et al., A novel coronavirus from patients with pneumonia in china, 2019, New England Journal of Medicine.

[2] F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei, et al., A new coronavirus associated with human respiratory disease in china, Nature 579 (7798) (2020) 265–269.

[3] H. A. Rothan, S. N. Byrareddy, The epidemiology and pathogenesis of coronavirus disease (covid-19) outbreak, Journal of autoimmunity (2020) 102433.

[4] Z.-W. Ye, S. Yuan, K.-S. Yuen, S.-Y. Fung, C.-P. Chan, D.-Y. Jin, Zoonotic origins of human coronaviruses, International journal of biological sciences 16 (10) (2020) 1686.

[5] P. Zhou, X.-L. Yang, X.-G. Wang, B. Hu, L. Zhang, W. Zhang, H.-R. Si, Y. Zhu, B. Li, C.-L. Huang, et al., A pneumonia outbreak associated with a new coronavirus of probable bat origin, nature 579 (7798) (2020) 270–273.

[6] R. Lu, X. Zhao, J. Li, P. Niu, B. Yang, H. Wu, W. Wang, H. Song, B. Huang, N. Zhu, et al., Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding, The Lancet 395 (10224) (2020) 565–574.

[7] WHO, Who timeline - covid-19, `http://precog.iiitd.edu.in/people/anupama` (April 2020).

[8] CDC, Coronavirus (covid-19), `https://www.cdc.gov/coronavirus/2019-ncov/index.html` (2020).

[9] M. L. Holshue, C. DeBolt, S. Lindquist, K. H. Lofy, J. Wiesman, H. Bruce, C. Spitters, K. Ericson, S. Wilkerson, A. Tural, et al., First case of 2019 novel coronavirus in the united states, New England Journal of Medicine.

[10] I. Ghinai, T. D. McPherson, J. C. Hunter, H. L. Kirking, D. Christiansen, K. Joshi, R. Rubin, S. Morales-Estrada, S. R. Black, M. Pacilli, et al., First known person-to-person transmission of severe acute respiratory syndrome coronavirus 2 (sars-cov-2) in the usa, The Lancet.

[11] R. M. Burke, Active monitoring of persons exposed to patients with confirmed covid-19—united states, january–february 2020, MMWR. Morbidity and mortality weekly report 69.

[12] C. W. Yancy, Covid-19 and african americans, JAMA. Published online.

[13] S. S. Coughlin, J. X. Moore, V. George, J. A. Johnson, J. Hobbs, Covid-19 among african americans: From preliminary epidemiological surveillance data to public health action, American Journal of Public Health (2020) 1157–1159.

[14] S. Kodidela, A. Kumar, K. Gerth, S. Kumar, C. Walker, Lessons learned from health disparities among african ameri-cans in the hiv epidemic: What to expect for covid-19 and potential approaches to mitigate health disparity, Emerg Infect Dis Diag J: EIDDJ-100021 2 (03).

[15] S. Bandi, M. Z. Nevid, M. Mahdavinia, African american children are at higher risk for covid-19 infection, Pediatric Allergy and Immunology.

[16] D. B. G. Tai, A. Shah, C. A. Doubeni, I. G. Sia, M. L. Wieland, The disproportionate impact of covid-19 on racial and ethnic minorities in the united states, Clinical Infectious Diseases.

[17] E. E. Wiemers, S. Abrahams, M. AlFakhri, V. J. Hotz, R. F. Schoeni, J. A. Seltzer, Disparities in vulnerability to severe complications from covid-19 in the united states, Tech. rep., National Bureau of Economic Research (2020).

[18] N. N. Pettit, E. L. MacKenzie, J. Ridgway, K. Pursell, D. Ash, B. Patel, M. T. Pho, Obesity is associated with increased risk for mortality among hospitalized patients with covid-19, Obesity.

[19] J. Lighter, M. Phillips, S. Hochman, S. Sterling, D. Johnson, F. Francois, A. Stachel, Obesity in patients younger than 60 years is a risk factor for covid-19 hospital admission, Clinical Infectious Diseases.

[20] C. Covid, C. COVID, C. COVID, N. Chow, K. Fleming-Dutra, R. Gierke, A. Hall, M. Hughes, T. Pilishvili, M. Ritchey, et al., Preliminary estimates of the prevalence of selected underlying health conditions among patients with coronavirus disease 2019—united states, february 12–march 28, 2020, Morbidity and Mortality Weekly Report 69 (13) (2020) 382.

[21] M. Lipsitch, D. L. Swerdlow, L. Finelli, Defining the epidemiology of covid-19—studies needed, New England journal of medicine 382 (13) (2020) 1194–1196.

[22] M. G. Baker, T. K. Peckham, N. S. Seixas, Estimating the burden of united states workers exposed to infection or disease: a key factor in containing risk of covid-19 infection, PLoS One 15 (4) (2020) e0232452.

[23] J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, W. M. Getz, Superspreading and the effect of individual variation on disease emergence, Nature 438 (7066) (2005) 355.

[24] M. M. Sajadi, P. Habibzadeh, A. Vintzileos, S. Shokouhi, F. Miralles-Wilhelm, A. Amoroso, Temperature and latitude analysis to predict potential spread and seasonality for covid-19, Available at SSRN 3550308.

[25] N. Altieri, R. Barter, J. Duncan, R. Dwivedi, K. Kumbier, X. Li, R. Netzorg, B. Park, C. Singh, Y. S. Tan, et al., Curating a covid-19 data repository and forecasting county-level death counts in the united states, arXiv preprint arXiv:2005.07882.

[26] J. O. Ferstad, A. J. Gu, R. Y. Lee, I. Thapa, A. Y. Shin, J. A. Salomon, P. Glynn, N. H. Shah, A. Milstein, K. Schulman, et al., A model to forecast regional demand for covid-19 related hospital beds, medRxiv.

[27] K. Desmet, R. Wacziarg, Understanding spatial variation in covid-19 across the united states, Tech. rep., National Bureau of Economic Research (2020).

[28] N. Altieri, R. L. Barter, J. Duncan, R. Dwivedi, K. Kumbier, X. Li, R. Netzorg, B. Park, C. Singh, Y. S. Tan, T. Tang, Y. Wang, B. Yu, Curating a COVID-19 data repository and forecasting county-level death counts in the United States, arXivarXiv:2005.07882.
URL http://arxiv.org/abs/2005.07882

[29] B. D. Killeen, J. Y. Wu, K. Shah, A. Zapaishchykova, P. Nikutta, A. Tamhane, S. Chakraborty, J. Wei, T. Gao, M. Thies, M. Unberath, A County-level Dataset for Informing the United States' Response to COVID-19, arXivarXiv:2004.00756.
URL http://arxiv.org/abs/2004.00756

[30] E. C. Polley, M. J. van der Laan, Super Learner in Prediction, U.C. Berkeley Division of Biostatistics Working Paper (2010) 1–19.
URL http://biostats.bepress.com/ucbbiostat/paper266/

[31] M. J. Van Der Laan, E. C. Polley, A. E. Hubbard, Super learner, Statistical Applications in Genetics and Molecular Biology 6 (1). doi:10.2202/1544-6115.1309.

[32] E. LeDell, M. J. Van Der Laan, M. Peterson, AUC-Maximizing Ensembles through Metalearning, International Journal of Biostatistics 12 (1) (2016) 203–218. doi:10.1515/ijb-2015-0035.

[33] S. Dudoit, M. J. V. D. Laan, Asymptotics of Cross-Validated Risk Estimation in Estimator Selection and Performance Assessment Asymptotics of Cross-Validated Risk Estimation in Estimator Selection and Performance Assessment, U.C. Berkeley Division of Biostatistics Working Paper Series Year.

[34] M. A. Hernán, J. M. Robins, Estimating causal effects from epidemiological

data, Journal of Epidemiology and Community Health 60 (7) (2006) 578–586. doi:10.1136/jech.2004.029496.

[35] J. M. Robins, Addendum To 'a New Approach To Causal Inference in Mortality Studies With a Sustained Exposure Period - Application To Control of the Healthy Worker Survivor Effect'., Computers & mathematics with applications 14 (9-12) (1987) 923–945. doi:10.1016/0898-1221(87)90238-0.

[36] Maya L Petersen, Kristin E Porter, Susan Gruber, Yue Wang, M. J. van der Laan, Diagnosing and responding to violations in the positivity assumption, Stat Methods Med Res 21 (1) (2012) 1–13. doi:10.4018/978-1-4666-2661-4.ch001.

[37] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria (2013).
URL http://www.R-project.org/

[38] T. J. Hastie, R. J. Tibshirani, Generalized additive models, Chapman and Hall.

[39] R. S. Loomba, G. Aggarwal, S. Aggarwal, S. Flores, E. G. Villarreal, J. S. Farias, C. J. Lavie, Disparities in case frequency and mortality of coronavirus disease 2019 (covid-19) among various states in the united states, medRxiv.

[40] M. W. Tenforde, E. B. Rose, C. J. Lindsell, N. I. Shapiro, D. C. Files, K. W. Gibbs, M. E. Prekker, J. S. Steingrub, H. A. Smithline, M. N. Gong, et al., Characteristics of adult outpatients and inpatients with covid-19—11 academic medical centers, united states, march–may 2020, Morbidity and Mortality Weekly Report 69 (26) (2020) 841.

[41] A. C. Paulo, M. Correia-Neves, T. Domingos, A. G. Murta, J. Pedrosa, Influenza infectious dose may explain the high mortality of the second and third wave of 1918–1919 influenza pandemic, PLoS One 5 (7) (2010) e11655.

Figures

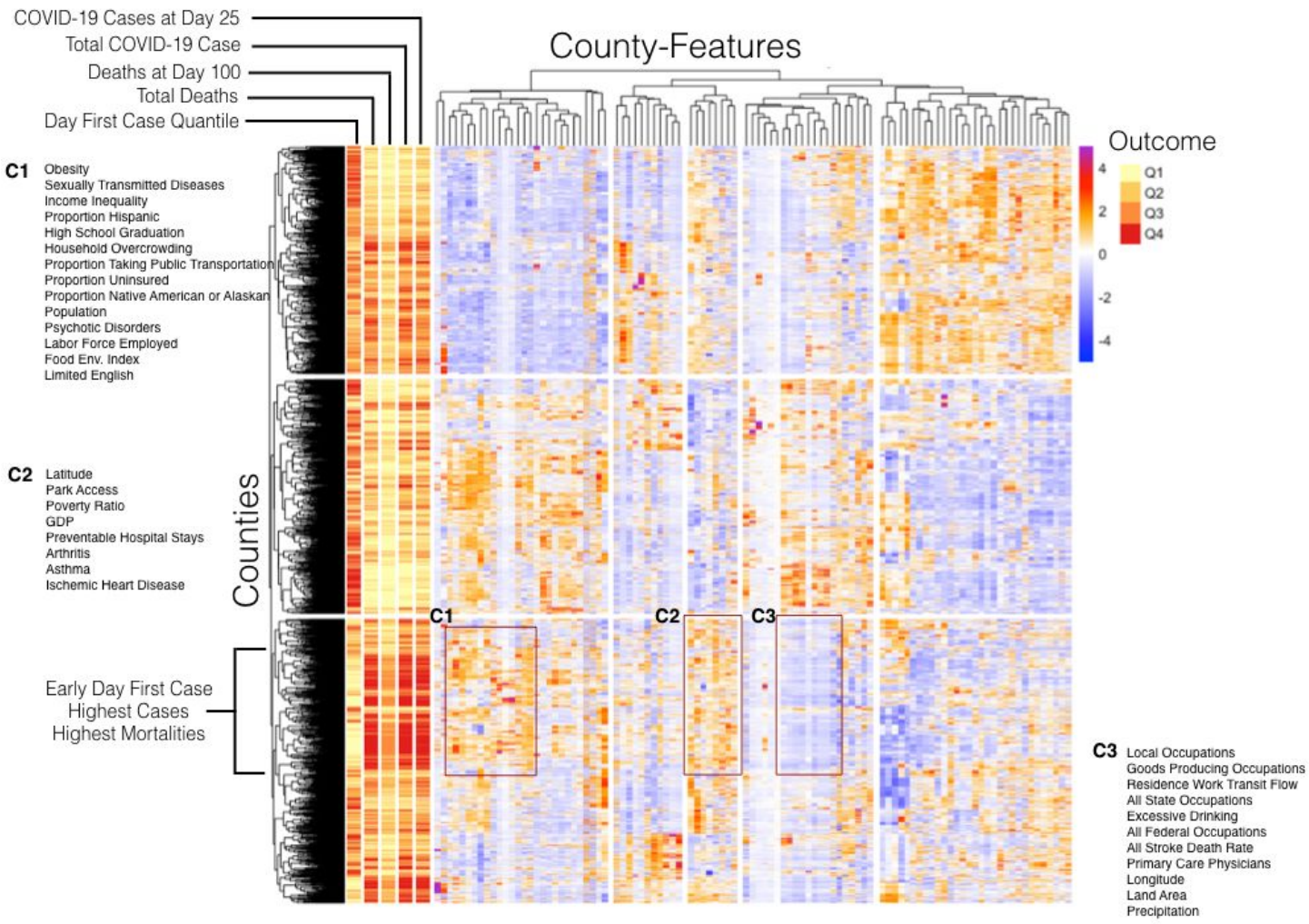

Figure 1

COVID-19 heatmap visualization of the distribution of county-level data.

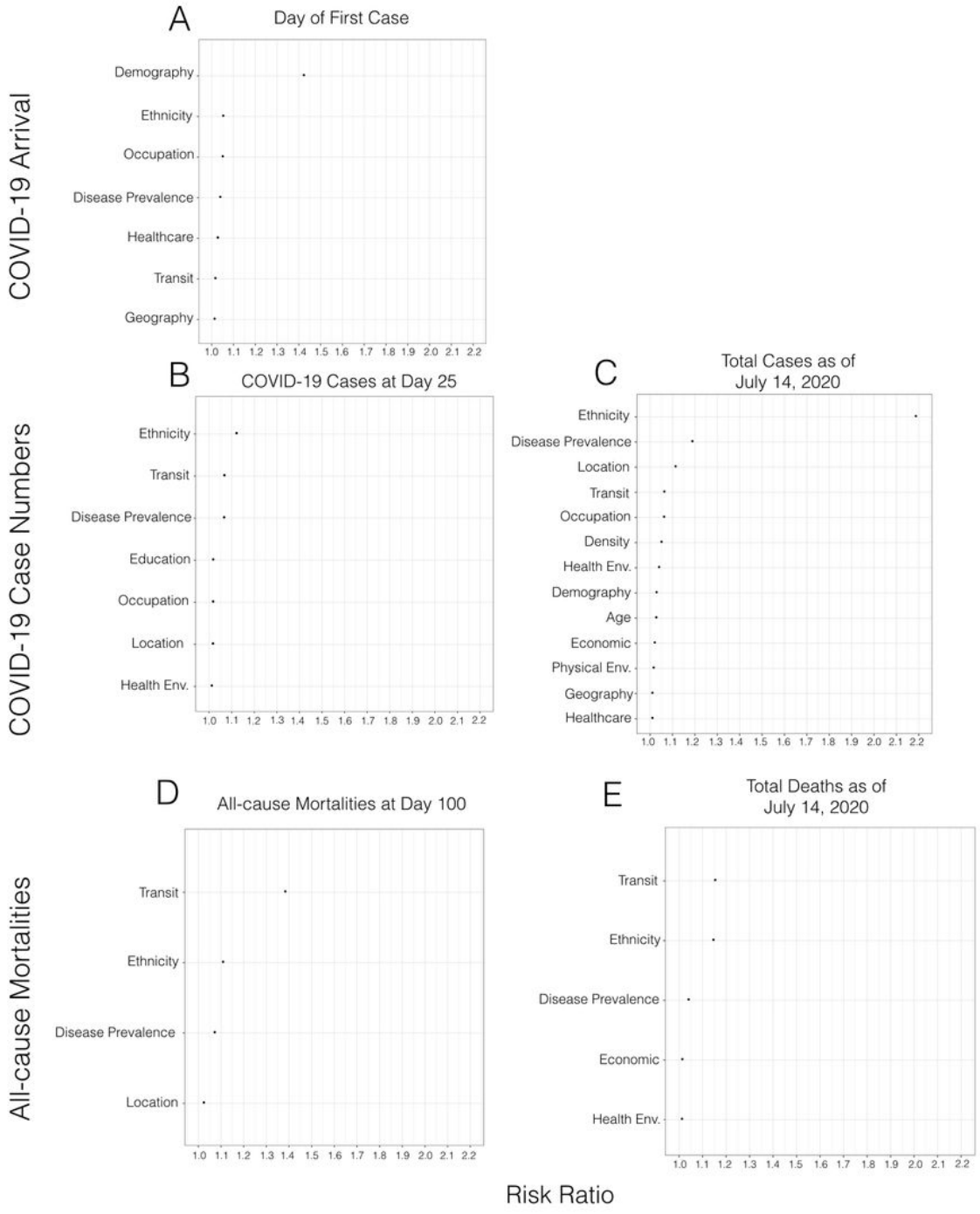

Figure 2

Variable importance as indicated by the relative increase of mean-squared error when the block of variables is removed.

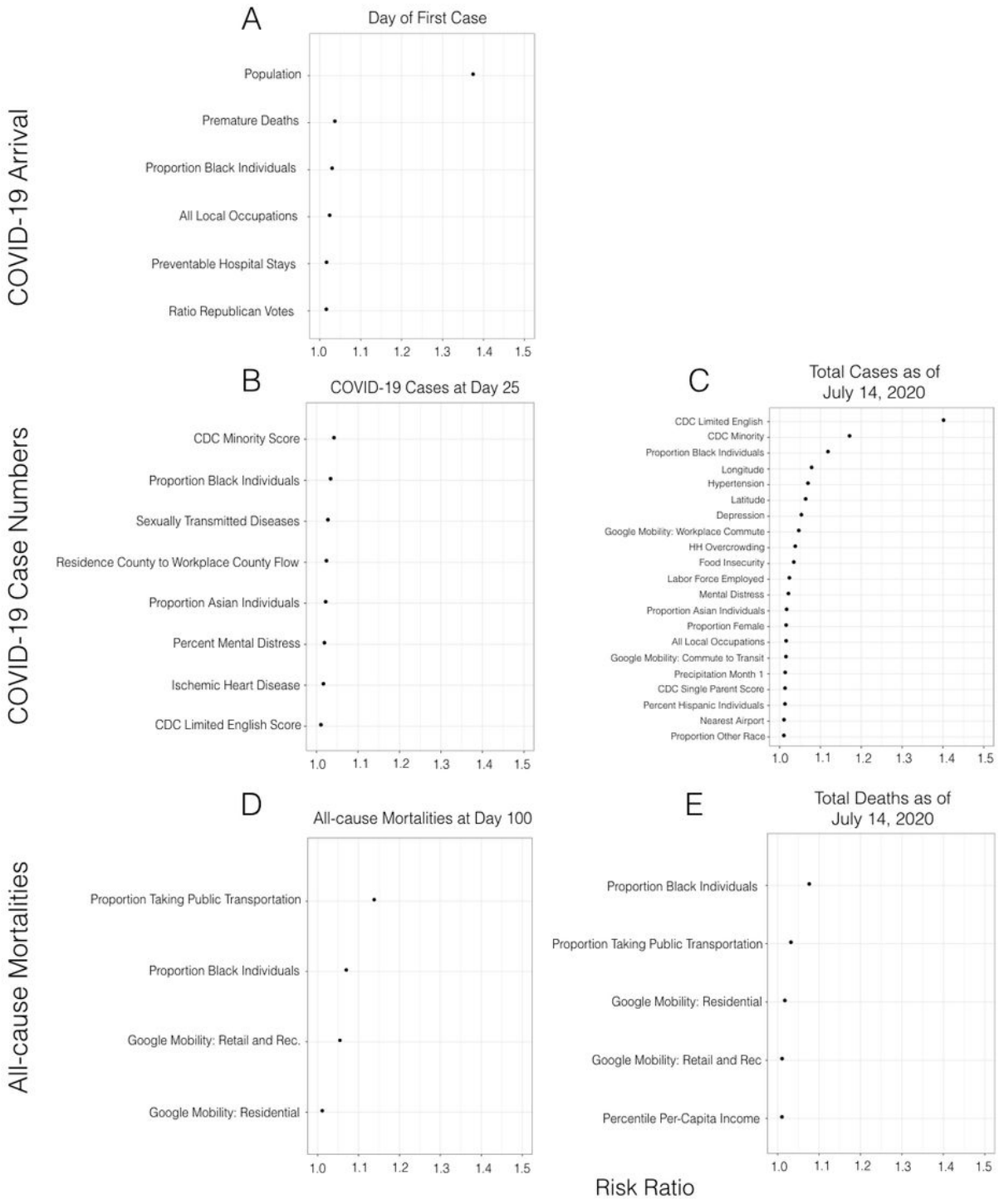

Figure 3

Variable importance as indicated by the relative increase of mean-squared error when a single variable is removed.

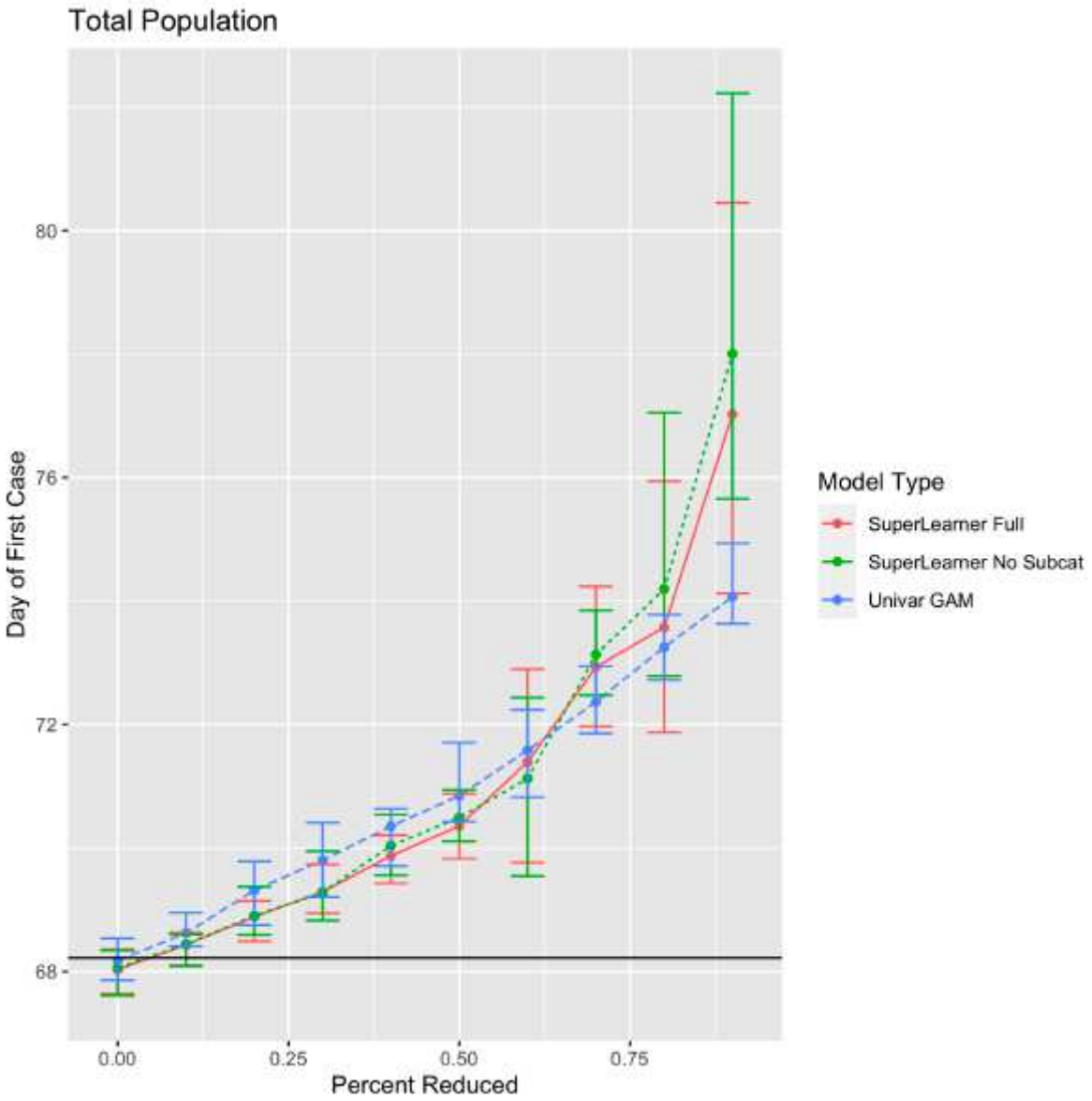

Figure 4

Marginal predictions of day of first case (relative to index time) for different proportional reductions of total population size for models adjusting for all other covariates, only covariates not in sub-category (see supplement table 1) and unadjusted.

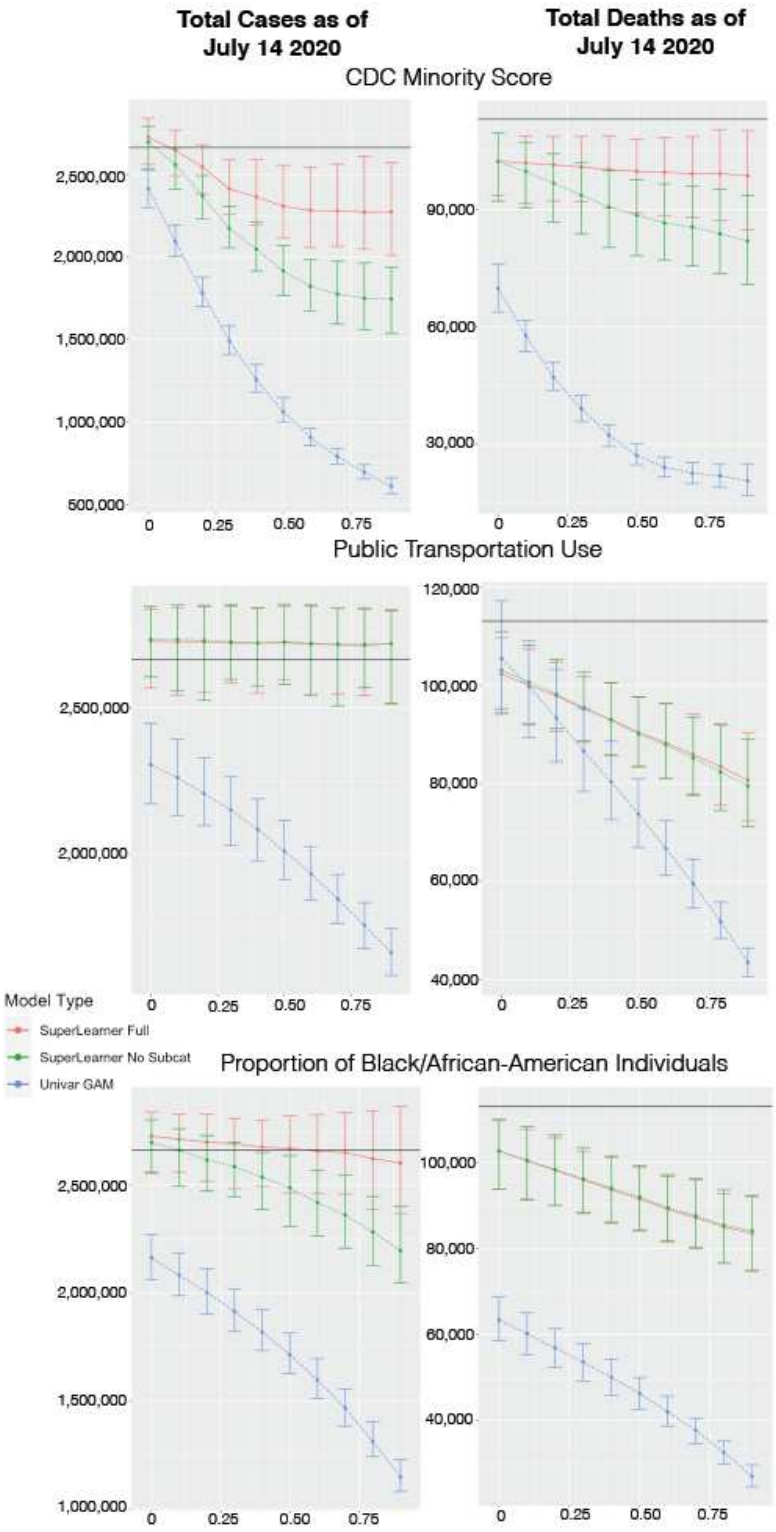

Figure 5

Marginal predictions of total cases and deaths by July 14, 2020) for three of the most consistently important variables in predicting the count outcomes: CDC minority score, proportion of Black individuals and a metric of public transportation use. X-axis is different proportional reductions of each of the three predictors, the Y-axis is the marginal predicted total counts for models adjusting for all other covariates, only covariates not in sub-category (see supplement table 1) and unadjusted.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- GetzHubbardTransportation4.pdf