Gathering, evaluating, and aggregating social scientific models

Miriam A. Golden[1,*] Tara Slough[2,*] Haoyu Zhai[1] Alexandra Scacco[3]
Macartan Humphreys[3] Eva Vivalt[4] Alberto Diaz-Cayeros[5] Kim Yi Dionne[6]
Sampada KC[7] Eugenia Nazrullaeva[8] P. M. Aronow[9] Jan-Tino Brethouwe[10]
Anne Buijsrogge[10] John Burnett[6] Stephanie DeMora[11] José Ramón Enríquez[12]
Robbert Fokkink[10] Chengyu Fu[12] Nicholas Haas[13] Sarah Virginia Hayes[14]
Hanno Hilbig[15] William Hobbs[16] Dan Honig[17] Matthew Kavanagh[14]
Roy Lindelauf[18] Nina McMurry[3] Jennifer Merolla[6] Amanda Lea Robinson[19]
Julio S. Solís Arce[12] Marijn ten Thij[20] Fulya Felicity Türkmen[6]
Stephen M. Utych[21]

13 September 2023

Abstract

On what basis can we claim a scholarly community understands a phenomenon? Social scientists generally propagate many rival explanations for what they study. How best to discriminate between or aggregate them introduces myriad questions because we lack standard tools that synthesize discrete explanations. In this paper, we assemble and test a set of approaches to the selection and aggregation of predictive statistical models representing different social scientific explanations for a single outcome: original crowd-sourced predictive models of COVID-19 mortality. We evaluate social scientists' ability to select or discriminate between these models using an expert forecast elicitation exercise. We provide a framework for aggregating discrete explanations, including using an ensemble algorithm (model stacking). Although the best models outperform benchmark machine learning models, experts are generally unable to identify models' predictive accuracy. Findings support the use of algorithmic approaches for the aggregation of social scientific explanations over human judgement or ad-hoc processes.

[1] European University Institute
[2] New York University
[3] Wissenschaftszentrum Berlin für Sozialforschung
[4] University of Toronto
[5] Stanford University
[6] University of California, Riverside
[7] University of British Columbia
[8] London School of Economics and Political Science
[9] Yale University
[10] Delft University of Technology
[11] University of Pennsylvania Annenberg Public Policy Center
[12] Harvard University
[13] Aarhus University
[14] Georgetown University
[15] Princeton University
[16] Cornell University
[17] University College London
[18] Netherlands Defense Academy
[19] Ohio State University
[20] Maastricht University
[21] Independent Researcher

[*] Corresponding author. Email: miriam.golden@eui.eu, tara.slough@nyu.edu

On what basis can we claim that a scholarly community understands some phenomenon well? Social scientists generally propagate discrete rival explanations for phenomena they study. To create collective knowledge, we need to survey rival accounts, evaluate them, and aggregate their best insights. How best to discriminate among or aggregate competing explanations represents an important but largely neglected meta-scientific question in the social sciences. There are not formal methods for explanatory aggregation in the social sciences. Outside of meta-analysis — which aggregates estimates for common explanations of common outcomes in different samples — formal aggregation is rarely attempted.

We field test tools for evaluating, filtering, and aggregating social scientific explanations for a single outcome. To do so, we *gather*, *evaluate*, and then *aggregate* rival explanations, assessing individual models and model aggregations in terms of predictive accuracy. We report the gains from aggregation as well as how expert human evaluations that draw on theoretical expertise compare with algorithmic analogues.

We conduct this field testing in the context of social scientific predictive models of COVID-19 mortality. In the gathering stage, we crowd-sourced statistical models incorporating political and social variables, asking researchers to predict future crossnational and sub-national COVID-19 mortality. We provided a simple web interface for model submissions, giving independent teams of participants access to common data but not to other submissions. The Covid Model Challenges (MCs) received 88 submissions from 60 different individuals based at 32 institutions in 10 countries (see Table S6).

In the evaluation stage (outlined in our Pre-Analysis Plan; see S8), we assess the performance of submitted models on outcome data not yet available at the time of model construction and submission, and identify the models with greatest predictive accuracy. To measure predictive accuracy, we rank submitted models according to their out-of-sample pseudo-R^2. The pseudo-R^2 of the best model is 0.483 while that of the median model is only 0.171, indicating wide variation in model quality. A workhorse machine learning (ML) model, constructed using Lasso, is also fit on all common predictors and generates an out-of-sample pseudo-R^2 of 0.377. Thus, although most contributions were not very accurate, ome contestants submitted highly predictive models.

We use two additional methods to evaluate models. First, we implement a stacking estimator that generates a meta-model based on all submitted models (*1*). The stacking estimator allocates weights to the predictions of each constituent model to maximize the meta-model's predictive accuracy. Second, we evaluate 175 expert forecasts of the predictive accuracy of the submitted models. These forecasts predicted the either a pseudo-R^2 ranking or the stacking weights for subsets of models. Results suggest stark limits to social scientists' abilities to filter explanations of common outcomes when synthesizing multiple accounts.

In the aggregation stage, we implement six separate methods (detailed below) to filter and combine models. The ensemble algorithm we use (stacking) outperforms the median model by a large margin and, in out-of-sample data, it outperforms the best model by 4 percent. This finding is consistent with a recent evaluation of probabilistic public health forecasts of COVID-19 across U.S. states, which also reports high accuracy in ensemble methods and high variation among stand-alone models (*2*). We further show that stacking greatly outperforms aggregate predictions drawn from the expert forecasts, suggesting benefits from algorithmic aggregation. In sum, our results show the benefits that come from aggregating insights from rival explanations over selecting among them.

The Problem: One Outcome, Many Explanations

In the social sciences, scholars simultaneously develop many explanations for important political and social outcomes, including the causes of economic growth, government corruption, political democratization, and collective violence. Yet, when we attempt to advance understanding of outcomes, we tend to construct new explanations rather than (re-)evaluating or synthesizing existing ones. While developing new theories is clearly important, this individualistic novelty-seeking process means we generate many, often disjointed, explanations for core outcomes, to the exclusion of building on what we already know (*3*, *4*). Such a process ignores the importance of assessing the merits of competing explanations. But absent standardized strategies for knowledge aggregation, we may place undue weight on findings from early studies or, more worryingly, from higher status researchers (*5*).

Social science research currently uses one of two distinct approaches when aggregating and synthesizing knowledge in a particular area, as outlined in Columns A and B of Table 1. First, experts write analytic review essays in the form of comprehensive literature reviews. Reviews are useful for systematically enumerating existing explanations for a particular outcome. They can also identify gaps in theory or in evidence underpinning claims in a literature. But their

Table 1: Characteristics of methods for aggregating social scientific evidence

	A. Review Essay	B. Meta-Analysis	C. Multiple Selection/Aggregation Strategies
N treatments	Any	One	Many
N outcomes	Any	≥ 1 (each measured in each study)	One
Sample	Any	Multiple	Common
Quantity(-ies) of interest	Unclear	Common structural parameters across studies or samples	Metrics of predictive accuracy

abilities to assess the merits of competing (or unrelated) explanations are unknown. The results we report below cast doubt on the ability of analytic reviews to synthesize knowledge.

Meta-analysis provides a second, more formal, approach to aggregation. In general, meta-analyses hone in on the relationship between a single cause (or treatment) and a set of one or more outcomes, measured in multiple studies conducted in different settings. Recent examples in the social sciences include (*6–10*). When constituent studies are internally valid, measure the effects of a common externally valid mechanism, and utilize harmonized study designs, meta-analysis can provide an estimate of a common treatment effect (or average outcome) across studies (*11*). However, meta-analysis does not offer a framework to combine different explanations for a single common outcome.

We propose to use multiple strategies to evaluate competing models. Some focus on selecting a single model according to predictive accuracy whereas others use ensemble methods to aggregate across models (*12*). Among ensemble methods, "stacking" (*13*) is a common approach; it runs models in parallel and aggregates based on how models combine to predict outcomes. (*1*) shows that stacking is a particularly applicable ensemble method when — as in the present study — the universe of models is open.

We implement these various methods algorithmically and also implement a forecasting exercise that generates analogues using expert judgement. We compare the predictive accuracy of these different methods on multiple explanations, where each explanation is represented as a statistical model. Models differ in their functional forms and in their predictors.

Research Design: The COVID-19 Model Challenges

The COVID-19 pandemic sparked a large body of social scientific work on its political, social, and behavioral determinants and outcomes (*14–18*). This wave of topically-focused research has reproduced, at breakneck speed, known pathologies of knowledge accumulation in the social sciences, as researchers have produced a bevy of largely disconnected arguments on common, critically important outcomes related to COVID-19. Although independent teams of researchers have articulated a large number of distinct arguments, there have been few attempts to synthesize emerging evidence (exceptions are (*19*, *20*)). As a result, we do not know what we know.

That said, the rapid proliferation of social scientific studies of COVID-19 provides an opportunity to field test strategies to combine research findings. Drawing on social science research featuring similar open challenges (*21*, *22*), we designed and implemented a set of COVID-19 Model Challenges. Participants designed models in the period running from December 2020 through January 2021 (Figure S1). The MCs encouraged researchers to develop and submit statistical models that used political and/or social variables to predict logged cumulative COVID-19 mortality per million people as of August 31, 2021 on specified data samples. We incentivized submissions by granting co-authorship to those who submitted the most predictive models (defined as models receiving non-zero weight (≥ 0.001) in the stacking exercise). The P.I.s provided an interactive web platform with harmonized covariates and outcome data on COVID-19 mortality through November 16, 2020 that modelers could use in design and submission (see Figure S3 for the web platform and https://osf.io/pgydn for lists of covariates available). Details regarding the research design appear in S1.

We elicited models predicting COVID-19 mortality over four separate samples: a crossnational sample of 166 countries and sub-national samples of for all states in India, Mexico, and the United States of America (USA), respectively. India, Mexico, and the USA are federal countries where some public-health policy is made at the state level. For each sample, we elicited both "general" and "parameterized" models. General models each specify the model's functional form but not parameter values, whereas parameterized models specify both the functional form and model parameter values. In

the body of this paper, we focus on crossnational general models, reporting findings from the seven other MCs in the Supplementary Information (SI).

We compare the performance of the models received in each challenge with (1) a model with standard "epidemiological" covariates and (2) a model generated by a Lasso (least absolute shrinkage and selection operator) algorithm on the full set of assembled predictor variables. Lasso provides a widely-used machine learning algorithm that produces interpretable models akin to the MC submissions (*23*).

We evaluate models on the basis of their predictive power. Our primary metric of predictive accuracy, given by Equation (1) in Methods and Materials, resembles R^2 but is evaluated for general models using leave-one-out predictions.[1] We also compare the correlation between leave-one-out predictions and observed outcomes (using Equation (2)), which abstracts from the levels (or intercepts) of the predictions. In addition to comparing the performance of separate models, we aggregate them using a model-stacking estimator (see Equation (3)). The estimator combines predictions from models by putting a weight on each. It generates an aggregate prediction as a convex combination of leave-one-out model predictions weighted by the stacking weights allocated to each model.

How does human forecasting of model performance compare with an algorithmic comparison or aggregation of models? Forecasting has become increasingly common across the social sciences (*24*, *25*). It provides a way to access expertise about social processes. Our interest focuses on expert evaluations of models, which is, informally, how models are routinely evaluated, for instance via peer review processes. In February and March 2021, we used the Social Science Prediction Platform (https://socialscienceprediction.org/) to elicit expert forecasts of the performance of the models earlier submitted to the MCs. We received 175 expert forecasts, 83 of which focused on how the crossnational models would perform on future data.

We randomly assigned experts into two groups to elicit two sets of forecasts: a horserace and a stacking forecast. In the horserace, experts saw a subset of six randomly-selected general models that were submitted to a given challenge and were asked to guess the probability that a model would be the most predictive in the set. In the stacking exercise, forecasters were asked to allocate weights across models — analogous to those generated through an algorithmic stacking analysis — over a subset of seven randomly-selected models. We compare the resulting rankings of models (in terms of either predictive ability or stacking weight) from this expert forecasting to those generated by the analogous algorithm. We also construct the aggregate prediction implied by the forecasts of the "representative expert" and of the "wisdom of the crowds." We compare the forecast predictive performance of models to that of the "representative expert" — proxied by the median-performing stacking forecast — and to the "wisdom of the crowds" assessment — proxied by the average stacking weights made by all forecasters. We implement these analyses to evaluate what an algorithmic ensemble method can add over how experts process competing models.

Results

We describe the collection of models gathered, their performance individually and comparatively, and the results of the various human and statistical procedures we use to aggregate models.

Gathering Models

Of the 88 models received across the four challenges, 26 addressed the crossnational challenge. Participants had varying levels of expertise; the modal participant holds a Ph.D. We cannot easily establish how representative participants are relative to any specific scholarly community. Model submitters may differ along multiple dimensions from "typical" social scientific researchers: they may be more likely to volunteer, more interested in COVID-19, or more interested in public health generally. However, this is true of any body of social science literature, so the MCs do not differ in this respect from other processes that feed scholarly explanations into the public domain. A particular advantage of the kind of open challenge we sponsored is that it levels the playing field for submissions, erasing standard professional hierarchies.

[1] Note that even though models make predictions about future (out-of-sample) COVID-19 mortality, parameters of general models are estimated using August 2021 outcome data. As such, we pre-specified the use of leave-one-out predictions as an out-of-sample test for the general (but not the parameterized) models.

Most model submissions used the maximum allowable three predictive variables. Ten models introduced user-submitted predictors absent from the datasets made available by the P.I.s. Each model was accompanied by a short text offering theoretically-motivated justifications ("logics") for the inclusion of particular variables and the functional form specified. Model submitters were encouraged to reference relevant scholarly literature in these texts. The logics were concise by design, but they ranged considerably in depth and in the degree to which they engaged existing literature. Descriptive statistics about the MCs and the models submitted are provided in Tables S4-S5.

Figure 1 summarizes the most common predictors found in crossnational models. The figure first orders political and social variables according to their frequency of use and then orders other — mostly health and demographic — variables by how commonly they appeared. Color coding indicates how frequently pairs of variables were entered together. The data depicted in the figure shows that the most common variables are trust in others, trust in government, and government effectiveness. The most common pairing of variables is trust in government and healthcare access — a coupling used in three of 26 submissions. Across all four challenges, 30 percent of models include measures of trust in society or trust in government in predicting COVID-19 mortality. The frequency with which distinct submissions used common predictors is distinguishable from random selection of predictors from the MC-provided data set (p-value = 0.004), which suggests that participants considered common arguments from the literature or shared intuitions when constructing their models.

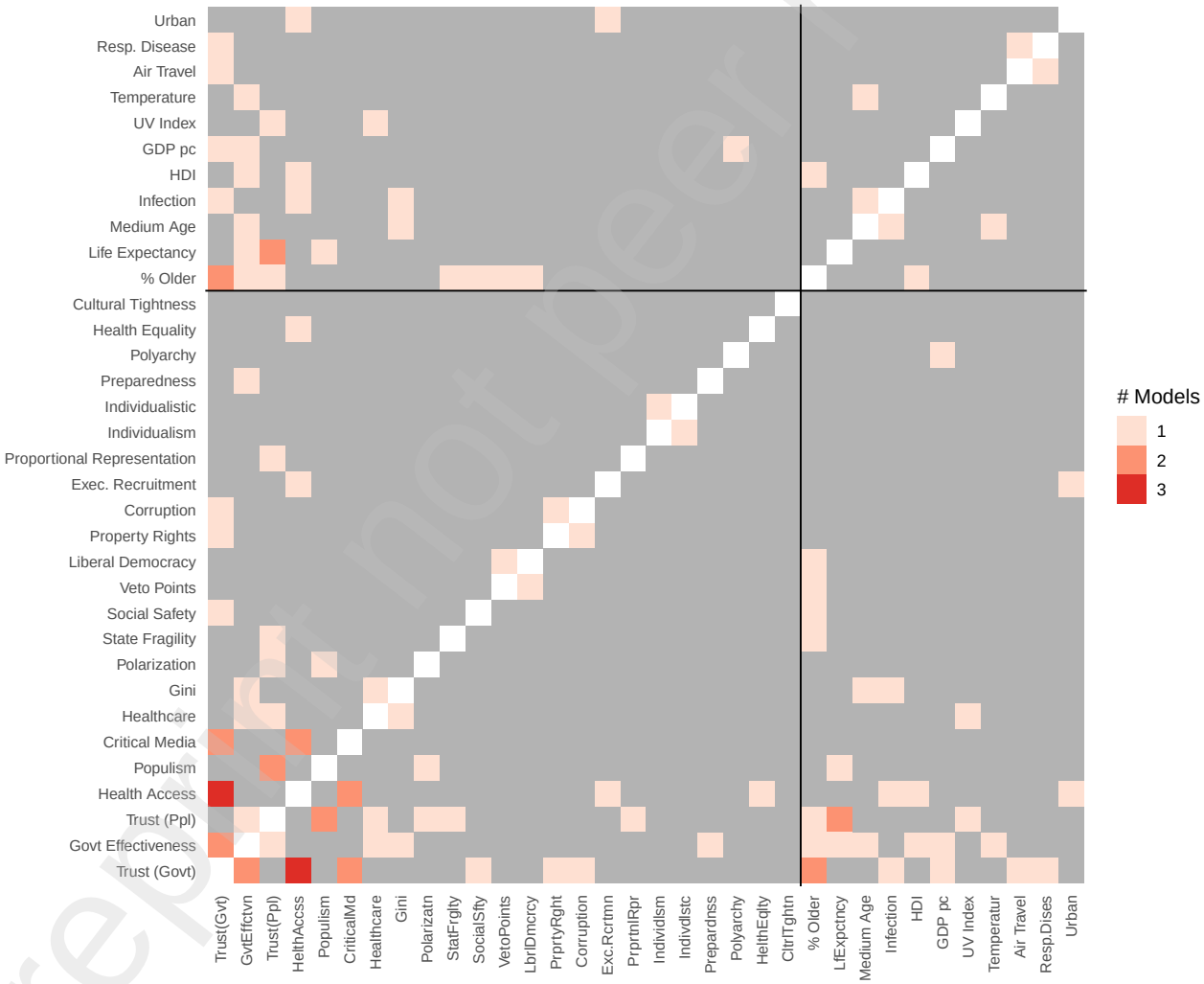


Figure 1: Pairwise combinations of variables submitted to the crossnational MC. The lower left quadrant shows social and political variables provided in the MC. The upper right quadrant shows other variables provided in the MC. Variable definitions available at https://osf.io/pgydn.

Evaluating Models

To evaluate the performance of individual models, we focus on out-of-sample predictive performance. (Details regarding statistical methods used for evaluation and aggregation appear in S2.) Figure 2 depicts the leave-one-out predictions arising from each crossnational general model that was submitted on the x-axis. On the y-axis we plot the outcome: logged cumulative COVID-19 mortality per million as of August 31, 2021. Each point represents one country. Our measure of predictive performance is calculated using Equation (1).[2] Following the interpretation of the R^2 measure, a perfectly predictive model would have a pseudo-$R^2 = 1$, where higher values indicate greater predictive power. However, the pseudo-R^2 is penalized when the leave-one-out predictions vary substantially from the predictions made using the full sample. This allows the pseudo-R^2 to be arbitrarily negative. In Figure 2, models are ordered from best to worst performing according to this metric.

Inspection of the data displayed in Figure 2 reveals that models vary substantially in their predictive power. The pseudo-R^2 of the best model is 0.483 but only 0.171 for the median model. Interpreting these metrics on an absolute scale is more challenging than making relative comparisons. Because the Lasso model is fit on all common predictors, it provides one possible benchmark. The pseudo-R^2 of the Lasso model is 0.377 and it ranks fourth (out of 28 models) in predictive power. The models that outperform Lasso are all theoretically motivated in that their authors provided reasons justifying inclusion of each variable.

Table S8 reports the estimated coefficients of the three best performing models in a more typical format. The best performer, "Trust in Authoritarian Government," is a simple linear model that combines three variables: a measure of trust in government, the presence or absence of a critical mass media, and access to sanitation. The logic submitted with this model states that the presence of a critical media is intended as a proxy for data manipulation by the government. The second best performing model, "Government Capacity and Social Inequality," includes measures of governmental effectiveness, the quality of healthcare, and economic inequality, and includes quadratic terms. One additional theoretically-motivated model outperforms the Lasso model. The "Perverse Development" model includes measures of access to sanitation and the human development index (HDI), both of which were intended, according to the logic supplied, to capture a country's level of economic development. The model predicted higher levels of COVID mortality in more developed contexts.

The MCs deliberately excluded post-treatment policy variables that measured government responses to COVID-19. Nonetheless, it is interesting that none of the top three models include political institutional variables (such as regime type) or measures of implied political priorities (such as the presence of a populist party in government) that have been widely cited in the popular press as predictive of COVID-19 mortality.

Given the many models that were submitted to the MC, how can we select the better models? We first evaluate models using two types of contests, mirroring the methods used in forecasting: a horserace between models and a stacking algorithm that weights them. The horserace orders models according to predictive performance, whereas stacking weights models according to their contribution to an aggregate meta-model. We then compare the two sets of results to the results from expert-elicited versions of the same two contests collected in the forecasting exercises. This gives us four metrics with which to assess relative model performance.

Figure 3 displays plots of the results of these assessments. Comparing the algorithmic implementations of the horserace and stacking contests, we see that few models receive non-zero stacking weights. The stacking meta-model draws on three constituent models despite minimal differences in the pseudo-R^2 across all the models. Many models utilize similar predictors. Nonetheless, the stacking estimates suggest that much of the collective predictive power of the models that we assess is concentrated in only a few of the best-performing models. The skew of estimated stacking weights towards the two top-performing models is striking.[3] These weights, combined with the coefficient estimates in Table S8, show that crossnational COVID-19 mortality is increasing in health access and decreasing in trust in government, etc.

The horserace and stacking contests implemented during the forecasting exercise generate top performing models that are different than those generated by their algorithmic cousins. In the horserace, there is no overlap between the top five

[2] If we were to rely on model predictions rather than leave-one-out predictions, the measure captured by Equation (1) would be equivalent to each model's R^2.

[3] Note that the weights that are estimated on each model by stacking are relative to the set of models that are evaluated.

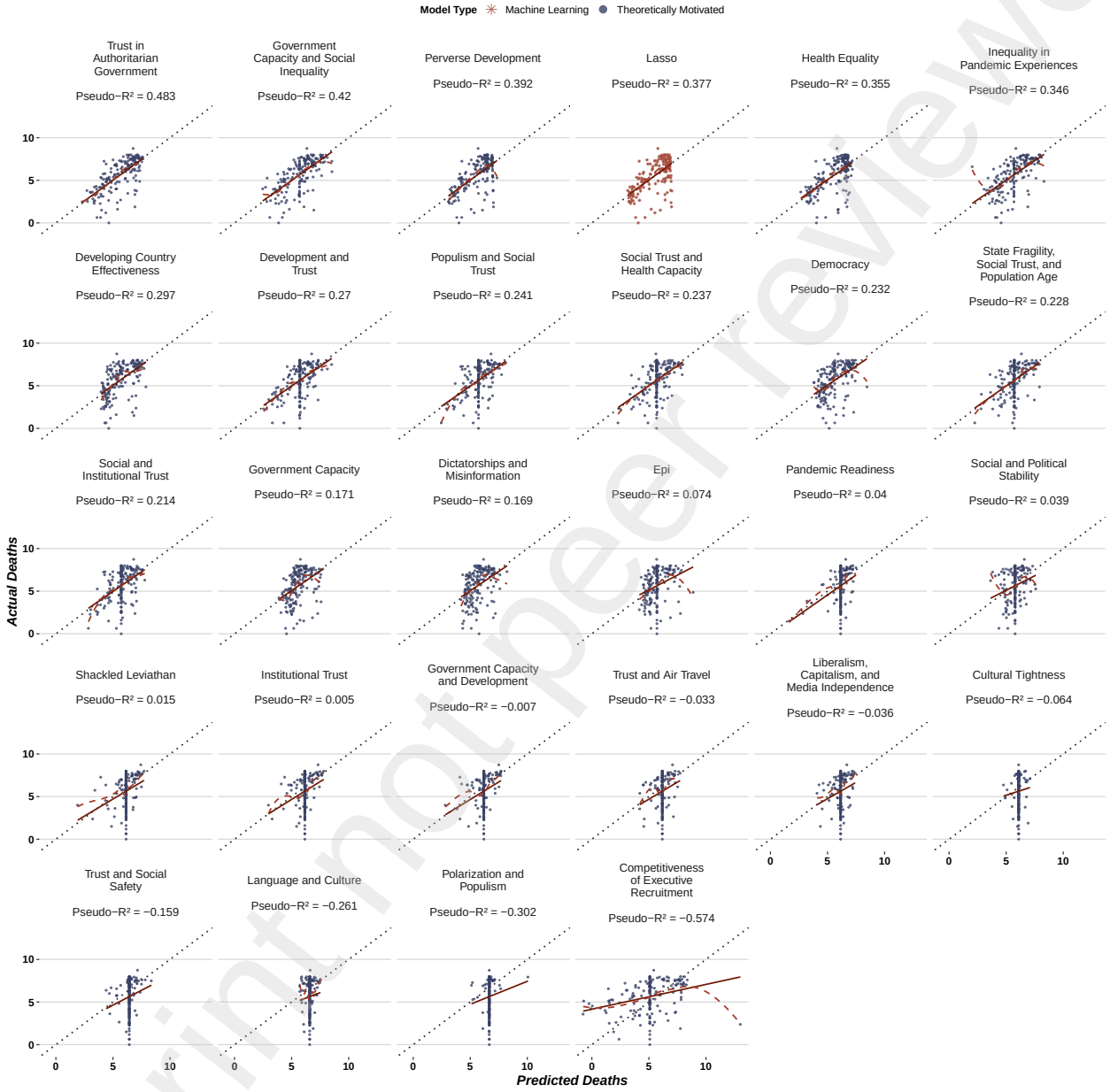


Figure 2: Evaluating: actual versus predicted deaths. Leave-one-out predictions of general models submitted to the crossnational MC and observed COVID-19 mortality as of August 31, 2021. Facets are ordered from highest to lowest pseudo-R^2. Dotted diagonal lines are 45 degree lines and fitted lines are estimated by OLS and LOESS. Non-machine learning models are theoretically justified user submissions whereas machine learning models were generated using a known (or reported) machine-learning algorithm.

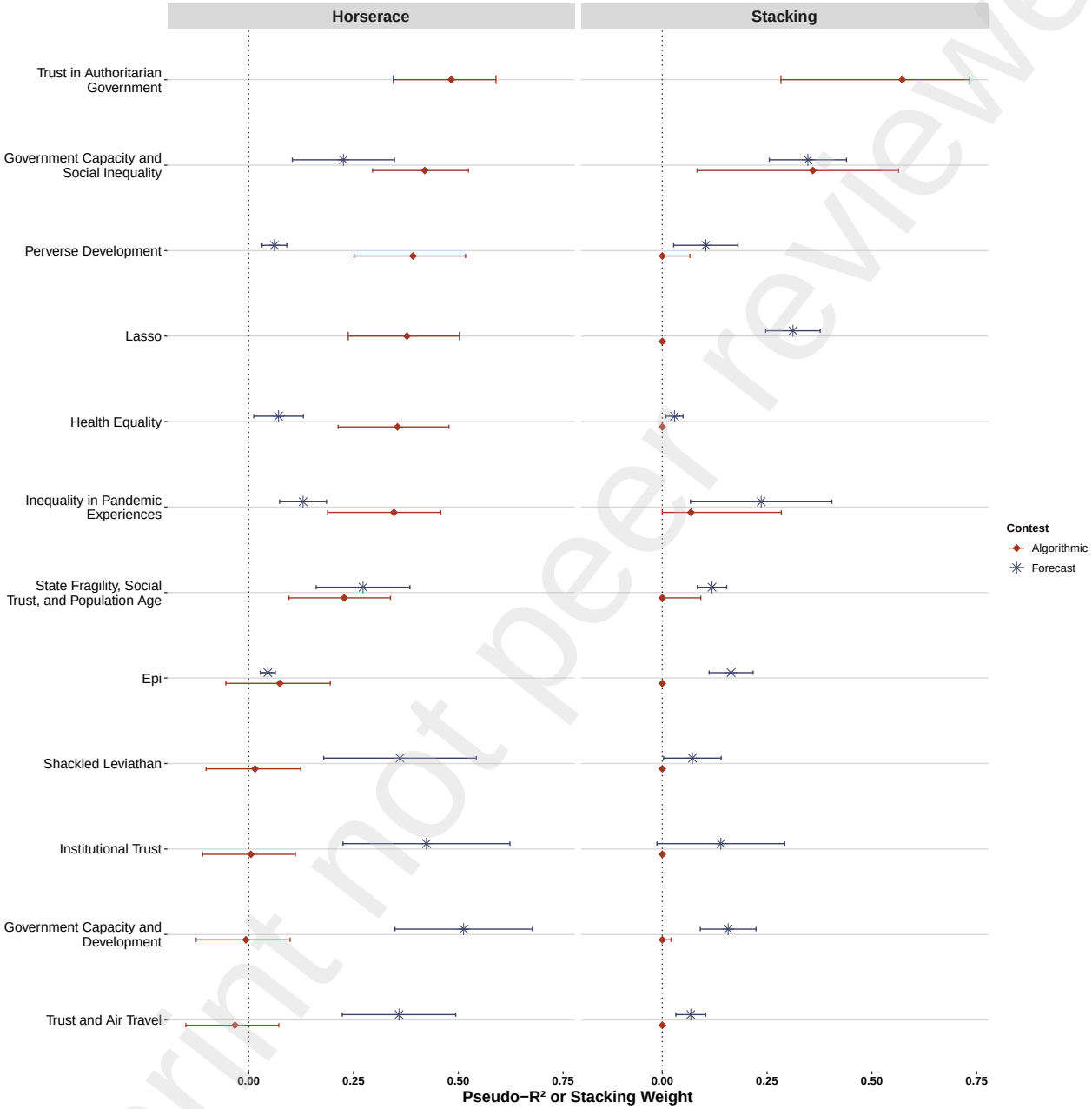


Figure 3: Comparison of models selected by the horserace and/or by stacking weights. The figures compare the results of the expert forecasts with the results of the algorithmic implementation in the same contest. The models included are among the top-five performers in any contest. 95% confidence intervals are generated by bootstrapping. Models with built-in procedures for improving fit (e.g. 'Trust in Authoritarian Government') that were not standard across models were not included in the forecasting exercises.

models determined by expert forecasters and those identified using an algorithm.[4] As the results depicted in Figure 3 show, there is some overlap between the algorithmic and forecast evaluations for the stacking contest. The weights elicited through forecasting are less skewed than those that arise from algorithmic stacking. However, weights are set-dependent in any model stacking. Due to concerns of tractability, the weights we elicited through forecasting were relative to smaller sets of models. By averaging over forecasts in different sets, we may observe regression toward the mean.

This analysis of model selection yields two central findings. First, comparison (horserace) and aggregation (stacking) prioritize different sets of models. In the general challenges, stacking heavily favors very few models, putting much lower weights on the others. This occurs even when differences of the pseudo-R^2's of individual models are actually quite minimal. Application of these metrics in other contexts is necessary to establish the generality of these patterns, though they hold across the four MCs for which we conducted forecasting. Second, we show that the average expert has limited ability to accurately identify the most predictive models. Even when experts are provided baseline performance metrics, as in our forecasting exercise, they do not do very well. However, to the best of our knowledge, the forecasting exercise we designed required novel cognitive tasks; perhaps with practice, forecasting performance would improve. To the extent that traditional methods for organizing and synthesizing knowledge produced by an existing literature ask researchers to identify the strongest arguments, our findings provide grounds for skepticism about their abilities to do so.

Aggregating Models

We now aggregate the predictions generated by the models and the expert forecasts. We compare different aggregation methods. In Figure 4, we show the results of six different aggregation methods. The rows depict two different ways of assessing predictive performance. The top row evaluates predictions of observed outcomes. The second normalizes both model predictions and outcomes, providing information about the correlations between them. The two columns show predictions for different time periods. The left column presents estimates of predictions of cumulative COVID-19 mortality as of August 31, 2021, which is the date for which MC participants were asked to predict. The second column presents out-of-sample predictions, which are evaluated as of June 20, 2022.[5]

We provide two benchmarks for each method. First, the "intercept only" metrics reflect the fact that both measures of variation explained normalize by a model that fits only the intercept (or mean) of the outcome. Thus, model performance measures that are above zero indicate that the leave-one-out predictions of a general model outperform a model consisting of only an intercept.[6] Because of normalization, an intercept-only model takes the value of zero for all analyses. Second, we benchmark model performance against a Lasso model selected on the basis of 2021 data to make out-of-sample 2022 predictions.

Our first two algorithmic measures of predictive performance — the best- and median-performing models in the MC — follow directly from the discussion in Section "Evaluating Models". Point estimates in the top row report the pseudo-R^2 of each model. The third algorithmic prediction examines the outcomes using the stacking meta-model. For purposes of out-of-sample predictions for 2022 in the righthand panel, we use the best, median, and stacking models that were selected on the basis of the 2021 data. For all algorithmic methods, we construct a sampling distribution of model performance by bootstrapping the data (resampling 166 countries with replacement).

The remaining three methods aggregate expert forecasts. The first metric examines the predictive power of the expert-favored model. As we documented earlier in Figure 3, the model that experts deem most likely to be the most predictive does not align with the model that is objectively found to be most predictive. The next two methods aggregate expert stacking forecasts. The "representative expert" forecast depicts the median aggregate stacking forecast. The final metric presents a "wisdom of the crowds" stacking model that aggregates over forecasters' stacking weights.

Having described what each part of Figure 4 represents, we now highlight two key findings that emerge from it. First,

[4]The horserace forecast measures the probability that experts believe a model will explain the most variation. Its algorithmic counterpart measures the variation explained by the leave-one-out predictions of the model.

[5]In the forecasting exercise, we asked participants for forecasts for model performance as of August 31, 2021 and also asked them to forecast as of August 31, 2022. However, because some governments stopped collecting COVID-19 mortality data before the latter date, we calibrate forecasts against the last date for which we were able to locate appropriate crossnational and subnational mortality data, June 20, 2022.

[6]Note that, as shown in Equation (1) and Equation (2) (in Methods and Materials), the intercept is a constant for all units in each sample; that it, it is not based on a leave-one-out approach.

the stacking estimator outperforms the other algorithmic and forecast-based alternatives that we use. In the left panel, this should not be a surprise: by construction, the stacking estimator outperforms all other convex combinations of model predictions in-sample, and all other aggregation methods can be represented as convex combinations of these predictions. However, stacking superiority persists in the right panel, where we use the aggregate stacking model fit on 2021 data to predict out-of-sample mortality in 2022. The latter out-of-sample result does not occur by construction.

Second, the estimates reported in Figure 4 indicate a substantial drop-off in predictive performance between the best and median models both in- and out-of-sample. This is by construction. However, the estimates also document the poor performance of forecasters in selecting predictive models. This reconfirms the analyses reported above (Section 4.2) and demonstrates that ensemble aggregation — here, model stacking — can be a particularly useful tool in light of experts' apparently limited ability to accurately discriminate between models or to forecast predictive performance.

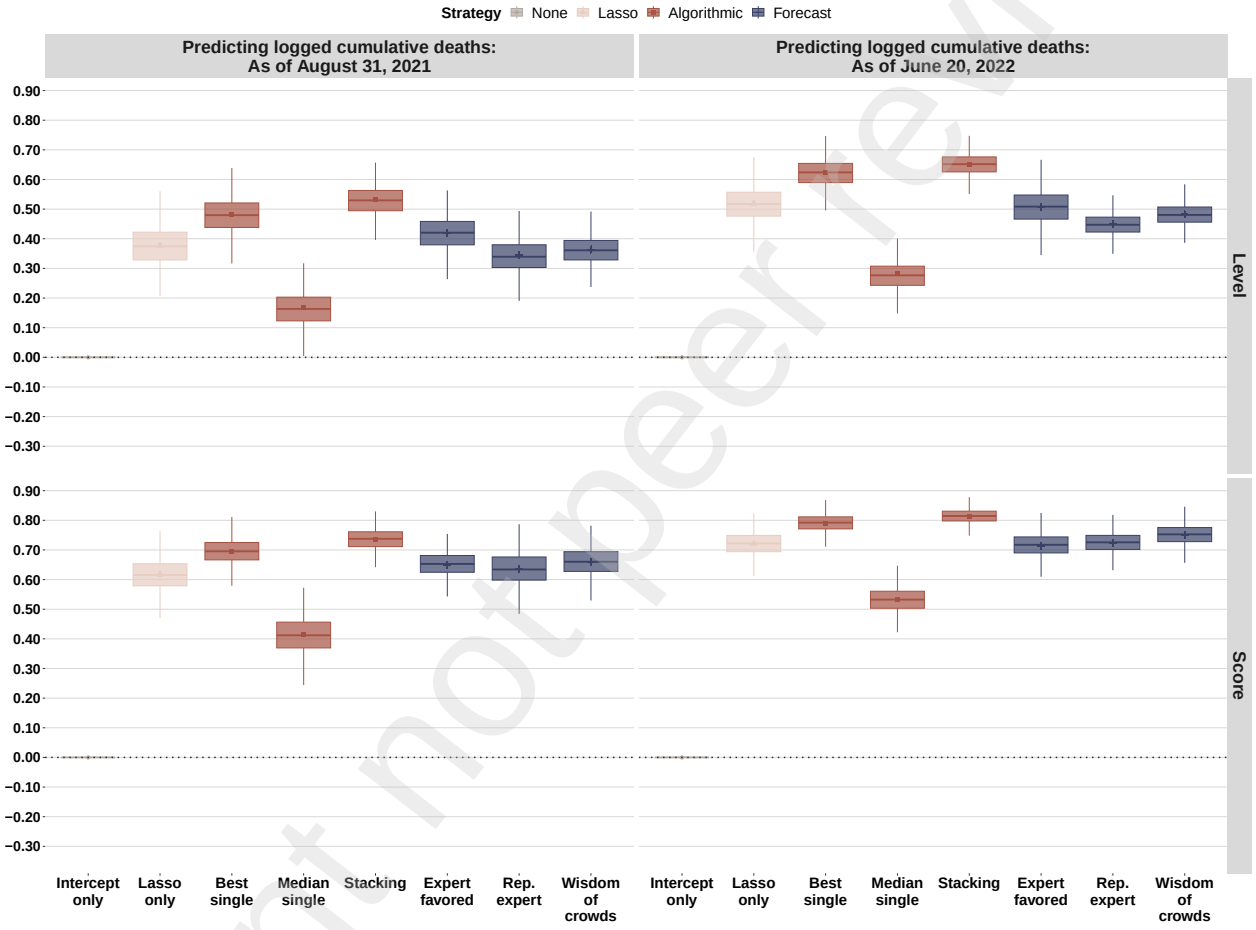
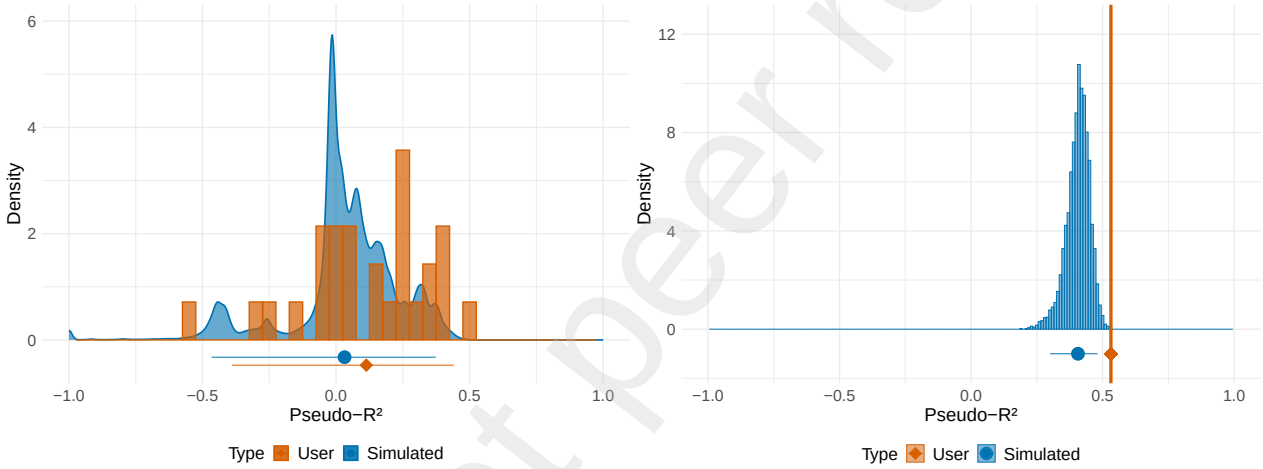


Figure 4: Comparison of predictive accuracy using different metrics. The top row of boxplots shows predictive performances (pseudo-R^2) of cumulative COVID-19 deaths per million and the bottom row shows correlations between predictions and actual mortality. The left column of boxplots assesses predictive accuracy on cumulative mortality as of August 31, 2021. The right column evaluates out-of-sample predictions of cumulative mortality through June 2022 using the models selected on the basis of the August 31, 2021 data. Each boxplot shows the interquartile range; whiskers are two standard deviations above and below. Interquartile ranges and 95% confidence intervals are generated by bootstrapping.

Discussion

The tools we have employed allow us to take stock of the predictive power of different methods in contexts where there are many explanations for a single outcome.

When we evaluate models submitted to the Model Challenges, we find that the more successful theoretically-motivated models out-perform a Lasso-generated model. However, the performance of the "typical" model — that is, the model exhibiting median performance — is far worse than Lasso. In Panel (a) of Figure 5, we compare the user-submitted models to a random sample of 130,000 $(5,000 \times 26)$ models generated from the MC-provided data and Shiny application. Specifically, we randomly sample permutations of three-predictor models from the MC-provided data and then randomly select the functional form of the model (linear, quadratic, or with interactions). Among the quadratic and interaction models, we randomly select which parameters are included in the model (see Appendix S6 for our sampling algorithm). Results show that the strongest of the submitted models clearly fall in the top percentiles of all possible models; thus, eliciting models from experts provides an advantage over any algorithmic production of models. However, many weaker submitted models do not perform well relative to the distribution of all possible models. In Panel (b), we compare the stacking prediction to stacking predictions generated from the identical 5000 random samples of 26 random three-predictor models generated from the crossnational MC data. The stacking prediction outperforms all of the predictions from a "null" distribution of stacking models ($p = 0$). By aggregating expert models via stacking, we can substantially enhance the predictive performance of a set of models. Implementing an ensemble algorithm to combine features of multiple models adds predictive value, documenting the utility of aggregation, and specifically of algorithmic aggregation.



(a) Horserace R^2 distribution.

(b) Stacking R^2 distribution.

Figure 5: Panel (a) depicts the observed distribution of pseudo-R^2's and the density plot depicts the distribution of pseudo-R^2's from our sample of 130,000 linear, quadratic, and interactive three-predictor models using the common MC dataset. Panel (b) shows how the predictive performance of the stacking model compares to the predictive performance of randomly generated stacking models. All models are crossnational general models.

The best models submitted to the MCs outperform a Lasso machine-learning benchmark and are over-represented among top performers in the simulation. To a skeptic of social science, it may not be obvious that social scientists are capable of generating highly predictive models; that, in other words, they exhibit an ability to accurately explain the social world. However, the sharp drop-off in performance between the best and median models, combined with experts' limited abilities to accurately identify the best-performing models (Figure 3), is cause for concern. If the development of knowledge depends on the abilities of experts to assess the merits of multiple empirically-supported claims, social scientists should address issues of aggregation more systematically.

That experts do far less well than algorithms in model evaluation and aggregation is perhaps surprising. Success in combining intuitions generated by many scholars to explain a common outcome is often viewed as a subtle art requiring deep expertise and insight. We show that a statistical algorithm in fact performs better at this in our context. Our findings suggest that scholars could profitably devote more resources to systematizing the models characterizing their explanations so that these can be aggregated using statistical methods. Ensemble procedures seem likely to produce more credible meta-models than informal reasoning; social scientists do better when their expertise is combined than almost all of them do alone.

While the specific MCs that we implemented were made possible by early and sustained attention to a new outcome of

interest to social scientists — COVID-19 mortality — several features of our procedures may be worth replicating in more established social science literatures. In particular, there is a need to evaluate competing theories on common samples using common measures of an outcome. The algorithmic tools that we employ — model comparison based on predictive power and model stacking to generate an aggregate prediction — can easily be implemented in such settings. These forms of model assessment and combination harness the aggregate inputs of social scientists, documenting the strength of collective over individual knowledge.

Methods and Materials

Evaluating model performance

For a full description of the MCs, see S1. For a full description of statistical methods and quantities of interest, see S2. Our algorithmic measures of model performance used in the tables and figures are constructed as follows. In the following, let $\boldsymbol{x_i}$ denote a vector of explanatory variables for unit i and $\hat{f}^k_{-i}$ the predictive model k trained on data that excludes unit i. Then the leave-one-out prediction for unit i under model k is $\widehat{y}_{ik} = \hat{f}^k_{-i}(\boldsymbol{x_i})$.

For Figure 2 and the "level"-based model summaries in Figure 4, we measure the pseudo-R^2 using Equation (1). For the "score"-based model summaries in Figure 4, we measure the correlation using Equation (2). In these equations, y_i is the observed outcome in unit i. For the general models, the out-of-sample prediction, $\widehat{y}_{ik}$, for unit i under model k is given above. The Z superscript indicates to a z-score transformation of the outcome, y_i, or the prediction for model k, $\widehat{y}_{ik}$.

$$\text{Pseudo } R^2 = 1 - \frac{\sum_{i=1}^{N}(\widehat{y}_{ik} - y_i)^2}{\sum_{i=1}^{N}(\overline{y}_i - y_i)^2} \tag{1}$$

$$\text{Correlation} = 1 - \frac{\sum_{i=1}^{N}(\widehat{y}^Z_{ik} - y^Z_i)^2}{2\sum_{i=1}^{N}(\overline{y}^Z_i - y^Z_i)^2} \tag{2}$$

Stacking model

The stacking estimator takes the (leave-one-out) out-of-sample predictions of each model as inputs. It identifies the optimal weighting of these predictions, and selects a vector of non-negative weights summing to 1, w, to minimize the loss function:

$$L(w) = \sum_{i=1}^{N}\left(y_i - \sum_{k=1}^{K} w_k \widehat{y}_{ik}\right)^2$$

Intuitively, a vector of weights w placed on models, results in an aggregated prediction for the unit $y_i^{\text{stacking}}(w) = \sum_k w_k \widehat{y}_{ik}$ and loss is assessed by how far the vector $y^{\text{stacking}}(w)$ is from the observed outcomes y. We estimate the stacking weights employed in Figures 3 and 4 using Equation (3). A detailed description of the procedure appears in S2.3.

$$w = \arg\min_{w} \sum_{i=1}^{N}\left(y_i - \sum_{k=1}^{K} w_k \widehat{y}_{ik}\right)^2 \text{ s.t. } w_k \geq 0 \;\forall\; k, \sum_{k=1}^{K} w_k = 1 \tag{3}$$

References

1. Y. Yao, A. Vehtari, D. Simpson, A. Gelman, Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Analysis*. **13**, 917–1007 (2021).

2. E. Y. Cramer, E. L. Ray, V. K. Lopez, J. Bracher, A. Brennen, A. J. Castro Rivadeneira, A. Gerding, T. Gneiting, K. H. House, Y. Huang, others, Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the united states. *Proceedings of the National Academy of Sciences*. **119**, e2113561119 (2022).

3. D. J. Watts, Should social science be more solution-oriented? *Nature Human Behavior*. **1**, 1–4 (2017).

4. G. F. Davis, Editorial essay: What is organizational research for? *Administrative Science Quarterly*. **60**, 179–188 (2015).

5. A. van de Rijt, S. M. Kang, M. Restivo, A. Patil, Field experiments of success-breeds-success dynamics. *Proceedings of the National Academy of Sciences*. **111**, 6934–939 (2014).

6. Banerjee Abhijit, E. Duflo, N. Goldberg, D. Karlan, R. Osei, W. Parrienté, J. Shapiro, B. Thuysbaert, C. Udry, A multifaceted program causes lasting progress for the very poor: Evidence from six countries. *Science*. **348** (2015).

7. T. Dunning, G. Grossman, M. Humphreys, S. D. Hyde, C. Mcintosh, G. Nellis, Claire L. Adida, E. Arias, C. Bicalho, T. C. Boas, M. T. Buntaine, S. Chauchard, A. Chowdhury, J. Gottlieb, D. F. Hidalgo, M. Holmlund, R. Jablonski, E. Kramon, H. Larreguy, M. Lierl, J. Marshall, G. McClendon, M. A. Melo, D. L. Nielson, P. M. Pickering, M. R. Platas, P. QuerubÍn, P. Raffler, N. Sircar, Voter information campaigns and political accountability: Cumulative findings from a preregistered meta-analysis of coordinated trials. *Science Advances*. **5** (2019).

8. A. Coppock, S. J. Hill, L. Vavreck, The small effects of political advertising are small regardless of context, message, sender, or receiver: Evidence from 59 real-time randomized experiments. *Science Advances*. **6**, 1–6 (2020).

9. T. Slough, D. Rubenson, R. Levy, F. A. Rodriguez, M. B. del Carpio, M. T. Buntaine, D. Christensen, A. Cooperman, S. Eisenbarth, P. J. Ferraro, L. Graham, A. C. Hardman, J. Kopas, S. McLarty, A. S. Rigterink, C. Samii, B. Seim, J. Urpelainen, B. Zhang, Adoption of community monitoring improves common pool resource management across contexts. *Proceedings of the National Academy of Sciences*. **10.1073**, 1–10 (2021).

10. G. Blair, J. M. Weinstein, F. Christia, E. Arias, E. Badran, R. A. Blair, A. Cheema, A. Farooqui, T. Fetzer, G. Grossman, D. Haim, Z. Hameed, R. Hanson, A. Hasanain, D. Kronick, B. S. Morse, R. Muggah, F. Nadeem, L. L. Tsai, M. Nanes, T. Slough, N. Ravanilla, J. N. Shapiro, B. Silva, P. C. L. Souza, A. M. Wilke, Community policing does not build citizen trust in police or reduce crime in the global south. *Science*. **374**, eabd3446 (2021).

11. T. Slough, S. A. Tyson, External validity and meta-analysis. *American Journal of Political Science*. **Forthcoming** (2022).

12. J. M. Montgomery, F. M. Hollenbach, M. D. Ward, Improving prediction using ensemble Bayesian model averaging. *Political Analysis*. **20**, 271–91 (2012).

13. D. H. Wolpert, Stacked generalization. *Neural networks*. **5**, 241–259 (1992).

14. A. Acharya, J. Gerring, A. Reeves, Is health politically irrelevant? Experimental evidence during a global pandemic. *BMJ global health*. **5**, e004222 (2020).

15. O. Bargain, U. Aminjonov, Trust and compliance to public health policies in times of COVID-19. *Journal of Public Economics*. **192**, 104316 (2020).

16. F. J. Elgar, A. Stefaniak, M. J. A. Wohl, The trouble with trust: Time-series analysis of social capital, income inequality, and COVID-19 deaths in 84 countries. *Social Science & Medicine*. **263**, 113365 (2020).

17. Q. Han, B. Zheng, M. Cristea, M. Agostini, J. J. Belanger, B. Gutzkow, J. Kreienkamp, PsyCorona Collaboration, N. P. Leander, 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. *Psychological Medicine*, 1–32 (2021).

18. J. Min, Does social trust slow down or speed up the transmission of COVID-19? *PLoS ONE*. **15**, e0244273 (2020).

19. A. R. Piquero, W. G. Jennings, E. Jemison, C. Kaukinen, F. M. Knaul, Domestic violence during the COVID-19 pandemic – Evidence from a systematic review and meta-analysis. *Journal of Criminal Justice*. **74**, 101806 (2021).

20. E. Robinson, A. Jones, I. Lesser, M. Daly, International estimates of intended uptake and refusal of COVID-19 vaccines: A rapid systematic review and meta-analysis of large nationally representative samples. *Vaccine*. **39**, 2024–2034 (2021).

21. J. Bennett, S. Lanning, "The Netflix prize" in *Proceedings of KDD Cup and Workshop 2007* (2007), pp. 1–6.

22. M. J. Salganik, I. Lundberg, A. T. Kindel, C. E. Ahearn, K. Al-Ghoneim, A. Almaatouq, D. M. Altschul, J. E. Brand, N. B. Carnegie, R. J. Compton, D. Datta, T. Davidson, A. Filippova, C. Gilroy, B. J. Goode, E. Jahani, R. Kashyap, A. Kirchner, S. McKay, A. C. Morgan, A. Pentland, K. Polimis, L. Raes, D. E. Rigobon, C. V. Roberts, D. M. Stanescu, Y. Suhara, A. Usmani, E. H. Wang, M. Adem, A. Alhajri, B. AlShebli, R. Amin, R. B. Amos, L. P. Argyle, L. Baer-Bositis, M. Büchi, B.-R. Chung, W. Eggert, G. Faletto, Z. Fan, J. Freese, T. Gadgil, J. Gagné, Y. Gao, A. Halpern-Manners, S. P. Hashim, S. Hausen, G. He, K. Higuera, B. Hogan, I. M. Horwitz, L. M. Hummel, N. Jain, K. Jin, D. Jurgens, P. Kaminski, A. Karapetyan, E. H. Kim, B. Leizman, N. Liu, M. Möser, A. E. Mack, M. Mahajan, N. Mandell, H. Marahrens, D. Mercado-Garcia, V. Mocz, K. Mueller-Gastell, A. Musse, Q. Niu, W. Nowak, H. Omidvar, A. Or, K. Ouyang, K. M. Pinto, E. Porter, K. E. Porter, C. Qian, T. Rauf, A. Sargsyan, T. Schaffner, L. Schnabel, B. Schonfeld, B. Sender, J. D. Tang, E. Tsurkov, A. van Loon, O. Varol, X. Wang, Z. Wang, J. Wang, F. Wang, S. Weissman, K. Whitaker, M. K. Wolters, W. L. Woon, J. Wu, C. Wu, K. Yang, J. Yin, B. Zhao, C. Zhu, J. Brooks-Gunn, B. E. Engelhardt, M. Hardt, D. Knox, K. Levy, A. Narayanan, B. M. Stewart, D. J. Watts, S. McLanahan, Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*. **117**, 8398–8403 (2020).

23. R. Tibshirani, Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*. **58**, 267–288 (1996).

24. S. DellaVigna, D. Pope, Predicting experimental results: Who knows what? *Journal of Political Economy*. **126**, 2410–2456 (2018).

25. S. DellaVigna, D. Pope, E. Vivalt, Predict science to improve science. *Science*. **366**, 428–429 (2019).

26. S. E. Bokemper, G. A. Huber, A. S. Gerber, E. K. James, S. B. Omer, Timing of COVID-19 vaccine approval and endorsement by public figures. *Vaccine*. **39**, 825–829 (2021).

27. N. Basta, E. Moodie, "COVID-19 vaccine tracker" (2021).

28. WHO, "Draft landscape and tracker of COVID-19 candidate vaccines" (2021).

29. O. J. Wouters, K. C. Shadlen, M. Salcher-Konrad, A. J. Pollard, H. J. Larson, Y. Teerawattananon, M. Jit, Challenges in ensuring global access to COVID-19 vaccines: Production, affordability, allocation, and deployment. *The Lancet*. **397**, 1023–1034 (2021).

30. L. R. Baden, H. M. El Sahly, B. Essink, K. Kotloff, S. Frey, R. Novak, D. Diemert, S. A. Spector, N. Rouphael, C. B. Creech, J. McGettigan, S. Khetan, N. Segall, J. Solis, A. Brosz, C. Fierro, H. Schwartz, K. Neuzil, L. Corey, P. Gilbert, H. Janes, D. Follmann, M. Marovich, J. Mascola, L. Polakowski, J. Ledgerwood, B. S. Graham, H. Bennett, R. Pajon, C. Knightly, B. Leav, W. Deng, H. Zhou, S. Han, M. Ivarsson, J. Miller, T. Zaks, Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine. *New England Journal of Medicine*. **384**, 403–416 (2021).

31. P. M. Folegatti, K. J. Ewer, P. K. Aley, B. Angus, S. Becker, S. Belij-Rammerstorfer, D. Bellamy, S. Bibi, M. Bittaye, E. A. Clutterbuck, C. Dold, S. N. Faust, A. Finn, A. L. Flaxman, B. Hallis, P. Heath, D. Jenkin, R. Lazarus, R. Makinson, A. M. Minassian, K. M. Pollock, M. Ramasamy, H. Robinson, M. Snape, R. Tarrant, M. Voysey, C. Green, A. D. Douglas, A. V. S. Hill, T. Lambe, S. C. Gilbert, A. J. Pollard, Safety and immunogenicity of the ChAdOx1 nCoV-19 vaccine against SARS-CoV-2: A preliminary report of a phase 1/2, single-blind, randomised controlled trial. *The Lancet*. **396**, 1979–1993 (2021).

32. D. Y. Logunov, I. V. Dolzhikova, D. V. Shcheblyakov, A. I. Tukhvatulin, O. V. Zubkova, A. S. Dzharullaeva, A. V. Kovyrshina, N. L. Lubenets, D. M. Grousova, A. S. Erokhova, A. G. Botikov, F. M. Izhaeva, O. Popova, T. A. Ozharovskaya, I. B. Esmagambetov, I. A. Favorskaya, D. I. Zrelkin, D. V. Voronina, D. N. Shcherbinin, A. S. Semikhin, Y. V. Simakova, E. A. Tokarskaya, D. A. Egorova, M. M. Shmarov, N. A. Nikitenko, V. A. Gushchin, E. A. Smolyarchuk, S. K. Zyryanov, S. V. Borisevich, B. S. Naroditsky, A. L. Gintsburg, Safety and efficacy of an rAd26 and rAd5 vector-based heterologous prime-boost COVID-19 vaccine: An interim analysis of a randomised controlled phase 3 trial in Russia. *The Lancet*. **397**, 671–681 (2021).

33. M. J. Mulligan, K. E. Lyke, N. Kitchin, J. Absalon, A. Gurtman, S. Lockhart, K. Neuzil, V. Raabe, R. Bailey, K. A. Swanson, P. Li, K. Koury, W. Kalina, D. Cooper, C. Fontes-Garfias, P.-Y. Shi, Ö. Türeci, K. R. Tompkins, E. E. Walsh, R. Frenck, A. R. Falsey, P. R. Dormitzer, W. C. Gruber, U. Şahin, K. U. Jansen, Phase i/II study of COVID-19 RNA vaccine BNT162b1 in adults. *Nature*. **586**, 589–593 (2021).

34. F. P. Polack, S. J. Thomas, N. Kitchin, J. Absalon, A. Gurtman, S. Lockhart, J. L. Perez, G. Pérez Marc, E. D. Moreira, C. Zerbini, R. Bailey, K. A. Swanson, S. Roychoudhury, K. Koury, P. Li, W. V. Kalina, D. Cooper, R. W. Jr. Frenck, L. L. Hammitt, Ö. Türeci, H. Nell, A. Schaefer, S. Ünal, D. B. Tresnan, S. Mather, P. R. Dormitzer, U. Şahin, K. U. Jansen, W. C. Gruber, Safety and efficacy of the BNT162b2 mRNA covid-19 vaccine. *New England Journal of Medicine*. **385**, 1761–1773 (2021).

35. M. Voysey, S. A. C. Clemens, S. A. Madhi, L. Y. Weckx, P. M. Folegatti, P. K. Aley, B. Angus, V. L. Baillie, S. L. Barnabas, Q. E. Bhorat, S. Bibi, C. Briner, P. Cicconi, A. M. Collins, R. Colin-Jones, C. L. Cutland, T. C. Darton, K. Dheda, C. J. A. Duncan, K. R. W. Emary, K. J. Ewer, L. Fairlie, S. N. Faust, S. Feng, D. M. Ferreira, A. Finn, A. L. Goodman, C. M. Green, C. A. Green, P. T. Heath, C. Hill, H. Hill, I. Hirsch, S. H. C. Hodgson, A. Izu, S. Jackson, D. Jenkin, C. C. D. Joe, S. Kerridge, A. Koen, G. Kwatra, R. Lazarus, A. M. Lawrie, A. Lelliott, V. Libri, P. J. Lillie, R. Mallory, A. V. A. Mendes, E. P. Milan, A. M. Minassian, A. McGregor, H. Morrison, Y. F. Mujadidi, A. Nana, P. J. O'Reilly, S. D. Padayachee, A. Pittella, E. Plested, K. M. Pollock, M. N. Ramasamy, S. Rhead, A. V. Schwarzbold, N. Singh, A. Smith, R. Song, M. D. Snape, E. Sprinz, R. K. Sutherland, R. Tarrant, E. C. Thomson, M. E. Török, M. Toshner, D. P. J. Turner, J. Vekemans, T. L. Villafana, M. E. E. Watson, C. J. Williams, A. D. Douglas, A. V. S. Hill, T. Lambe, S. C. Gilbert, A. J. Pollard, Safety and efficacy of the ChAdOx1 nCoV-19 vaccine (AZD1222) against SARS-CoV-2: An interim analysis of four randomised controlled trials in Brazil, South Africa, and the UK. *The Lancet*. **397**, 99–111 (2021).

36. A. de Figueiredo, C. Simas, E. Karafillakis, P. Paterson, H. J. Larson, Mapping global trends in vaccine confidence and investigating barriers to vaccine uptake: A large-scale retrospective temporal modelling study. *The Lancet*. **396** (2021).

37. J. V. Lazarus, S. C. Ratzan, A. Palayew, L. O. Gostin, H. J. Larson, K. Rabin, S. Kimball, A. El-Mohandes, A global survey of potential acceptance of a COVID-19 vaccine. *Nature medicine*. **27**, 225–228 (2021).

38. J. S. Solís Arce, S. S. Warren, N. F. Meriggi, A. Scacco, N. McMurry, M. Voors, G. Syunyaev, A. A. Malik, S. Aboutajdine, O. Adeojo, D. Anigo, A. Armand, S. Asad, M. Atyera, B. Augsburg, M. Awasthi, G. E. Ayesiga, A. Bancalari, M. B. Nyqvist, E. Borisova, C. M. Bosancianu, M. R. C. García, A. Cheema, E. Collins, F. Cuccaro, A. Z. Farooqi, T. Fatima, M. Fracchia, M. L. G. Soria, A. Guariso, A. Hasanain, S. Jaramillo, S. Kallon, A. Kamwesigye, A. Kharel, S. Kreps, M. Levine, R. Littman, M. Malik, G. Manirabaruta, J. L. H. Mfura, F. Momoh, A. Mucauque, I. Mussa, J. A. Nsabimana, I. Obara, M. J. Otálora, B. W. Ouédraogo, T. B. Pare, M. R. Platas, L. Polanco, J. A. Qureshi, M. Raheem, V. Ramakrishna, I. Rendrá, T. Shah, S. E. Shaked, J. N. Shapiro, J. Svensson, A. Tariq, A. M. Tchibozo, H. A. Tiwana, B. Trivedi, C. Vernot, P. C. Vicente, L. B. Weissinger, B. Zafar, B. Zhang, D. Karlan, M. Callen, M. Teachout, M. Humphreys, A. M. Mobarak, S. B. Omer, COVID-19 vaccine acceptance and hesitancy in low- and middle-income countries. *Nature medicine*. **27**, 1385–1394 (2021).

39. T. Jones, A coefficient of determination for probabilistic topic models (2019), doi:10.48550/ARXIV.1911.11061.

Acknowledgments

We thank Rens Chazottes and Julian Vierlinger for expert research assistance. We are grateful to P. M. Aronow, Jasper Cooper, Alex Coppock, Chad Hazlett, Kosuke Imai, Maggie Penn, and Cyrus Samii for comments on an early version of the Model Challenge research design and Evidence in Government and Politics for the opportunity to present a draft

of results. **Funding:** We thank the European University Institute and the Tinker Emergency Fund from the Center for Latin American Studies at Stanford University. For infrastructure support, we thank the Wissenschaftszentrum Berlin für Sozialforschung. **Authors contributions:** MG performed Conceptualization, Data curation, Funding acquisition, Investigation, Project administration, Resources, Supervision, Visualization, Writing – review & editing; TS performed Conceptualization, Formal Analysis, Investigation, Methodology, Supervision, Validation, Visualization, Writing – original draft, Writing - review & editing; HZ performed Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Resources, Software, Validation, Visualization; AS performed Conceptualization, Data curation, Investigation, Project administration, Resources, Supervision, Visualization, Writing – review & editing; MH performed Conceptualization, Formal Analysis, Investigation, Methodology, Resources, Validation, Visualization, Writing – review & editing; EV performed Conceptualization, Investigation, Methodology, Resources, Software, Supervision; ADC performed Conceptualization, Data curation, Funding acquisition, Resources, Validation; KD performed Conceptualization, Data curation, Investigation, Resources; SK performed Data curation, Investigation, Software, Resources; EN performed Data curation, Software; PA performed Investigation, Resources, Methodology; JB, AB, JB, SD, RF, SH, LR, JM, AR, MT, FT, and SU performed Investigation, Resources; WH performed Investigation, Methodology; and JE, CF, NH, JJ, DH, MK, NM, and JSA performed Investigation. **Competing interests:** None of the authors has any competing interests. **Data and materials availability:** The codebooks are available at Open Science Framework, where the data will be deposited once the paper is accepted for publication.

Supplementary materials

Table of Contents

S1 Research Design

In this project, we study the aggregation of social scientific knowledge. We study aggregation in the context of social scientists' predictions about the trajectory of deaths during the COVID-19 pandemic. We created the COVID-19 Model Challenge to emulate two stages of the development of broader social scientific research agendas:

1. **Model generation**: We invited researchers to develop models that use social and political variables to predict cumulative COVID-19 deaths, as measured by logged deaths per million, on of August 31, 2021. We asked researchers to include verbal explanations for why selected socio-political variables would predict COVID-19 mortality. We created four challenges: (1) across countries; (2) across states in the USA; (3) across Mexican states; and (4) across Indian states.

   Researchers contributed models of COVID-19 mortality between December 1, 2020 and January 20, 2021. When making predictions, participants were provided cumulative COVID-19 mortality rates as of November 16, 2020. We refer to the researchers who submitted models as *modelers*.

2. **Model assessment by other researchers**: We invited social scientists to assess the predictive capability of the models amassed in stage one. Forecasters were asked to evaluate the predictive performance of models as of August 31, 2021 and August 31, 2022.

   Forecasters evaluated models on the Social Science Prediction Platform during May 2021. To aid their assessments, we provided predictive metrics for each model as of February 2021.

We depict the sequence of the research design in Figure S1.

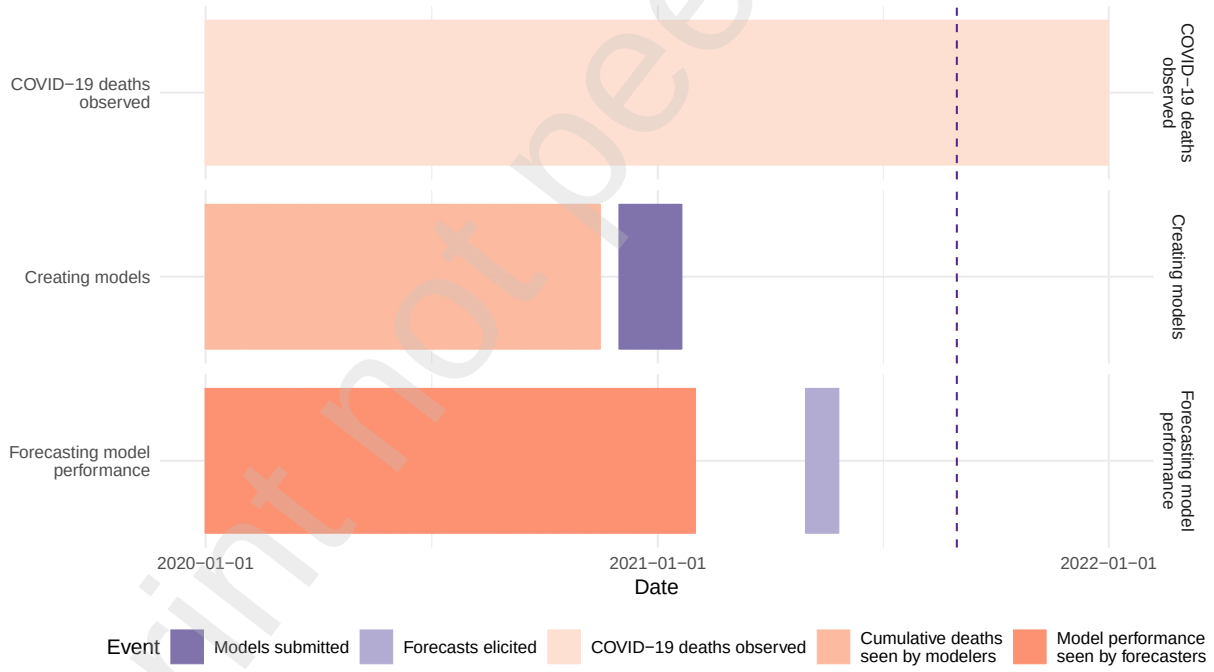


Figure S1: A schematic depiction of our research timeline.

S1.1 The Outcome: Cumulative Covid Mortality

The outcome for all challenges is logged COVID-19 deaths per million residents on August 31, 2021. We collect COVID-19 outcome data from the following sources:

- **Crossnational challenge**: European Centre for Disease Prevention and Control (ECDC), accessed November 16, 2020; March 3, 2022; and October 18, 2022.

- **India challenge**: Government of India (https://www.mygov.in/corona-data/covid19-statewise-status/), accessed November 16, 2020; March 23, 2021; September 7, 2021; March 2, 2022; and October 18, 2022.
- **Mexico challenge**: Government of Mexico (https://coronavirus.gob.mx/datos/#DownZCSV), accessed November 16, 2020; March 23, 2021; September 8, 2021; March 4, 2022; and October 2, 2022.
- **United States challenge**: The COVID Tracking Project at *The Atlantic* (https:/covidtracking.com/data/download/all-states-history.csv), accessed November 16, 2020 and March 23, 2021; The COVID-19 Response at the Centers for Disease Control and Prevention (CDC) (https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36), accessed September 14, 2021; March 3, 2022; and October 18, 2022.

When participants entered the COVID-19 model challenge, modelers had access cumulative COVID-19 mortality data as of November 16, 2020. They were asked to predict cumulative mortality as of August 31, 2021. Figure S2 shows our outcome measure for the crossnational challenge. The left panel shows the evolution of logged deaths per million. The vertical lines denote the data shown to modelers during the Model Challenge and the date at which we evaluate predictions (August 31, 2021). Each line represents a country. To illustrate the changes in COVID-19 mortality that participants predicted, we depict the three countries at the 10^{th}, 50^{th}, and 90^{th} percentiles in (percent) change in COVID-19 mortality between November 16, 2020 and August 30, 2021. The countries are Spain, Romania, and Uganda, respectively.

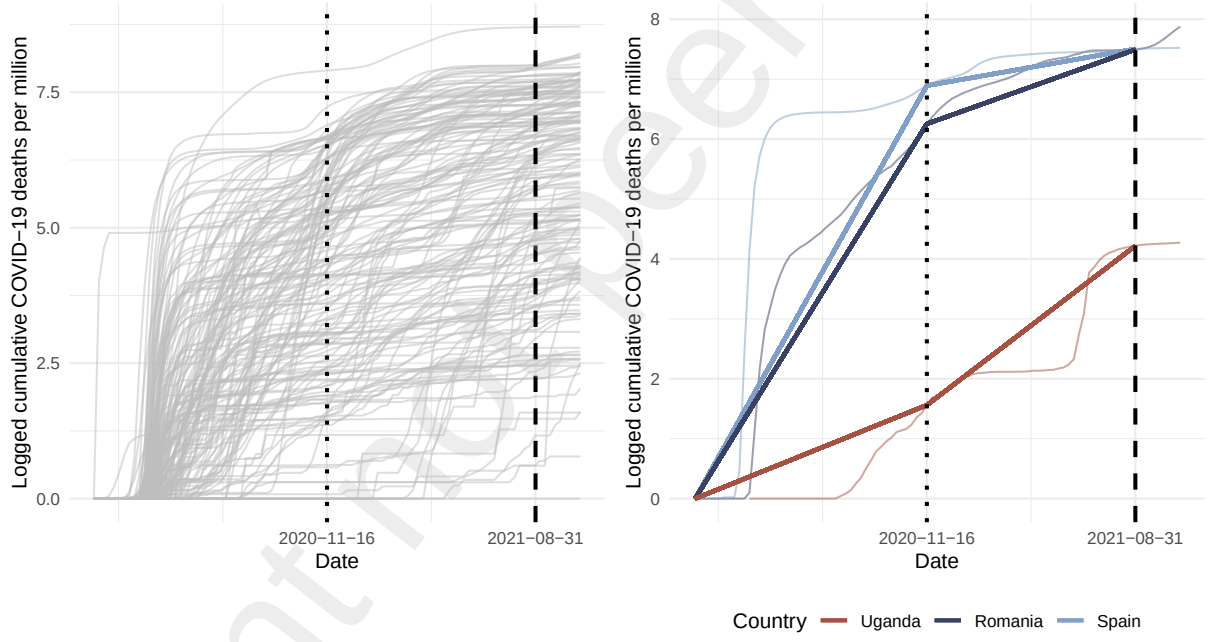


Figure S2: Outcome data for the crossnational challenge. The left panel depicts the growth logged cumulative COVID-19 deaths per million. Each line represents a country. Vertical lines reflect the data provided in the MC and the main outcome, mortality as of August 31, 2021. The right panel shows countries at the first decile (Spain), median (Romania), and top decile (Uganda) in terms of change in the outcome between November 16, 2020 and August 31, 2021.

Outcome data for the other tasks is analogous. Table S1 reports summary statistics for the outcome — logged cumulative COVID-19 deaths per million — for each challenge. Unsurprisingly, there is greater between- than within-country variation.

Challenge	# of obs.	Logged cumulative COVID-19 deaths per million, as of August 31, 2021				
		Median	Mean	Minimum	Maximum	Std. Dev.
Crossnational	166	6.12	5.64	0	8.73	1.87
India	31	5.87	5.87	4.5	7.7	0.84
Mexico	32	7.62	6.55	5.82	8.59	0.41
USA	50	7.56	7.41	6.03	8.02	0.47

Table S1: Summary statistics for our outcome measure by challenge. We add 1 to our cases per million prior to logging, such that 0 is interpretable as no deaths. (There were no reported COVID-19 deaths in the Solomon Islands as of August 31, 2021.)

S1.2 Model Challenges Overview

Between December 2020 and January 20, 2021, we solicited statistical models from social scientists, asking them to predict cumulative numbers of COVID-19 deaths per million as of August 31, 2021. To provide context, the period was one when questions about vaccine availability (*26–29*), efficacy beyond clinical trials (*30–35*), and public willingness to accept vaccination (*36–38*) were particularly salient. New variants (including Delta) emerged only after predictions had been made. Thus, uncertainty over the trajectory of COVID-19 pandemic at the time of the challenges complicated the task for participants of making out-of-sample predictions of mortality.

Individuals or teams were encouraged to submit models to a website showing the cumulative number of COVID-19 deaths as of November 16, 2020 as well as data we had assembled on many possible predictors, including measures of state capacity, political priorities, political institutions, and social structures. (See https://osf.io/pgydn.) Submission of additional predictors was also permitted. The interface let users provide models to predict mortality across countries (global challenge) or across states (national challenges) in India, Mexico, and the United States.

The platform was publicly available. We advertised to social scientists through social media (Twitter), via professional listservs (the American Political Science Association, the European Political Science Association, the Society for Political Methodology, Evidence in Governance and Politics, and others). In addition, we sent individual emails directed at researchers at the top 100 research institutions globally as well as specifically in the USA, Mexico, and India.

The interface, depicted in Figure S3, allowed researchers to:

1. Choose a model challenge to enter — crossnational, India, Mexico, or USA (see Figure S3b).
2. Select up to three predictors and see the performance of a linear bivariate model that uses each predictor on COVID-19 mortality data as of November 16, 2020 (see Figure S3b).
3. Optionally upload new regressors not already available in our data repository (see Figure S3b).
4. Optionally change functional form of the models to allow interaction, polynomial, or custom model submissions (see Figure S3c).
5. Optionally predict parameter values for models, enabling submission of "parameterized models" (see Figure S3d).
6. Provide a logic to explain the model (required). We encouraged researchers to describe why the set of predictors they chose mattered for the outcome, with references to relevant literatures (see Figure S3e).
7. Enter the Model Challenge by submitting (a) model(s) (see Figure S3f).

As participants developed their models, they could explore how their models performed on “current data” (cumulative mortality counts up to 16 November 2020). They could also examine bivariate plots representing the relationship between each of their chosen predictors and the outcome variable (logged cumulative deaths per million). We report the codebooks that were available on the interface in https://osf.io/87mku. These codebooks provided information on the definition of and data sources for all predictors.

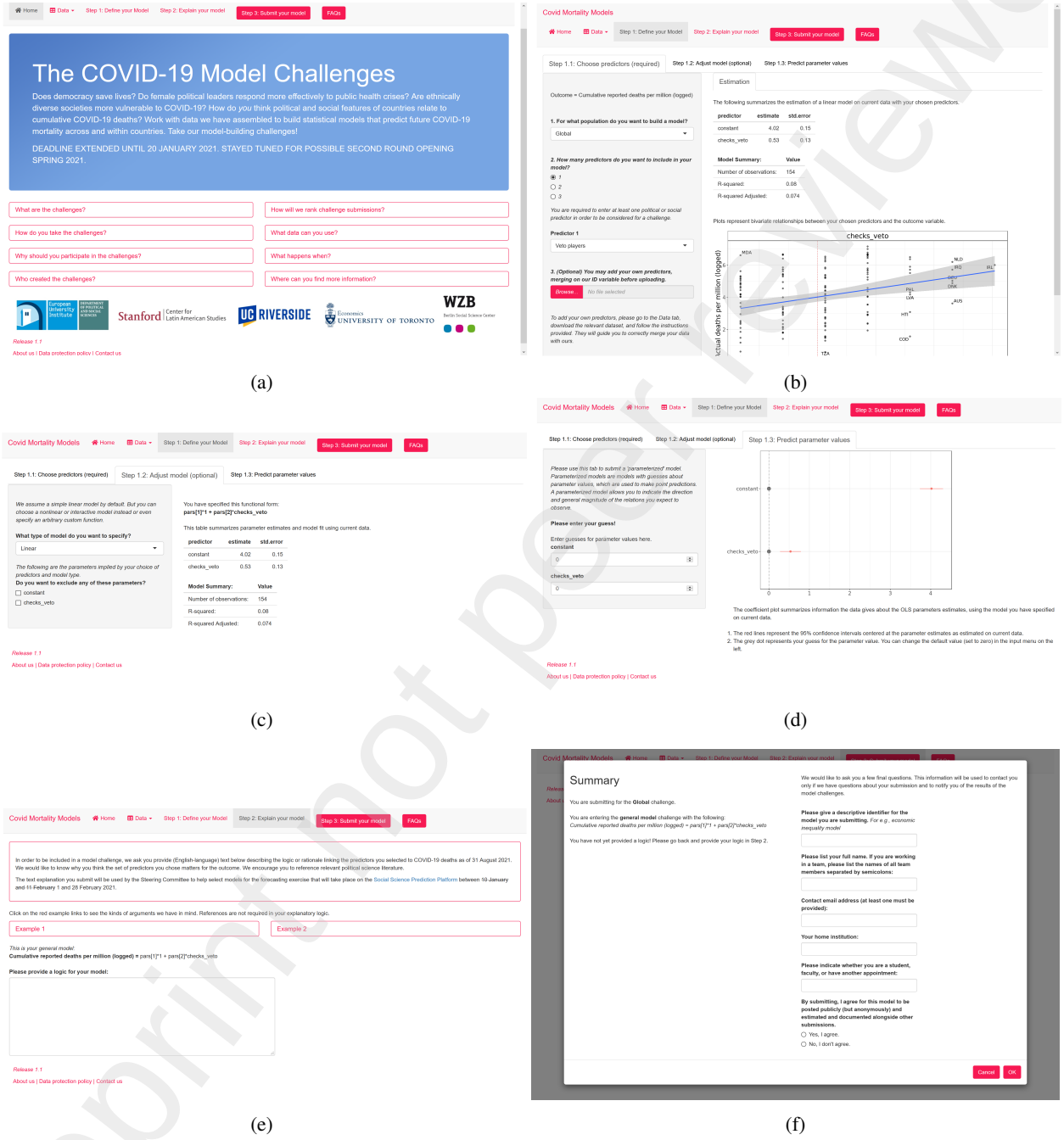


Figure S3: Screenshots from the MC interface. Plots and reported statistics were presented dynamically.

S1.3 Classifying Models

S1.3.1 General and Parameterized Models

We distinguish between general and parameterized models throughout our analyses. All models, indexed by k, take the form:

$$y_{ik} = f(\mathbf{x}_i, \boldsymbol{\theta}_k).$$

In our setup, a model is defined by the predictors it includes, $\mathbf{x}_i$, and its parameters, $\boldsymbol{\theta}_k$. We call a model "general" if its parameters, $\boldsymbol{\theta}_k$, are estimated from the data (as of August 31, 2021). We call a model "parameterized" if its parameters were specified as part of the predictive model. Some of the parameterized models were fit on the COVID-19 mortality data provided at baseline.

S1.3.2 Theory-driven and Machine Learning Models

We further distinguish between "theory-driven" and "Machine Learning" (ML) models. In theory-driven models, modelers submitted predictors along with an argument or logic for why these variables might predict COVID-19 mortality. In ML models, modelers used some automated process or algorithm to select predictors and or/the functional form of the model. The Model Challenge encouraged submission of theory-driven models. We received substantially more theory-driven than ML models.

S1.3.3 Benchmark Models

In addition to user-submitted models, we analyze two other predictive models for each challenge: a model with standard epidemiological predictors (denoted the "usual suspects model") and a model with predictors selected by Lasso ("Lasso model"). The usual suspects model allows us to assess the additional explanatory power of social and political variables beyond basic epidemiological predictions.

The usual suspects models for each challenge are reported in Table S2 and the Lasso models in Table S3.

In total, we received 88 distinct model submissions. Table S4 reports the breakdown of submissions for each of the four challenges (crossnational, India, Mexico, and the USA) disaggregated by model type (general or parameterized). For ten of the models submitted, participants also uploaded their own data. Collectively, the two types of additional models (Lasso and usual suspects) take both a general and parameterized form for each of the four challenges, yielding an additional 16 models. Thus, Table S4 includes 104 models: 88 submissions and 16 benchmark models.

S1.4 Forecasting Details

In the forecasting stage of this project, we used the Social Science Prediction Platform to elicit expert assessments of models. Forecasting took place in May 2021. We recruited subjects through the platform, disciplinary email listservs, and personalized email invitations to a pool of scholars with relevant expertise. Respondents were randomly assigned to one of two types of forecasts: a horserace elicitation or a stacking weights elicitation. Each participant was first asked to complete a forecast for one of the general crossnational models. Conditional on completion of the crossnational forecast, respondents could opt to provide a forecast for country-specific challenge.

Figure S4 reports the number of respondents who entered and completed each forecasting exercise. There is substantial dropoff in the initial, crossnational forecasts. Only 42.6 percent and 30 percent of individuals who entered the system completed the horserace and stacking forecasts, respectively. As a result, the difference in completion rates for the initial forecasts was 12.6 percentage points ($p = 0.06$). The ratio of additional forecasts to initial forecast completion is 43/35 and 49/48 for the horserace and stacking exercises, respectively. Collectively these completion rates suggest that the stacking exercise may have been more challenging or taxing than the horserace forecasts. We detail the procedures for each elicitation in the following subsections.

Data	General Form	Parameterized Form
Crossnational	deaths_per_mio_log ∼ gdp_pc + share_older + resp_disease_prev + hosp_beds_pc + precip + urban + pop_density_log	deaths_per_mio_log ∼ 4.316 + 0.1928*gdp_pc + 0.8683*share_older + 0.1824*resp_disease_prev - 0.3077*hosp_beds_pc -0.2592*precip + 0.2703*urban -0.1345*pop_density_log
India	deaths_per_mio_log ∼ gdp_pc + share_older + resp_disease_prev + hosp_beds_pc + precip + urban_pct + pop_density	deaths_per_mio_log ∼ 4.3110 + 0.5937*gdp_pc + 0.1085*share_older -0.4212*resp_disease_prev + 0.2625*hosp_beds_pc -0.1048*precip + 0.1679 *urban_pct + 0.0446*pop_density
Mexico	deaths_per_mio_log ∼ gdp_pc + share_older + irag_rate + hosp_beds_pc + precip + urban_pct + pop_density	deaths_per_mio_log ∼ 6.6278 + 0.0593*gdp_pc + 0.0869*share_older + 0.1166*irag_rate + 0.1416*hosp_beds_pc + 0.0681*precip + 0.1717*urban_pct -0.1373*pop_density
USA	deaths_per_mio_log ∼ gdp_pc + share_older + resp_disease_prev + hosp_beds_pc + precip + urban_pct + pop_density	deaths_per_mio_log ∼ 6.3422 -0.1673*gdp_pc + 0.0522*share_older -0.1489*resp_disease_prev + 0.3919*hosp_beds_pc + 0.1217*precip + 0.2476*urban_pct + 0.212*pop_density

Table S2: "Usual suspects" models (referred to in the text as epidemiological models) include the following predictors: GDP per capita (gdp_pc), share of population over 65 (share_older), respitory disease prevalence (resp_disease_prev), hospital beds per capita (hosp_beds_pc), precipitation in millimeters per month (precip), share of population living in urban areas (urban_pct), and population density (pop_density). The parameterized form was fit on outcome data as of November 16, 2020.

Challenge	General Form	Parameterized Form
Crossnational	deaths_per_mio_log ∼ acc_sanitation + healthcare_qual	deaths_per_mio_log ∼ 3.9815 + 0.5718*acc_sanitation + 0.588*healthcare_qual
India	deaths_per_mio_log ∼ gdp_pc + hosp_beds_pc + pct_poor + reserve_proportion + urban_pct	deaths_per_mio_log ∼ 4.3503 + 0.0382*gdp_pc + 0.278*hosp_beds_pc -0.0649*pct_poor - 0.4854*reserve_proportion + 0.2783*urban_pct
Mexico	deaths_per_mio_log ∼ health_expendpc + pct_poor + pct_tertiaryemp	deaths_per_mio_log ∼ 6.6278 + 0.12*health_expendpc - 0.1461*pct_poor + 0.0813*pct_tertiaryemp
USA	deaths_per_mio_log ∼ gini + hosp_beds_pc + pct_religious + pop_density + urban_pct	deaths_per_mio_log ∼ 6.2735 + 0.2325*gini + 0.2491*hosp_beds_pc + 0.1534*pct_religious + 0.2374*pop_density + 0.2517*urban_pct

Table S3: Lasso models for each challenge. The parameterized form was fit on outcome data as of November 16, 2020.

	Theoretically Motivated		Machine Learning		Total
Challenge	General	Parameterized	General	Parameterized	
Crossnational	27	15	1	1	44
India	8	6	1	1	16
Mexico	8	5	1	1	15
USA	18	6	3	2	29
Total	61	32	6	5	104

Table S4: Total number of models analyzed in each challenge, including eight Lasso and eight epidemiological benchmark models.

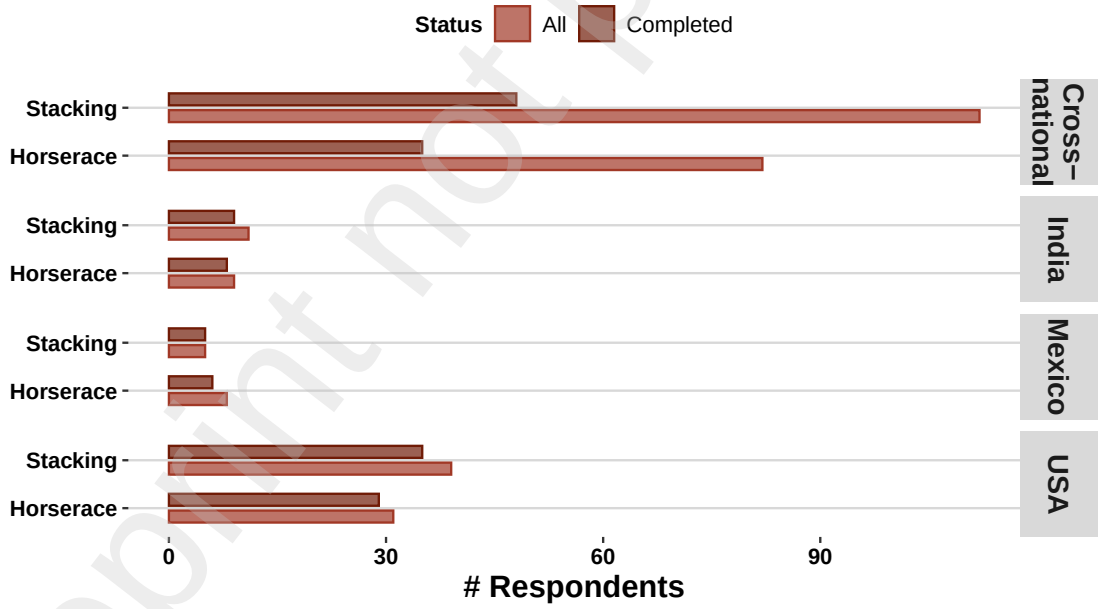


Figure S4: Selection into the forecasting activity.

S1.4.1 Horserace Elicitation

The goal of the horserace forecast was to elicit the **probability that a model would be the most predictive** out of a set of six models. The six models included: (1) five randomly selected models among the theoretical (non-ML) general models for a given challenge and (2) the epidemiological model. The set of models that forecasters viewed varied across respondents; different respondents saw different subsets of the general models.

Forecasters read the following instructions for the "horserace" elicitation for the crossnational challenge:

> "We now present six statistical models. Five were proposed by other researchers. The sixth model contains only a set of standard epidemiological predictors.
>
> We are interested in how well these models explain the **residual variance** in mortality. By this we mean the variance in mortality outcomes after accounting for a set of controls selected using a machine learning algorithm. For details on these controls and the selection process, click on or hover here.
>
> Your task is to assign the probability to each model that it will explain the most residual variance against the other models in the set in **cumulative COVID-19 deaths per capita** for all countries. You will be asked to do this for two future points in time: **31 August 2021** and **31 August 2022**. In other words, **how likely is it that each model will perform the best?**
>
> Please predict the **probability that each model will explain the most residual variance** as of 31 August 2021 and 31 August 2022. As you are putting your prediction on each model (i.e., the probability you assign to it), keep in mind that entries in each column must range between 0 and 100; **you should not enter negative probabilities**. In principle, the probabilities in each column should sum to 100 but we will rescale them if they do not.
>
> To inform your predictions, we show how much residual variance each model actually explained as of February 2021. Again, by residual variance we mean: how much of the crossnational variance in COVID-19 deaths the model explained over and above that explained by the controls. Remember that **you are not predicting the residual variance itself** but rather the probability that a model performs better than the other five.
>
> You can click on or hover over each model to view a summary of the logic that was submitted with it."

We provide a representative screenshot of the forecasting interface for a horserace forecast in Figure S5a.

S1.4.2 Stacking Forecasts

The goal of the stacking forecasts was to elicit the **stacking weights**, analogous to those that we estimate using Equation (8) over a subset of seven models. The six models included: (1) five randomly selected models among the theoretical (non-ML) general models for a given challenge; (2) the Lasso model for that challenge; and (3) the epidemiological model. The set of models that forecasters viewed varied across respondents; different respondents saw different subsets of the general models.

Forecasters read the following instructions for the "stacking" elicitation for the crossnational challenge:

> "We now present seven statistical models. The first five were proposed by other researchers. The sixth model contains epidemiological predictors and the last model a set of predictors selected by a machine learning algorithm. Click on or hover here for more details on the selection process.
>
> Your task is to provide a weight for each model. You should assign larger weights to models if you would pay relatively more attention to the predictions of those models when forming an **overall prediction**.
>
> For example, you might trust the predictions from only one model and put all weight on that model, or you might think the best prediction comes from a weighted average of the predictions of three or four different models.

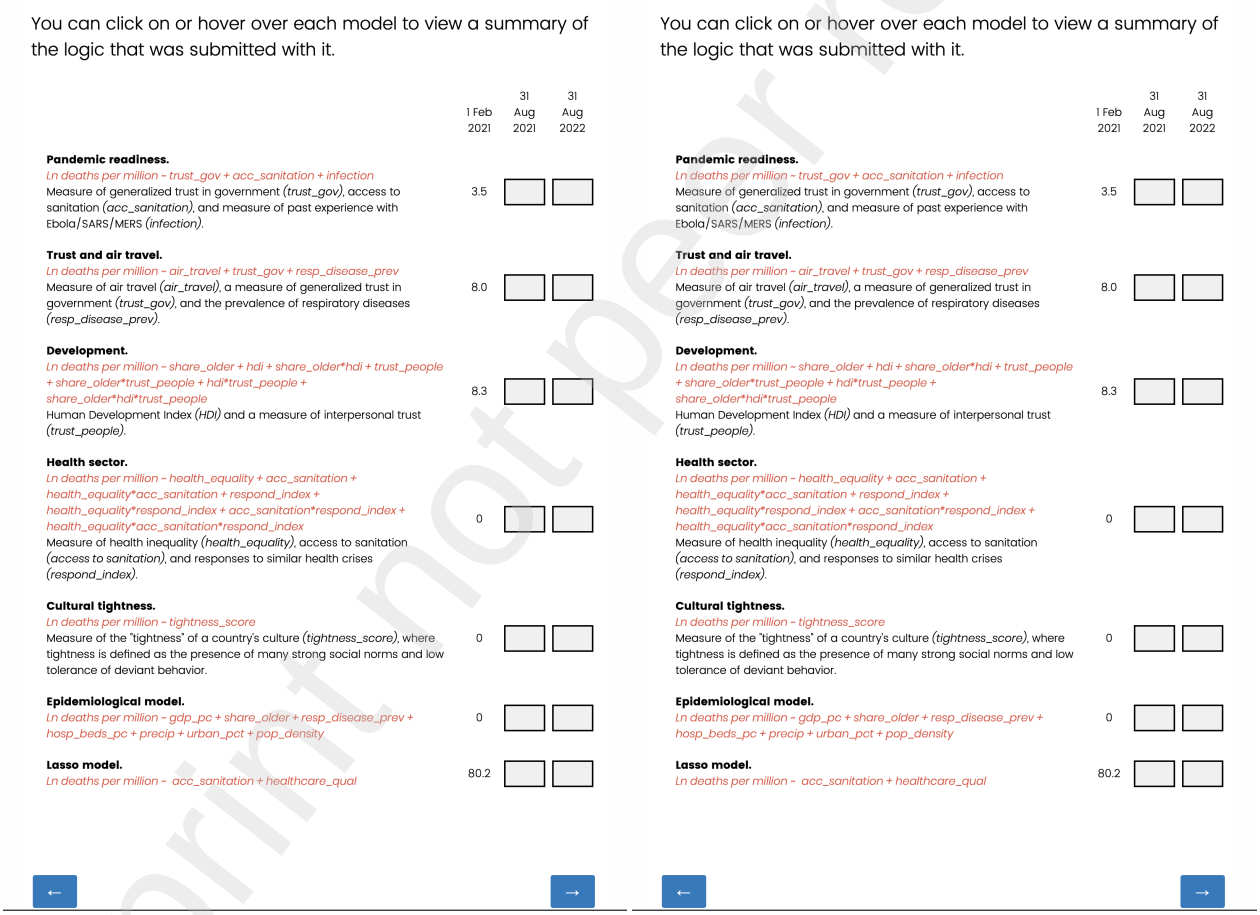
The outcome is **cumulative COVID-19 deaths per capita** for all countries at two future points in time: **31 August 2021** and **31 August 2022.**

Please enter weights for each model below. You should assign larger weights to models if you would pay relatively more attention to the predictions of those models when forming an overall prediction.

As you are assigning weights, keep in mind that your entries in each column must range between 0 and 100; **you should not enter negative weights**. In principle, the weights in each column should sum to 100 but we will rescale them if they do not.

To inform your predictions, in the first column we report the weight assigned to each model when they are combined via a **stacking model** with data from February 2021. Stacking is a statistical procedure that weights each model by its contribution when combined with the others in the set to generate a more accurate prediction. Your task is similar except that it relies on your expertise rather than an algorithm. You can click on or hover over each model to view a summary of the logic that was submitted with it."

We provide a representative screenshot of the forecasting interface for a stacking forecast in Figure S5b.



(a) Horeserace forecast interface

(b) Stacking forecast interface

Figure S5: Forecasting interface for two representative forecasts. Forecasters could hover over the models to read a description of the logic behind each model (as submitted by the modelers).

S2 Statistical Methods for Model Evaluation and Aggregation

In this section, we elaborate upon the metrics that we use to evaluate model performance. We first discuss how the models are fit.

S2.1 Fitting the Predictive Models

First, we denote models submitted as part of the COVID-19 Models Challenge as:

$$y_i = f_k(\mathbf{x}_i|\boldsymbol{\theta}_k) \tag{4}$$

We index observations by i and models by k. The outcome, logged cumulative COVID-19 deaths per million, is denoted y_i and $\mathbf{x}_i$ denotes a matrix of predictor variable(s). The parameters of the model are $\boldsymbol{\theta}_k$. In general models, $\boldsymbol{\theta}_k$ are estimated from the data. In parameterized models $\boldsymbol{\theta}_k$ are specified as part of the models.

We evaluate models on the basis of out-of-sample prediction since we use fixed models to predict future outcomes. Our approach to out-of-sample prediction is different for general models than for parameterized models.

For general models, the parameters of a model, $\boldsymbol{\theta}_k$, are estimated using the outcome data. We use leave-one-out (LOO) predictions of each model to emulate out-of-sample prediction in order to guard against overfitting. The leave-one-out prediction for unit i is:

$$\widehat{y}_{ik}^{\text{loo}} = f_k(\mathbf{x}_i|\widehat{\boldsymbol{\theta}_k^{-i}}) \tag{5}$$

where $\widehat{\boldsymbol{\theta}_k^{-i}}$ are model parameter(s) fit on all observations excluding unit i. When we examine the predictions of general models, $\widehat{y}_{ik} \equiv \widehat{y}_{ik}^{\text{loo}}$.

For parameterized models, we naturally have a form of out-of-sample prediction since we use fixed models to predict future outcomes. Our predictions are therefore given by:

$$\widehat{y}_{ik} = f_k(\mathbf{x}_i|\boldsymbol{\theta}_k) \tag{6}$$

where the parameters $\boldsymbol{\theta}_k$ are specified as part of the model (not fit on the outcome data).

S2.2 Defining Model Success: Individual Models

We focus on two measures of model success, one which examines levels of predicted and actual outcomes and one which examines scores of predicted and actual outcomes. Our analysis of levels considers $\widehat{y}_{ik}$ and y_i. Our analysis of scores examines Z-score transformations of $\widehat{y}_{ik}$ and y_i, which we will denote with the superscript Z (i.e., $\widehat{y}_{ik}^Z$ and y_i^Z).

Our metrics of model success are given by:

$$v_k = 1 - \alpha \frac{\sum_i (\widehat{y}_{ik} - y_i)^2}{\sum_i (\overline{y}_i - y_i)^2} \tag{7}$$

where α is a scale parameter and $\overline{y}_i$ denotes the mean of y_i.

For the level approach, we evaluate (7) by setting $\alpha = 1$ and using our (raw) predictions $\widehat{y}_{ik}$ and (raw) observed outcomes y_i. We refer to this measure as a pseudo-R^2. For general models, in the absence of LOO prediction, $v_k = R^2$ and, as such, $v_k \in [0, 1]$. With LOO prediction, $v_k \leq R^2$ since $(\widehat{y}_{ik}^{\text{loo}} - y_i)^2 \geq (\widehat{y}_{ik}^{\text{all}} - y_i)^2$, where $\widehat{y}_{ik}^{\text{all}}$ is the model fit on *all* observations (including i). When v_k measures the pseudo-R^2, $v_k \in (-\infty, 1]$. Higher values of v_k indicate more accurate predictions.

For the score approach, we evaluate (7) by setting $\alpha = \frac{1}{2}$ and using our normalized predictions $\widehat{y}_{ik}^Z$ and normalized outcomes y_i^Z. This measure is equivalent to the correlation between $\widehat{y}_{ik}$ and y_{ik}. Therefore, for the score approach, $v_k \in [-1, 1]$. Prediction accuracy is again increasing in v_k. Note that $\widehat{y}_{ik}$ are predictions of y_{ik}. Thus, a negative correlation — no matter how strong — indicates lower accuracy than a correlation of zero in this setting.

S2.3 Stacking

We use model stacking to aggregate the predictions of all user models, the epidemiological models, and the Lasso-selected models. The stacking approach estimates a set of k weights w so that the weighted average of model predictions has the smallest possible error.

Formally we estimate:

$$w = \arg\min_w \sum_{i=1}^{n} \left(y_i - \sum_k w_k \widehat{y}_{ik} \right)^2 \text{ s.t. } w_k \geq 0 \; \forall k, \sum_{k=1}^{K} w_k = 1 \tag{8}$$

As above, $\widehat{y}_{ik}$ refers to the $\widehat{y}_{ik}^{\text{loo}}$ for all general models. Larger weights provide a measure of the contribution of a model to an aggregate model and are taken here as a measure of unique predictive ability within the set of k models provided.

S2.4 Forecasting

We measure the accuracy of the elicited forecasts using several metrics, which we detail below.

Model-level metrics: Recall that for the horserace elicitation, forecasters predicted the probability that each model would be the best-performing model in a randomly-selected set. Our horserace measure of model performance is simply the mean expected probability that the model would be the top-performer in the set. We estimate the standard error as $\frac{\sigma_k}{\sqrt{n_k}}$, where σ is the standard deviation of forecasts for model k and n_k is the total number of forecasts elicited for model k. We average over forecasts elicited in different subsets of models.

For the stacking elicitation, forecasters predicted the stacking weights that would be assigned to each model. To calculate stacking weights, we again evaluate the mean stacking weight assigned to each model across different elicitations (and different sets of models). As in the horserace forecasts, we estimate the standard error as $\frac{\sigma_k}{\sqrt{n_k}}$, where σ is the standard deviation of elicited stacking weights for model k and n_k is the total number of forecasts elicited for model k.

Aggregate metrics: We examine three aggregate measures of elicited forecast accuracy, as described below:

1. **Expert-favored models**: We select the model with the largest average weight assigned to it by the experts as the experts' most favored model. As such, for model set c we select model k that maximizes the mean expert weight:

$$\hat{k}^c = \text{argmax}_k \left\{ \sum_j w_k^j \right\} \tag{9}$$

   This yields model predictions given by:

$$\widehat{y}_i^c = \widehat{y}_{i\widehat{k}^c} \tag{10}$$

2. **Representative expert**: As with the algorithmic stacking models, each expert's weighting of models generates an aggregate model with a prediction for unit i by expert j of:

$$\widehat{y}_i^j = \sum_k \widehat{w}_k^j \widehat{y}_{ik}^{\text{loo}} \tag{11}$$

   We use the leave-one-out designation here to remind readers that forecasts were only elicited over general models where we employ the leave-one-out predictions in all metrics of prediction accuracy. We can plug this into Equation (7) to measure the success of an expert's stacking model. The representative expert's aggregate model

set is defined by the elicited weights such that:

$$w^r = \{w^j | v^j = \text{median}(v^h)_{h \in H}\} \tag{12}$$

where H is the set of forecasters assigned to the stacking elicitation.

3. **Wisdom of the crowds**: To construct a wisdom of the crowds aggregate forecast, for each model set, we calculate the normalized average weight placed on a model by experts. As such, for model set c, we calculate:

$$w_k^c = \frac{\sum_j \widehat{w}_k^j}{\sum_k \sum_j \widehat{w}_k^j} \tag{13}$$

This yields model predictions given by:

$$\widehat{y}_i^c = \sum_k w_k^c \widehat{y}_{ik}^{\text{loo}} \tag{14}$$

S3 Gathering: Supplementary Results

In this section, we provide complementary results for the gathering stage of our analysis.

S3.1 Summary Statistics

Table S5 provides an overview of the general models submitted cross all four challenges, including information about (i) the functional form of the models; (ii) the number of predictors used and addition of predictors from outside the MC datasets; (iii) whether the models were theoretically motivated — i.e., whether they included a theoretical argument for why their selected variables should predict COVID-19 mortality or whether the model was generated using machine learning methods; (iv) whether the models included references to existing literature to justify inclusion of selected variables ("has model justification"); and (v) the number of modelers who submitted each model. Table S6 provides information about the model challenge participants.

S3.2 List of Model Submissions

Table S7 enumerates all of the submitted models. They are ranked, within challenge, by the pseudo-R^2 metric.

Feature	N	Mean	SD	Mode	Min	Max
Total						
Use of Non-Linear Function	58	0.414	0.497	0	0	1
Number of Unique Predictors	58	2.672	0.509	3	1	3
Has Theoretical Motivation	58	0.724	0.451	1	0	1
Has Model Justification	58	0.448	0.502	0	0	1
Submitted Own Data	58	0.155	0.365	0	0	1
Is Predictive Model	58	0.017	0.131	0	0	1
Team Size	58	1.879	1.983	1	1	8
Crossnational						
Use of Non-Linear Function	26	0.538	0.508	1	0	1
Number of Unique Predictors	26	2.615	0.571	3	1	3
Has Theoretical Motivation	26	0.692	0.471	1	0	1
Has Model Justification	26	0.462	0.508	0	0	1
Submitted Own Data	26	0.192	0.402	0	0	1
Is Predictive Model	26	0.000	0.000	0	0	0
Team Size	26	1.654	1.719	1	1	8
India						
Use of Non-Linear Function	7	0.286	0.488	0	0	1
Number of Unique Predictors	7	2.571	0.535	3	2	3
Has Theoretical Motivation	7	0.714	0.488	1	0	1
Has Model Justification	7	0.429	0.535	0	0	1
Submitted Own Data	7	0.286	0.488	0	0	1
Is Predictive Model	7	0.000	0.000	0	0	0
Team Size	7	2.571	2.820	1	1	8
Mexico						
Use of Non-Linear Function	7	0.429	0.535	0	0	1
Number of Unique Predictors	7	2.714	0.488	3	2	3
Has Theoretical Motivation	7	1.000	0.000	1	1	1
Has Model Justification	7	0.571	0.535	1	0	1
Submitted Own Data	7	0.143	0.378	0	0	1
Is Predictive Model	7	0.000	0.000	0	0	0
Team Size	7	2.286	2.563	1	1	8
USA						
Use of Non-Linear Function	18	0.278	0.461	0	0	1
Number of Unique Predictors	18	2.778	0.428	3	2	3
Has Theoretical Motivation	18	0.667	0.485	1	0	1
Has Model Justification	18	0.389	0.502	0	0	1
Submitted Own Data	18	0.833	0.383	1	0	1
Is Predictive Model	18	0.056	0.236	0	0	1
Team Size	18	1.778	1.833	1	1	8

Table S5: Overview of general models across all four challenges. The top panel shows results pooled across all challenges. The next four panels show results broken down by each challenge.

Table S6: Tally of Model Challenge Participants (Raw Numbers)

Challenge	Participants	Institutions	Countries
Crossnational	42	21	9
India	18	6	5
Mexico	15	6	3
USA	29	15	6
All	60	32	10

Table S7: Model composition and performance by challenge.

	Model Name	Variables included and functional form specified	Pseudo-R^2	Stacking weight
Crossnational, general				
1	Trust in Authoritarian Government	deaths_per_mio_log = β_0 + β_1*acc_sanitation + β_2*trust_gov + β_3*media_critical	0.483	0.572
2	Government Capacity and Social Inequality	deaths_per_mio_log = β_0 + β_1*gov_effect + β_2*healthcare_qual + β_2*gini + β_4*gov_effect^2 + β_5*gini^2	0.420	0.359
3	Perverse Development	deaths_per_mio_log = β_0 + β_1*acc_sanitation + β_2*hdi	0.392	0.000
4	Health Equality	deaths_per_mio_log = β_0 + β_1*health_equality + β_2*acc_sanitation + β_3*health_equality*acc_sanitation + β_4*respond_index + β_5*health_equality*respond_index + β_6*acc_sanitation*respond_index + β_7*health_equality*acc_sanitation*respond_index	0.355	0.000
5	Inequality in Pandemic Experiences	deaths_per_mio_log = β_0 + β_1*gini + β_1*infection + β_2*med_age_2013 + β_3*gini*med_age_2013 + β_4*infection*med_age_2013	0.346	0.068
6	Developing Country Effectiveness	deaths_per_mio_log = β_0 + β_1*med_age_2013 + β_2*gov_effect + β_3*temp_mean + β_4*temp_mean^2	0.297	0.000
7	Development and Trust	deaths_per_mio_log = β_0 + β_1*share_older + β_2*hdi + β_3*share_older*hdi + β_5*trust_people + β_6*share_older*trust_people + β_7*hdi*trust_people + β_8*share_older*hdi*trust_people	0.270	0.000
8	Populism and Social Trust	deaths_per_mio_log = β_0 + β_1*electoral_pop + β_2*life_exp_2017 + β_3*trust_people + β_3*electoral_pop*trust_people	0.241	0.000
9	Social Trust and Health Capacity	deaths_per_mio_log = β_0 + β_1*trust_people + β_2*healthcare_qual + β_3*UVINDEX	0.237	0.000
10	Democracy	deaths_per_mio_log = β_0 + β_1*share_older + β_2*checks_veto + β_3*vdem_libdem + β_4*share_older*vdem_libdem + β_5*share_older*checks_veto*vdem_libdem	0.232	0.000
11	State Fragility, Social Trust, and Population Age	deaths_per_mio_log = β_0 + β_1*trust_people + β_2*state_fragility + β_3*share_older	0.228	0.000
12	Social and Institutional Trust	deaths_per_mio_log = β_0 + β_1*life_exp_2017 + β_2*trust_people + β_3*gov_effect + β_4*trust_people*gov_effect	0.214	0.000
13	Government Capacity	deaths_per_mio_log = β_0 + β_1*pandemic_prep + β_2*gov_effect	0.171	0.000
14	Dictatorships and Misinformation	deaths_per_mio_log = β_0 + β_1*gdp_pc + β_2*v2x_polyarchy	0.169	0.000
15	Pandemic Readiness	deaths_per_mio_log = β_0 + β_1*trust_gov + β_2*acc_sanitation + β_3*infection	0.040	0.000

Table S7: Model composition and performance by challenge. Lasso and Epidemiological models are not shown in this table since they are already displayed in Tables S3 and S2, respectively. The Lasso model receives positive stacking weight in the India and Mexico general challenges. *(continued)*

ID	Model	Functional Form	Pseudo-R^2	Stacking Weight
16	Social and Political Stability	deaths_per_mio_log = β_0 + β_1*trust_people + β_2*pr + β_3*trust_people*pr + β_4*gini + β_5*trust_people*gini	0.039	0.000
17	Shackled Leviathan	deaths_per_mio_log = β_0 + β_1*hdi + β_2*gov_effect + β_3*hdi*gov_effect + β_4*trust_gov + β_5*hdi*trust_gov + β_6*gov_effect*trust_gov	0.015	0.000
18	Institutional Trust	deaths_per_mio_log = β_0 + β_1*gov_effect + β_2*share_older + β_3*trust_gov + β_4*gov_effect^2	0.005	0.000
19	Government Capacity and Development	deaths_per_mio_log = β_0 + β_1*gov_effect + β_2*gdp_pc + β_3*trust_gov + β_4*gov_effect^2	-0.007	0.000
20	Trust and Air Travel	deaths_per_mio_log = β_0 + β_1*air_travel + β_2*trust_gov + β_3*resp_disease_prev	-0.033	0.000
21	Liberalism, Capitalism, and Media Independence	deaths_per_mio_log = β_0 + β_1*property_rights + β_2*vdem_mecorrpt + β_3*trust_gov	-0.036	0.000
22	Cultural Tightness	deaths_per_mio_log = β_0 + β_1*tightness_score	-0.064	0.000
23	Trust and Social Safety	deaths_per_mio_log = β_0 + β_1*share_older + β_2*soc_safety + β_3*trust_gov	-0.159	0.000
24	Language and Culture	deaths_per_mio_log = β_0 + β_1*idv + β_2*inflectional_ftr	-0.261	0.000
25	Polarization and Populism	deaths_per_mio_log = β_0 + β_1*polar_rile + β_2*trust_people + β_3*electoral_pop	-0.302	0.000
26	Competitiveness of Executive Recruitment	deaths_per_mio_log = β_0 + β_1*urban + β_2*acc_sanitation + β_3*xrcomp_2018 + β_4*urban^2 + β_5*acc_sanitation^2 + 0	-0.574	0.000
Crossnational, parameterized				
1	Trust in Authoritarian Government	deaths_per_mio_log = 4.75 + 0.9 * acc_sanitation - 1.25 * trust_gov	0.141	0.437
2	Populism and Social Trust	deaths_per_mio_log = 4.8 + 0.9*electoral_pop + 0.9*electoral_pop*trust_people + 0.75*life_exp_2017 -0.9*trust_people	0.097	0.155
3	Liberalism, Capitalism, and Media Independence	deaths_per_mio_log = 5 + 0.7*property_rights -0.7*trust_gov -0.1*vdem_mecorrpt	0.008	0.000
4	Social and Institutional Trust	deaths_per_mio_log = 4.5 -0.5*gov_effect + life_exp_2017 -0.5*trust_people + 0.5*trust_people*gov_effect	-0.079	0.000
5	Government Capacity and Development	deaths_per_mio_log = 5 + 0.1*gdp_pc -0.5*gov_effect + 0.5*gov_effect^2 -0.9*trust_gov	-0.121	0.000
6	Government Capacity and Social Inequality	deaths_per_mio_log = 4 + 0.2*gini + 0.15*gini^2 -0.6*gov_effect -0.2*gov_effect^2 + 2*healthcare_qual	-0.356	0.168
7	Perverse Development	deaths_per_mio_log = 4 + 0.4*acc_sanitation + 0.6*hdi	-0.369	0.000
8	Health Equality	deaths_per_mio_log = 4 + 1.5*acc_sanitation + 0*acc_sanitation*respond_index -0.1*health_equality + 0*health_equality*acc_sanitation + 0.01*health_equality*acc_sanitation*respond_index + 0*health_equality*respond_index + 0*respond_index	-0.390	0.168
9	Competitiveness of Executive Recruitment	deaths_per_mio_log = 1.46*acc_sanitation + 0.85*acc_sanitation^2 -0.14*urban -0.04*urban^2 + 1.37*xrcomp_2018	-0.946	0.000

Table S7: Model composition and performance by challenge. Lasso and Epidemiological models are not shown in this table since they are already displayed in Tables S3 and S2, respectively. The Lasso model receives positive stacking weight in the India and Mexico general challenges. *(continued)*

ID	Model	Functional Form	Pseudo-R^2	Stacking Weight
10	Development and Trust	deaths_per_mio_log = 4 + 1*hdi + 0.5*hdi*trust_people + 0.25*share_older -0.5*share_older*hdi + 0.25*share_older*hdi*trust_people -1*share_older*trust_people -0.55*trust_people	-0.993	0.000
11	Pandemic Readiness	deaths_per_mio_log = 9 -2*acc_sanitation + 1.4*infection -1.2*trust_gov	-2.201	0.058
12	Social and Political Stability	deaths_per_mio_log = 1.5 -0.2*gini -0.3*pr -0.1*trust_people + 0.1*trust_people*gini + 0.1*trust_people*pr	-5.581	0.000
13	Polarization and Populism	deaths_per_mio_log = 1*electoral_pop -0.5*polar_rile + 0.5*trust_people	-7.557	0.000
14	Language and Culture	deaths_per_mio_log = 3.5 + 1.5*idv + 0.05*inflectional_ftr	-1322.027	0.013
India, general				
1	Health Sector Capacity	deaths_per_mio_log = β_0 + β_1*pandemic_prep + β_2*pct_poor + β_3*pandemic_prep*pct_poor + β_4*hosp_beds_pc + β_5*pandemic_prep*hosp_beds_pc + β_6*pct_poor*hosp_beds_pc + β_7*pandemic_prep*pct_poor*hosp_beds_pc	0.363	0.451
2	Interactions and Political Pressures	deaths_per_mio_log = β_0 + β_1*gdp_pc + β_2*urban_pct + β_3*election_margin	0.306	0.231
3	Urbanisation and Healthcare	deaths_per_mio_log = β_0 + β_1*gdp_pc + β_2*public.health.total.budget.2015 + β_3*urban_pct	0.301	0.034
4	Business and Density	deaths_per_mio_log = β_0 + β_1*minority_pct + β_2*gdp_pc + β_3*urban_pct	0.295	0.000
5	GDP, TB Prevalence, and Anti-immigration Attitudes	deaths_per_mio_log = β_0 + β_1*pct_anti_immig + β_2*tb_per_100k + β_3*gdp_pc	0.204	0.000
6	Minority Representation and Urbanization	deaths_per_mio_log = β_0 + β_1*reserve_proportion + β_2*urban_pct + β_3*reserve_proportion*urban_pct	0.094	0.260
7	Government Capacity	deaths_per_mio_log = β_0 + β_1*average_events_per_state + β_2*leader_experience	-0.145	0.000
India, parameterized				
1	Business and Density	deaths_per_mio_log = 4.7 + 0.2*gdp_pc -0.2*minority_pct + 0.5*urban_pct	-1.668	0.945
2	Urbanisation and Health Care	deaths_per_mio_log = 5.35 + 0.41*gdp_pc -13*public.health.total.budget.2015 + 0.44*urban_pct	-2.021	0.000
3	Health Sector Capacity	deaths_per_mio_log = 4.3 + 0.4*hosp_beds_pc + 1.8*pandemic_prep + 2*pandemic_prep*hosp_beds_pc + 1.5*pandemic_prep*pct_poor + 2.5*pandemic_prep*pct_poor*hosp_beds_pc -0.35*pct_poor -0.25*pct_poor*hosp_beds_pc	-2.154	0.055
4	Minority Representation and Urbanization	deaths_per_mio_log = 4.2 -0.6*reserve_proportion -0.6*reserve_proportion*urban_pct + 0.2*urban_pct	-3.267	0.000
5	Interactions and Political Pressures	deaths_per_mio_log = 4.25 + 0.05*election_margin + 0.4*gdp_pc + 0.4*urban_pct	-3.684	0.000
Mexico, general				

Table S7: Model composition and performance by challenge. Lasso and Epidemiological models are not shown in this table since they are already displayed in Tables S3 and S2, respectively. The Lasso model receives positive stacking weight in the India and Mexico general challenges. *(continued)*

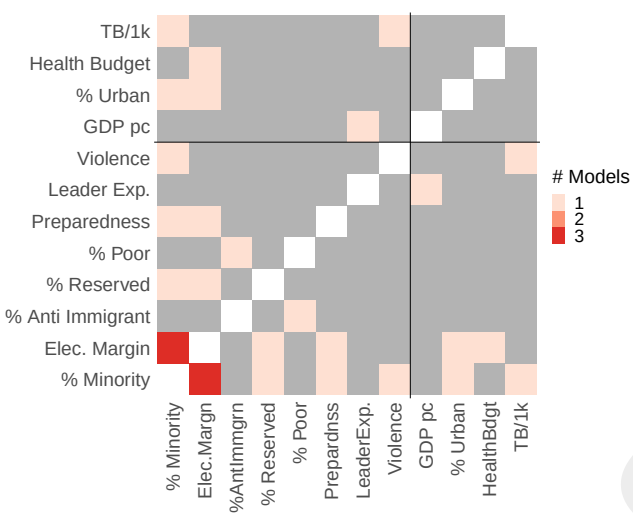
ID	Model	Functional Form	Pseudo-R^2	Stacking Weight
1	Political Leadership, Poverty, and Obesity	deaths_per_mio_log = β_0 + β_1*election_margin + β_2*pct_poor + β_3*obesity	0.371	0.000
2	Social Trust and Catholicism	deaths_per_mio_log = β_0 + β_1*pct_catholic + β_2*election_margin + β_3*trust_people + β_4*pct_catholic*trust_people	0.347	0.360
3	Trust, Poverty, and TB Prevalence	deaths_per_mio_log = β_0 + β_1*pct_poor + β_2*trust_people + β_3*tuberc_cases	0.345	0.072
4	Poverty, Electoral Competitiveness, and Public Goods	deaths_per_mio_log = β_0 + β_1*hosp_beds_pc + β_2*pct_poor + β_3*hosp_beds_pc*pct_poor + β_4*election_margin	0.040	0.000
5	Government Experience	deaths_per_mio_log = β_0 + β_1*pandemic_prep + β_2*leader_experience	-0.103	0.000
6	Interactions and Political Pressures	deaths_per_mio_log = β_0 + β_1*gdp_pc + β_2*election_margin + β_3*urban_pct	-0.300	0.000
7	Investment Inequality	deaths_per_mio_log = β_0 + β_1*hosp_beds_pc + β_2*gini + β_3*hosp_beds_pc*gini + β_4*health_expendpc + β_5*hosp_beds_pc*health_expendpc + β_6*gini*health_expendpc + β_7*hosp_beds_pc*gini*health_expendpc	-0.504	0.000
Mexico, parameterized				
1	Social Trust and Catholicism	deaths_per_mio_log = 7.6 + 0.19*election_margin -0.08*pct_catholic -0.197*pct_catholic*trust_people + 0.27*trust_people	0.619	1.000
2	Poverty, Electoral Competitiveness, and Public Goods	deaths_per_mio_log = 6.7 + 0.2*election_margin + 0.3*hosp_beds_pc + 0.25*hosp_beds_pc*pct_poor -0.05*pct_poor	-4.756	0.000
3	Investment Inequality	deaths_per_mio_log = 6.7 + 0.01*gini -0.01*gini*health_expendpc + 0.2*health_expendpc + 0*hosp_beds_pc + 0.35*hosp_beds_pc*gini + 0.05*hosp_beds_pc*gini*health_expendpc -0.3*hosp_beds_pc*health_expendpc	-5.074	0.000
4	Interactions and Political Pressures	deaths_per_mio_log = 6.6 + 0.15*election_margin + 0.1*gdp_pc + 0.25*urban_pct	-5.186	0.000
USA, general				
1	Inequality and Polarization	deaths_per_mio_log = β_0 + β_1*party_leg_right + β_2*pop_density + β_2*gini + β_3*party_leg_right*gini + β_4*pop_density*gini	0.549	0.389
2	Density, Inequality, and Religiosity	deaths_per_mio_log = β_0 + β_1*gini + β_2*pct_religious + β_3*pop_density + β_4*gini^2 + β_5*pct_religious^2 + β_5*pop_density^2	0.501	0.259
3	Inequality and Capacity	deaths_per_mio_log = β_0 + β_1*gini + β_2*urban_pct + β_2*hosp_beds_pc	0.500	0.328
4	Right Party Power and Income Inequality	deaths_per_mio_log = β_0 + β_1*gini + β_2*party_leg_right + β_3*pop_density	0.487	0.000
5	Religiosity	deaths_per_mio_log = β_0 + β_1*pop_density + β_2*pct_religious + β_3*gini	0.429	0.000
6	Women in Leadership	deaths_per_mio_log = β_0 + β_1*pop_density + β_2*percentage_of_women + β_2*gini	0.363	0.000
7	Ethnicity, Inequality, and Healthcare Capacity	deaths_per_mio_log = β_0 + β_1*gini + β_2*hosp_beds_pc + β_3*ethnic_frac_score	0.344	0.000

Table S7: Model composition and performance by challenge. Lasso and Epidemiological models are not shown in this table since they are already displayed in Tables S3 and S2, respectively. The Lasso model receives positive stacking weight in the India and Mexico general challenges. *(continued)*

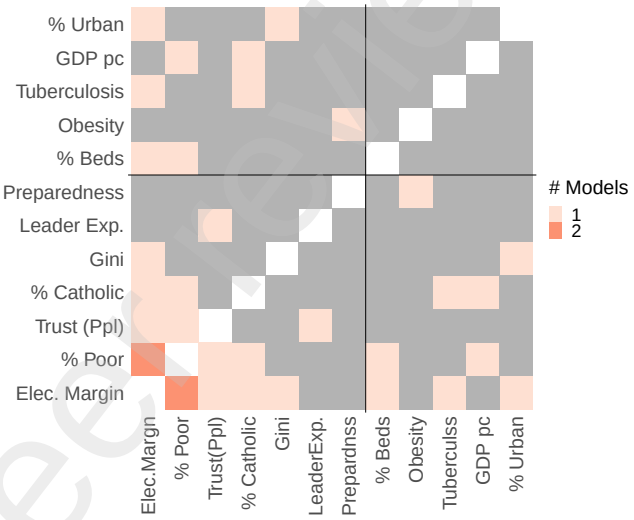
ID	Model	Functional Form	Pseudo-R^2	Stacking Weight
8	Social Contact	deaths_per_mio_log = β_0 + β_1*pct_religious + β_2*pct_poor + β_3*pop_density	0.332	0.000
9	Institutional and Social Trust	deaths_per_mio_log = β_0 + β_1*pop_density + β_2*trust_gov + β_3*gini	0.325	0.000
10	Community Equality and Trust	deaths_per_mio_log = β_0 + β_1*gini + β_2*civil_society + β_3*trust_people	0.317	0.000
11	Religion, Economic Inequality, and Minority Status	deaths_per_mio_log = β_0 + β_1*pct_religious + β_2*minority_pct + β_3*gini	0.299	0.000
12	Inequality and Urbanity	deaths_per_mio_log = β_0 + β_1*urban_pct + β_2*gini	0.227	0.000
13	Poverty and Social Exclusion	deaths_per_mio_log = β_0 + β_1*minority_pct + β_2*pct_poor + β_2*gini	0.203	0.000
14	Institutional Trust and Race	deaths_per_mio_log = β_0 + β_1*trust_gov + β_2*pop_density + β_3*minority_pct + β_4*minority_pct^2	0.201	0.024
15	Vaccination Coverage	deaths_per_mio_log = β_0 + β_1*s_diffs + β_2*share_older + β_3*Influenza_vaccination_age_65*share_older	0.020	0.000
16	Population Health, Religiosity, and Compliance	deaths_per_mio_log = β_0 + β_1*pct_religious + β_2*resp_disease_prev + β_3*share_older	-0.085	0.000
17	Government Experience	deaths_per_mio_log = β_0 + β_1*leader_experience + β_2*corrected_score	-0.121	0.000
USA, parameterized				
1	Vaccination Coverage	deaths_per_mio_log = 7.3 -0.31*Influenza_vaccination_age_65*share_older -0.23*s_diffs -0.17*share_older	-0.102	0.974
2	Inequality and Polarization	deaths_per_mio_log = 6 + 0.5*gini + 0.5*party_leg_right -0.4*party_leg_right*gini + 0.4*pop_density -0.2*pop_density*gini	-5.498	0.026
3	Social Contact	deaths_per_mio_log = 6.3 + 0.1*pct_poor + 0.25*pct_religious + 0.5*pop_density	-5.555	0.000
4	Population Differences	deaths_per_mio_log = 6.3 + 0.3*gini + 0*gini^2 + 0.3*pct_religious -0.1*pct_religious^2 + 0.2*pop_density + 0.01*pop_density^2	-6.166	0.000
5	Inequality and Urbanity	deaths_per_mio_log = 6.1 + 0.6*gini + 0.1*urban_pct	-8.265	0.000

S3.3 Pairwise Combinations of Predictors

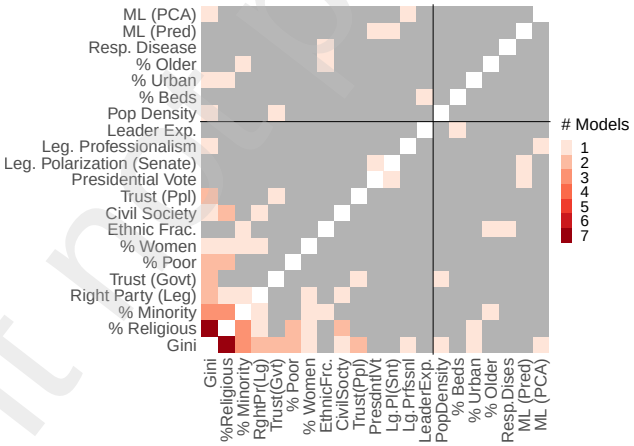
Figure S6 depicts the pairwise combinations of predictors in the country-specific challenges analogous to Figure 1. See the discussion of 1 in the main text for interpretation of these plots.



(a) Variable combination for India.



(b) Variable combination for Mexico.



(c) Variable combination for the USA.

Figure S6: Pairwise combinations of variables submitted to the country-specific challenges.

Table S8: Regression results for: Crossnational data (general models)

	Authoritarian Trust	Govt. Capacity and Inequality	Perverse Development
(Intercept)	5.70^{***}	5.88^{***}	5.61^{***}
	(0.10)	(0.16)	(0.11)
Health Access	1.14^{***}		0.72^{**}
	(0.10)		(0.25)
Trust (Govt)	-0.59^{***}		
	(0.15)		
Critical Media	0.24		
	(0.12)		
Govt Effectiveness		-0.34	
		(0.22)	
Healthcare		1.62^{***}	
		(0.23)	
Gini		0.05	
		(0.16)	
Govt Effectiveness^2		-0.56^{***}	
		(0.11)	
Gini^2		0.29^{**}	
		(0.09)	
HDI			0.54^{*}
			(0.24)
R^2	0.51	0.51	0.42
Adj. R^2	0.50	0.49	0.41
Num. obs.	166	144	162

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

S4 Evaluating: Supplementary Results

In this section, we report additional results relevant to the evaluating stage for all challenges.

S4.1 Regression table for top three crossnational models

Each general model is a regression model. Our fit statistics (pseudo-R^2 and correlation) summarize the predictive performance of each model. However, for illustrative purposes, in Table S8 we show the top three performing general models (ranked by pseudo-$\hat{R}^2$) in the crossnational challenge.

S4.2 Pseudo-R^2 Performance in All Challenges

Table S9 summarizes the performance of the best and median-performing models in each challenge.

Figures S7-S10 include scatter plots analogous to Figure 2 for the remaining seven challenges. For the country-specific challenges, we show only results for the general models to save space.

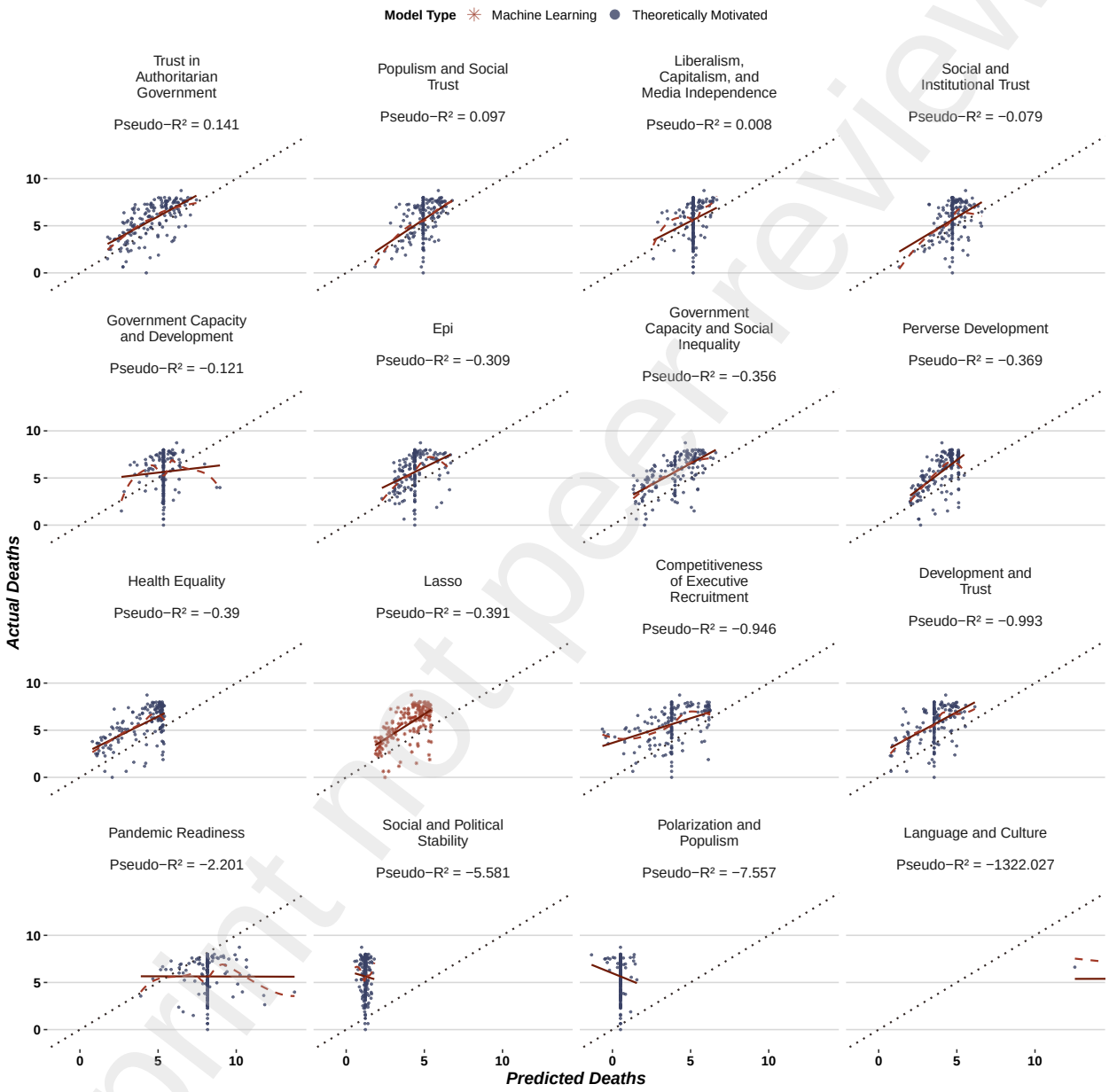


Figure S7: Summary of model predictions and observed COVID-19 mortality from crossnational parameterized models.

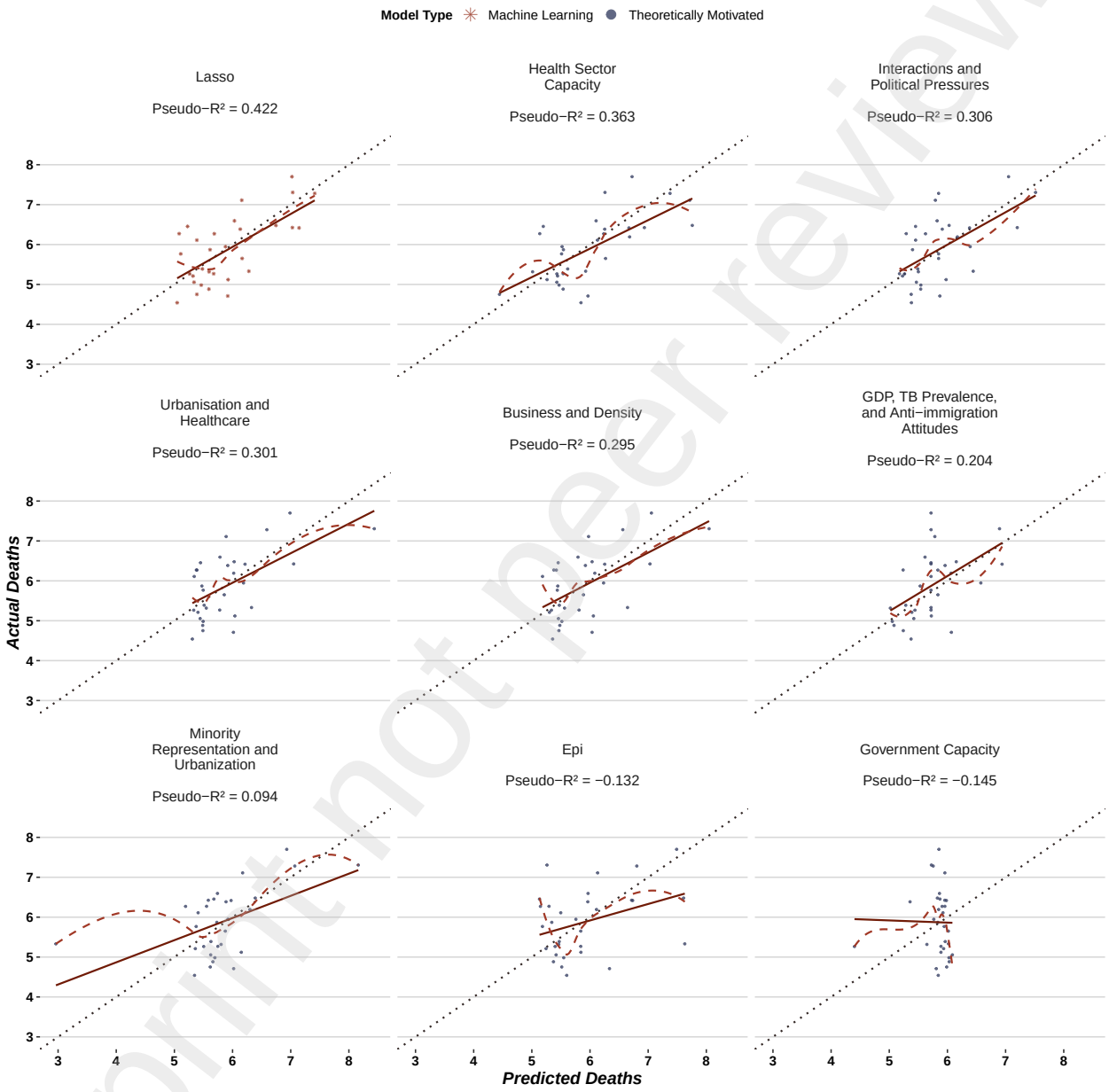


Figure S8: Summary of model predictions and observed COVID-19 mortality from general models for the India data.

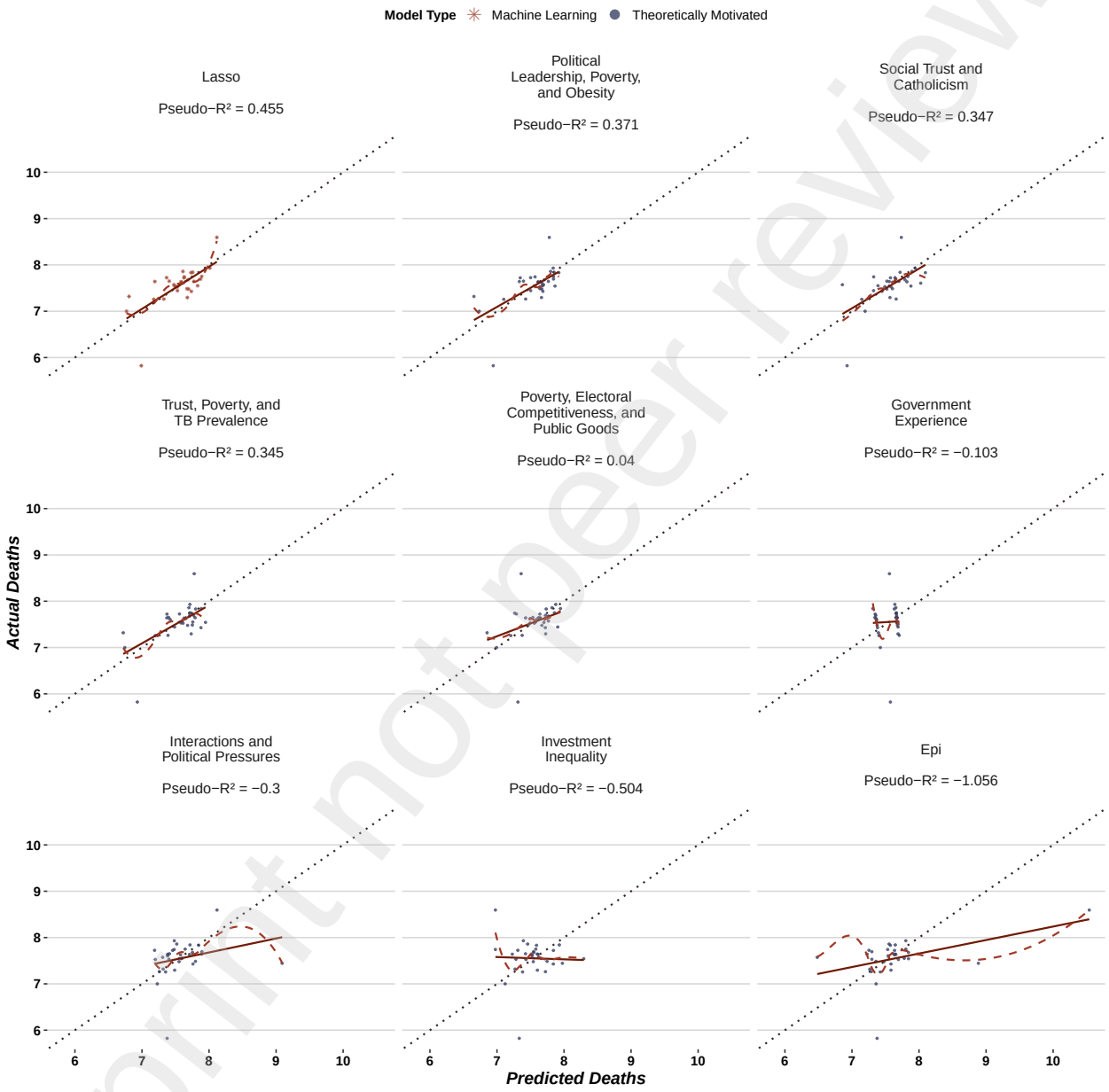


Figure S9: Summary of model predictions and observed COVID-19 mortality from general models for the Mexico data.

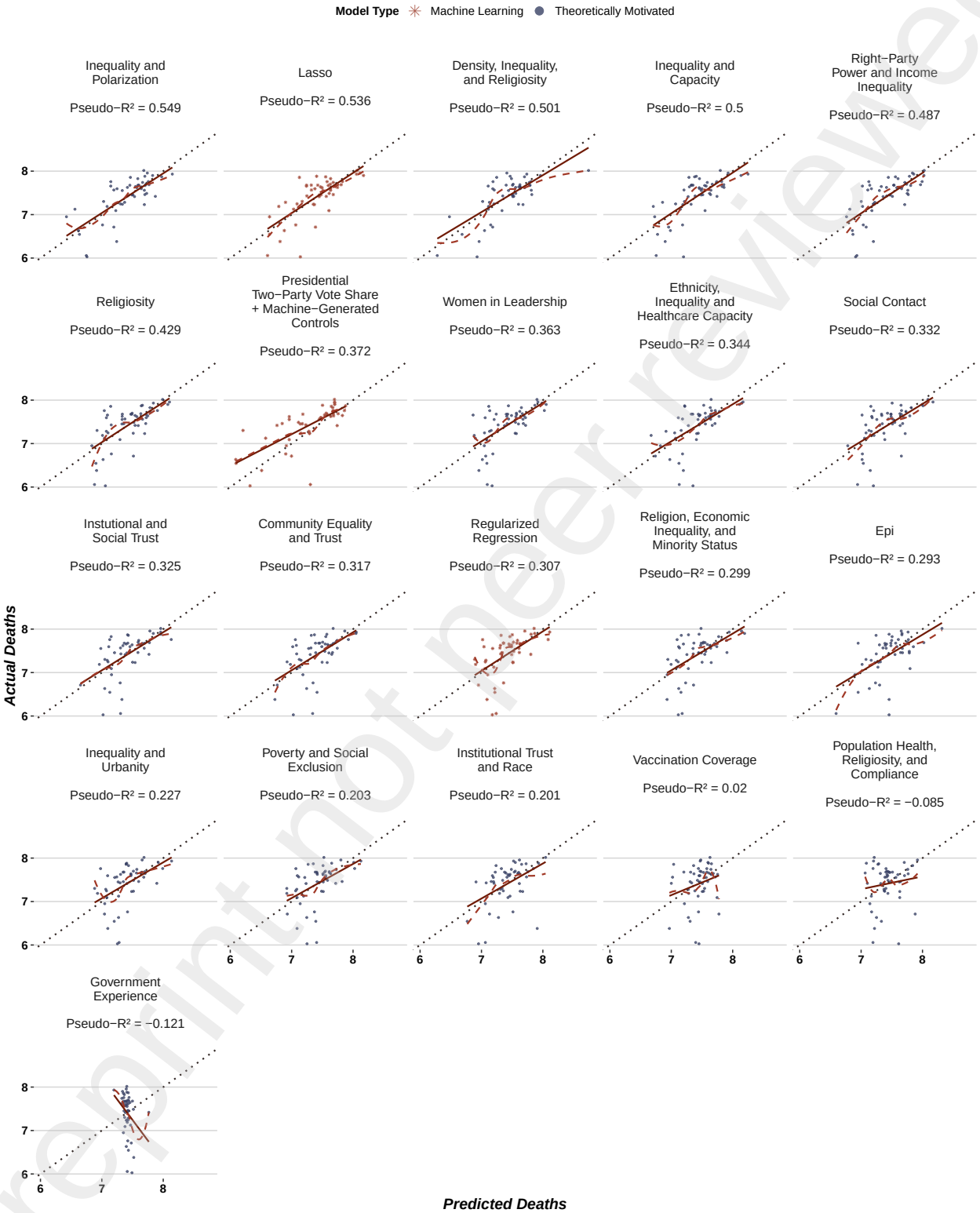


Figure S10: Summary of model predictions and observed COVID-19 mortality from general models using USA data.

Challenge	# of models	Pseudo-R^2	
		Best	Median
Crossnational, general	28	0.483	0.170
Crossnational, parameterized	16	0.141	-0.379
India, general	9	0.422	0.295
India, parameterized	7	-1.67	-3.12
Mexico, general	9	0.455	0.04
Mexico, parameterized	6	0.619	-4.72
USA, general	19	0.549	0.325
USA, parameterized	7	-0.102	-5.56

Table S9: Summary of results from the evaluating analysis of pseudo-R^2 for each challenge.

S4.3 Stacking and Model Selection

We now report the results of our model selection exercise. Table S10 summarizes the range of the top-five models using each selection metric. Recall that we only elicit expert opinion for the general models. As such, the expected horserace and expected stacking metrics are not applicable to the parameterized models. Figures S11-S14 are analogous to Figure 3 in the main text. note that we only show general model results for the country-specific challenges as before.

Challenge	# of models		Range (Top-5)			
	# Algorithmic	# Forecast	Horserace	Stacking	Horserace Forecast.	Stacking Forecast.
Crossnational, general	7	9	[0.355, 0.483]	[0, 0.572]	[0.273, 0.513]	[0.157, 0.347]
Crossnational, parameterized	8	-	[-0.121, 0.141]	[0.058, 0.437]	*Not applicable*	
India, general	7	6	[0.295, 0.422]	[0.0245, 0.451]	[0.060, 0.831]	[0.125, 0.292]
India, parameterized	6	-	[-3.267, -1.668]	[0, 0.945]	*Not applicable*	
Mexico, general	7	7	[0.040, 0.455]	[0, 0.456]	[0.080, 0.439]	[0.104, 0.394]
Mexico, parameterized	6	-	[-5.074, 0.619]	[0, 1]	*Not applicable*	
USA, general	7	8	[0.487, 0.549]	[0, 0.389]	[0.273, 0.493]	[0.161, 0.311]
USA, parameterized	6	-	[-5.633, -0.102]	[0, 0.974]	*Not applicable*	

Table S10: Summary of model selection results. For each selection method, we choose the top five performing models in each set. The number of models refers to the number of unique models selected across the horserace and stacking approaches for either the algorithmic or elicited model selection procedures.

S4.4 Stability of Model Evaluation Results Over Time

To assess the robustness of predictions to the specific date at which they are evaluated (August 31, 2021), Figure S15 plots the evolution of these metrics over time. To create this plot, we have estimated the pseudo-R^2 and stacking weights on weekly cumulative mortality data over the course of the pandemic. Both metrics evolve relatively smoothly. While there is more over-time variation in the weights afforded to each model than the pseudo-R^2's, two observations are of note. First, our top-two performing models ("Trust in Authoritarian Government" and "Government Capacity and Social Inequality") are the top two models throughout the post-prediction period. Second, after August 31, 2021 we see a notable decrease in the weight afforded the "Government Capacity and Social Inequality" model.

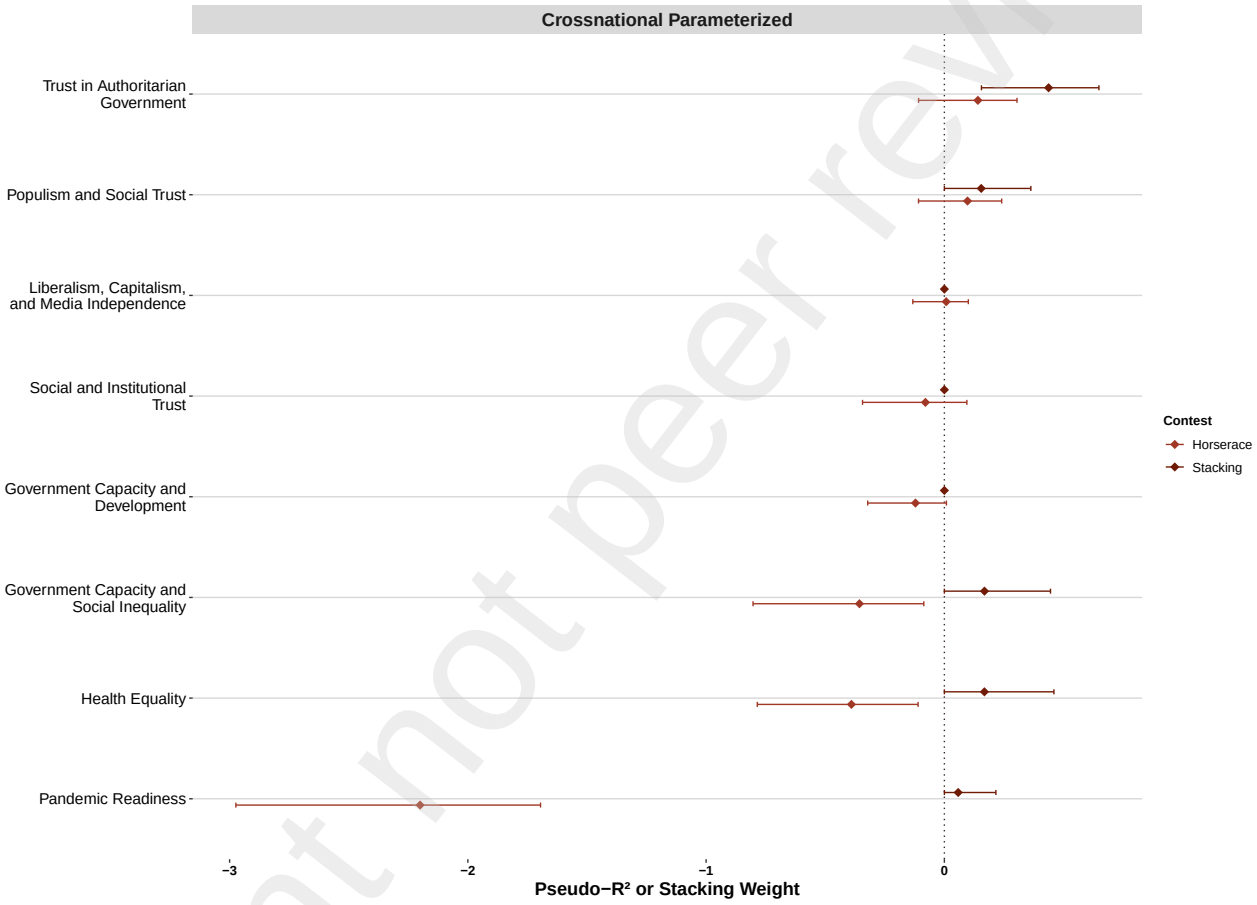


Figure S11: Summary of model predictions using two algorithmic methods for parameterized models from the crossnational challenge.

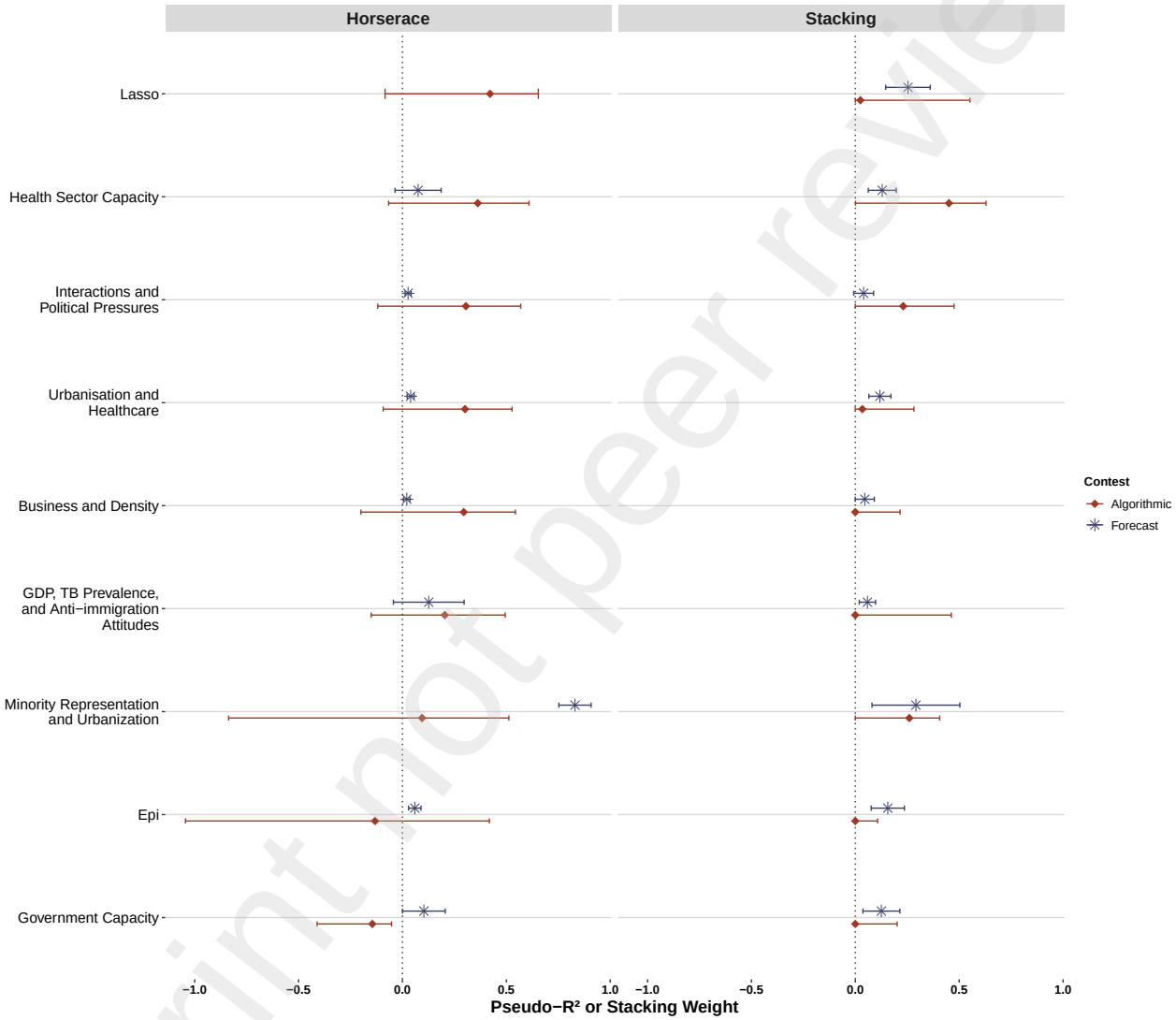


Figure S12: Model selection using four methods for the general models from India.

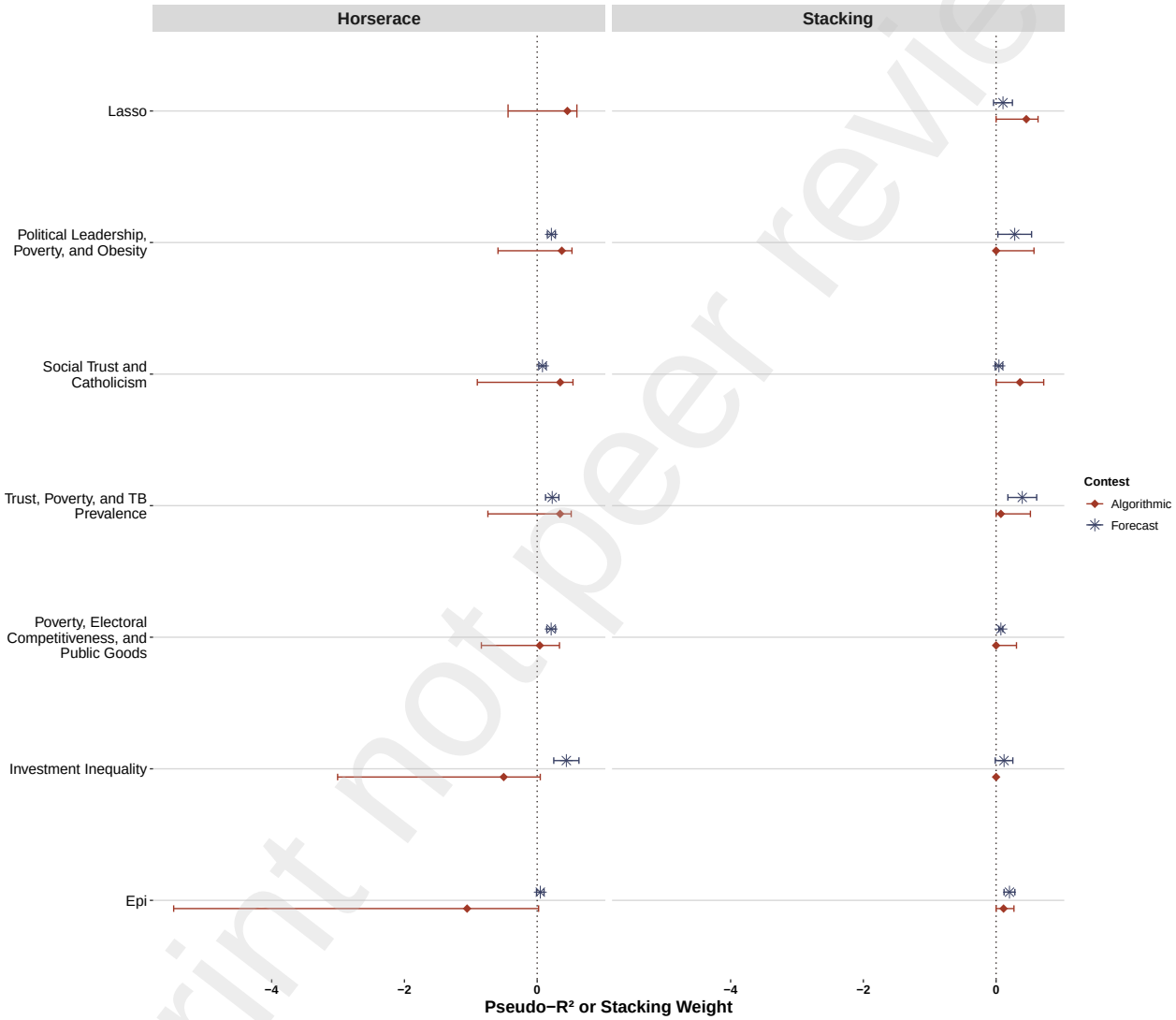


Figure S13: Model selection using four methods for the general models from Mexico.

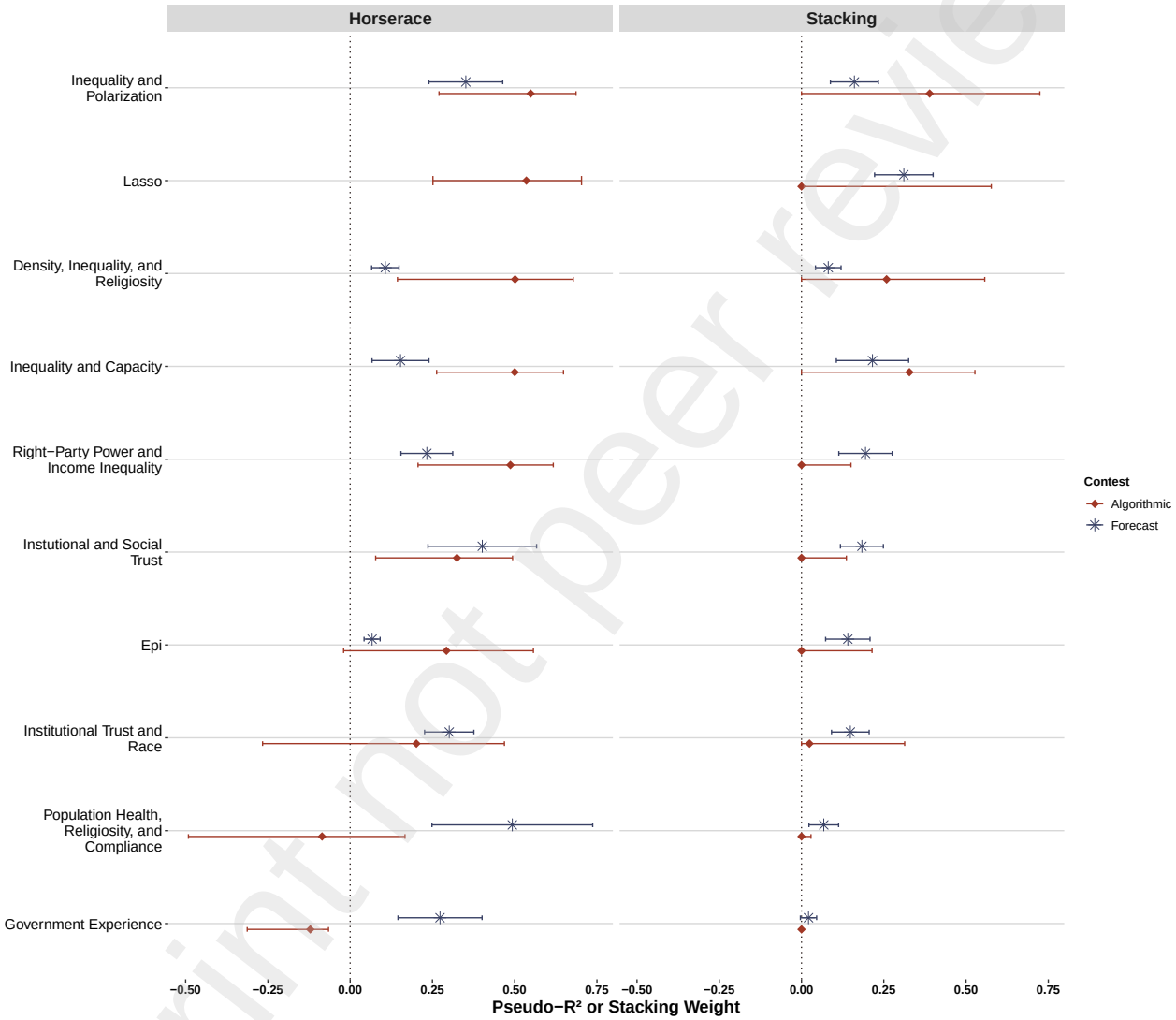
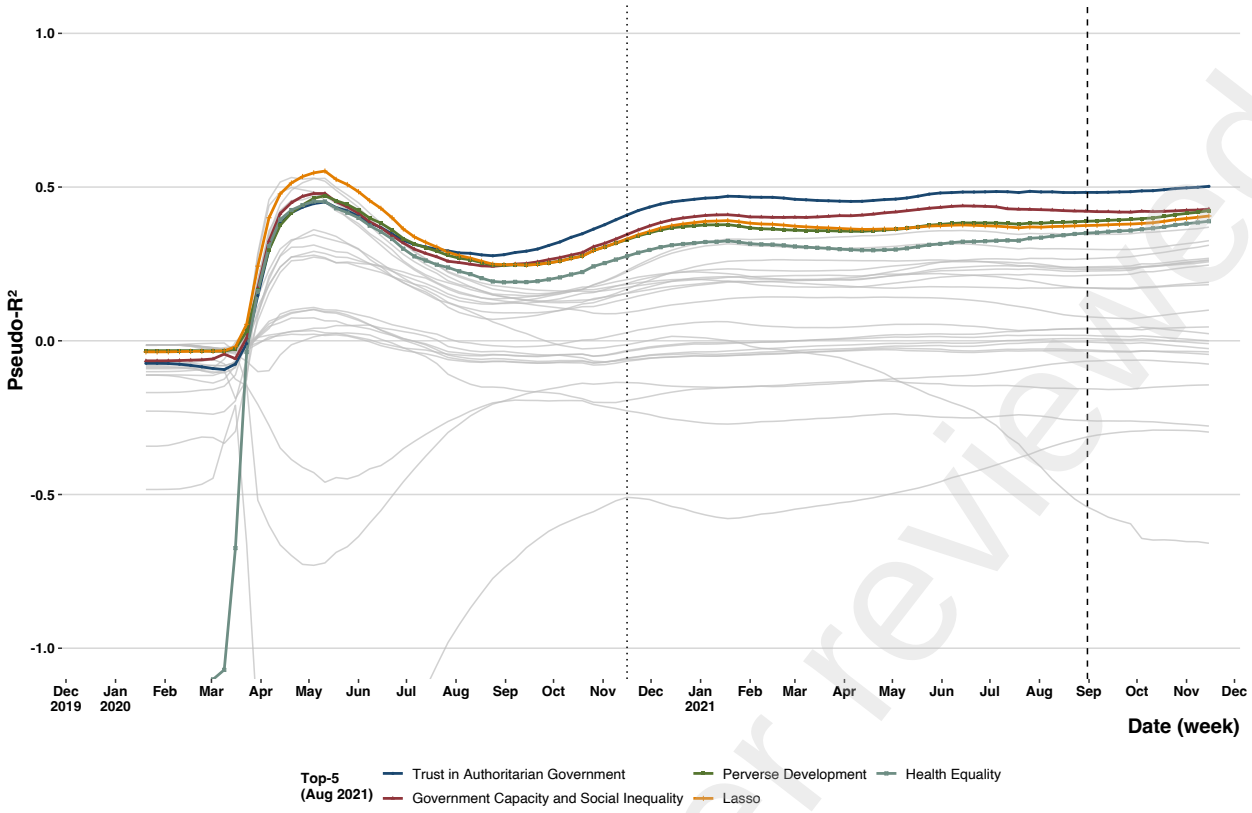
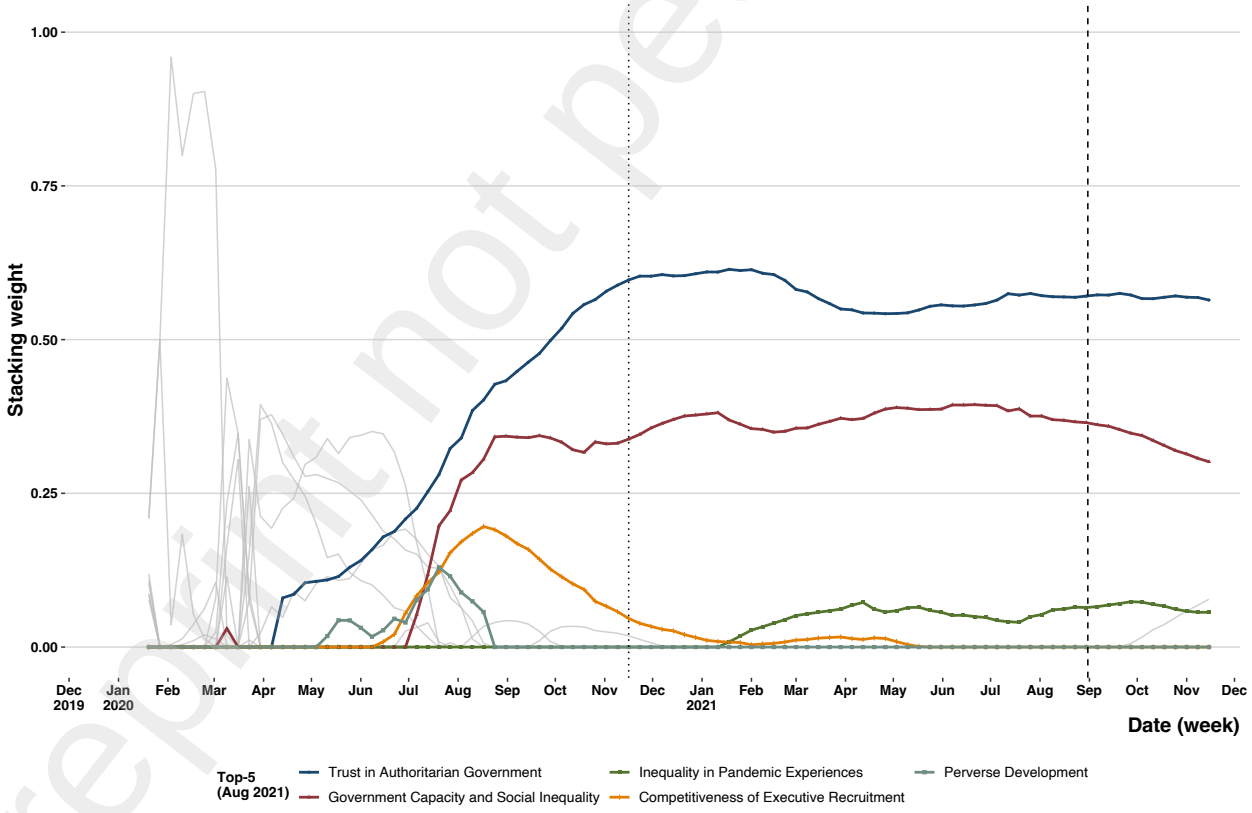


Figure S14: Model selection using four methods for the general models from the US.



(a) Evolution of pseudo-R^2 measure over time.



(b) Evolution of stacking weights over time.

Figure S15: Evolution of measures of model performance over time.

S4.5 Relationships among Models

In order to describe the relationship between the submitted models, we propose two metrics. For the input distance metric, we construct an average multivariate R^2 measure (MV-R^2) to measure the mean distance between two models' predictors. The multivariate R^2 generalizes the bivariate regression R^2 to the multivariate regression setting (*39*). For two models with predictor sets $\mathbf{x}_1, \mathbf{x}_2$, we measure their multivariate R^2 by (15).

$$\begin{aligned} \text{MV-}R^2_{2\to1} &= 1 - \frac{\sum_{i=1}^{N} d(\mathbf{x}_{1,i}, \hat{\mathbf{x}}_{1,i})}{\sum_{i=1}^{N} d(\mathbf{x}_{1,i}, \bar{\mathbf{x}}_1)} \\ \text{MV-}R^2_{1\to2} &= 1 - \frac{\sum_{i=1}^{N} d(\mathbf{x}_{2,i}, \hat{\mathbf{x}}_{2,i})}{\sum_{i=1}^{N} d(\mathbf{x}_{2,i}, \bar{\mathbf{x}}_2)} \\ \text{Avg. MV-}R^2 &= \frac{1}{2}\left(R^2_{2\to1} + R^2_{1\to2}\right) \end{aligned} \tag{15}$$

where $d(.)$ is the Euclidean distance between two vectors and $\bar{\mathbf{x}}, \hat{\mathbf{x}}$ are the vectors of mean of a predictor set and fitted values from a multivariate OLS of this predictor set upon the other.

For the output distance metric, we estimate an adjusted correlation measure to measure the mean distance between two models' predicted outcomes. For two models with predicted outcomes y_1, y_2, we measure their adjusted correlation by (16).

$$\text{Adj. Correlation} = \frac{1}{2}\left(\frac{\sum_{i=1}^{N}(y_{1,i} - \bar{y}_1)(y_{2,i} - \bar{y}_2)}{\sqrt{\sum_{i=1}^{N}(y_{1,i} - \bar{y}_1)^2 \sum_{i=1}^{N}(y_{2,i} - \bar{y}_2)^2}} + 1\right) \tag{16}$$

For ease of interpretation of the network graphs, we use the model numbers (which are descending pseudo-R^2 rankings) in Table S7. These numbers are written on each node in Figures S16 and S17. In Figure S16, we plot two measures of the distance between the models submitted in the crossnational general challenge.

Figure S17 replicates the analysis from Figure S16 for the other three national challenges.

In both graphs, the models that are given positive (non-zero) weights by the stacking estimators have nodes colored in brown. We note that these nodes are generally dispersed across the network graphs. This provides suggestive evidence that the stacking model is affording greater weights to models that are providing different information.

S5 Aggregating: Supplementary Results

In this section, we report the results from the aggregation analysis of the other challenges to complement Figure 4.

A. Input

B. Output

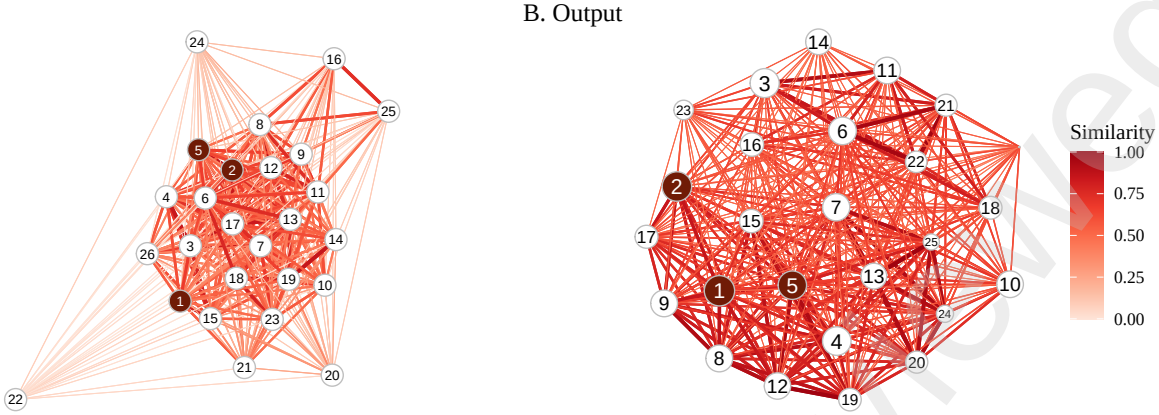


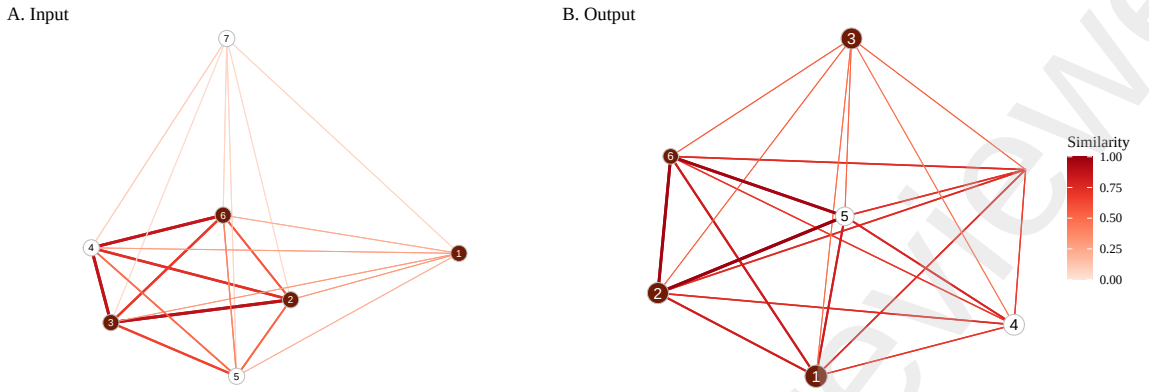
Figure S16: Network graphics of models submitted. The network in the left panel plots the multivariate R^2 between predictors of each of the submitted crossnational, general models. The network in the right panel plots the bivariate correlation between the leave-one-out predictions of each model and actual mortality (ρ), normalized to a 0-1 scale by the formula $\frac{\rho+1}{2}$. The numbering and, in the right panel, size, of the nodes corresponds to model performance according to the pseudo-R^2 metric we propose. The nodes corresponding to the models that were awarded positive weight in the stacking exercise are colored dark red.

S5.1 Difference betwen Stacking and Best Single Models

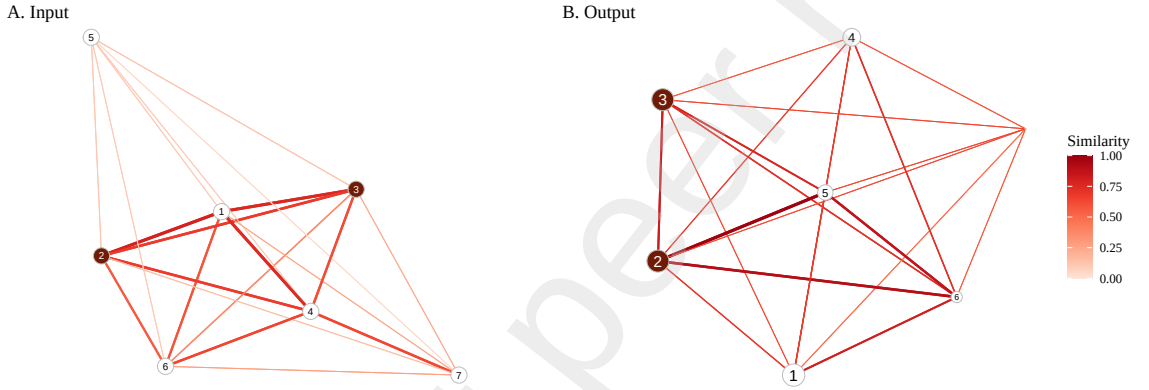
We first report the estimated differentials between the "best single" and "stacking" models in aggregate model performance. Figure S18 plots the gains in pseudo-$\hat{R}^2$ when we move from the best single to the stacking model for the crossnational analysis.

S5.2 Comparing Stacking Model to Best Individual Model

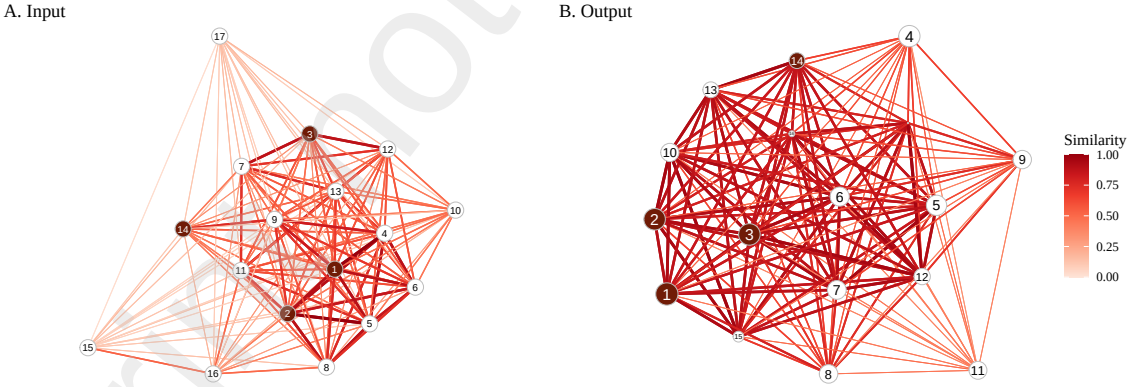
We quantify the gains in predictive accuracy of the stacking model over the best individual model in Figure S18. To conduct this analysis, we fix (a) the stacking model associated with each general challenge; and (b) the best-performing individual model associated with each general challenge. To conduct inference, we employ bootstrapping. Specifically, if there are N observations in a challenge, we resample N observations with replacement. With each bootstrapped sample, we compute: (a) the pseudo-R^2 using the stacking model associated with the general challenge and (b) the pseudo-R^2 of the best-performing individual model. Figure S18 plots the difference in these pseudo-R^2 statistics. The p-values test the one-sided null hypothesis that the pseudo-R^2 from stacking is less than the pseudo-R^2 from the best model. In sample, with the August 2021 COVID mortality outcomes, we reject this null hypothesis at the $\alpha = 0.1$ level in all challenges. Out of sample, we can only reject the null hypothesis at this level in the US challenge. However, we note that across challenges, most bootstrap iterations suggest predictive accuracy gains from stacking over the best individual model.



(a) Model networks for India.



(b) Model networks for Mexico.



(c) Model networks for the USA.

Figure S17: The network in the left panel plots the multivariate distance between predictors of each of the submitted crossnational, general models. The network in the right panel plots the bivariate distance between the leave-one-out predictions of each model. The numbering and, in the right panel, size, of the nodes corresponds to model performance according to the pseudo-R^2 metric we propose. The nodes corresponding to the models that were awarded positive weight in the stacking exercise are colored dark red.

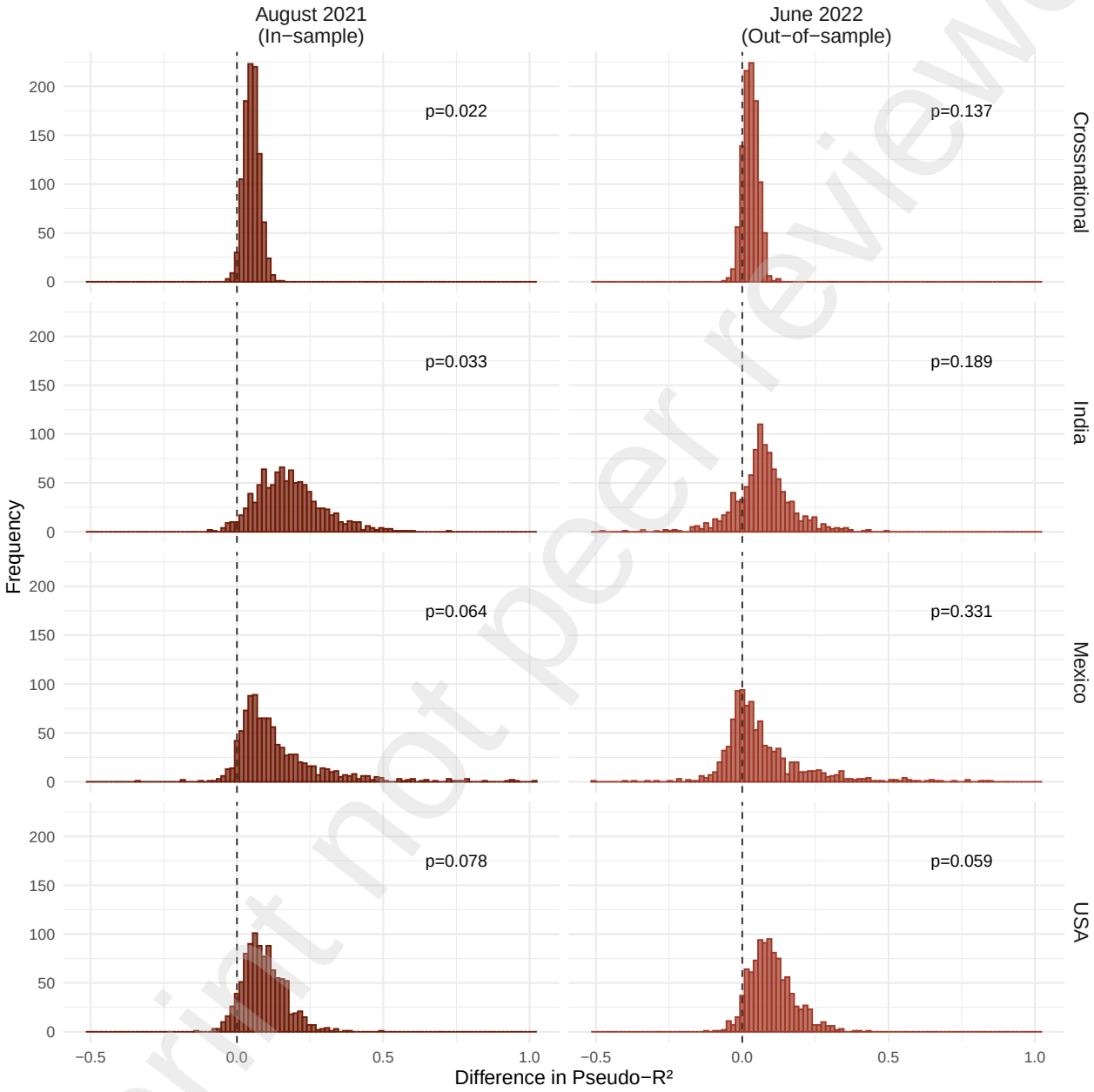


Figure S18: Performance differentials between stacking and best single models for the general models in each challenge. The histograms show the distribution of the difference in estimated pseudo-R^2's between the stacking (super-)model and the best single model in each challenge. We estimate this difference at the bootstrap step level, i.e., taking the difference between the two pseudo-$\hat{R}^2$'s for each of the 1,000 bootstrap iterations.

Table S11: Performance gains of stacked supermodel over the best single model.

		Performance (mean pseudo-R^2)		
Metric	Timepoint	Best Single	Stacking	Gain
Crossnational				
Level	Aug 2021	0.48	0.53	0.10
Score	Aug 2021	0.70	0.74	0.06
Level	June 2022	0.63	0.65	0.04
Score	June 2022	0.79	0.81	0.03
India				
Level	Aug 2021	0.42	0.58	0.38
Score	Aug 2021	0.67	0.77	0.15
Level	June 2022	0.10	0.16	0.68
Score	June 2022	0.37	0.41	0.10
Mexico				
Level	Aug 2021	0.45	0.54	0.19
Score	Aug 2021	0.68	0.74	0.10
Level	June 2022	0.47	0.50	0.06
Score	June 2022	0.69	0.71	0.02
USA				
Level	Aug 2021	0.55	0.63	0.15
Score	Aug 2021	0.74	0.80	0.07
Level	June 2022	0.47	0.56	0.19
Score	June 2022	0.69	0.75	0.09

S5.3 Aggregation Results: Other Challenges

We now summarize the results of aggregating across the four general and four parameterized challenges in Table S12. Figures S19-S22 report the full results for each of the challenge analogous to 4 in the main text. Note that we only show results for the general models for the country-specific challenges as before.

S5.4 Redefining Expert-Favored Models

To define the expert-favored model, we have used forecasters' stacking predictions as the basis for selecting their most preferred model. This is for consistency and comparability with the other approaches in aggregating used in this paper. Here we provide results using an alternative selection metric, namely the forecasters' horserace predictions (probability of each model of being the best-performing model). The most preferred models selected under the horserace and the stacking predictions are shown in Table S13 alongside their average weights for each model challenge. With the exception of India, experts selected different models in each forecast. Figure S23 shows the best models' performances in the actual aggregating analysis using each forecasting metric. Since the elicited stacking model also consistently outperforms the horserace best model, we consider the former as an upper bound for aggregating performance using the expert-favored strategy.

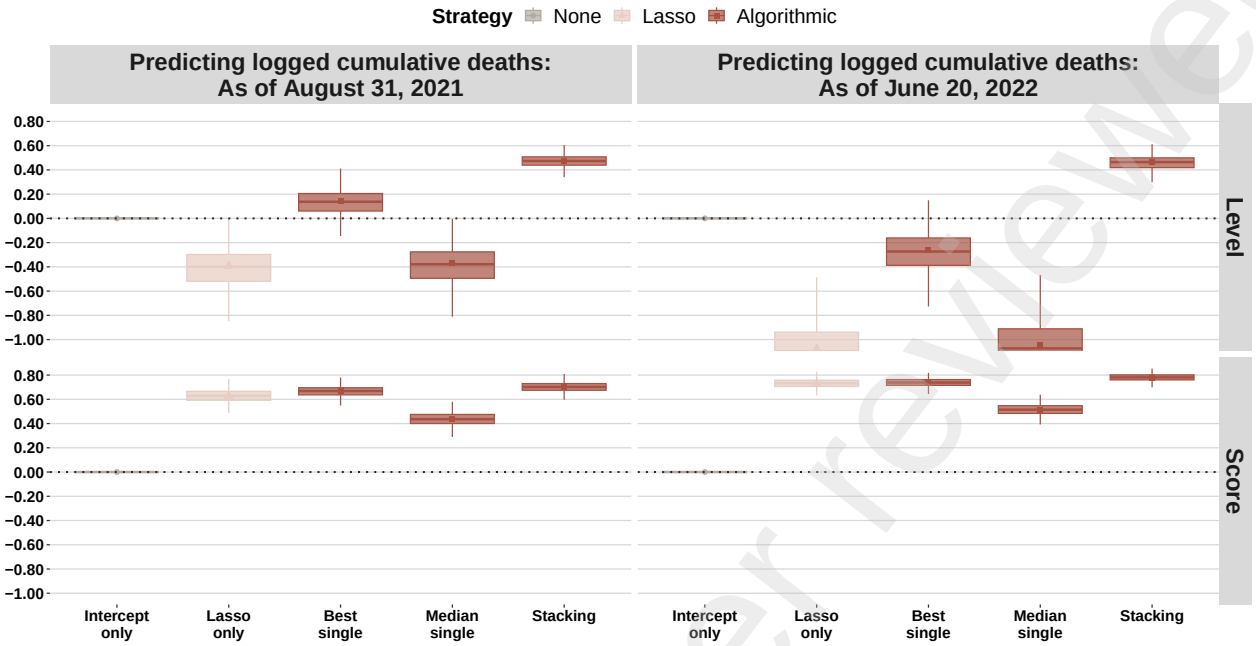


Figure S19: Prediction aggregation metrics for crossnational parameterized models. Because we did not elicit forecasts for parameterized models, we do not include the expert favored, representative expert, or wisdom of the crowds metrics.

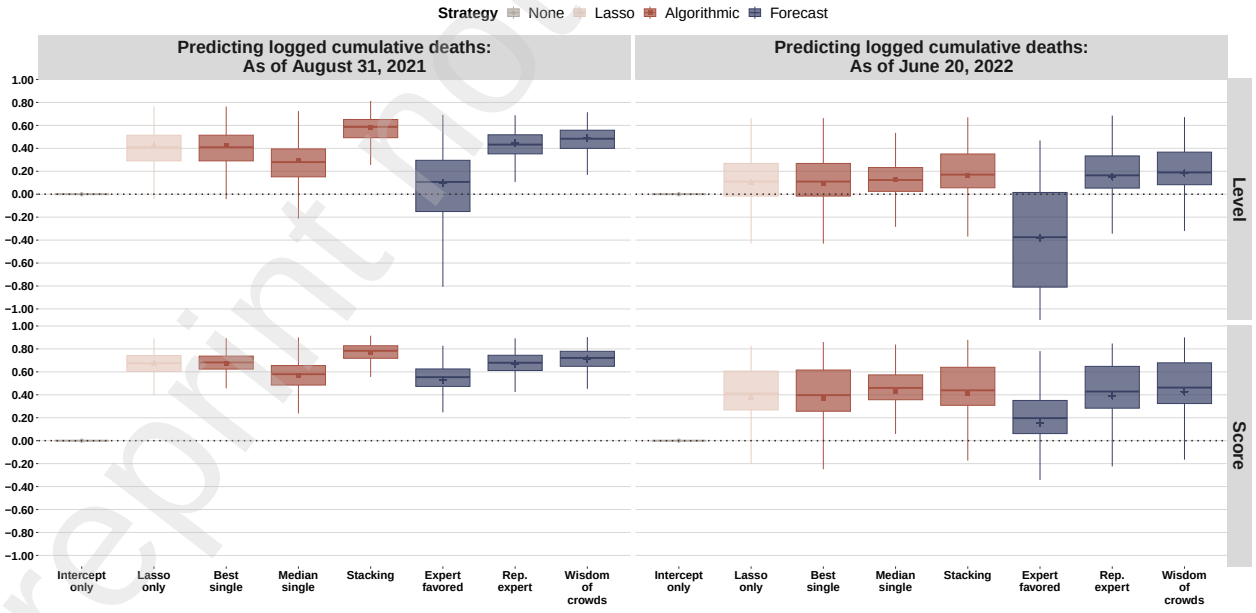


Figure S20: Prediction aggregation metrics for general models for India.

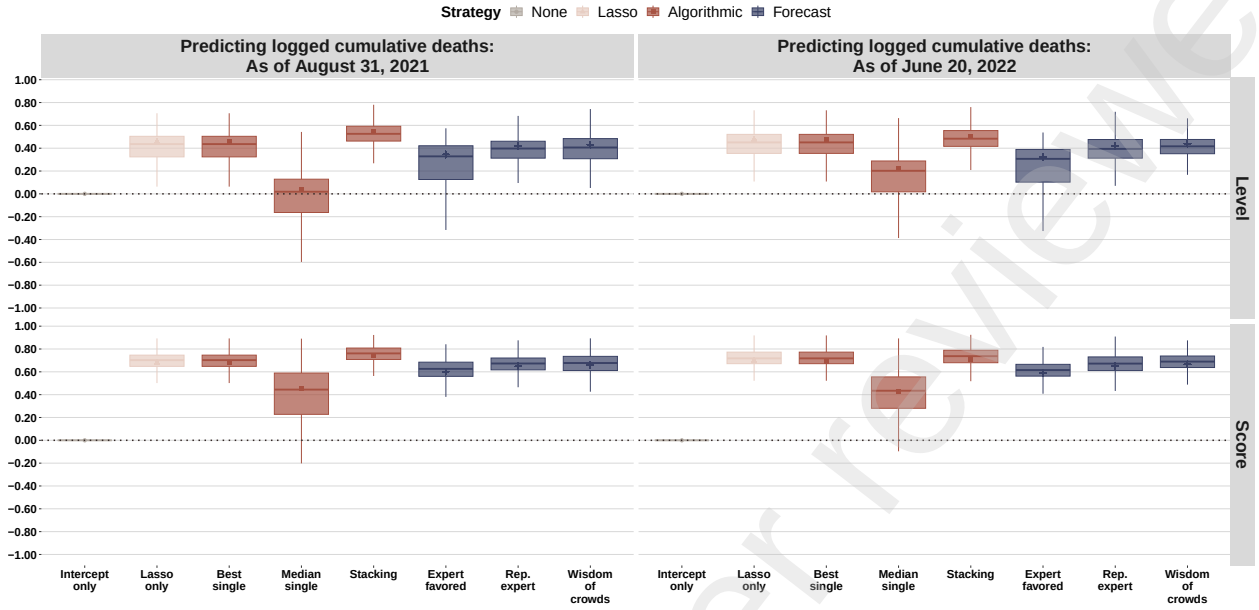


Figure S21: Prediction aggregation metrics for general models for Mexico.

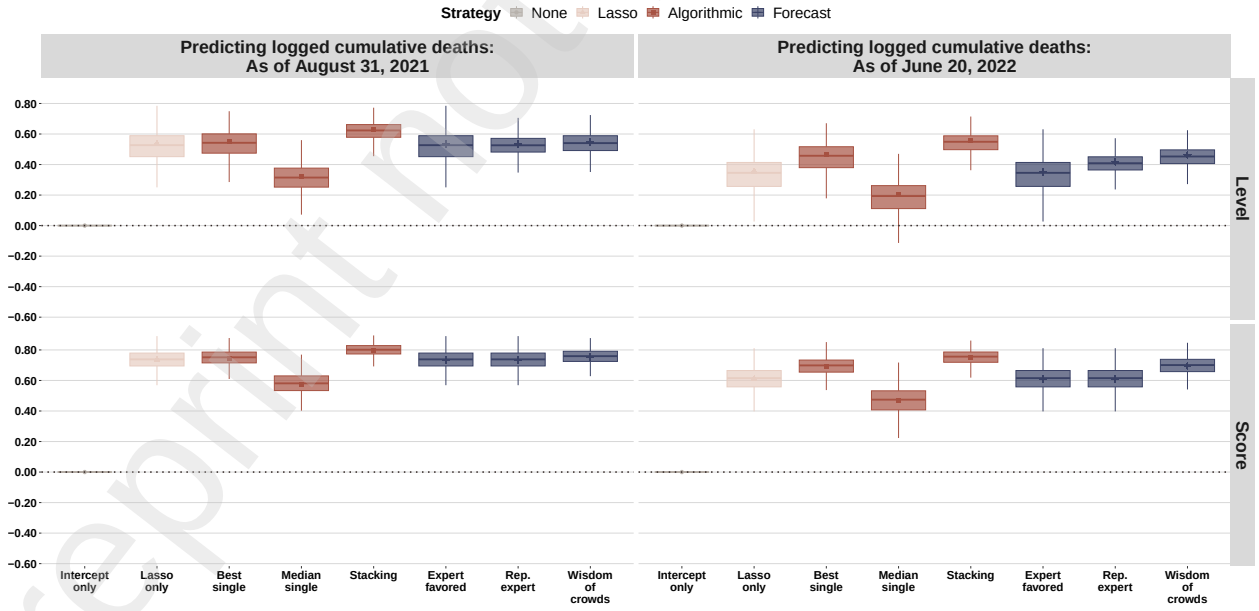


Figure S22: Prediction aggregation metrics for general models for the USA.

Challenge	# of models in		Estimate					
	Algorithmic metrics	Elicited metrics	Best single	Median single	Stacking	Expert favored	Rep. expert	Wisdom of crowds
Crossnational, general	28	26	0.483	0.169	0.532	0.420	0.345	0.364
Crossnational, parameterized	16	-	0.141	-0.369	0.477	*Not applicable*		
India, general	9	9	0.422	0.295	0.582	0.094	0.447	0.489
India, parameterized	7	-	-1.67	-3.11	-1.67	*Not applicable*		
Mexico, general	9	9	0.455	0.040	0.543	0.345	0.418	0.431
Mexico, parameterized	6	-	0.619	-4.76	0.619	*Not applicable*		
USA, general	19	18	0.549	0.325	0.631	0.536	0.535	0.547
USA, parameterized	7	-	-0.102	-5.56	-0.098	*Not applicable*		

Table S12: Summary of model aggregation metrics across challenges. Because we did not elicit expert forecasts over parameterized models, we do not include those metrics for them.

Forecast	August 2021 Forecast		August 2022 Forecast	
	Favored Model	Forecast Weight	Favored Model	Forecast Weight
Crossnational				
Horserace	Trust, Development, and State	0.513	Trust, Development, and State	0.512
Stacking	Government Capacity and Social Inequality	0.347	Government Capacity and Social Inequality	0.365
India				
Horserace	Minority Representation and Urbanization	0.831	Minority Representation and Urbanization	0.751
Stacking	Minority Representation and Urbanization	0.292	Minority Representation and Urbanization	0.259
Mexico				
Horserace	Investment Inequality	0.439	Investment Inequality	0.414
Stacking	Trust Poverty and TB	0.394	Trust Poverty and TB	0.409
USA				
Horserace	Health	0.493	Health	0.469
Stacking	Lasso	0.311	Lasso	0.302

Table S13: Expert-favored models using two different forecast weights.

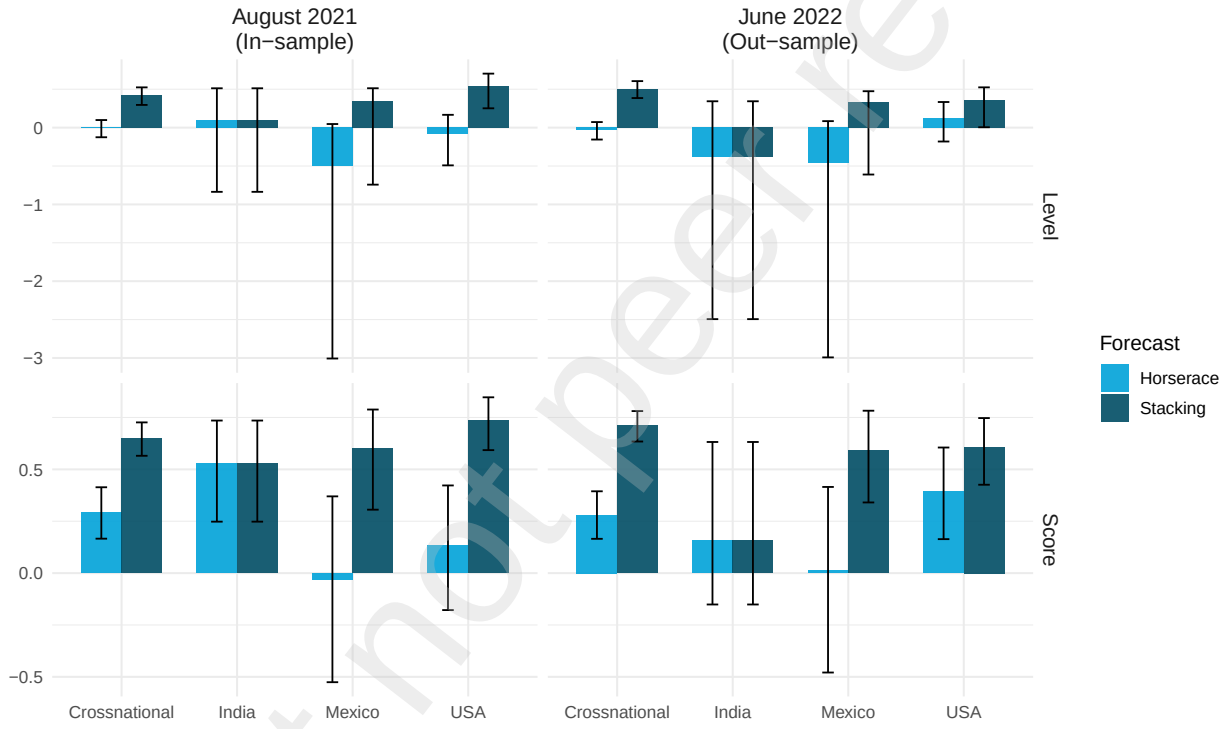


Figure S23: Comparison of performances for the expert-favored models selected using two types of forecast weights.

Table S14: Comparing model performances between user-submitted and simulated three-predictor models.

	User-submitted						Simulated					
Challenge	Mean	SD	95% CI (L)	95% CI (U)	Skew	Kurtosis	Mean	SD	95% CI (L)	95% CI (U)	Skew	Kurtosis
Crossnational	0.11	0.24	-0.39	0.44	-0.83	3.55	0.03	0.21	-0.47	0.37	-1.34	6.69
India	0.19	0.21	-0.14	0.41	-0.74	2.08	-0.13	0.35	-1.00	0.35	-1.02	3.53
Mexico	-0.04	0.49	-0.90	0.44	-0.79	2.51	-0.22	0.40	-1.00	0.37	-0.72	2.71
USA	0.30	0.20	-0.11	0.54	-0.80	2.90	-0.06	0.19	-0.50	0.29	-1.72	10.39

S6 Simulating Model Selection by Machine

We extend the analysis from Figure 5 to the other challenges in this section. First, we outline our algorithm for sampling of models. The sampling strategy parallels the format of the MCs and the Shiny app that was provided to modelers. For each challenge we:

1. Randomly sample three predictors from the MC predictors.
2. Randomly select one type of model: polynomial (quadratic), interaction, or neither, each with probability 1/3.
3. For polynomial or interaction models, we follow the Shiny menu of options to select terms to be excluded from the statistical model. We do so by generating a Bernoulli random variable (with $p = 0.5$) for each term and including the term if the draw takes the value 1 and omitting the term if the draw takes the value 0.

Following this algorithm, we sample $5000 \times M_c$ (M_c is the total number of user-submitted models in each challenge c) models per challenge. Figure S24 shows the performance of the randomly generated models relative to the user submitted models in each MC. Table S14 presents the corresponding summary statistics for the two types of models per MC. Figure S25 compares the performance of the stacking model estimated on the user submitted relative to the stacking model estimated on equivalent-sized sets of randomly generated models. In three of four challenges, our estimated stacking models outperform every simulated model. In the Mexico challenge, 6.9 percent of the simulated models outperform our estimated stacking models ($p = 0.069$). Consistent with our interpretation of Figure 5, this suggests that the best user-submitted models outperform machine-selected models. These highly-predictive models yield performance gains of the stacking meta-model.

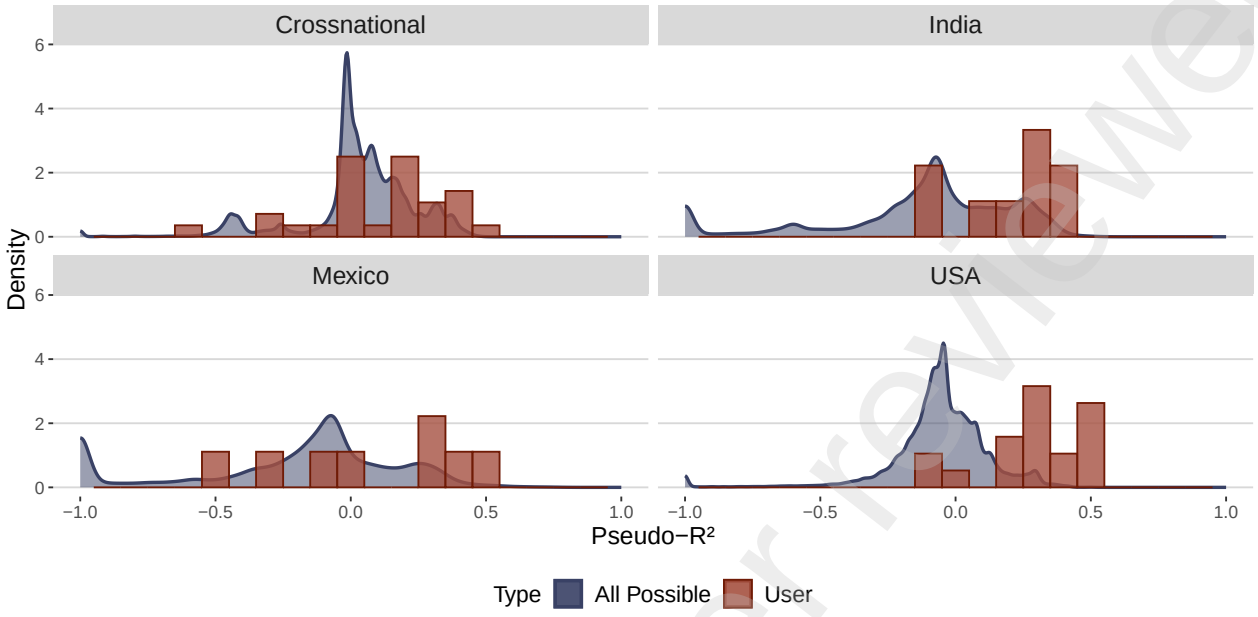


Figure S24: Horserace simulation. The density plots represent the distribution of pseudo-R^2's from 5,000 sets of simulated three-predictor models in the common MC datasets for each challenge.

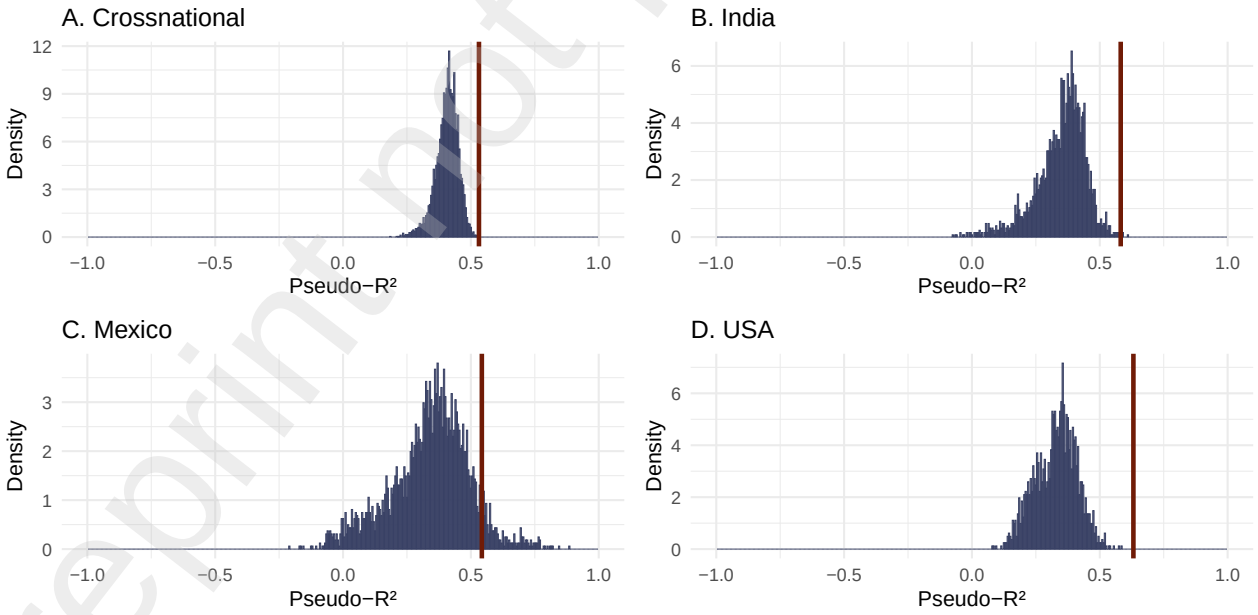


Figure S25: Stacking simulation. The histograms represent the distribution of pseudo-R^2's from 5,000 simulated stacking models in the common MC datasets for each challenge.

Challenge	% of observations remaining	
	General	Parameterized
Crossnational	1.81%	7.8%
India	64.5%	87.1%
Mexico	100%	100%
USA	92%	92%

Table S15: Percent of observations remaining after listwise deletion for any missing predictor.

S7 Missing Predictors and Imputation

In this section, we examine the robustness of our model evaluation to alternative treatments of missing predictors. In the model challenge, instructions read:

> "If a proposed model uses data with missing observations, the model will be fit without any imputation (unless an imputation procedure is provided). The predicted values from the model will then have missing values. Missing predictions will then be implemented using average predictions from the model. The idea is that models are assessed on how they predict for the full set of cases and so a full set of predictions should be provided, even if these are based on incomplete data."

We can see the implementation of this approach in Figures 2 and S8-S10 where there is a mass of points in a vertical line at the mean. We prespecified two alternate approaches to missingness. First, we prespecified evaluating models on only cases without missing data (i.e., listwise deletion). This strategy is not viable for the crossnational challenge given high levels of missingness, as shown in Table S15.

In addition to inherent uncertainty regarding the future trajectory of the COVID-19 phenomenon, the variables that we provided were not all fully available and some covariates had more missingness than others. Our primary approach to missingness, which was communicated to MC participants, was to impute the sample mean for observations with missing predictors. If a submission included an imputation algorithm with a model, we treat the algorithm as part of the model. In Figure 2, we depict observations with missing data for any predictor as points that appear on vertical lines. As the data in the figure show, the predictive accuracy of many weaker-performing models is limited by missing data. We view inaccurate predictions stemming from missing data as an artifact of the prediction exercise.

Due to the excessive level of listwise missingness in the crossnational data we have adopted a multiple imputation procedure to deal with missingness in our model data. We use the Multivariate Imputation by Chained Equations (MICE) algorithm to impute missing observations in each variable according to its level of measurement and as a function of synthetic values in other variables in the same dataset. For each challenge, we obtain a fully imputated challenge datasets from this procedure, re-run our models on the new data and replicate our main analyses on the updated models without missing inputs. We depict the results in Figures S26 to S28 where pre- and post- imputation results are shown side-by-side.

Generally speaking, input imputation has lead to an improvement in model performance across all challenges in all analyses. Unsurprisingly, the greatest improvement occurs in cases most affected by missingness. In particular, the pre-prediction imputation led to a reshuffling of mode performance rankings in the evaluating analyses, where previously worst-affected models now usually achieve higher pseudo-R^2 and stacking weights than before, compared to their less-affected competitors. In contrast, the aggregating analysis experiences less significant improvement in terms of aggregate model performances.

In the Mexican case, there are no missing predictors among those used in parameterized models, so its results are the same before and after imputation. For concise display, we omit outputs for the Mexican challenge in this section.

S7.1 Evaluating

This section compares the results of the evaluation analysis with mean-imputation to those with multiple-imputation. We report results for the crossnational general challenges where missingness was most severe (per Table S15). Figure S26 shows that model performance with respect to pseudo-R^2 improves substantially with multiple- over mean-imputation. Figure S27 shows that top-performing models with respect to both pseudo-R^2 and stacking weights changes also changes considerably in their composition after multiple imputation.

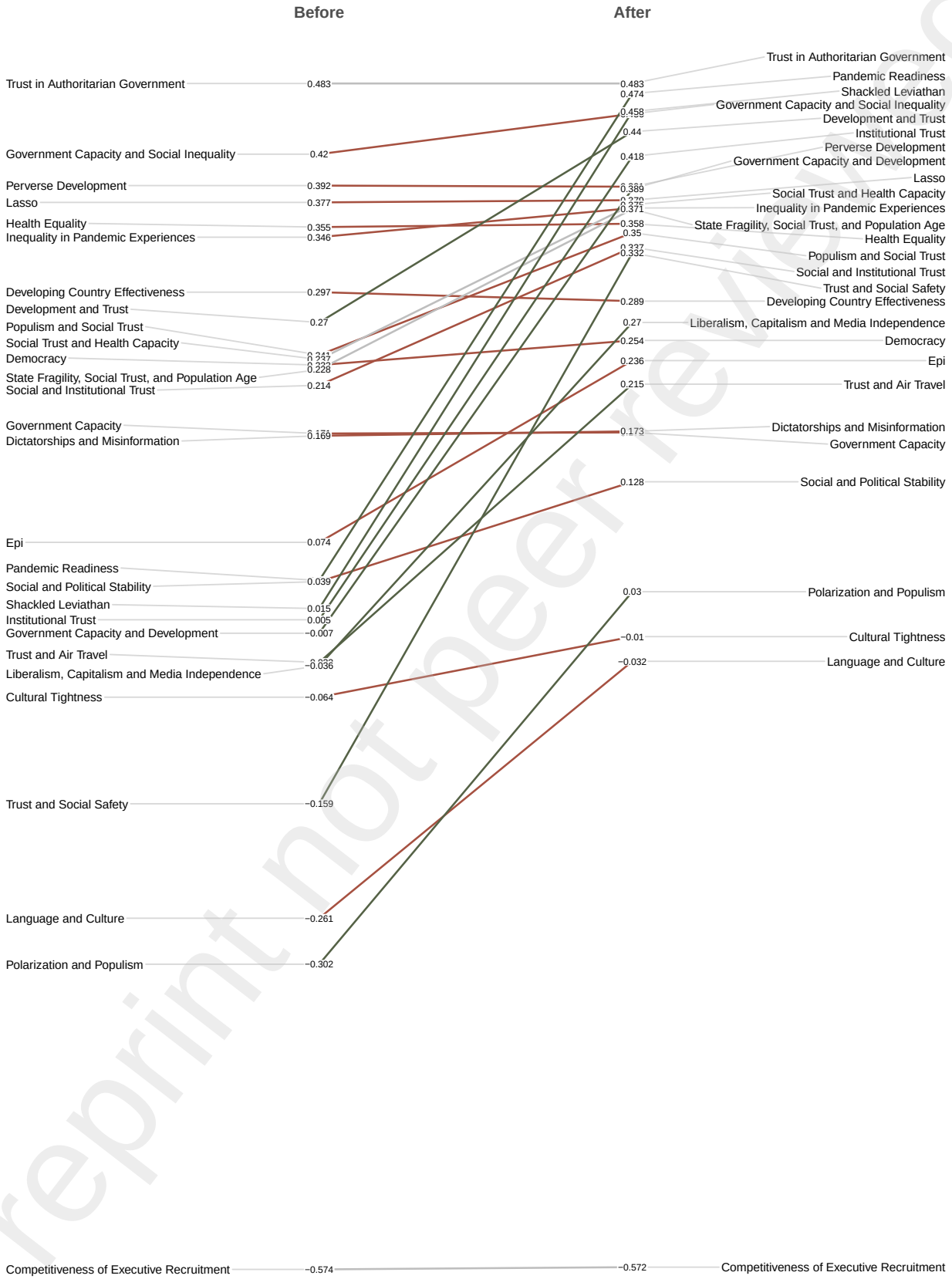


Figure S26: Pre- and post-imputation performances of the general models for crossnational data.

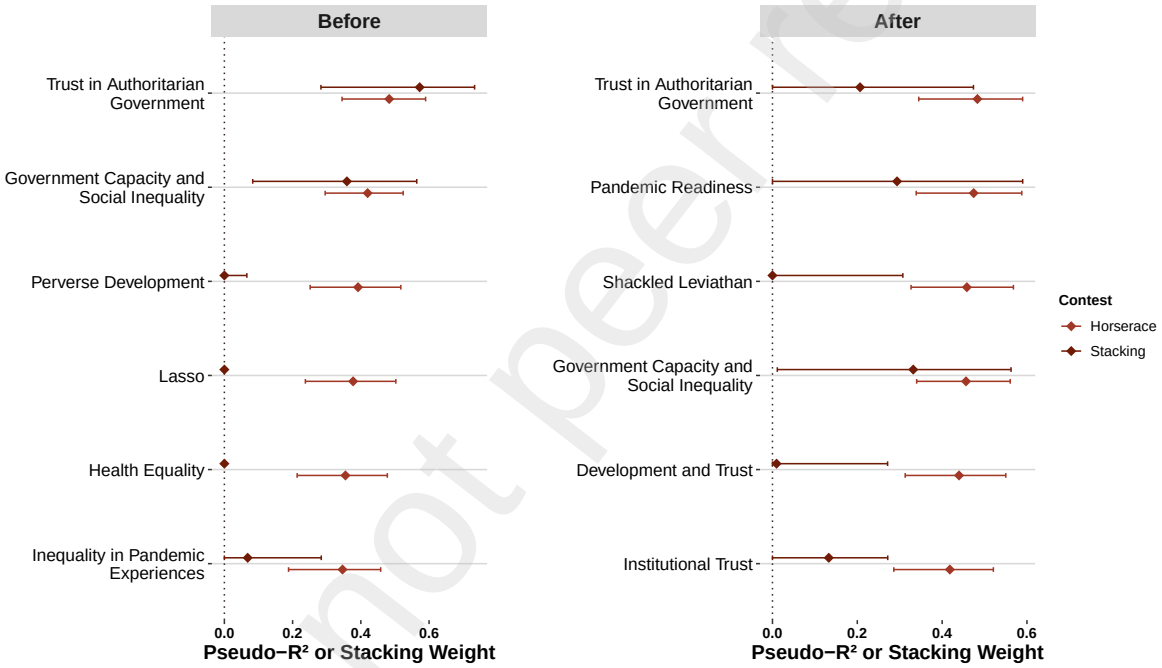


Figure S27: Model selection before and after imputation for crossnational general models.

S7.2 Aggregating

Figure S28 compares the results reported in Figure 4 to those generated with multiple imputation. Imputation improves predictive performance of all models. However, the main results described in the main text are maintained.

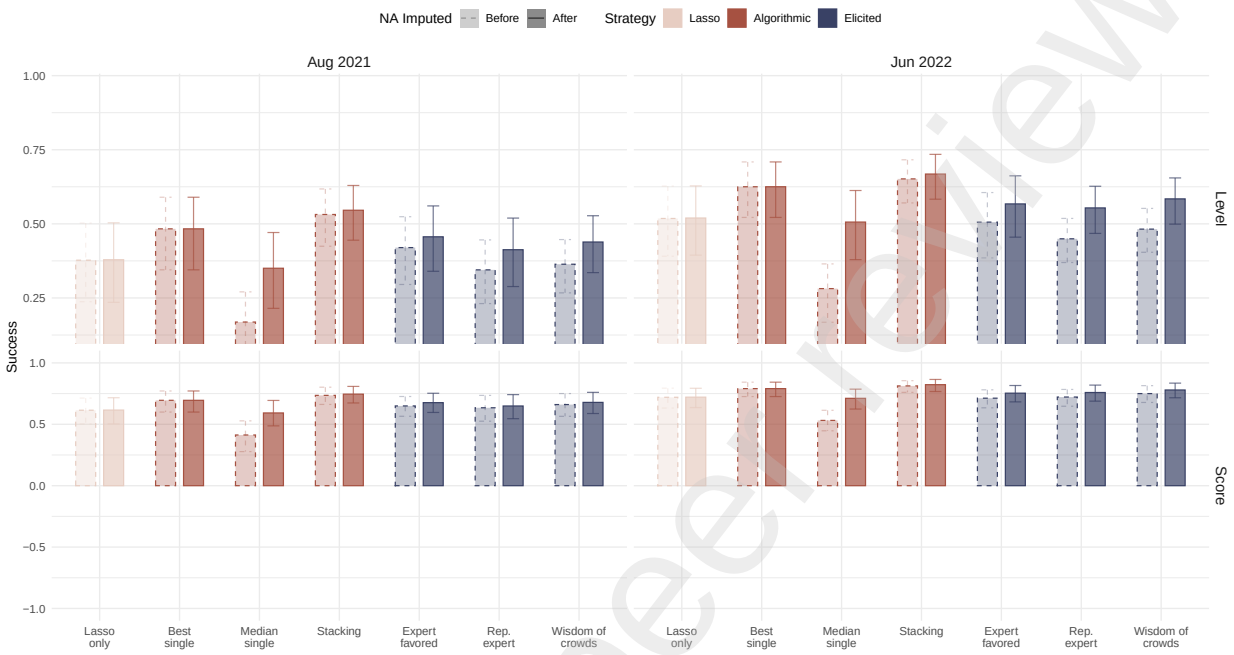


Figure S28: Prediction aggregation metrics for crossnational general models before and after imputation.

S8 Deviations from Pre-Analysis Plan

1. In our metric of model success for the level approach, we prespecified:

$$v_k = -\sum_i (\widehat{y}_{ik} - y_i)^2$$

   To facilitate cleaner interpretation of v_k, we have normalized this expression to (7).

2. The pre-analysis plan suggests three steps: gathering, selecting, and aggregating. We have reconceptualized the gathering stage in this paper to focus on the content of submitted models and not predictive performance. Analyses that were pre-specified as "gathering" and "selecting" have now been combined and are reported as "evaluating."

3. In our evaluating analysis, we pre-specified examining the Lasso-residualized pseudo-R^2 measure for the horserace. However, in the forecasting, respondents were not asked to make model predictions relative to the Lasso model. To improve comparability across the two types of horserace evaluating exercises, we have not residualized the models in the algorithmic approach.

4. Due to the high level of missingness with listwise deletion of observations with missing covariates (see Table S15), we do not examine the robustness of our metrics of predictive accuracy on this subset of observations. We do implement multiple imputation to assess robustness.

5. Due to one model challenge participant submitting two pairs of identical models in the crossnational challenge (one general and one parameterized) we have removed one model from each model type. This reduces the total number of submitted models from 90 reported in the PAP to the 88 analyzed in this paper. All figures reported in the paper reflect the actual number of models evaluated in the analyses discussed above.

6. The pre-analysis plan underspecified the source of the elicited predictions in the aggregation step. Figure 4 uses predictions from the stacking forecasts only. We include Figure S23 to show that this indicates that our approach (using stacking forecasts), if anything, overstates expert abilities to aggregate.