scientific reports

OPEN

Public health factors help explain cross country heterogeneity in excess death during the COVID19 pandemic

Min Woo Sun[1,3]✉, David Troxell[2,3] & Robert Tibshirani[1,2]

The COVID-19 pandemic has taken a devastating toll around the world. Since January 2020, the World Health Organization estimates 14.9 million excess deaths have occurred globally. Despite this grim number quantifying the deadly impact, the underlying factors contributing to COVID-19 deaths at the population level remain unclear. Prior studies indicate that demographic factors like proportion of population older than 65 and population health explain the cross-country difference in COVID-19 deaths. However, there has not been a comprehensive analysis including variables describing government policies and COVID-19 vaccination rate. Furthermore, prior studies focus on COVID-19 death rather than excess death to assess the impact of the pandemic. Through a robust statistical modeling framework, we analyze 80 countries and show that actionable public health efforts beyond just the factors intrinsic to each country are important for explaining the cross-country heterogeneity in excess death.

The World Health Organization (WHO) estimates that the COVID-19 pandemic has led to 14.9 million excess deaths worldwide in 2020 and 2021[1]. Excess death is defined as the difference between actual reported death counts and what was expected under "normal" conditions based on data from earlier years[2]. While excess death can result from non-COVID causes, excess death has shown to be a more accurate measure assessing the true impact and death toll of the COVID-19 pandemic[3–5]. Undertesting and underreporting of COVID-19 cases, for instance due to limited testing capabilities, account for some of this excess death[6,7]. Indirect death resulting from COVID-19 due to hospital strain and overwhelmed public health infrastructures has also been shown to be associated with greater excess deaths[8].

Although the world collectively experienced a massive number of these excess deaths during the COVID-19 pandemic, the impact across different countries varies widely. While some countries suffered around 450 excess deaths per 100, 000 people, other countries saw nearly 0 excess deaths per 100, 000[1]. What factors drive this change? Were certain countries simply better suited to handle the pandemic due to robust healthcare infrastructure, a healthier population, and a younger population? Or can these pre-existing traits not explain such variations alone, implying that some COVID-19 related public health efforts lessened the impact of the pandemic? In this paper, we split the set of all variables into two distinct subsets: "intrinsic" features that countries inherited before COVID-19 was declared a pandemic in March 2020, and "actionable" features that can potentially be modified within the timeframe of the pandemic. By explicitly investigating governmental policies and public health efforts in conjunction with ingrained economic and health factors through this lens, we can better understand and help explain the drivers of excess deaths and the large variations of fatalities between countries.

Much of the past work regarding COVID-19 death at the country-level focuses on the impact of the pandemic on groups in different income levels (such as Refs.[9,10]). Another major area of study is the thorough analysis of COVID-19 death drivers in specific, single countries (such as Ref.[11,12]). In regard to studying the variation in COVID-19 deaths across countries, articles like[13] reveal that being strict in regards to certain policies - such as testing and contact tracing policies - are highly associated with countries that initially mitigated the spread and effect of the disease. However, this study and many others do not place a quantitative comparison among these factors, and no statistical model is created to estimate these features' impacts. Other studies purely investigate intrinsic population risk factors. For example, the findings from Ref.[14] state that Alzheimer's disease and some

[1]Department of Biomedical Data Science, Stanford University, 450 Serra Mall, Stanford, CA 94305, USA. [2]Department of Statistics, Stanford University, 450 Serra Mall, Stanford, CA 94305, USA. [3]These authors contributed equally: Min Woo Sun and David Troxell. ✉email: minwoos@stanford.edu

lung-related illnesses are associated with higher case mortality rates of COVID-19. The study performed in Ref.[15] incorporates both population risk factors such as lung-related illness prevalence and governmental policies like COVID-19 testing strategies. However, these models predict case fatality rate (CFR) rather than excess deaths. Additionally, like many of the other aforementioned studies, the analysis was performed in 2020 or early 2021 making it impossible to estimate and incorporate the effects of the COVID-19 vaccine.

In our paper, we incorporate and add to many of the characteristics of the studies previously mentioned. Our primary contributions to the study of COVID-19 death variations are as follows:

1. We build a gradient boosting model that allows for the quantification of how much various factors drive excess death. Additionally, we use these models to discuss which combinations of factors have especially deleterious effects on relatively lower-GDP countries and countries with obesity rates in the highest quartile of all countries in the study.
2. We view the analysis through the lens of intrinsic features and actionable features, and we specifically demonstrate the prediction power gained from considering actionable public health efforts. We formalize this notion via a bootstrapped hypothesis testing procedure. Also, we fit two models – one with only intrinsic features and one including both intrinsic and actionable features. We use these two models to construct confidence intervals regarding the change in excess death estimates for specific countries after adding actionable features.
3. We make steps to ensure that our analysis is robust. For example, we use excess death as our target variable which provides a more encompassing view of the pandemic's mortality burden on a given country. We also incorporate many countries in our study rather than focusing on a particular country or region. Lastly, we incorporate newer, measurable factors such as vaccination rate.

Results

In this section we report the insights from our statistical analyses. We demonstrate that actionable factors—particularly COVID-19 vaccination rates and trust in official COVID-19 advice from governmental entities—are important in helping explain the cross-country heterogeneity of excess death.

Key drivers of excess death

We show in the "Methods" section that our final model successfully captures some of the intrinsic structure and signal in the dataset. The two questions of interest when dissecting how this model makes predictions are "which features do the final model deem as useful?" and "how exactly do these useful features contribute to the model?". Feature importance methods implemented via the scheme detailed in the "Methods" section are one way to answer the first question. Figure 1 reveals the six features with the highest relative importance in the final model.

Two of the six features are "actionable" features, motivating the bootstrapped hypothesis test framework in the subsection "Increase in Predictive Performance from Actionable Features" which studies the improvement in predictive performance from including actionable variables. The percent of citizens with at least one dose of a COVID-19 vaccine in the study's time frame and whether citizens base their pandemic-related decisions on advice from their government are deemed "actionable" factors since they could be altered via public health efforts. We note that in our study, we analyze how these variables impact predictive modeling results and behavior rather than make direct causal claims. Each variable in our study, though, was included because of its natural relation to health and COVID-19 or because of its existence in prior research works. For example, higher trust in governmental COVID-19 advice has been associated with higher degrees of health measures such as self-quarantine and hand-washing[16]. Moreover, higher trust in doctors has been shown to be associated with higher treatment adherence and higher fulfillment of medical needs[17]. This association led to the inclusion of confidence in a nation's hospitals as it too measures trust in public health providers.

Before discussing how these important variables specifically contribute to the modeling process, we note that stability is a crucial aspect to feature importance rankings and algorithms[18]. Past works have analyzed the relative stability of feature importance rankings for certain tree-based methods[19] and neural networks[20]. Feature importance methods lose an extreme degree of validity and interpretability if small perturbations to the dataset drastically change importance values and rankings. To test the stability of our rankings, we randomly remove 5% of our dataset, perform 5-fold cross validation, and compute the feature importance values. We repeat this process 500 times. Figure 2 displays the average ranking for the variables along with bars depicting the standard deviation for each variable's ranking across trials.

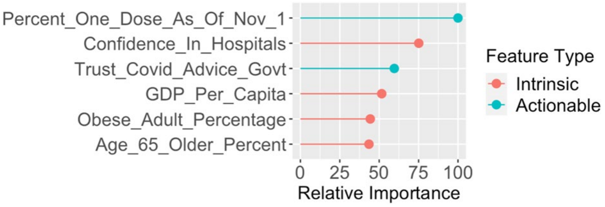

Figure 1. Six features in the final gradient boosting model with the highest relative feature importance. Feature definitions can be found in Supplementary Table S1.

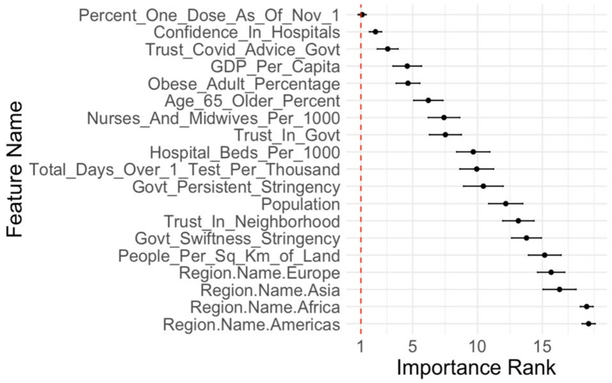

Figure 2. Feature importance rankings after 500 trials of randomly removing 5% of the dataset before performing 5-fold cross validation and calculating importance. For each variable, the dot represents the mean ranking, and the bars on each side represent the standard deviation across trials.

We see in Fig. 2 that the standard deviation in rankings resulting from dataset perturbations is typically around 1 to 1.5. For the top few variables, the standard deviation is less than 1. This suggests that our feature importance rankings are relatively stable in that they do not drastically change after small dataset changes.

After noting the relative stability of the feature importance rankings, one may ask "How do these important features specifically contribute to the model? Figure 3 depicts the partial dependency graphs for the six features with highest relative importance in the gradient boosting model.

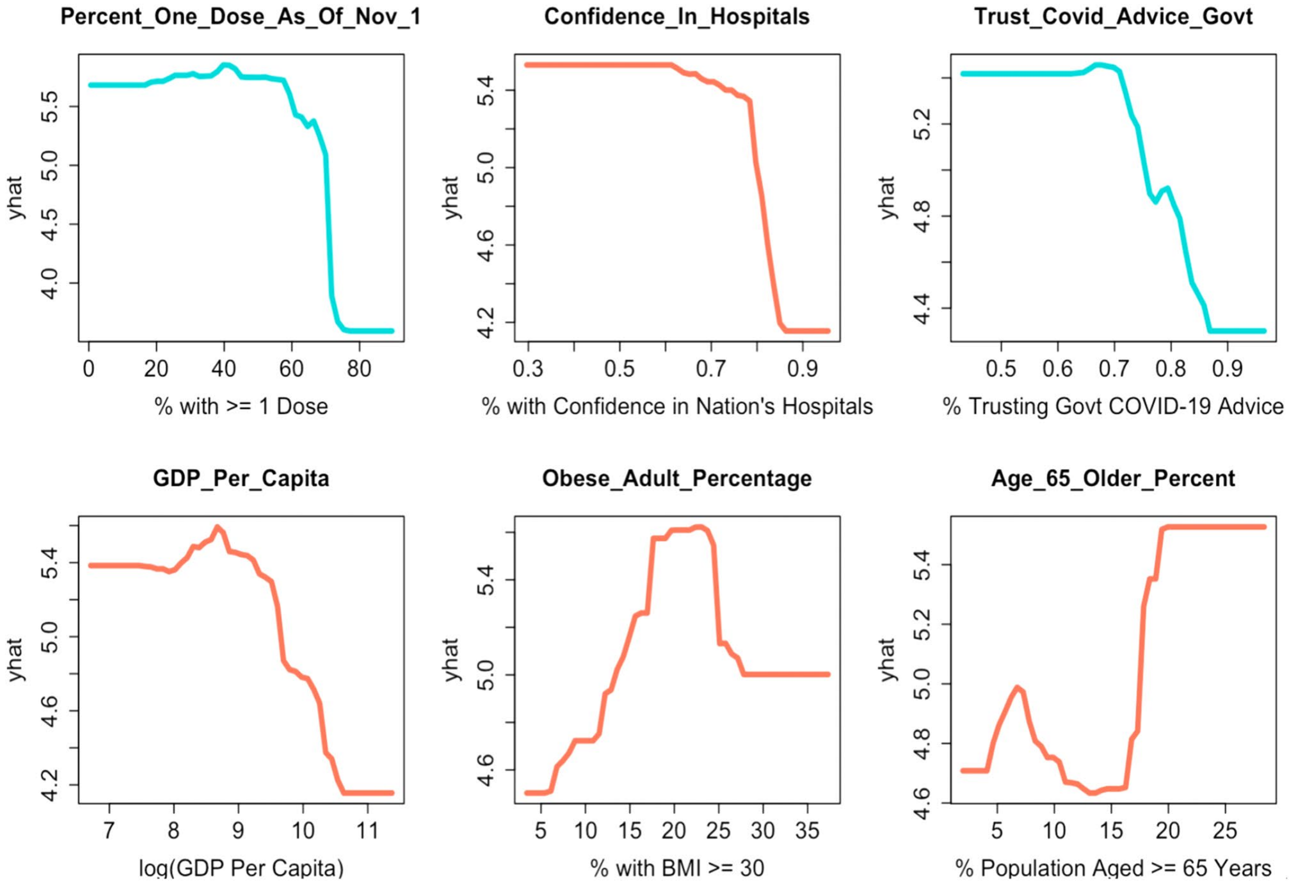

Figure 3. Partial dependency graphs for the six features with the highest relative importance in the final gradient boosting model. Each (x,y) pair on a given graph displays that the average estimate of excess death is y if all observations had the value of x for the variable of interest. The scale of the y-axis is predicted excess death standardized via a cube-root transformation.

We see that after marginalizing over all other variables, the estimate of excess death in the model decreases as certain variables increase. These variables include the percent of citizens with at least one COVID-19 vaccine dose, the percent of citizens basing their pandemic-related decisions on official advice from their national government, the percent of citizens who have confidence in their nation's hospitals, and the GDP per capita. Conversely, the model's estimate of excess death increases as other features increase. These features include the obesity percentage in the country and the percent of people aged 65 and over. Through these partial dependency plots, we can also look at countries that are similar in numerous aspects yet exhibit a large difference between their excess death. For example, we study the differences in excess death in Canada (54 excess deaths per 100,000 people) and the United States (265 excess deaths per 100,000 people). We see in Fig. 4 some explanation as to why the boosting procedure predicts lower excess death estimates for Canada.

After knowing which features are most important to the model and how they contribute to the estimation process, the next question to investigate is "how do variables in the model interact with one another?" Partial dependency plots can again provide some insights to this question. For our final gradient boosting model, we see in Fig. 5 an interaction between GDP per capita and how much the citizens of a country trust and base pandemic-related decisions on their government's advice regarding COVID-19.

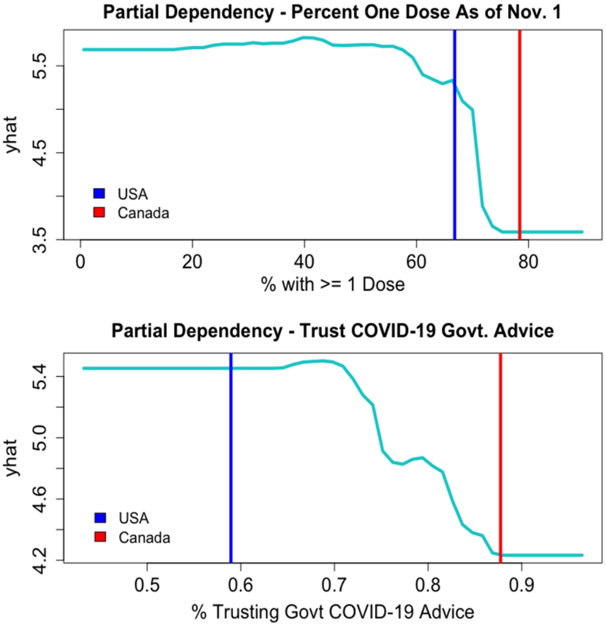

Figure 4. Partial dependency plots for two actionable features The vertical lines represent the observed values for the United States in blue and Canada in red. The y-axis represents the (cube-root) estimate of excess death after marginalizing over all other variables.

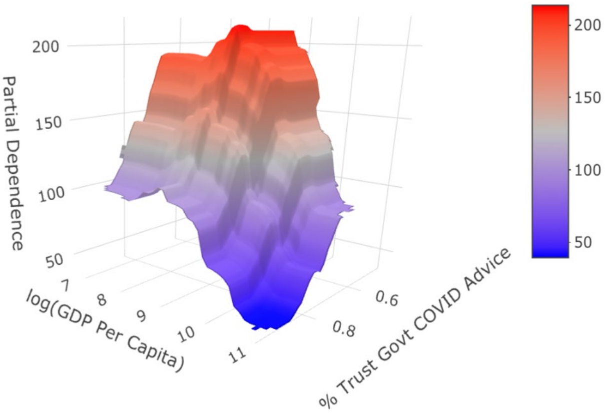

Figure 5. Partial dependency graph for GDP per capita and percent of citizens who base pandemic-related behavior decisions on official advice from their national government. The Partial Dependence z-axis is on the scale of excess deaths per 100,000 people.

Effectively, we see that the "penalty"—or increase in excess death estimate—for not trusting governmental COVID-19 advice is 43.7% greater for lower GDP countries than higher GDP countries. This is according to this class of models that successfully captures some inherent structure in the dataset as seen in Table 1 and Fig. 10. To see this relationship, one can imagine slicing the figure along a (log) GDP per capita of 10.5. Along this curvature, we note the increase in partial dependence as we move from around 85% government trust to around 60%. We see a similar increase exist when slicing the figure at a (log) GDP per capita of 7.5. If no interaction were present, these increases would be nearly equivalent. However, we see the penalty for moving from 85% trust (or higher) to 60% trust (or lower) is greater for countries with 3500 (USD) GDP per capita or less compared to countries with 36000 (USD) per capita or greater. This implies that following official, governmental advice regarding COVID-19 becomes particularly more imperative for lower-GDP countries to receive lower excess death estimates in our modeling process.

Next, we can analyze and quantify the interaction between vaccination rate and obesity percentage in the same way. Figure 6 depicts this relationship.

Following the same logic as the previous example, we see that countries in roughly the highest obesity quartile experience about a 11.2% greater penalty than countries in the lowest quartile when moving from "very vaccinated" (about 72% vaccination rate or better, which is the top third) to "relatively unvaccinated" (about 50% vaccination rate or worse, which is around the bottom half in our study). Therefore, we see via this quantification how much more imperative distributing vaccines becomes to get lower predicted excess death estimates for countries with higher obesity rates.

These partial dependency plots provide a view into the inner-workings of our gradient boosting model. Not only do we see the degree to which certain actionable features decrease excess death estimates, but we also identified particularly harmful combinations of actionable and intrinsic country characteristics. This provides some insights as to which policies and metrics become increasingly important for countries that meet certain criteria.

Increase in predictive performance from actionable features

Since the final model suggests actionable features are important in predicting country-level excess death, we attempt to formalize this importance via the hypothesis test detailed in the "Methods" section. The results from the test indicate that the increase in predictive performance from the set of actionable features is statistically significant (p-value = 0.0001 and observed RMSE difference = 11.7 excess deaths per 100,000 people).

	Gradient boosting	Random forest	LASSO
Best hyperparameters	Number of trees = 1600	Variables sampled per split = 3	$\lambda = 0.01$
	Shrinkage = .007		
	Interaction depth = 2		
	Min obs per node = 10		
CV RMSE for best hyperparameter combination	145.2	154.3	179.6

Table 1. Results of the repeated cross validation scheme for 3 model classes. Hyperparameter grids had similar number of combinations and granularity level across classes.

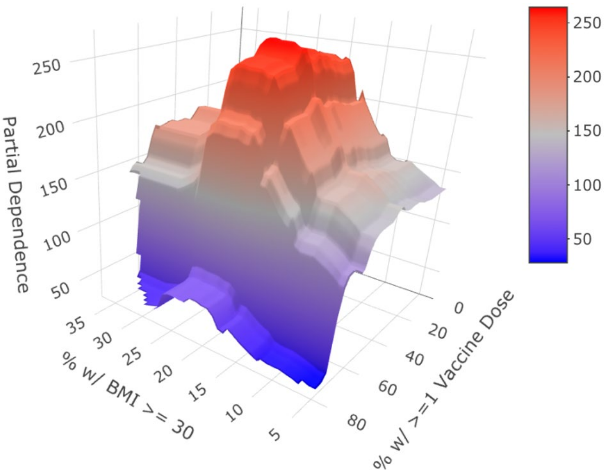

Figure 6. Partial dependency graph for percent of people with at least one dose of COVID-19 vaccine and and percent of citizens with BMI of at least 30. The Partial Dependence z-axis is on the scale of excess deaths per 100,000 people.

As formalized in the bootstrapped hypothesis test detailed in the "Methods" section, the gradient boosting modeling framework experiences a statistically significant decrease in RMSE (RMSE units are excess deaths per 100,000 people) when including actionable features as compared to using intrinsic variables only or intrinsic variables plus other randomly generated variables. In other words, the modeling framework inherently captures some signal via feature set expansion with actionable features. Because of this gain in overall predictive power, one may be interested in how specific predictions change after the feature set expansion (despite no guarantees whether a given prediction itself decreases in error). In addition to understanding how the model used these features to decrease overall RMSE, another naturally arising question is "which specific countries see increases in excess death estimates after adding actionable features"? To answer these questions through methods other than partial dependency figures, we create a measure called the "delta value". This value measures the difference in predicted excess death between the model with the actionable features and the model without the actionable features, further described in the "Methods" section. We construct confidence intervals for the delta value based on 100 bootstrap iterations. Figure 7 displays the effect of adding actionable features in the model for the 51 countries in which the direction predicted excess death improved, meaning that the model successfully predicted whether the given country's estimate of excess death should rise or fall after considering actionable features. In other words, if $\hat{Y}_I < Y$ then $\hat{Y}_I < \hat{Y}_{IA}$. Or, if $\hat{Y}_I > Y$ then $\hat{Y}_I > \hat{Y}_{IA}$, where Y is the observed excess death, $\hat{Y}_I$ is the predicted excess death from intrinsic-only model, and $\hat{Y}_{IA}$ is the predicted excess death from the full model including actionable features. We show the confidence intervals for all 80 countries in Supplementary Fig. S1. Here, we focus on the top two actionable features identified in the subsection "Key Drivers of Excess Death", desiring to understand how the model used these features to decrease RMSE. We see that countries to the right of the vertical zero line—i.e. countries that have increased predicted excess death after including actionable features—tend to have lower vaccination rates and less trust in COVID-19 advice from governments. This is displayed in the kernel density estimation plots—and while not shown - the trend persists not just for the 51 countries displayed, but for all 80 countries studied. This can also be found in Supplementary Fig. S1. The opposite effect holds for countries to the left of the vertical zero line. These countries had lower predicted excess death after including the actionable features and had higher vaccination rate and trust in COVID-19 advice from governments. This further demonstrates the importance of considering the actionable features in understanding the heterogeneity in excess death at the country level.

Discussion

In December 2019, the novel severe acute respiratory syndrome coronavirus 2 (SARS-Cov-2) led to the beginning of a devastating pandemic that has disrupted the global economy and resulted in millions of deaths around the world[21]. Beyond the direct death resulting from COVID-19 infections, the pandemic has had a tremendous indirect impact due to overwhelmed healthcare systems and limited medical resources. Over two years have passed since the beginning of the pandemic, yet we do not fully understand the effects of public health mandates and policies on COVID-19 death and whether the vast difference in death counts between countries are attributable to regional intrinsic factors. In this study, we take a data-driven approach to understand this problem by developing a statistical modeling framework to identify key drivers of excess death and assess the effect of actionable factors like COVID-19 vaccination and policies. We show that intrinsic factors like age distribution and comorbidity risks that have been shown to be positively associated with COVID-19 mortality in prior studies are important in explaining the cross-country variation in excess death. We also show in accordance with past literature that certain intrinsic trust metrics are important when estimating excess death. Particularly, the trust the population has in its nation's hospitals—potentially signifying a robust system more capable of handling the pandemic's burden—is an important feature.

We further demonstrate the importance of actionable features via a bootstrapped hypothesis testing framework. Our tests reveal that expanding the dataset beyond just the intrinsic country characteristics obtained in this study can aid in modeling large variation in excess death we see across countries, and that adding actionable features aids prediction quality more than adding random noise variables to a statistically significant degree. Specifically, proportion of population with one COVID-19 vaccine dose and trust in COVID-19 advice from governments emerge at the top of the list for factors contributing to the predictive performance of our final model. Both of these variables show negative association with excess death. In addition, we detail the particularly deleterious effect that poor performance in these actionable variables can have on certain groups of countries in the modeling process. For instance, we show that low GDP countries can experience a disproportionate rise in estimated excess death if its citizens also do not trust their government's pandemic-related advice. Similarly, we display that countries with obesity rates in the highest quartile of all countries in the study can experience a disproportionate rise in estimated excess death if the country is relatively unvaccinated, and we quantify this effect. We also note that the two principle component features signifying the swiftness in the government's response to the pandemic and the duration of such policy implementations did not appear near the top drivers of excess death. One explanation for this relatively low feature importance could be that there is a difference between policy implementation and policy compliance. This is perhaps too why the amount to which citizens trust their government's pandemic-related advice appears as one of the most significant variables in explaining excess death.

While our study aims to be holistic, a number of limitations are present. For example, our analysis does not exist in a causal framework. Rather than make causal claims, we study how various features impact modeling performance when the goal is to explicitly predict excess death across countries. Additionally, it is difficult to discern whether a government's pandemic-related policy implementations were reactive or proactive in terms of excess death without including time in the analysis. Future work for this study may include a temporal analysis that enables for a quantification of how different policies and intrinsic variables become significant or insignificant over time. One could also focus on a specific region of the world, which would lead to more complete data.

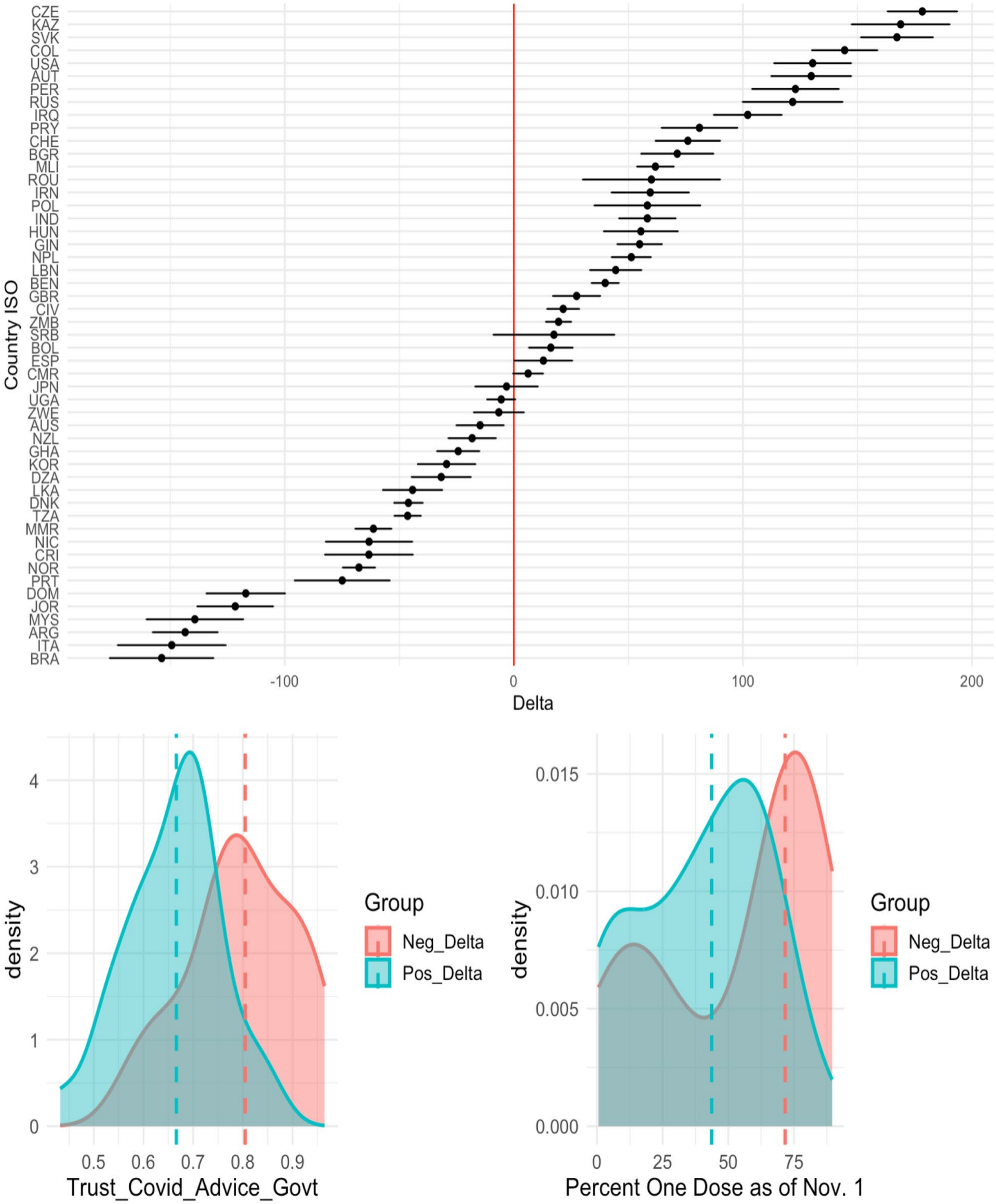

Figure 7. In the top plot, we show the 95% bootstrap confidence intervals of the "delta value", which measures the difference in excess death prediction between the model fit on both intrinsic and actionable features, and the model fit only on intrinsic features. The bottom two figures are density plots of "trust in COVID-19 advice form governments" and "percent of population with one vaccination dose as of Nov. 1" respectively, colored by whether a country's delta value lies to the left or to the right of the vertical zero line. The dashed lines represent the median values. Countries with delta to the left of the vertical zero line (i.e. countries with lower predicted excess death) tend to have higher vaccination rate and trust in government COVID-19 advice.

Regardless, understanding the cross-country heterogeneity of excess death is key to gaining a comprehensive understanding of the true impact of the pandemic and can lead to actionable guidance for government and public health institutions in preventing future deaths. This study helps elucidate the factors contributing to excess death at the population level during the COVID-19 pandemic and can help guide governments in improving their response to pandemics in the future, ultimately saving human lives.

Methods

In this section, we give details of our overall modeling framework for our analysis and how we arrive at our "final model". We build two models: one with only intrinsic features and one with both intrinsic and actionable features. We use and evaluate the predictions from these two models to assess the effects of actionable features such as vaccination rate and government policies during the COVID-19 pandemic.

Data collection and feature descriptions

First, we note that the time frame of our analysis is January 1, 2020 to November 1, 2021. This period is chosen for a number of reasons, including that the Omicron variant of the coronavirus SARS-CoV-2 was detected in early November 2021[22], and different variants can contain varying patterns of spread and mortality. Additionally, this time frame still allows for the analysis of COVID-19 vaccination policies. For this study, we use the excess death estimates from WHO as the dependent variable. We obtain 34 different initial variables for our analysis. Obesity percentage, age distribution, GDP per capita, population density, and hospital beds per 1000 people are a few of the "intrinsic" variables that were collected. Conversely, some of the collected features that are seen as at least moderately controllable over the course of the pandemic include number of days with comprehensive contact tracing, masking policies, workplace closures, days with available public testing, and official vaccination policies. The policy variables at the country-level were obtained from Ref.[23], while the socio-demographic, health, and government-spending habit information was obtained from World Bank Open Data. In addition to these data sources, we also include thorough survey information regarding trust of various entities among citizens for each country. We include these trust metrics since a number of studies report an association between trust and compliance to policy measures. For example, the study in Ref.[24] reports greater compliance for high-trust nations in the beginning of the pandemic, while[25] found an association between interpersonal trust and physical distancing adherence. Some of the trust metrics included in our analysis - such as overall confidence in the country's hospitals and trust in the national government - are classified as intrinsic variables. Others - such as the degree to which citizens trust their government's advice relating to COVID-19 measures - are classified as actionable. The survey data was obtained from Ref.[26]. COVID-19 related questions were typically asked in 2020, while other information to gauge citizen trust in various entities was obtained from 2018 survey data. For each survey, each country typically has approximately 1000 respondents. See Appendix A for the complete dataset description.

Final model construction for COVID-19 excess deaths

Pre-processing

First, we note that some countries in our dataset did not have every variable available. For countries that only have a few missing features, we impute these missing values via nuclear norm regularization. This method iteratively finds the soft-thresholded SVD and the algorithm in Ref.[27] is employed. After removing countries with many variables missing and filtering for countries with over 5,000,000 people due to COVID-19 data quality concerns, we continue our analysis with 80 countries from around the world. To simultaneously make all variables in our final data matrix exist on roughly the same scale while lowering skewness in some variables, a log transformation is performed for the GDP per capita, people per square kilometer, and population variables. No further standardization was performed since our feature importance techniques involve permuting values of a variable rather than simply use magnitude of linear coefficients (where scaling would matter), our remaining variables exist on roughly the same scale anyway (10^2 or less), and un-transformed variables more closely allow for interpretable modeling. Next, we note that interpretations of feature importance can be obfuscated when multi-collinearity is present, as it becomes challenging to rank variables in relative importance if a collection of such variables have similar effects. To avoid obscuring such interpretations and ensure drivers of excess death in the model can be reliably interpreted, we perform Principal Component Analysis (PCA) on the set of highly correlated features. The first principal component can be used in the model as a single variable representing the set of features as a linear combination. After creating various correlation plots, we decide to perform this method for variables falling into two groups: i) how quickly the country's government responded to COVID-19 and ii) how long the country's government held strict COVID-19 policies in place. An example of this dimension-reduction method for our dataset can be found in Fig. 8.

Critically, we note in Fig. 8 that interpretations after this method are still clear. Since each feature contributes in the same direction along the first principal component, the newly created feature can be interpreted as "a greater, positive magnitude indicates a longer-lasting strictness in COVID-19 policy implementations". We note too that when applying PCA to the second group of features with high multicollinearity - i.e. features relating to how quickly the government responded to the pandemic and put policies in place - all features also contribute in the same direction. This leads to the interpretation "a greater, positive magnitude indicates the government was swift in implementing stricter COVID-19 policies".

Hyperparameter tuning

After preprocessing the data, a number of models are employed using both the intrinsic and actionable features in the dataset. Specifically, Least Absolute Shrinkage and Selection Operator (LASSO) regression[28], Random Forest regression[29], and Gradient Boosting regression[30] models are created and analyzed. However, the limited

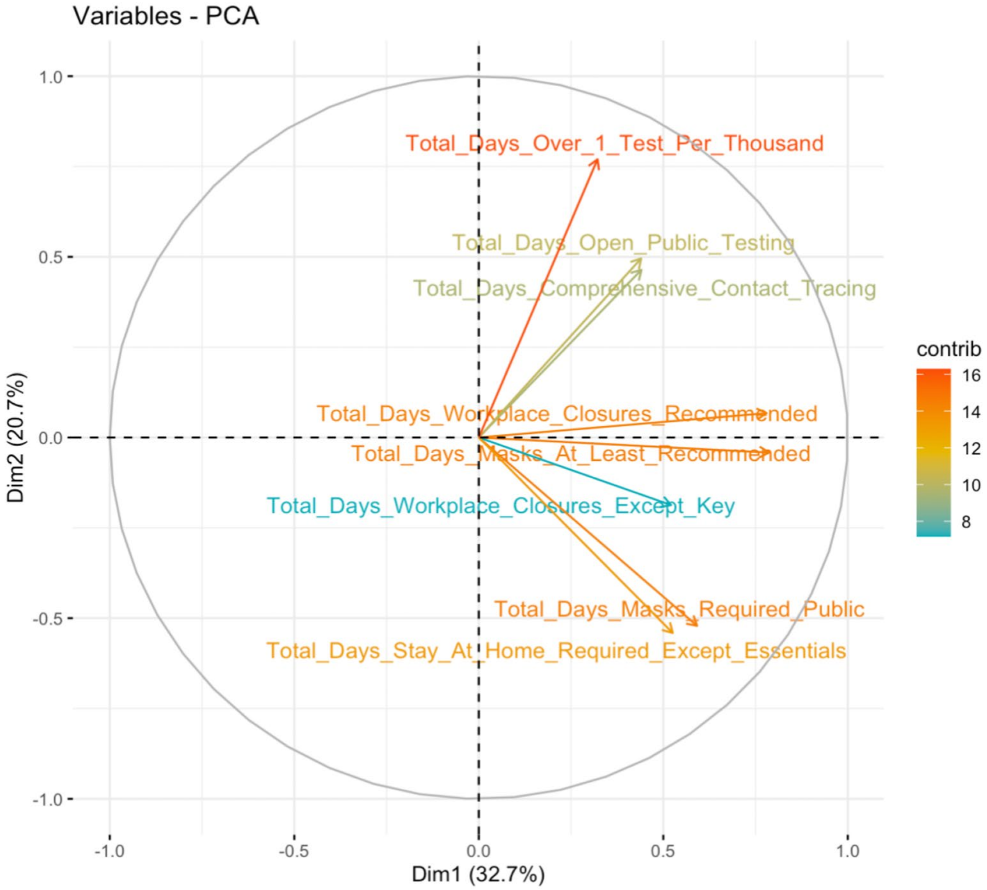

Figure 8. Bi-plot for one example of PCA in the employed dataset. All variables relate to how long the government's COVID-19 policy was kept in place.

sample size present in our dataset prevents the typical training and testing data split. Therefore, in order to estimate test accuracy for each model class and to pick hyperparameter combinations, we construct a repeated cross-validation scheme.

We collect 100 different CV predictions for each observation to obtain a more robust sense of predictions that could follow from slightly varying training sets. Figure 9 depicts the flow of the CV scheme.

Cross-validation results and model selection

Table 1 details the output from the 5-fold cross-validation scheme.

We see that across the 100 repetitions, gradient boosting is the class of models that result in the lowest cross-validation (CV) error. The CV root mean squared error values (with units being excess deaths per 100,000 people)

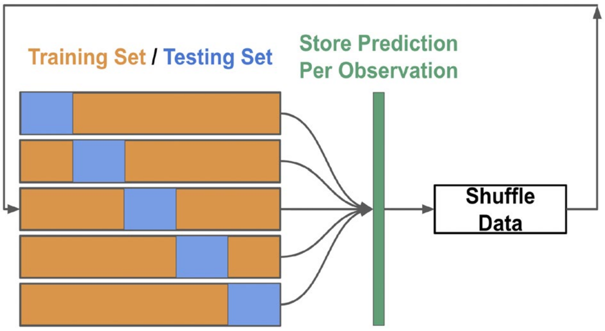

Figure 9. Schematic of the employed cross-validation algorithm. 100 predictions per observation were made for each hyperparameter combination.

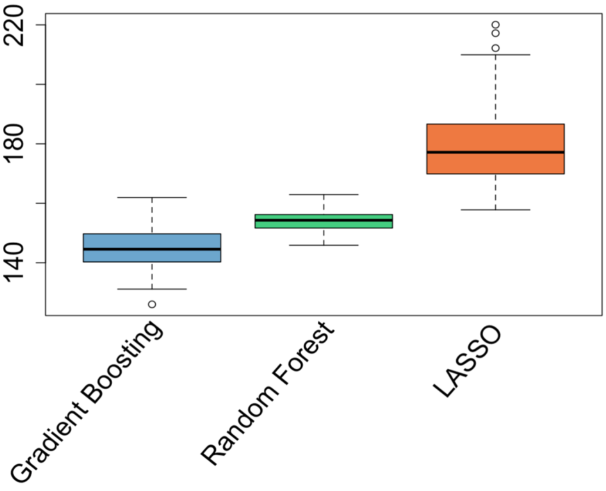

Figure 10. Distribution of CV RMSE. Each boxplot consists of the 100 different, individual CV RMSE values.

in Table 1 involve all $N * 100$ values (resulting from 100 repetitions of 5-fold cross validation). Figure 10 details the distributions of the RMSE for each individual repetition of the repeated CV process.

In Fig. 10 we see that the gradient boosting model class with the specified hyperparameters consistently provides lower RMSE values for individual 5-fold CV trials. Due to this overall performance and consistency demonstrated through the repeated CV procedure, we choose to employ gradient boosting with the tuned hyperparameters as the "final model".

After seeing how the gradient boosting model class characterized by the hyperparameter values specified in Table 1 leads to lower CV error, the next step is to investigate how this type of model makes predictions. Therefore, we train a "final model" using all of the available data to get a more holistic sense of the model's capturing of the inherent structure of the dataset. This "final model" leads to the insights discussed in the Results section.

Feature importance and partial dependencies

The parameters and estimation process of the final model detailed in "Cross-Validation Results and Model Selection" are dissected in numerous ways. Namely, feature importance and partial dependencies provide insights regarding the model. In our analysis, the importance of a feature is calculated via the increase in prediction error resulting from randomly permuting the values of the feature. The algorithm is similar to that in Ref.[29] but uses the entire dataset rather than out-of-bag observations[31]. Next, to define partial dependency, consider a sub-vector of input variables of interest denoted $\mathcal{X}_S$, its complement with other inputs $\mathcal{X}_c$, and an estimator $\hat{f}$. Partial dependency is defined as the marginal average[32]:

$$\tilde{f}_s(\mathcal{X}_s) = E_{\mathcal{X}_c}\hat{f}(\mathcal{X}_s, \mathcal{X}_c) \tag{1}$$

$$\tilde{f}_s(\mathcal{X}_s) \approx \frac{1}{N}\sum_{i=1}^{N}\hat{f}\left(\mathcal{X}_s, \mathcal{X}_c^{(i)}\right) \tag{2}$$

This allows for the expression of the estimator $\hat{f}$ as a function of the variables of interest $\mathcal{X}_s$.

Bootstrap hypothesis testing

In order to evaluate the impact of the actionable features, we build two models following the same hyperparameter tuning procedure. One model is only fit on the set of intrinsic features while the other model is fit on the set of both the intrinsic and actionable features as shown in Fig. 11. This dichotomy of models generates two sets of excess death predictions and RMSE. However, supervised learning algorithms can improve their predictive performance by increasing the number of features. In order to demonstrate that the set of actionable features are improving the model performance beyond the effect of simply adding more variables, we test the following hypothesis:

$$\mathrm{H}_0 : \mathrm{A} \perp \mathrm{Y} \tag{3}$$

$$\mathrm{H}_1 : \mathrm{A} \not\perp \mathrm{Y} \tag{4}$$

where A is the set of actionable features and Y is the response variable, excess death. To construct the null distribution, we perform the following bootstrapping procedure. We first fit a gradient boosting model with just the intrinsic features and compute the residuals. We bootstrap these residuals to generate $\{r_1^*, r_1^*, \cdots, r_B^*\}$. We add these residuals to the predicted excess death from the intrinsic-only model to get $y^* = \hat{y}_I + r^*$. We fit another gradient boosting model with both intrinsic and actionable features and the corrupted response variable y^* and

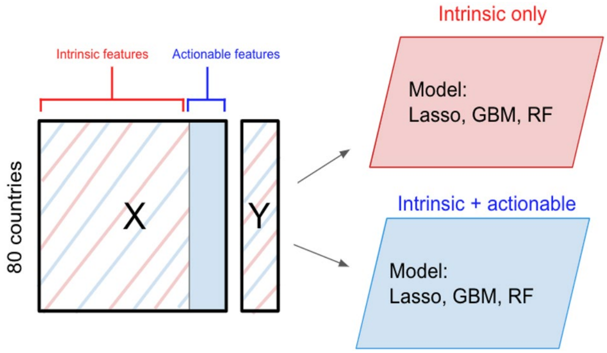

Figure 11. Diagram of the intrinsic vs. actionable feature modeling framework. Red represents the intrinsic features while blue represents the actionable features.

generate the residuals $\{r'_1, r'_1, \cdots, r'_B\}$. Finally we compute the difference of RMSE from the residuals r^* and r' as our test statistic.

We repeat this procedure for 1000 iterations to generate 1000 bootstrapped RMSE differences. The p-value is given by,

$$\frac{\textbf{card}\ (\text{Bootstrapped RMSE differences} > \text{Observed RMSE difference})}{B} \tag{5}$$

We attain a p-value of 0.0001 which suggests that the improvement in model performance through the inclusion of actionable features is statistically significant. The null distribution and observed RMSE difference of excess deaths per 100,000 people is shown in Fig. 12.

Measuring the effect of actionable features

As shown in "Bootstrap Hypothesis Testing", the overall modeling process inherently gains signal and predictive power when actionable variables are included in the feature set. Therefore, one may be interested in analyzing how specific predictions for any given country change after this feature set expansion, and in analyzing how the model uses the actionable features to decrease RMSE. To assess the effect of adding actionable features in the model for each country, we measure how much the prediction changes between the model including and the model excluding the set of actionable features. Here, we compute the quantity for each country i:

$$\delta^{(i)} = \hat{Y}^{(i)}_{IA} - \hat{Y}^{(i)}_{I} \tag{6}$$

where $\hat{Y}_{IA}$ are the excess death predictions from the model trained with both the intrinsic and actionable variables, and $\hat{Y}_I$ are the excess death predictions from the model without the actionable variables.

We further construct 95% confidence intervals for delta estimates. We estimate the variance of the deltas by bootstrapping - i.e. sampling with replacement - the countries in the pre-processed data matrix and then computing the difference of excess death predictions from the two models $\hat{Y}_I$ and $\hat{Y}_{IA}$. We repeat this procedure

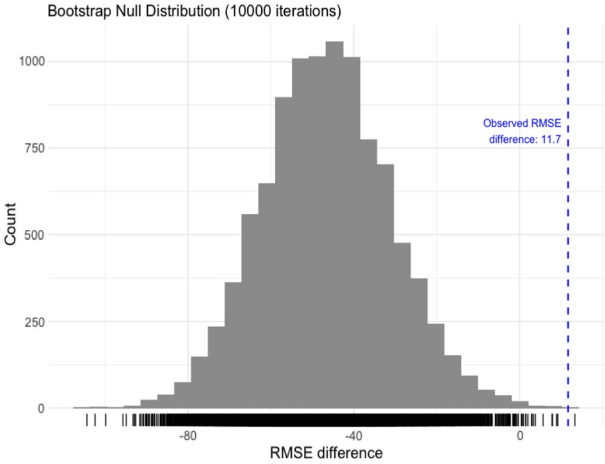

Figure 12. The gray histogram represents the null distribution constructed from 10,000 permuted RMSE differences and the red line represents the observed RMSE difference. This procedure results in a p-value of 0.0001.

for B iterations, generating B deltas $\delta^{(i)} = \{\delta_1^{(i)}, \cdots, \delta_B^{(i)}\}$ for each country i. We compute the sample standard deviation as:

$$\hat{s}^{(i)} = \sqrt{\frac{\sum_{b=1}^{B}\left(\delta_b^{(i)} - \bar{\delta}^{(i)}\right)^2}{B-1}} \tag{7}$$

Using this estimate, we construct the confidence intervals for each country i as follows:

$$\mathrm{CI}_i = \bar{\delta}^{(i)} \pm z_{1-\alpha}\frac{\hat{s}^{(i)}}{\sqrt{B}} \tag{8}$$

In practice due to sampling with replacement, some of the bootstrapped data matrix will have missing countries, so we bootstrap until each country has at least B number of δ.

Data availibility

All data can be found in https://github.com/minwoosun/covid-mortality/tree/main/analysis/data and detailed descriptions of the data are available in https://github.com/minwoosun/covid-mortality/blob/main/analysis/data/Country_Data_Descriptions.pdf.

Code availability

All code supporting this work is available on: https://github.com/minwoosun/covid-mortality.

Received: 30 November 2022; Accepted: 23 September 2023
Published online: 27 September 2023

References

1. Knutson, V., Aleshin-Guendel, S., Karlinsky, A., Msemburi, W. & Wakefield, J. Estimating global and country-specific excess mortality during the covid-19 pandemic (arXiv:2205.09081) (2022).
2. Checchi, F. & Roberts, L. Interpreting and using mortality data in humanitarian emergencies. Humanitarian Practice Network (2005)
3. Zimmermann, L. V., Salvatore, M., Babu, G. R. & Mukherjee, B. Estimating Covid-19-related mortality in India: An epidemiological challenge with insufficient data. *Am. J. Public Health* **111**(S2), 59–62. https://doi.org/10.2105/AJPH.2021.306419 (2021).
4. Ioannidis, J. P. A. Over- and under-estimation of Covid-19 deaths. *Eur. J. Epidemiol.* **36**(6), 581–588. https://doi.org/10.1007/s10654-021-00787-9 (2021).
5. Wang, H. *et al.* Estimating excess mortality due to the Covid-19 pandemic: A systematic analysis of covid-19-related mortality, 2020–21. *Lancet* **399**(10334), 1513–1536. https://doi.org/10.1016/S0140-6736(21)02796-3 (2022).
6. Lau, H. *et al.* Evaluating the massive underreporting and undertesting of Covid-19 cases in multiple global epicenters. *Pulmonology* **27**(2), 110–115. https://doi.org/10.1016/j.pulmoe.2020.05.015 (2021).
7. Li, R. *et al.* Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (sars-cov-2). *Science* **368**(6490), 489–493. https://doi.org/10.1126/science.abb3221 (2020).
8. French, G. *et al.* Impact of hospital strain on excess deaths during the Covid-19 pandemic-united states, July 2020–July 2021. *Am. J. Transplant. Off. J. Am. Soc. Transplant. Am. Soc. Transplant Surg.* **22**(2), 654–657. https://doi.org/10.1111/ajt.16645 (2022).
9. Bong, C.-L., Brasher, C., Chikumba, E., McDougall, R., Mellin-Olsen, J. & Enright, A. The covid-19 pandemic: effects on low-and middle-income countries. Anesthesia and analgesia (2020).
10. Miguel, E. & Mobarak, A. M. The economics of the Covid-19 pandemic in poor countries. *Ann. Rev. Econ.* **14**, 253–285 (2021).
11. Boccia, S., Ricciardi, W. & Ioannidis, J. P. What other countries can learn from Italy during the Covid-19 pandemic. *JAMA Intern. Med.* **180**(7), 927–928 (2020).
12. Liu, W., Yue, X.-G. & Tchounwou, P.B. Response to the COVID-19 epidemic: The Chinese experience and implications for other countries. MDPI (2020)
13. Baniamin, H. M., Rahman, M. & Hasan, M. T. The covid-19 pandemic: Why are some countries coping more successfully than others?. *Asia Pacific J. Public Admin.* **42**(3), 153–169 (2020).
14. Hashim, M. J., Alsuwaidi, A. R. & Khan, G. Population risk factors for Covid-19 mortality in 93 countries. *J. Epidemiol. Global Health* **10**(3), 204 (2020).
15. Sorci, G., Faivre, B. & Morand, S. Explaining among-country variation in Covid-19 case fatality rate. *Sci. Rep.*https://doi.org/10.1038/s41598-020-75848-2 *(2020).*
16. Han, Q. *et al.* Trust in government regarding Covid-19 and its associations with preventive health behaviour and prosocial behaviour during the pandemic: A cross-sectional and longitudinal study. *Psychol. Med.* **53**(1), 149–159. https://doi.org/10.1017/S0033291721001306 (2023).
17. Krot, K. & Rudawska, I. How public trust in health care can shape patient overconsumption in health systems? The missing links. *Int. J. Environ. Res. Public Health* **18**, 3860. https://doi.org/10.3390/ijerph18083860 (2021).
18. Calle, M. L. & Urrea, V. Letter to the editor: Stability of random forest importance measures. *Brief. Bioinform.* **12**(1), 86–89. https://doi.org/10.1093/bib/bbq011 (2010).
19. Wang, H., Yang, F. & Luo, Z. An experimental study of the intrinsic stability of random forest variable importance measures. *BMC Bioinform.* **17**(1), 60. https://doi.org/10.1186/s12859-016-0900-5 (2016).
20. Ghorbani, A., Abid, A. & Zou, J. Interpretation of Neural Networks is Fragile (2018).
21. Zhu, N. *et al.* A novel coronavirus from patients with pneumonia in China, 2019. *N. Engl. J. Med.* **382**(8), 727–733. https://doi.org/10.1056/NEJMoa2001017 (2020).
22. Mallapaty, S. Where did omicron come from? Three key theories. *Nature* **602**, 26–28 (2022).
23. Ritchie, H., Mathieu, E., Rodés-Guirao, L., Appel, C., Giattino, C., Ortiz-Ospina, E., Hasell, J., Macdonald, B., Beltekian, D. & Roser, M. Coronavirus pandemic (covid-19). Our World in Data (2020).
24. Devine, D., Gaskell, J., Jennings, W. & Stoker, G. Trust and the coronavirus pandemic: What are the consequences of and for trust? An early review of the literature. *Polit. Stud. Rev.* **19**(2), 274–285. https://doi.org/10.1177/1478929920948684 (2021).
25. Petherick, A. *et al.* A worldwide assessment of changes in adherence to Covid-19 protective behaviours and hypothesized pandemic fatigue. *Nat. Hum. Behav.* **5**(9), 1145–1160. https://doi.org/10.1038/s41562-021-01181-x (2021).
26. Gallup: Wellcome Global Monitor. UK Data Service (2022).

27. Mazumder, R., Hastie, T. & Tibshirani, R. Spectral regularization algorithms for learning large incomplete matrices. *J. Mach. Learn. Res.* **11**(80), 2287–2322 (2010).
28. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. Roy. Stat. Soc. Ser. B (Methodol.)* **58**(1), 267–288 (1996).
29. Breiman, L. Random forests. *Mach. Learn.* **45**(1), 5–32 (2001).
30. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **29**(5), 1189–1232. https://doi.org/10.1214/aos/1013203451 (2001).
31. Greenwell, B., Boehmke, B. & Cunningham, J. Generalized boosted regression models [R package GBM version 2.1.8.1]. Comprehensive R Archive Network (CRAN) (2022).
32. Hastie, T., Tibshirani, R., Friedman, J. H. & Friedman, J. H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* Vol. 2 (Springer, 2009).

Acknowledgements

MWS is funded by the National Library of Medicine (2T15LM00703339). RT is funded by the National Institutes of Health (5R01 EB001988-16) and the National Science Foundation (19 DMS1208164).

Author contributions

M.W.S. and D.T. are co-first authors and equally contributed to this work. R.T. supervised the work. All authors have read and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-43407-0.

Correspondence and requests for materials should be addressed to M.W.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.