Gathering, evaluating, and aggregating social scientific models of COVID-19 mortality

22 March 2022; This version 11 September, 2022

Abstract

The social sciences generate multiple explanations for outcomes but lack formal procedures to assess competing explanations. We gather, evaluate, and aggregate explanations for COVID-19 mortality within and across countries. To gather models, we sponsored an open challenge. Models submitted vary in their predictive accuracy but the best predict mortality more accurately than a machine learning benchmark. We then use three algorithms and expert forecasts of model performance to aggregate models. A stacking estimator combines models to offer the most accurate predictions. Expert forecasters are less accurate than the stacking algorithm, suggesting limits to social scientists' ability to evaluate multiple explanations. The main contribution of our paper is to provide evidence favoring a disciplined method for aggregating multiple explanations for common social scientific outcomes. [123 words]

1 Introduction

On what basis can we claim that a scholarly community understands some phenomenon well? Do, for instance, social scientists understand the stark crossnational or subnational variation in COVID-19 mortality?

Social scientific contributions for any particular outcome generally come in the form of discrete rival explanations. To assess collective knowledge, we need a way to survey these accounts, evaluate them, and then aggregate the best features of rival explanations. The aggregation challenge represents an important and largely neglected problem in social science methodological discussions.

There are currently no standardized procedures available to filter and aggregate various explanations made by a community of scholars. In some cases — when there are both a common cause and a common outcome — meta-analysis represents a viable approach (Borenstein et al. 2021). But there is no methodological equivalent to meta-analysis when the question of interest relates to a single outcome but accounts differ in the types of explanations (or variables and functional forms) they provide (Debray et al. 2014). The latter case encompasses much literature in social scientific disciplines. How can we filter and aggregate knowledge across different models or explainations of the same outcome?

In this paper, we address these questions by field testing tools for filtering and aggregating social scientific explanations for a single prominent outcome: national and subnational levels of COVID-19 mortality. We *gather*, *evaluate*, and then *aggregate* rival explanations, studying the effectiveness of different methods that filter and aggregate.

In the gathering stage, we invited researchers to take part in a a set of four related open challenges and to propose statistical models to predict future outcomes for our variable of interest; namely, crossnational and subnational patterns of COVID-19 mortality. In contrast to more interactive approaches, such as that developed in Abernethy and Frongillo (2011), we provided a simple web interface to crowdsource model submissions, giving independent teams of is participants access to common data but not to other submissions. We received 88 submissions from 60 different individuals based at 32 institutions in 10 countries (see Table S12).

In the evaluation stage, we implemented and assessed the performance of the submitted models on future data (data, that is, that had not yet been available at the time of model construction and submission) and identified the best performing models according to predictive accuracy. To assess predictive accuracy, we first rank all the submitted models according to their out-of-sample pseudo-R^2. The pseudo-R^2 of the best model is 0.483 whereas that of the median model is only 0.171, indicating wide variation in the quality of the submitted models. We also compare the performance on pseudo-R^2 of submitted models to that of a Lasso model — a workhorse machine learning (ML) model — fit on all of the common predictors. The pseudo-R^2 of the Lasso model is 0.377 and it ranks behind three of the submitted models in the main

(crossnational) challenge. All of the models that outperform Lasso are theoretically- or substantively-motivated rather than generated by machine learning algorithms.

We then evaluate models using two additional methods. First, we implement a stacking estimator to generate a meta-model comprising all submitted models (Yao et al. 2021).[1] The stacking estimator allocates weights to the predictions of each constituent model so as to maximize the predictive accuracy of the meta-model. We find that the stacking estimator generally places non-zero weights on only a handful of models (just three in the crossnational challenge), effectively filtering out information from other submitted models. The second method that we use to evaluate models uses expert forecasting. We ask each volunteer forecaster to predict the pseudo-R^2 ranking or the stacking weights of a subset of models. Comparing the estimated pseudo-R^2's and stacking weights generated algorithmically to those provided by expert forecasters demonstrates that experts are unable to identify accurately the most predictive models and are also unable to allocate stacking weights that resemble those that are estimated algorithmically. This comparison suggests stark limits to social scientists' abilitites to filter theories and explanations of common outcomes when synthesizing a literature: expert forecasts of the predictive accuracy of statistical models do not resemble the actual predictive accuracy of these models.

In the aggregation stage, we implemented six separate methods to filter and combine models, methods which we detail below. The stacking method outperforms the alternatives by construction because stacking makes a composite prediction computed from a weighted sum of model predictions. The composite prediction generated by stacking outperforms the best single submitted model (by construction), but more importantly, we show that it outperforms the "typical" (median) model by a large margin. In other words, the median model submitted to the challenge is not very good at predicting COVID-19 mortality. Stacking also greatly outperforms the aggregate predictions derived using three different metrics based on the expert forecasts.

This scaffolding of analytic procedures allows us to address several meta-scientific questions that are central to social science: How can we assess the strength of competing arguments? Can distinct explanations be filtered and combined to help us better explain outcomes we care about? Are expert consumers of research capable of accurately characterizing or prioritizing explanations that better predict outcomes? Our procedures also provide an answer to a central substantive question: how can we characterize the best collective views of social scientists about the determinants of variation in COVID-19 mortality? Does social scientific expertise improve significantly our understanding of variations in COVID-19 mortality?

Our study provides proof of concept of the utility of harnessing collective knowledge in the social sciences to respond to an empirical question of substantive importance. Although the stacking estimator we implement has often been used in the literature to evaluate competing models, to our knowledge it has not been used in a meta-scientific context to filter and combine rival explanations of a common outcome. Strikingly, the best prediction that stacking produces draws directly on statistical models provided by only a few scholars, models that use a small but diverse set of predictors that include not only features of government and the political system but also aspects of social structure. Substantively, we find that COVID-19 mortality across and within countries reflects deep structural features of society that are likely resistant to short-term modification.

2 The Problem

In the social sciences, scholars simultaneously develop and test many explanations for important political and social outcomes. As a scholarly community, we have theorized about the causes of economic growth, government corruption, political democratization, and collective violence, among other outcomes. Yet, when we attempt to advance our understanding of these outcomes, we tend to search for new explanations rather than to (re-)evaluate or to synthesize existing ones. While development of new theories and arguments is clearly important, this individualistic, competitive process means we confront many, often disjointed, explanations for core outcomes in our disciplines rather than building on and accumulating what we know (Watts 2017; Davis 2015).

There are many reasons why the process of knowledge aggregation and accumulation is so fragmented and piecemeal. Part of the problem stems from the simple fact that professional incentives favor novelty, which leads researchers to

[1] Stacking simultaneously evaluates and aggregates; in this step, we discuss how we use stacking to filter models that receive zero weight in the meta-model.

search for new answers instead of evaluating those previously proposed by others (Koole and Lakens 2012; Nosek, Spies, and Motyl 2012; Galiani, Gertler, and Romero 2017). Furthermore, disciplinary norms encouraging the simultaneous development of theoretical and empirical knowledge arguably lead to multiple relatively "thin" but (possibly) testable theories. A deeper reason may be that for many problems of interest, social scientists hold out little hope for arriving at complete explanations in the first place and so focus attention less on explaining outcomes — "causes of effects" — than on explaining the effects of given causes (Pearl 2015; Gelman and Imbens 2013). While focusing on effects of causes is of clear value — it is essential, for instance, for determining whether to implement a given policy — it risks leaving the basic explanatory question unanswered.[2]

In an environment of many explanations for the same phenomenon, it is hard to assess the merits of competing explanations or to evaluate the potential complementarities that may exist between different explanations of the same outcome. This can lead to the use of heuristics when attempting to understand or review a given literature, heuristics that may in turn limit our understanding of the social world (Jolly and Chang 2019). For example, absent clear and professionally accepted strategies for knowledge aggregation, we may, by default, tend to trust or accept findings from early studies of a particular outcome or, more worryingly, accept findings based on the identities or professional affiliations of researchers over an unbiased evaluation of the theoretical or empirical merits of the work. In what follows, we implement both algorithmic and expert evaluations of explanations. This allows us to compare how well social scientists do at evaluating and combining rival explanations with the results of formal statistical methods.

Social scientists typically follow a few common approaches when attempting to aggregate and synthesize knowledge in a particular area. First, experts in a given field may write analytic reviews of existing research, in the form of handbooks organized around a particular outcome or comprehensive literature reviews that are published in journals such as the *Annual Reviews* series. Analytic reviews can identify important gaps in a theory or in evidence that underpins central claims in a literature. Reviews may also propose new connections between previously disconnected explanations. But despite the value of this form of synthesis, it also has limitations. Absent a common set of data or a more standardized and systematic analytic approach, it is difficult to compare the merits of competing (or unrelated) explanations.

A more formal approach to aggregation within a field of study is meta-analysis. Meta-analysis combines evidence from multiple studies that estimate the same relationship of interest. In general, recent meta-analyses hone in on the relationship between a single cause (or treatment) Z and a set of outcomes Y, which are measured in multiple studies conducted in different settings. Prominent recent examples in the social sciences include Banerjee et al. (2015), Dunning et al. (2019), Coppock, Hill, and Vavreck (2020), Slough et al. (2021), and Blair et al. (2021). When constituent studies are internally valid, measure the effects of a common externally valid mechanism, and utilize harmonized study designs, meta-analysis provides an estimate of a common treatment effect (or average outcome) across studies (Slough and Tyson 2022). Given a sufficient number of studies and estimates, a meta-analysis can also provide estimates of heterogeneity in effect size across settings or subpopulations. However, meta-analysis is not a suitable method to assess or combine multiple explanations of a common outcome.

Faced with many potential determinants of a common outcome — in this case, mortality from COVID-19 — we implement and compare various strategies for evaluating and aggregating evidence. Our strategies focus on evaluating the predictive accuracy of different explanations that all use common data. In our work, each explanation is represented as a statistical model. We compare models according to their out-of-sample predictive performance. In addition to comparing models, we implement a stacking estimator that generates a combined prediction across all models. We provide evidence in what follows that the stacking estimator performs better than other methods in generating a filtered and combined model with good predictive accuracy.

Table 1 summarizes the differences between the three frameworks for aggregation that we have just described.

3 The COVID-19 Model Challenges

We apply our approach to aggregating multiple explanations of a common outcome in the context of the COVID-19 pandemic. The pandemic rapidly sparked a large body of social scientific work on its political, social, and behavioral determinants and outcomes (Acharya, Gerring, and Reeves 2020; Bargain and Aminjonov 2020; Bosancianu et al. 2021;

[2] Another deeper reason is that there is no reason to expect that even "complete" explanations would be unique: for instance, an explanation of a phenomenon that focuses on a given cause can be replaced by an explanation that replaces the cause with a cause of that cause (Strevens 2011). We return to this point below.

Table 1: Characteristics of research designs that aggregate evidence in social science

	Review Essay	Meta-Analysis	Our Approach
# Treatments	Any	One	Many
# Outcomes	Any	≥ 1 (each measured in each study)	One
Sample	Any	Multiple	Common
Quantity of interest	Unclear	Common structural parameters across studies or samples.	Metrics of predictive accuracy

Cepaluni, Dorsch, and Branyiczki 2020; Cepaluni, Dorsch, and Dzebo 2021; Elgar, Stefaniak, and Wohl 2020; Han et al. 2021; Min 2020). This unprecedented wave of research has reproduced many general features of the social science literature at breakneck speed as researchers have produced a bevy of, at times, 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 the evidence that has accumulated (two exceptions are Piquero et al. (2021) and Robinson et al. (2021)). The rapid production of COVID-19 research arguably exacerbates the problem we described above concerning the proliferation of disconnected arguments and empirical findings.

However, understandably broad interest in COVID-19 also provides a convenient opportunity to develop a more general strategy for aggregating research findings. To this end, we designed and implemented a set of COVID-19 Model Challenges (MCs). The MCs encouraged researchers to develop and submit statistical models that use political and/or social variables to predict logged cumulative COVID-19 mortality per million people as of August 31, 2021 on a specified sample of data (see Figures S1-S2). Submitters designed their models in December 2020 and January 2021. We incentivized submissions by offering co-authorship to those who submitted the most predictive models. These "modelers" comprise co-authors of the current paper. The P.I.s provided an interactive web platform with clean, harmonized covariates and outcome data on COVID-19 mortality through November 16, 2020 that modelers could use in the design and submission of their models (see Figure S3).

We elicited models predicting COVID-19 mortality for four separate data samples: crossnationally for 168 countries around the world and subnationally for states in India, Mexico, and the United States of America (USA). India, Mexico, and the USA are federal countries in which (some) public-health policy is made by the states. For each of these samples, we elicited both "general" and "parameterized" models. General models specify the functional form — but not the value of parameters — of the statistical models whereas parameterized models specify both the functional form and model parameters. We focus on results of the crossnational general models in this paper and report findings from all eight MCs in the Supplementary Information.

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. These two models benchmark submitted models; the Lasso model in particular provides a widely-used machine learning algorithm that produces highly interpretable models akin to the MC submissions (Tibshirani 1996). We evaluate models on the basis of their predictive power. Our primary metric of predictive accuracy, given by Equation (1), resembles R^2 but is evaluated for the general models using leave-one-out predictions. We also evaluate the correlation between leave-one-out predictions and observed outcomes, given by 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)). This estimator combines predictions from models by putting weights on each model. We then generate an aggregate prediction as a convex combination of leave-one-out model predictions weighted by the estimated stacking weights.

How much do we gain from the formal comparison of models and from the stacking aggregation? To answer these questions, we elicit expert forecasts of the performance of the models submitted to the Model Challenges using the Social Science Prediction Platform.[3] Forecasting results has become an increasingly common practice across the social

[3]The platform is accessed at https://socialscienceprediction.org/.

sciences (DellaVigna and Pope 2018; DellaVigna, Pope, and Vivalt 2019; DellaVigna, Otis, and Vivalt 2020). In our context, forecasting allows us to evaluate the correspondence between algorithmic approaches to comparing and combining social science analyses and the assessments of experts, who in turn constitute the consumers (and reviewers) of social science research. In February and March 2021, we elicited expert forecasts of the performance of the models that had been submitted to the MCs. We compare estimated predictive performance of models to the assessment of a "representative expert" — proxied by the median-performing stacking forecast — as well to the assessment made by the "wisdom of the crowds" — proxied by the average stacking weights made by all forecasters.

We randomly assigned expert forecasters into two groups to elicit two forms of forecasts: a horserace and a stacking forecast. In the horserace forecast, experts see a subset of six randomly-selected examples of the general models that were submitted to a given challenge and guess the probability that a model will be the most predictive in the set. In the stacking exercise, forecasters allocate weights across models — analogous to those generated through a stacking analysis — over a subset of seven randomly-selected models. We compare the implied rankings of models (in terms of predictive ability or stacking weight) that are forecast to those generated by the analogous algorithm. We also construct the aggregate prediction implied by the forecasts of a representative expert (from the stacking arm only) and by the "wisdom of the crowds" (also from the stacking arm). We implement these analyses to evaluate what our approach to aggregation adds over how experts process the literature.

4 Results

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

4.1 Gathering Models

We received a total of 88 model submissions across the model challenges. More models were submitted to the crossnational challenge than to any of the country-specific challenges: 42 percent of all models were submitted to the crossnational challenge. Sixty-four percent of models (across challenges) are general models and the other 36 percent are parameterized models (Table S5). Among the models submitted to the crossnational challenge, most exhaust the budget that allowed submissions to include no more than three unique predictive variables (Table S11). Ten models introduce user-submitted predictors not included in the MC dataset that had been assembled by the P.I.s. A majority (57 percent) of models use only linear functional forms rather than interactions between predictors or polynomial terms. The three machine learning-based models that were submitted were more likely to use more complex functional forms. Each model was accompanied by a theoretical justification, though the justifications range in specificity and coherence.

A majority (75 percent) of models were submitted by solo researchers. About a quarter of submissions came from teams of up to eight researchers, although most commonly teams of two. Overall, the submissions reflect the work of 60 modelers based at 32 institutions in 10 countries (Table S12). We cannot easily establish the representativeness of these researchers relative to any specific scholarly community. There may be many dimensions upon which model submitters are different than the broader community of social scientific researchers: they be more likely to volunteer or more interested in COVID-19 or in public health generally. However, it is unlikely that any body of literature is representative of a community of scholars as individuals, so in this respect the MCs do not differ from other processes which place explanations in the public domain.

Figure 1 provides a summary of the most common predictors employed across models in the crossnational MCs. 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 were employed. The color coding indicates how frequently pairs of variables were entered together. We see that the most common variables in the submitted models focus on trust and government effectiveness; the most common pairing of variables couples trust in government with health access — a coupling used in three separate submissions. Although most submissions came from scholars with expertise in political science, variables capturing political institutions — such as measures of democracy or measures of political corruption — do not appear frequently. Across all four challenges, 30 percent of models include measures of trust. In contrast, just one general model uses a measure of democracy and no general models reference polyarchy, a standard political science term for polities where power is vested in multiple people (Dahl 1972).[4]

[4]Note that democracy and polyarchy are measured at the country (not subnational) level and were thus only provided in the crossnational

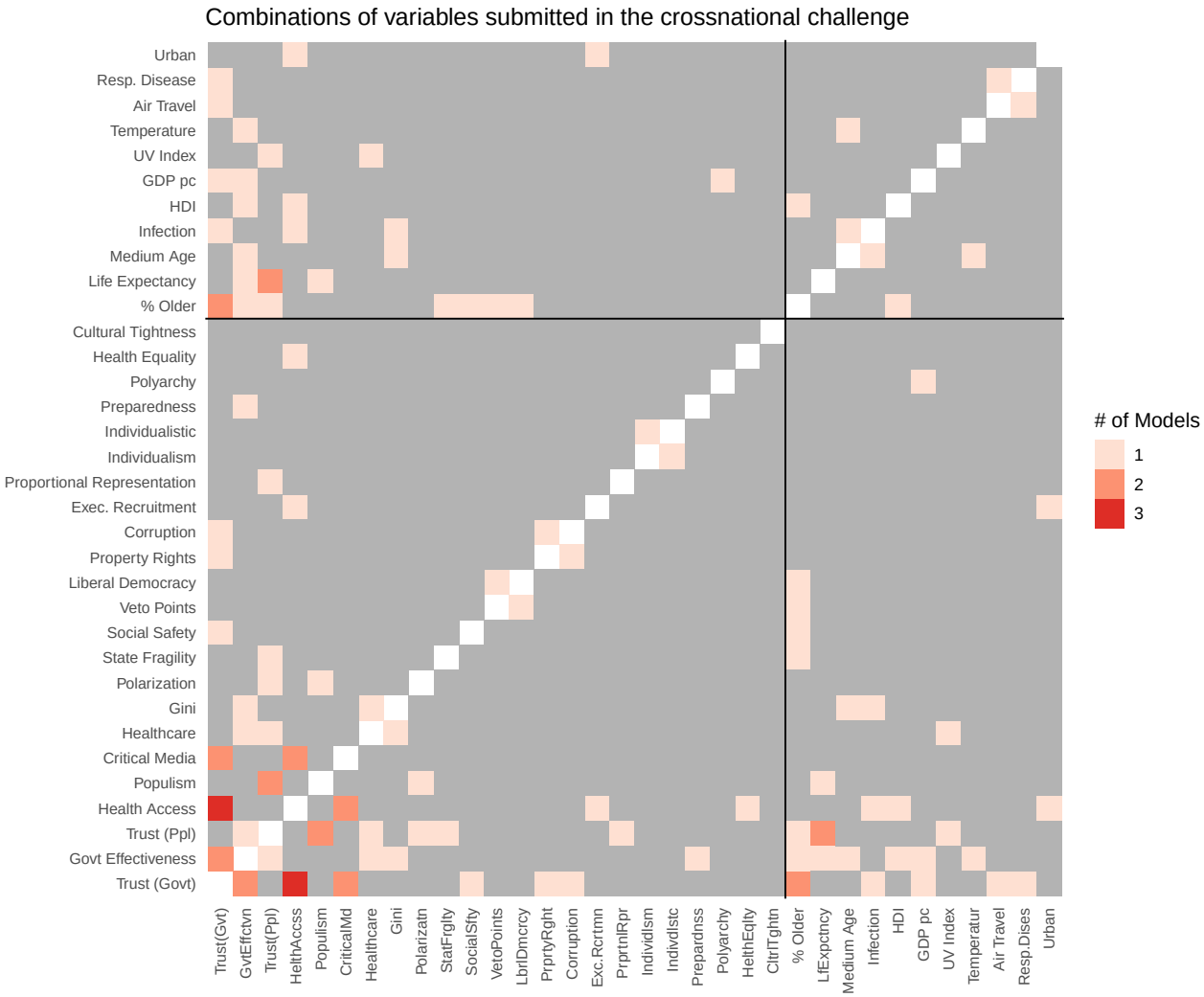


Figure 1: Pairwise combinations of variables submitted to the crossnational MC. *Notes*: The lower left quadrant includes the social and political variables provided to submitters. The upper right quadrant includes other variables provided.

Submitters were asked to provide theoretical justifications of their models and were encouraged to reference relevant scholarly literature. The justifications were concise but not necessarily very specific. Information provided with the submissions shows that individuals who engaged with the MCs had varying levels of social scientific expertise. The modal individual who submitted holds a Ph.D.

4.2 Evaluating Models

We begin our assessment of the models gathered in the crossnational MC by examining the predictive performance of individual submissions.

Figure 2 depicts the leave-one-out predictions from each of the crossnational general models submitted. 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), which specifies leave-one-out predictions for the outcome. If we were to rely on model predictions rather than leave-one-out predictions, this measure would be equivalent to each model's R^2. Following the interpretation of the R^2 measure, a perfectly predictive model would have a pseudo-$R^2 = 1$; and higher values indicate greater predictive power. However, when leave-one-out predictions depart substantially from the predicted values using all observations, the pseudo-R^2 is penalized. This allows the pseudo-R^2 to be arbitrarily negative. The models are ordered from best performing to worst performing according to this metric.

It is clear from inspection of the data shown in Figure 2 that leave-one-out predictions for all models correlate positively with actual COVID-19 mortality. This is mechanical for general models, since parameters are estimated using the outcome data. However, it is also clear 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 rather than making relative comparisons is more challenging. Because the Lasso model is fit on all of the 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. All of the models that outperform Lasso are theoretically- or substantively-motivated rather than generated by machine learning methods. While the pseudo-R^2 metric summarizes the performance of each model, the models can also be represented in regression tables. Table S15 illustrates this by depicting the top three models.

The best performing model is a simple linear model that combines three variables: a measure of trust in government, the presence of a critical mass media, and access to sanitation. The next best performing model combines a measure of governmental effectiveness, the quality of healthcare, and economic inequality, and includes non-linear terms but no interactions among variables. In both cases, the substantive logics accompanying the submissions were simple and included an independent logic justifying the inclusion of each of the three variables.

What kinds of barriers interfere with accurate predictions? It is useful to separate modelers' uncertainty from mechanical artifacts present in the prediction exercise. Models were created and submitted over an eight-week period running from December 2020 through January 2021. The period was one when questions about vaccine availability (Bokemper et al. 2021; Team 2021; Organization 2021; Wouters et al. 2021), efficacy beyond clinical trials (Baden et al. 2021; Folegatti et al. 2021; Logunov et al. 2021; Mulligan et al. 2021; Polack et al. 2021; Voysey et al. 2021), and public willingness to accept vaccination (Figueiredo et al. 2021; Lazarus et al. 2021; Solis Arce et al. 2021) 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.

In addition to this inherent uncertainty regarding the future trajectory of the underlying 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.

challenge.

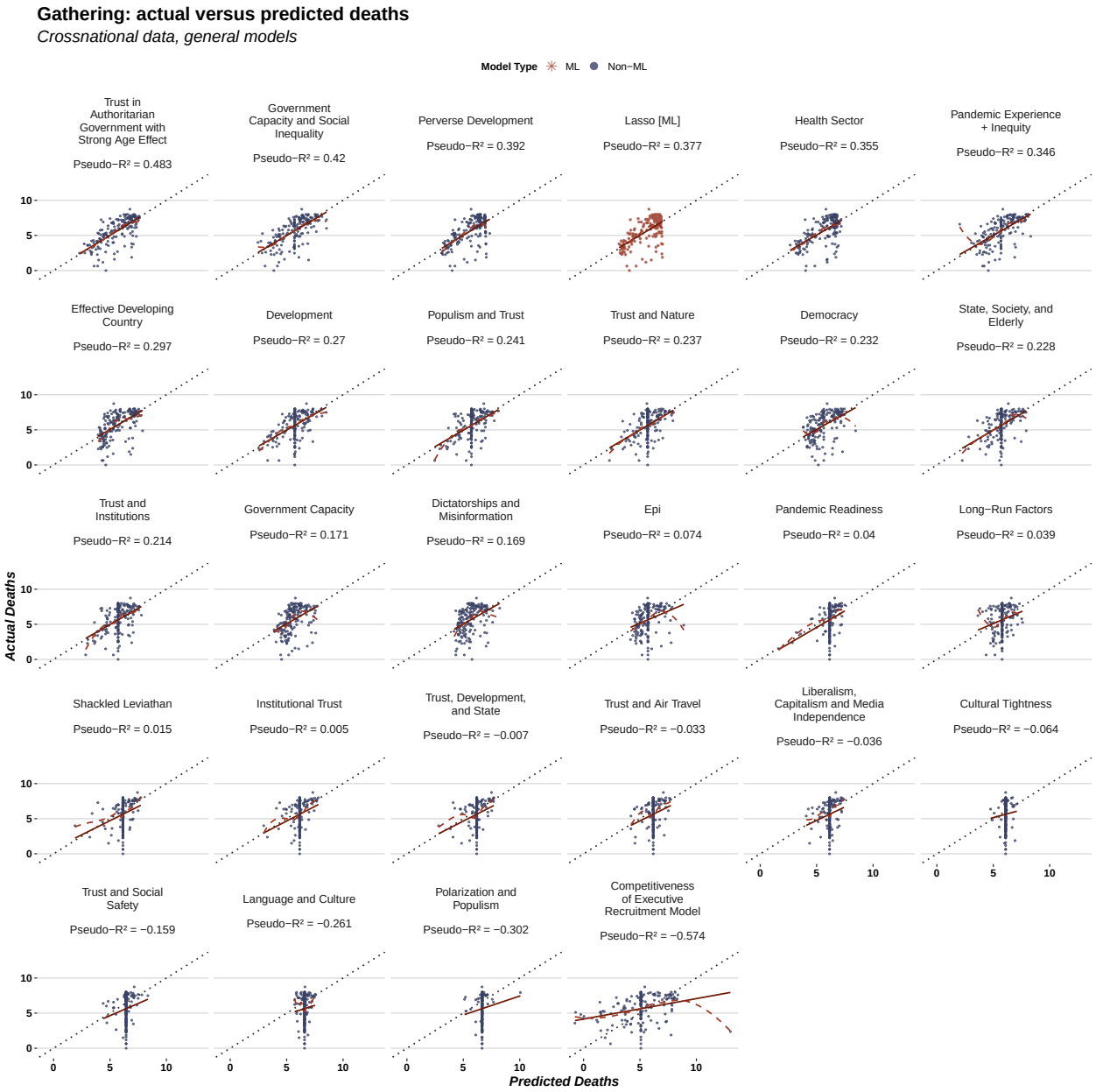


Figure 2: Evaluating: actual versus predicted deaths. *Notes*: Leave-one-out predictions of general models submitted to the crossnational challenge and observed COVID-19 mortality as of August 31, 2021. Facets are ordered from highest to lowest pseudo-R^2.

General models that were submitted to the crossnational challenge exhibit a range of predictive performances. In Figures S10–S16, we present plots analogous to those in Figure 2 for each of the challenges. To summarize the predictive performance of the models, Figure 3 presents plots of the distribution of pseudo-R^2 statistics. The plots show striking differences between general and parameterized models. The general models perform similarly across the three subnational and the crossnational challenges and in every case, they perform better than their parameterized model counterparts (see Table S13). The modal prediction in Figure 3 is censored at -1 in every parameterized challenge, suggesting poor predictive accuracy. Although the best model in terms of pseudo-R^2 across all challenges is a parameterized model for Mexico, it is very much an outlier among the parameterized models generally.

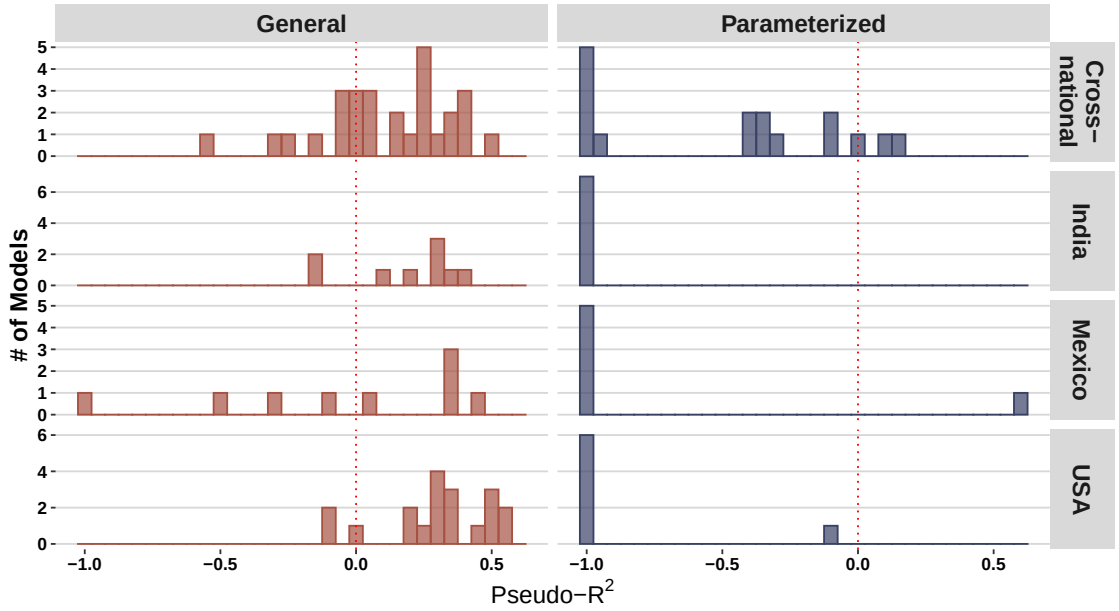


Figure 3: Distribution of pseudo-R^2 statistics for each challenge. *Notes*: Plots in the left column depict general models and plots in the right panel depict parameterized models. We censor observations at -1 and we report scatterplots and pseudo-R^2 statistics for each model in Figures S10-S16.

In the case of parameterized models, a negative pseudo-R^2 may be a function of inaccuracy in the predicted *levels* of mortality (an intercept shift) and/or predictions that correlate negatively with actual COVID-19 mortality.[5] We see strong evidence that most models incorrectly estimate the level of COVID-19 deaths. In the crossnational challenge, 15 of 17 parameterized models underpredict COVID-19 mortality; in the three subnational challenges, all parameterized models underpredict mortality. This is evident from the information displayed in Figures S10, S12, S14, and S16, where the vast majority of observations are situated above the 45^o line. Thus, modelers appear to have been nearly uniformly overly optimistic, in the sense of predicting fewer deaths than the numbers that actually transpired. The median model underpredicts mortality by the greatest degree in India, where cumulative mortality increased 185 percent between the November 2020 data supplied to MC participants and August 31, 2021. By contrast, deaths increased by 66 percent in Mexico and 46 percent in the USA over the same period. In these challenges, underprediction was less severe than in India.

In contrast to the poor prediction of levels (or intercepts) made by the parameterized models, almost all of them correlate positively with actual mortality. We observe a positive correlation between predicted and observed outcomes in 14 of 17 parameterized crossnational models and all country-specific ones. Many of the correlations are quite strong. The correlation between the parameterized predictions and the outcomes for the median model in each challenge ranges from 0.43 in the crossnational challenge to 0.75 in the Mexico challenge. Note that most social science theories make directional rather than point predictions about outcomes. Modelers' predictions of correlations may have been better than their predictions of levels of COVID-19 because the former is a more familiar empirical practice in the social sciences.

[5]Because the parameterized predictions are, by definition, out-of-sample predictions, we do not use leave-one-out predictions for these models.

Given the many models that were submitted in the MCs, how can we prioritize some explanations over others? We propose four metrics to assess model performance. We first evaluate models using two types of contests: a horserace between models and a stacking algorithm. We then compare these results to an expert-elicited implementation of the contests that was collected in the the forecasting exercise.

Figure 4 plots the five top models submitted to the general crossnational challenge that are selected by each metric. In the upper left panel, we report the top five performing models among non-ML models by the pseudo-R^2 metric. The 95% confidence intervals are generated by bootstrapping the data. As we have shown is generally true for crossnational general models in the data depicted in Figure 2, differences between the top five models are limited. The pseudo-R^2 varies from 0.346 to 0.483 across models. In contrast to the algorithmic horserace results, when we aggregate models into a single model very few models receive any non-zero weight. The stacking model places weights of 0.524 and 0.335 on the top two models, distributing the remaining weights that total 0.141 across 26 models. The top two models emerge consistently across selection methods, though the remainder of the models on the lists do not overlap. The Lasso model, our workhorse machine learning model, receives zero weight in the crossnational stacking exercise. The skew of estimated stacking weights towards the two top-performing models is striking.[6] Many models utilize similar predictors; recall that Figure 1 indicates that multiple submissions include measures of trust in government, inequality, development, and democracy. When multiple models include identical predictors, adding positive weights to similar models provides little additional predictive power. It is possible, of course, that some models that receive zero weight would have received a positive weight had they been compared to a different set of other models. Nevertheless, the stacking approach suggests 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 other challenges exhibit similar biases toward a few models, though to a lesser extent than the crossnational challenge (see Figures S18–S21). However, the estimated weights are relatively noisy. This is particularly the case — and unsurprising — for the subnational challenges that have fewer units.[7]

We also examine the stability in the over-time performance of the five top models to ensure that performance is not an artifact of the specific date when we evaluate predictions. Figure S22 depicts the evolution of pseudo-R^2s and estimated stacking weights for each model on a weekly basis throughout 2020 and 2021. Prior to the global spread of COVID-19 mortality in March 2020, no individual model accurately predicts crossnational variation in COVID-19 mortality. Only as COVID-19 spread around the world do models acquire predictive power. Two features of the over-time pseudo-R^2s stand out. First, the pseudo-R^2s for each model move quite smoothly. Second, the top five models as of August 31, 2021 were the same top five models at the time of prediction (albeit in a slightly different order) and they remain in the top seven models throughout the entire post-prediction period.

Analysis of the evolution of stacking weights suggests that the top two models that collectively receive 85 percent of the weights are the best performing models throughout the post-prediction period. After August 31, 2021, however, the weight assigned to the second model — "Government Capacity and Social Inequality" — begins to taper off and the weight assigned to the "Development" model begins to increase. The relatively smooth evolution of these weights offers reassurance that the weights assigned to models on August 31 are generally consistent with the weights attributed to the models throughout 2021.

The expert-elicited horserace and stacking contests implemented during the forecasting exercise generate different sets of top-performing models than those generated by their algorithmic cousins. The horserace forecasts produce no overlap between the top five models identified in the algorithmic and forecast challenges. It is worth noting that the horserace forecast measures the probability that experts believe a model would explain the most variation. The algorithmic horserace measures the variation explained by the leave-one-out predictions of the model. Nevertheless, the lack of overlap among the best-ranking models is notable. This suggests that, at least with regard to distinguishing between different explanations or models, experts — in the aggregate — are not particularly adept at identifying the best-performing model.

With respect to the horserace, there is some overlap between the algorithmic and forecast model evaluations. The "Government Capacity and Social Inequality" model ranks second in the algorithmic stacking and first in the forecast stacking. The "Pandemic Experience and Inequality" model ranks third in both implementations. The weights in the stacking that were elicited in forecasting are less skewed than in the algorithmic stacking. However, weights are

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

[7] In the crossnational sample, there are 168 countries; in the subnational challenges, there are 31 Indian states, 32 Mexican states, and 50 states in the USA.

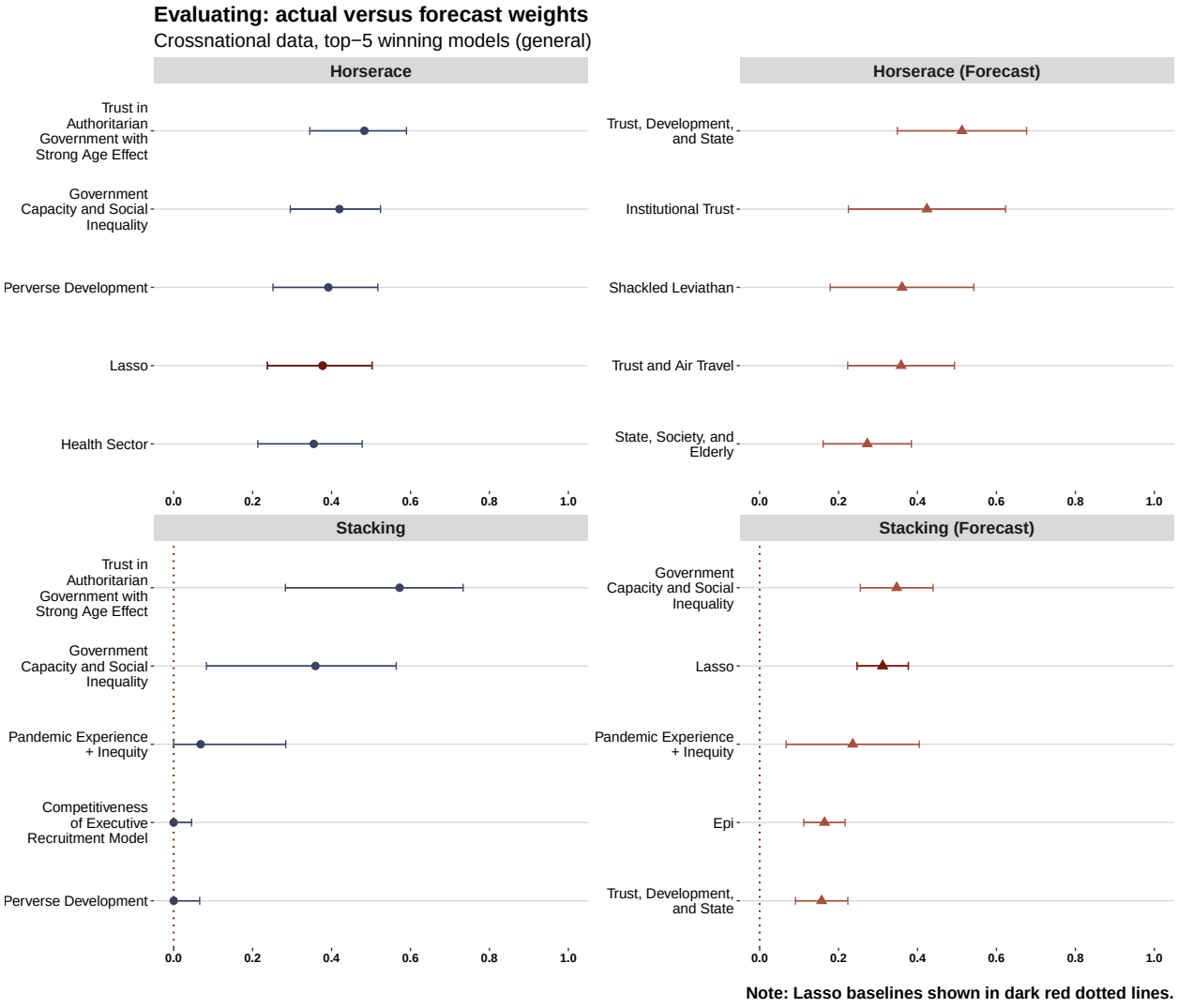


Figure 4: Top five models selected by each contest. The left column reports the results of the algorithmic implementation of each contest. The right column reports the results of the expert forecasts of the same contest.

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) versus aggregation (stacking) of models — or, in the broader analogy, explanations — prioritizes different sets of models. In the general challenges, stacking often heavily favors one to three models, putting much lower weights on the others. This is the case even when differences in the horserace of the predictive capacities of individual models are actually quite minimal. Application of these metrics in other contexts is necessary to establish the generality of these patterns. Second, we show that experts, on average, have limited abilities to identify accurately the most predictive models. Capacity is limited even when experts are provided baseline performance metrics, as in our forecasting exercise. To the extent that traditional approaches to organizing and synthesizing knowledge drawn from an existing literature ask researchers to identify the strongest arguments, our findings provide grounds for skepticism about their abilities to do so.

4.3 Aggregating Models

Finally, we turn aggregate predictive knowledge from the MCs and from expert forecasts. We compare different approaches to aggregating information provided by different arguments or statistical models. In Figure 5, we show the results of six methods to aggregate the findings of the general models. The rows depict two assessments of predictive performance. The first examines predictions of observed outcomes. The second normalizes both model predictions and the outcomes to measure correlations between predictions and outcomes. The panels look at predictions over different time periods. The left column presents our main estimates, which measure cumulative COVID-19 mortality as of August 31, 2021. The second column presents estimates of the changes in COVID-19 mortality outcomes between the data provided to MC participants that was produced in November 2020 and August 2021. The third column extends our estimates through January 2022.

We provide two benchmarks for each method. First, the "intercept only" metrics reflect the fact all the measures of variation explained normalize by a model that fits only the intercept (or mean) of the outcome. As such, 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.[8] Given this normalization, an intercept model takes the value of zero for all analyses. Second, we benchmark model performance against the Lasso model for each challenge.

Our first two measures of predictive performance — the best and median-performing models in each challenge — follow directly from our discussion in the past two sections. The point estimates reflect the pseudo-R^2 of each model. By definition, the single best performs better than the median model, though we note substantial differences in predictive performance. The sharp drop-off in performance between the best and median models, combined with experts' limited abilities to identify accurately the best-performing models (Figure 4), may be cause for concern when literatures depend on expert ability to assess the merits of different empirically-supported claims. As we observed in Figure 2, the best model performs better than our Lasso benchmark, though the difference in predictive power between the Lasso and best-performing models is not statistically significant at conventional thresholds.

The final algorithmic prediction examines the outcomes using the stacking estimator. In addition to using stacking to select models, stacking allows us to make a composite (or aggregate) prediction by taking the weighted sum of model predictions. Stacking outperforms the best model but the difference in performance between the two is small. The advantages of the stacking model are in part mechanical. For example, a stacking estimator could assign a weight of 1 to the best-performing model and 0 to all other models. To the extent that the optimal weights depart from this allocation, it must be the case that the stacking estimator outperforms the best model. We show in Figures 5 and S23–S29 that the relative benefits of the stacking estimator over the best model are limited. Nevertheless, stacking in every case represents the most predictive aggregation method that we estimate in every challenge. This suggest that implementing a stacking estimator could be a useful addition in literatures that are characterized by multiple models.

We next turn to predictions that come from the aggregation of expert forecasts. Our first metric examines the predictive power of the expert-favored model in each challenge. As in Figure 4, the expert-favored model — the one that experts deemed the most likely to be the most predictive — does not align with the model that is actually found to be the most predictive. It is thus unsurprising that the expert-favored model predicts less variation than the best model. Nevertheless,

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

in three of the four challenges the expert-favored model predicts mortality better than the median model. This may reflect reliance on interim model performance, information about which was shown to forecasters. Given the scale of India's COVID wave in mid-202, the interim performance data that was shown to forecasters is less prognostic than the other challenges of outcomes observed on August 31, 2021, making the task for India inherently harder. Thus, it is not surprising that for India, the expert-favored model fails to predict mortality better than the median model.

We now assess the performance of implied stacking models from the expert stacking forecasts. The "representative expert" forecast represents the median individual-level aggregate forecast. We also examine a "wisdom of the crowds" stacking model that aggregates over forecasters' stacking weights. These forecasts — which incorporate the predictions of multiple models — outperform the median and the expert-favored models. However, they underperform the algorithmic stacking model. This is not surprising. The algorithmic stacking model minimizes prediction error, meaning that it gives the upper bound on performance for any elicited stacking model.

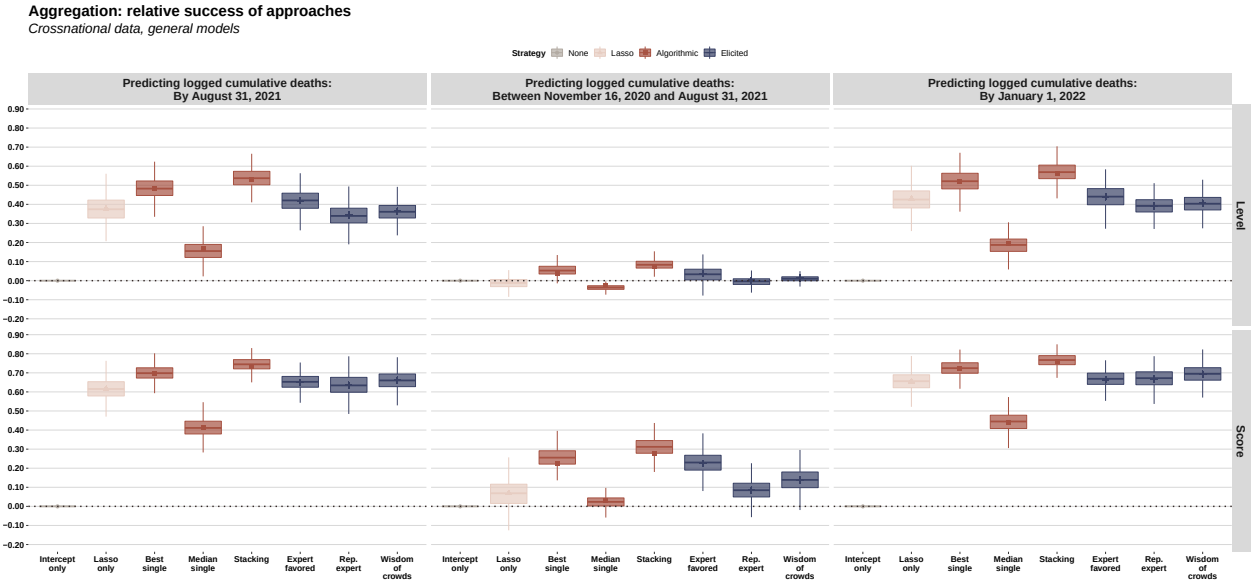


Figure 5: Comparison of predictive accuracy of different metrics that we propose. The top row of plots examines predictions of the level (or value) of cumulative COVID-19 deaths per million, while the bottom row of plots examines the correlation between predictions and actual mortaility. The left column of plots assesses predictive accuracy on cumulative mortality as of August 31, 2021. The middle column examines mortality *after* the data shared with modelers. The right column extends predictions through January 2022. Interquartile ranges and 95% confidence intervals are generated by bootstrapping.

The bottom panel of Figure 5 reports the same metrics except on the basis of Z-score transformations of the predictions and outcomes. This allows us to abstract from concerns about levels (or intercept shifts). The ranking of different models is quite similar between both score and level approaches. As in our discussion of the pseudo-R^2 of parameterized models, we show that these models — and consequently, the associated stacking model — tend to substantially underestimate COVID-19 mortality. However, predicted and actual deaths are correlated at rates that are generally indistinguishable from the correlation between general model predictions and actual deaths. We observe persistently similar levels of correlations between the general and parameterized models when we seek to predict changes in cumulative mortality rates after the models were submitted. We plot the correlation in predictive performance across general and parameterized versions of models in Figure S17.

Finally, Figure 6 shows the distance in the prediction space between submitted models, highlighting the models that received positive weight in model stacking. We see that the small number of selected models are on the edges of the collection of proposed models, though models 1 and 2 are relatively close in the prediction space they both differ markedly from model 5. This is broadly consistent the intuition that the stacking model prioritizes models that add new information. Here, Model #5 ("Pandemic Experience + Inequality") receives weight in stacking because it is so different than the best-performing models.

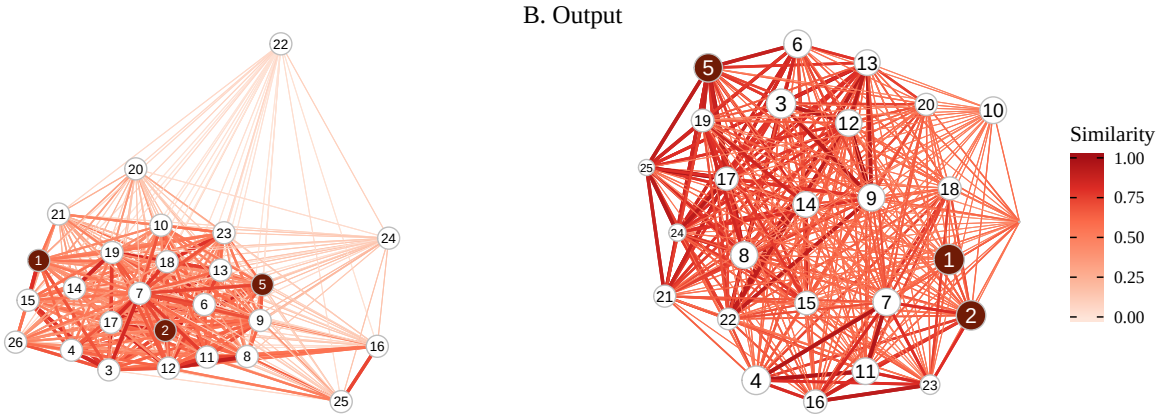


Figure 6: The network in the left panel plots the multivariate R^2 between predictors of each of the submitted crossnational, general models. The network in the right panel plots the bivariate correlation between the leave-one-out predictions of each model and actual mortality (ρ), normalized to a 0-1 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.

5 Discussion

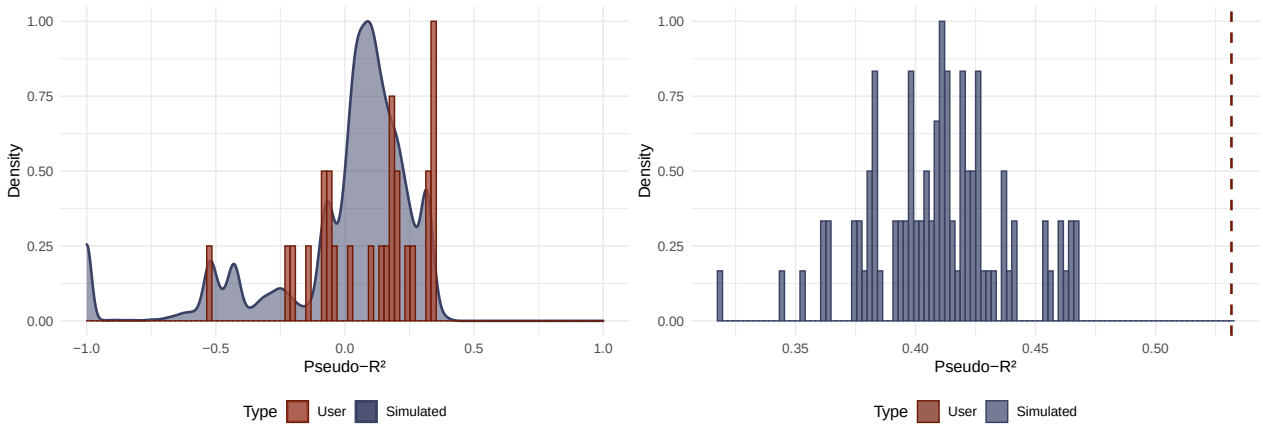
In this paper, we implement an approach to gathering, evaluating, and aggregating social scientific explanations with a focus on the problem of accounting for variation in COVID-19 mortality. The tools we have employed allow us to take stock of the predictive power of arguments in contexts where we have many explanations for a single outcome. In the COVID-19 Model Challenges, we aimed to predict COVID-19 mortality rates, an outcome that has attracted substantial scholarly attention since the inception of the pandemic. The design we advance could productively be applied to other outcomes for which social scientists have offered many explanations.

Substantively, we find a remarkable convergence in the explanations offered by social scientists. The most common accounts emphasize the role of social trust. Models that incorporate measures of social trust perform well and achieve prominence in the best performing stacked model. Most models that were submitted, including the more successful ones, rely on quite simple logics and did not, for instance, presuppose any interactions between different variables included in the models. They are, in this sense, more "variable centered" than "model centered."

In assessing performance of the submissions, we have shown that the more successful models out-perform the Lasso-generated model in three of four general challenges. However, the "typical" model — that is, the model with the median performance — performs far worse than Lasso. In Panel (a) of Figure 7, we further compare the user-submitted models to all permutations of three-predictor, linear models that can be generated from the Models Challenge-provided data. We show that while the strongest models are clearly in the top percentiles of all possible models, many of the weaker models do not perform particularly well relative to the distribution of all possible models. In Panel (b), we compare our stacking prediction to stacking predictions generated from 100 random samples of 27 random three-predictor models generated from the crossnatonal MC data. Our observed prediction far out-performs all of predictions from this "null" distribution of stacking models ($p = 0$). In this sense, by aggregating expert models through stacking, we can enhance the predictive performance of a set of models or, more broadly, a literature.

While the specific MCs that we implemented were facilitated by early and sustained interest in a new outcome of interest to social scientists — COVID-19 mortality — several features of our approach may be worth replicating in more established social science literatures. In particular, we emphasize the need to evaluate competing theories or arguments on common samples with common operationalizations of an outcome. This setting contrasts directly with settings in which meta-analyses are increasingly employed in the social sciences. Thus, our approach complements meta-analysis. Our approach also relies on the predictive power of theories, rather than structural assumptions.

The algorithmic tools that we employ — model comparison based on predictive power and model stacking to generate an aggregate prediction — are easily implementable in such settings. We show that these forms of model assessment and combination are useful precisely because experts are limited in their capacity to assess arguments in this way. We



(a) Horserace R^2 distribution.

(b) Stacking R^2 distribution.

Figure 7: Results of simulation checks. In panel (a) the histogram depicts the observed distribution of pseudo-R^2's and the density plot depicts the distribution of all pseudo-R^2's from all, linear three-predictor models in the common MC dataset. In panel (b) we show how the predictive performance of our stacking model compares to the predictive performance of randomly generated stacking models. All models are crossnational general models.

note that our design of the MC and forecasting may limit the application of the usual heuristics that scholars use to sift explanations in the literature. Whether these heuristics would improve the concurrence between the algorithmic and expert assessment remains an important open question. We advocate wider adoption of the approach to aggregation that we advance in literatures for which we have many explanations for a common outcome.

6 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, and Cyrus Samii for comments on an early version of the Model Challenge research design.

7 Methods and Materials

See SI for a full description of the MC and the analyses employed. Our algorithmic measures of model performance used in tables and figures are as follows. For Figures 2, 3 and the "level"-based model summaries in Figure 5, we measure the pseudo-R^2 using (1), For the "score"-based model summaries in Figure 5, we measure the correlation using (2). We estimate the stacking weights employed in Figures 4 and 5, we estimates the weights using (3).

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

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

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

8 References

Abernethy, Jacob D, and Rafael Frongillo. 2011. "A Collaborative Mechanism for Crowdsourcing Prediction Problems." *Advances in Neural Information Processing Systems* 24.

Acharya, Arnab, John Gerring, and Aaron Reeves. 2020. "Is Health Politically Irrelevant? Experimental Evidence During a Global Pandemic." *BMJ Global Health* 5 (10): e004222.

Baden, Lindsey R, Hana M El Sahly, Brandon Essink, Karen Kotloff, Sharon Frey, Rick Novak, David Diemert, et al. 2021. "Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine." *New England Journal of Medicine*.

Banerjee, Abhijit, Esther Duflo, Nathanael Goldberg, Dean Karlan, Robert Osei, William Parrienté, Jeremy Shapiro, Bram Thuysbaert, and Christopher Udry. 2015. "A Multifaceted Program Causes Lasting Progress for the Very Poor: Evidence from Six Countries." *Science* 348 (6236).

Bargain, Olivier, and Ulugbek Aminjonov. 2020. "Trust and compliance to public health policies in times of COVID-19." *Journal of Public Economics* 192: 104316. https://doi.org/10.1016/j.jpubeco.2020.104316.

Blair, Graeme, Jeremy M. Weinstein, Fotini Christia, Eric Arias, Emile Badran, Robert A. Blair, Ali Cheema, et al. 2021. "Community Policing Does Not Build Citizen Trust in Police or Reduce Crime in the Global South." *Science* 374 (6571): eabd3446.

Bokemper, Scott E, Gregory A Huber, Alan S Gerber, Erin K James, and Saad B Omer. 2021. "Timing of COVID-19 Vaccine Approval and Endorsement by Public Figures." *Vaccine*.

Borenstein, Michael, Larry V Hedges, Julian PT Higgins, and Hannah R Rothstein. 2021. *Introduction to Meta-Analysis*. John Wiley & Sons.

Bosancianu, Manuel, Hanno Hilbig, Macartan Humphreys, Sampada KC, Nils Lieber, and Alexandra Scacco. 2021. "Political and Social Correlates of Covid-19 Mortality."

Cepaluni, Gabriel, Michael Dorsch, and Reka Branyiczki. 2020. "Political Regimes and Deaths in the Early Stages of the COVID-19 Pandemic." {SSRN Scholarly Paper} ID 3586767. Rochester, NY: Social Science Research Network. https://doi.org/10.2139/ssrn.3586767.

Cepaluni, Gabriel, Michael Dorsch, and Semir Dzebo. 2021. "Populism, Political Regimes, and COVID-19 Deaths." {SSRN Scholarly Paper} ID 3816398. Rochester, NY: Social Science Research Network. https://doi.org/10.2139/ssrn.3816398.

Coppock, Alexander, Seth J. Hill, and Lynn Vavreck. 2020. "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 (eabc4046): 1–6.

Dahl, Robert A. 1972. *Polyarchy: Participation and Opposition*. New Haven: Yale University Press.

Davis, Gerald F. 2015. "Editorial Essay: What Is Organizational Research For?" *Administrative Science Quarterly* 60 (2): 179–88.

Debray, Thomas PA, Hendrik Koffijberg, Daan Nieboer, Yvonne Vergouwe, Ewout W Steyerberg, and Karel GM Moons. 2014. "Meta-Analysis and Aggregation of Multiple Published Prediction Models." *Statistics in Medicine* 33 (14): 2341–62.

DellaVigna, Stefano, Nicholas Otis, and Eva Vivalt. 2020. "Forecasting the Results of Experiments: Piloting an Elicitation Strategy." *AEA Papers and Proceedings* 110: 75–79.

DellaVigna, Stefano, and Devin Pope. 2018. "Predicting Experimental Results: Who Knows What?" *Journal of Political Economy* 126 (6): 2410–56.

DellaVigna, Stefano, Devin Pope, and Eva Vivalt. 2019. "Predict Science to Improve Science." *Science* 366 (6464): 428–29.

Dunning, Thad, Guy Grossman, Macartan Humphreys, Susan D. Hyde, Craig Mcintosh, Gareth Nellis, Claire L. Adida, et al. 2019. "Voter Information Campaigns and Political Accountability: Cumulative Findings from a Preregistered Meta-Analysis of Coordinated Trials." *Science Advances* 5 (7).

Elgar, Frank J., Anna Stefaniak, and Michael J. A. Wohl. 2020. "The trouble with trust: Time-series analysis of social capital, income inequality, and COVID-19 deaths in 84 countries." *Social Science & Medicine* 263: 113365. https://doi.org/10.1016/j.socscimed.2020.113365.

Figueiredo, Alexandre de, Clarissa Simas, Emilie Karafillakis, Pauline Paterson, and Heidi J. Larson. 2021. "Mapping Global Trends in Vaccine Confidence and Investigating Barriers to Vaccine Uptake: A Large-Scale Retrospective Temporal Modelling Study." *The Lancet*.

Folegatti, Pedro M, Katie J Ewer, Parvinder K Aley, Brian Angus, Stephan Becker, Sandra Belij-Rammerstorfer, Duncan Bellamy, et al. 2021. "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*.

Galiani, Sebastian, Paul Gertler, and Mauricio Romero. 2017. "Incentives for Replication in Economics."

Gelman, Andrew, and Guido Imbens. 2013. "Why Ask Why? Forward Causal Inference and Reverse Causal Questions." National Bureau of Economic Research.

Han, Qing, Bang Zheng, Mioara Cristea, Maximilian Agostini, Jocelyn J. Belanger, Ben Gutzkow, Jannis Kreienkamp, PsyCorona Collaboration, and N. Pontus Leander. 2021. "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. https://doi.org/10.1017/S0033291721001306.

Jolly, Eshin, and Luke J. Chang. 2019. "The Flatland Fallacy: Moving Beyond Low-Dimensional Thinking." *Topics in Cognitive Science* 11: 433–54.

Jones, Tommy. 2019. "A Coefficient of Determination for Probabilistic Topic Models." arXiv. https://doi.org/10.48550/ARXIV.1911.11061.

Koole, Sander L., and Daniël Lakens. 2012. "Rewarding Replications: A Sure and Simple Way to Improve Psychological Science." *Perspectives on Psychological Science* 7 (6): 608–14.

Lazarus, Jeffrey V, Scott C Ratzan, Adam Palayew, Lawrence O Gostin, Heidi J Larson, Kenneth Rabin, Spencer Kimball, and Ayman El-Mohandes. 2021. "A Global Survey of Potential Acceptance of a COVID-19 Vaccine." *Nature Medicine*.

Logunov, Denis Y, Inna V Dolzhikova, Dmitry V Shcheblyakov, Amir I Tukhvatulin, Olga V Zubkova, Alina S Dzharullaeva, Anna V Kovyrshina, et al. 2021. "The Lancet." *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*.

Min, Jungwon. 2020. "Does social trust slow down or speed up the transmission of COVID-19?" *PLoS ONE* 15 (12): e0244273. https://doi.org/10.1371/journal.pone.0244273.

Mulligan, Mark J, Kirsten E Lyke, Nicholas Kitchin, Judith Absalon, Alejandra Gurtman, Stephen Lockhart, Kathleen Neuzil, et al. 2021. "Phase i/II Study of COVID-19 RNA Vaccine BNT162b1 in Adults." *Nature*.

Nosek, Brian A., Jeffrey R. Spies, and Matt Motyl. 2012. "Scientific Utopia II: II. Restructuring Incentives and Practices to Promote Truth over Publishability." *Perspectives on Psychological Science* 7 (6): 615–31.

Organization, World Health. 2021. "Draft Landscape and Tracker of COVID-19 Candidate Vaccines."

Pearl, Judea. 2015. "Causes of Effects and Effects of Causes." *Sociological Methods & Research* 44 (1): 149–64.

Piquero, Alex R., Wesley G. Jennings, Erin Jemison, Catherine Kaukinen, and Felicia Marie Knaul. 2021. "Domestic Violence During the COVID-19 Pandemic - Evidence from a Systematic Review and Meta-Analysis." *Journal of Criminal Justice* 74: 101806. https://doi.org/https://doi.org/10.1016/j.jcrimjus.2021.101806.

Polack, Fernando P, Stephen J Thomas, Nicholas Kitchin, Judith Absalon, Alejandra Gurtman, Stephen Lockhart, John L Perez, et al. 2021. "Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine." *New England Journal of Medicine*.

Robinson, E, A. Jones, I. Lesser, and M. Daly. 2021. "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 (15): 2024–34.

Slough, Tara, Daniel Rubenson, Ro'ee Levy, Francisco Alpizar Rodriguez, María Bernedo del Carpio, Mark T. Buntaine, Darin Christensen, et al. 2021. "Adoption of Community Monitoring Improves Common Pool Resource Management Across Contexts." *Proceedings of the National Academy of Sciences* 10.1073: 1–10.

Slough, Tara, and Scott A Tyson. 2022. "External Validity and Meta-Analysis." *American Journal of Political Science* Forthcoming.

Solis Arce, Julio S, Shana S Warren, Niccollo F Meriggi, Alexandra Scacco, Nina McMurry, Maarten Voors, Georgiy Syunyaev, and Amyn Abdul Malik. 2021. "COVID-19 Vaccine Acceptance and Hesitancy in Low- and Middle-Income Countries." *Nature Medicine*.

Strevens, Michael. 2011. *Depth: An Account of Scientific Explanation*. Harvard University Press.

Team, McGill COVID19 Vaccine Tracker. 2021. "COVID-19 Vaccine Tracker." *Unknown Journal*.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society B* 58 (1): 267–88.

Voysey, Merryn, Sue Ann Costa Clemens, Shabir A Madhi, Lily Y Weckx, Pedro M Folegatti, Parvinder K Aley, Brian Angus, et al. 2021. "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*.

Watts, Duncan J. 2017. "Should Social Science Be More Solution-Oriented?" *Nature Human Behavior* 1 (15): 1–4.

Wouters, Olivier J, Kenneth C Shadlen, Maximilian Salcher-Konrad, Andrew J Pollard, Heidi J Larson, Yot Teer-

awattananon, and Mark Jit. 2021. "Challenges in Ensuring Global Access to COVID-19 Vaccines: Production, Affordability, Allocation, and Deployment." *The Lancet*.

Yao, Yuling, Aki Vehtari, Daniel Simpson, and Andrew Gelman. 2021. "Using Stacking to Average Bayesian Predictive Distributions (with Discussion)." *Bayesian Analysis*.

9 Supplementary materials

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. In so doing, we asked researchers to make arguments about why selected socio-political variables would predict COVID-19 mortality. We include four challenges in which researchers could predict variation in COVID-19 mortality: (1) across countries; (2) across US states; (3) across Mexican states; and/or (4) across Indian states.

   Researchers contributed models of COVID-19 mortality between December 1, 2020 and January 20, 2021. They were provided with cumulative COVID-19 mortality rates as of November 16, 2020 when making predictions. We refer to the researchers that 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 #1. They were asked to evaluate the predictive performance of models as of August 31, 2021 and August 31, 2022.

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

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

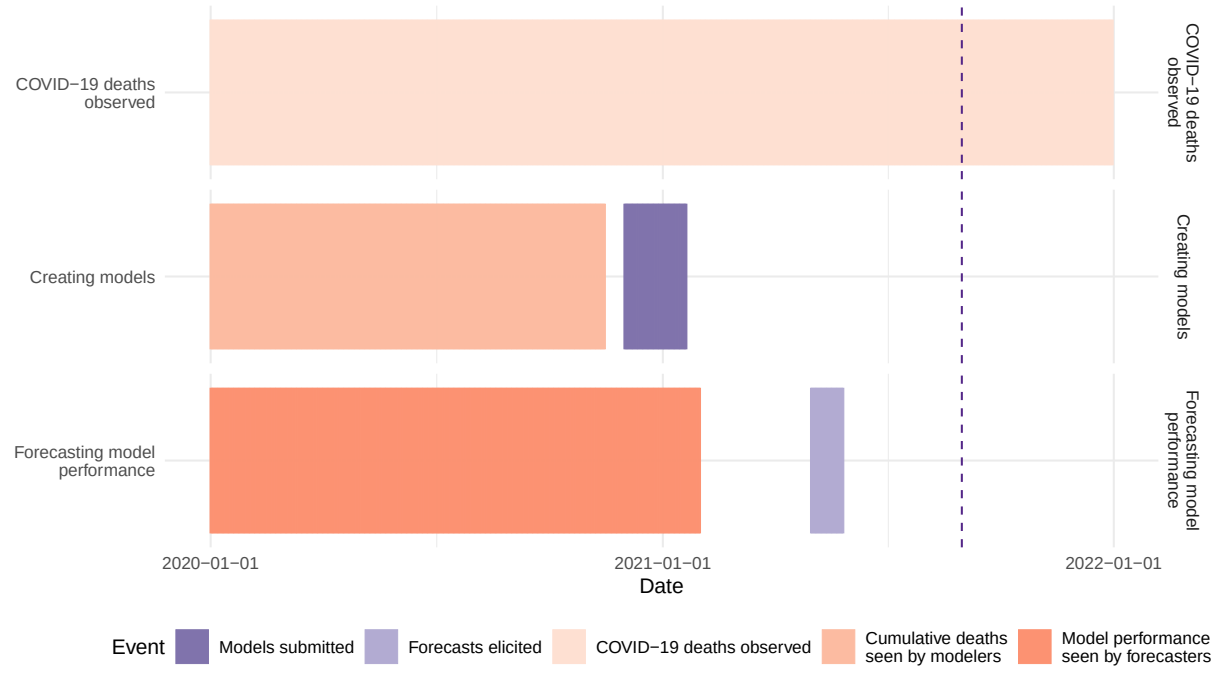


Figure S1: Timeline of COVID-19 Model challenge. The dashed vertical line is August 31, 2021. We assess models on their predictive ability at that date.

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 outcome data from the following sources:

- **Crossnational challenge**: COVID-19 mortality data comes from the European Centre for Disease Prevention and Control (ECDC).
- **India challenge**: COVID-19 mortality data, by state, comes from the Government of India.
- **Mexico challenge**: COVID-19 mortality data, by state, comes from the Government of Mexico.
- **United States challenge**: COVID-19 mortality data, by state, comes from the COVID Tracking Project at *The Atlantic*.

When participants entered the COVID-19 model challenge, modelers had access cumulative COVID-19 mortality data as of November 16, 2020. The 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 10th, 50th, and 90th percentiles in (percent) *change* in COVID-19 mortality between November 16, 2020 and August 30, 2021. These countries are Spain, Romania, and Uganda, respectively.

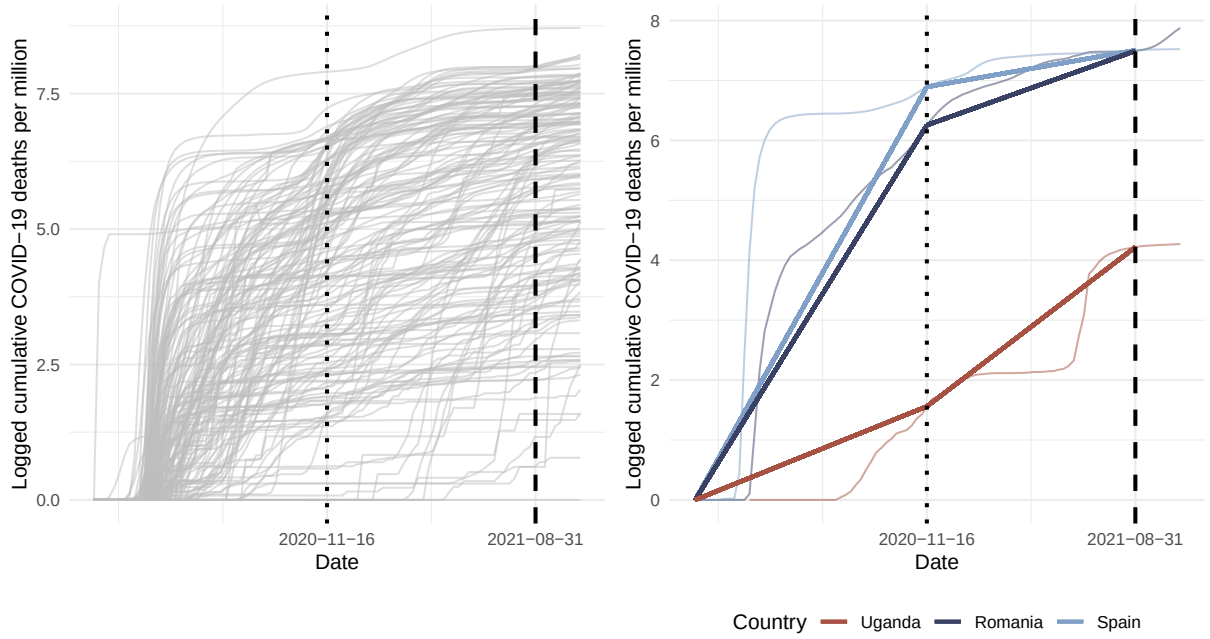


Figure S2: Outcome data for the crossnational challenge. The left panel depicts the growth logged cumulative COVID-19 deaths per million. Each line is a country The vertical lines reflect the data provided in the modeling challenge and the main outcome, mortality as of August 31, 2021. The right panel 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.

Our 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.

S2 Model Generation Details

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
US	50	7.56	7.41	6.03	8.02	0.47

Table S1: Summary statistics for our outcome measure, by challenge. Note that 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.)

S2.1 Challenge Overview

Between December 2020 and January 20, 2021, we solicited statistical models from political and other social scientists asking them to predict cumulative numbers of COVID-19 deaths as of August 31, 2021. 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 Tables S6-S9 for a list of predictors and data sources.) 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.

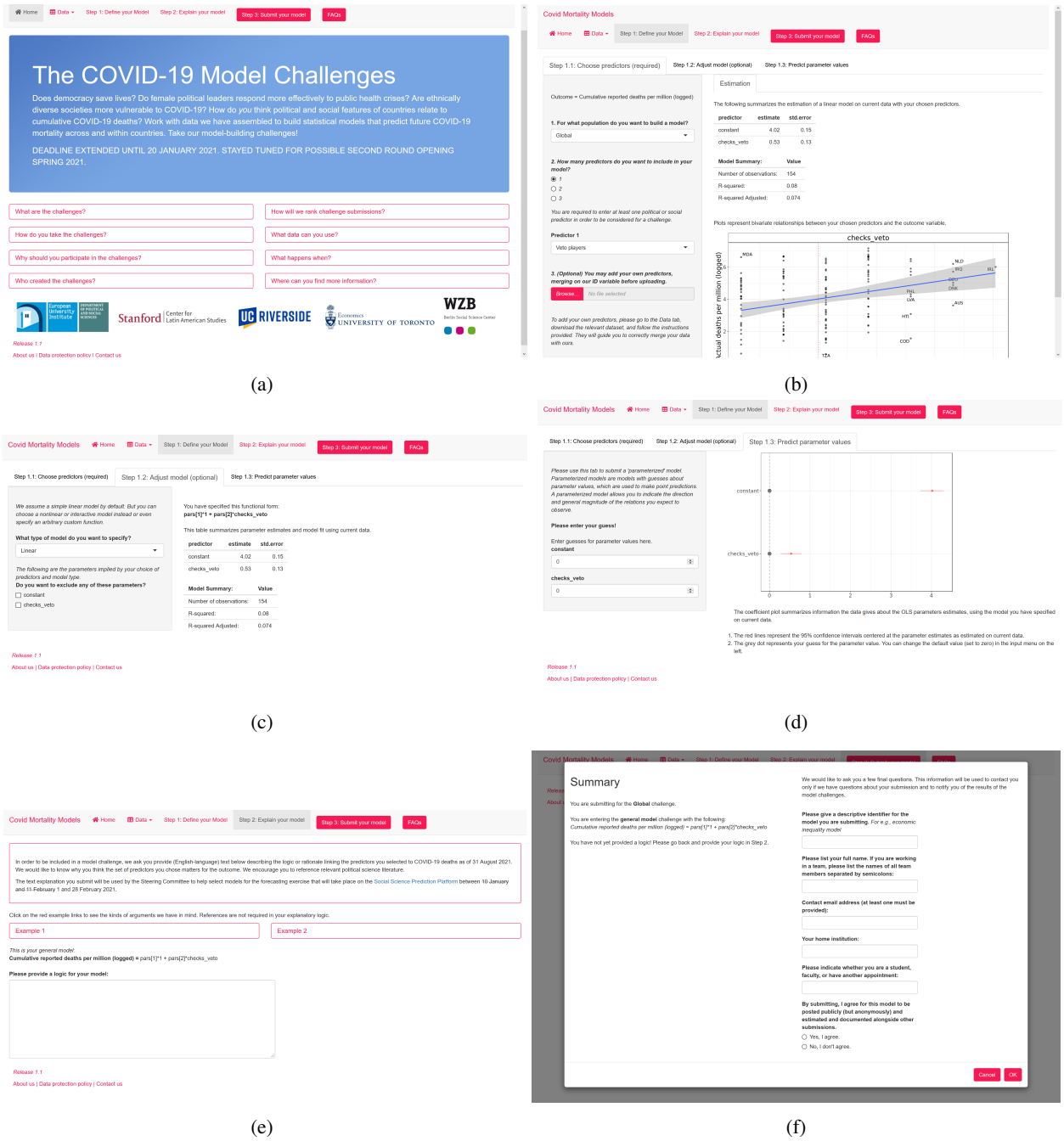
The platform was open to all researchers (and non-researchers). We advertised through social media (Twitter), via professional listservs (the American Political Science Association, the European Political Science Association, the Society for Political Methodology, Evidence in Governance and Politics, and others). In addition, we sent individual emails directed at a list of researchers from the top 100 research institutions globally as well as specifically in the US, Mexico, and India.

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

1. Choose a model challenge to enter — Global, India, Mexico, or US (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 matters 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 Tables S6-S9. These codebooks provided information on the definition of and data sources for all predictors.

In total, we received 88 distinct model submissions. Table S2 reports the breakdown of submissions for each of the four challenges (Global, India, Mexico, and the US) disaggregated by model type (general or parameterized). For ten of the models submitted, participants also uploaded their own data.



(a) (b) (c) (d) (e) (f)

Figure S3: Screenshots from model challenge interface. All plots and reported statistics were dynamic.

Challenge	General	Parameterized	Total
Crossnational	26	14	40
India	7	5	12
Mexico	7	4	11
US	19	6	25
Total	59	29	88

Table S2: Number of models submitted in each challenge.

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*1 + 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*1 + 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*1 + 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
US	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*1 + -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 S3: "Usual suspects" 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 the November 16, 2020 outcome data.

S2.2 General and parameterized models

We distinguish between general and parameterized 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 baseline COVID-19 mortality data provided at baseline.

S2.3 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. As such, we have substantially more theory-driven models than ML models.

S2.4 Additional Models

In addition to user-submitted models, we include 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"). We include a usual suspects model containing standard epidemiological predictors in each challenge 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 S3 and the Lasso models are reported in Table S4.

Challenge	General Form	Parameterized Form
Crossnational	deaths_per_mio_log ∼ acc_sanitation + healthcare_qual	deaths_per_mio_log ∼ 3.9815*1 + 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*1 + 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*1 + 0.12*health_expendpc + -0.1461*pct_poor + 0.0813*pct_tertiaryemp
US	deaths_per_mio_log ∼ gini + hosp_beds_pc + pct_religious + pop_density + urban_pct	deaths_per_mio_log ∼ 6.2735*1 + 0.2325*gini + 0.2491*hosp_beds_pc + 0.1534*pct_religious + 0.2374*pop_density + 0.2517*urban_pct

Table S4: Lasso models for each challenge. The parameterized form was fit on the November 16, 2020 outcome data.

	Theoretical		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
US	18	6	3	2	29
Total	61	32	6	5	104

Table S5: Total number of models analyzed in each challenge.

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. As such our full disaggregation of the models in Table S5 includes 104 models. Eighty-eight were submitted by entrants to the Model Challenge, as in Table S2 and there are 8 of each the usual suspects and Lasso models.

S2.5 Predictors provided by the COVID-19 Model Challenge

Tables S6 to S9 display the codebooks for each of the four challenge datasets. All of the models are constructed from this base set of covariates except when researchers submitted their own covariates.

Table S6: Codebook for: crossnational data

Variable Name	Variable Label	Definition	Source
acc_sanitation	Access to sanitation (GHSI)	Percentage of homes with access to at least basic sanitation facilities as reported in the 2019 Global Health Security Index (GHSI).	GHSI
air_travel	Air travel (passengers carried)	Air passengers carried include both domestic and international aircraft passengers of air carriers registered in the country. (IS.AIR.PSGR) (logged)	World Bank
al_etfra	Ethnic fractionalization	The variables reflect the probability that two randomly selected people from a given country will not be from the same ethnic group; the higher the number, the greater the degree of fractionalization. The indicator comes originally from Alesina et al. (2003)	Alesina et al. (QoG)
al_religfra	Religious fractionalization	Same as directly above, but this is based on the Alesina et al. (2003) data, and refers to the probability that two randomly drawn individuals are from different religious groups	Alesina et al. (QoG)

Table S6: Codebook for: crossnational data *(continued)*

Variable Name	Variable Label	Definition	Source
bureaucracy_corrupt	Public sector corruption	Question: How pervasive is political corruption? "The corruption index includes measures of six distinct types of corruption that cover both different areas and levels of the polity realm, distinguishing between executive, legislative and judicial corruption. Within the executive realm, the measures also distinguish between corruption mostly pertaining to bribery and corruption due to embezzlement. Finally, they differentiate between corruption in the highest echelons of the executive at the level of the rulers/cabinet on the one hand, and in the public sector at large on the other. The measures thus tap into several distinguished types of corruption: both "petty" and "grand"; both bribery and theft; both corruption aimed and influencing law making and that affecting implementation." The raw indicators that comprise the components and sub-components of the index are arrived at via expert assessments, aggregated through a Bayesian IRT measurement model (Pemstein et al. 2020). Scale: Interval, 0 to 1, with higher values denoting higher levels of political corruption.	V-Dem v10
checks_veto	Veto players	DPI checks measure of veto points. Definition varies depending on type of system. Generally, higher values denote contexts where there is electoral competitiveness in the legislature, and the two branches are controlled by opposing political forces. In presidential systems, higher values are produced by the existence of parties in the legislature that are allied with the president, but have a position on the economy that is closer to that of the main opposition party. In parliamentary systems, higher values are produced by the existence of parties in the legislature that are in the governing coalition, but have a position on the economy that is closer to that of the main opposition party. For further details, please see DPI 2017 codebook (pages 14, 15, 18, 19).	DPI 2017
count_powerless	Count of how many powerless ethnic groups are in country	Count of number of groups that are defined as "powerless" in the EPR data. Similar variable is used in Wegenast and Basedau (2014) as well.	EPR data
detect_index	Health data quality	Index of early detection and reporting of epidemics with potential international concern.	GHSI
dist_anyelec	Electoral pressure	Time to next election captures the number of days to the next parliamentary, presidential or senate election counted from the day the WHO declared Covid-19 a pandemic (March 11, 2020). The time to next election is constructed by using next election dates obtained from the IEFS election guide, the IPU Parline data and complemented by the Wikipedia list of next general elections.	IFES, IPU, Wikipedia list
electoral_pop	Electoral populism	"Populist" countries (=1) are those in which a leader is elected within a democratic setting (PolityIV>=6 in year of election) running a populist campaign. This exclude leaders that only become populist while in office. Autocrats that deploy populism to hold onto power are not included (eg Mugabe). All others coded 0, including NAs in original data. The measure we use from Kyle and Meyer (2020) focuses speci?cally on electoral populism and classi?es 17 states currently as having governments led by electoral populist parties, including a number of cases with signi?cant early deaths from Covid-19, including Italy, the US, Brazil, and Turkey.	Populism in Power
elf_epr	ELF index	ELF index computed based on EPR data (IMPORTANT: because population shares tended not to add to 100%, a residual "other" category was added to each country, and assigned the remaining population share up to 100. The index is computed taking this group into consideration as well)	EPR data
fdi	FDI (net inflows, USD)	Foreign direct investment, net inflows (BoP, current US$). BX.KLT.DINV.CD.WD	World Bank
fe_etfra	Ethnic fractionalization	Same as directly above, but this indicator is sourced from Fearon (2003): Restricting attention to groups that had at least 1 percent of country population in the 1990s, Fearon identifies 822 ethnic and "ethnoreligious" groups in 160 countries. This variable reflects the probability that two randomly selected people from a given country will belong to different such groups. The variable thus ranges from 0 (perfectly homogeneous) to 1 (highly fragmented). The values are assumed to be constant for all years.	QoG data
federal_ind	Index of federalism	Index combining measures of federalism from the Database of Political Institutions (DPI). Generally, higher values denote contexts where power and decision-making are more decentralized. More specifically, the index combines values for the DPI measures AUTON (are there autonomous regions?), MUNI (are municipal governments locally elected?), STATE (are there state/province governments locally elected?), AUTHOR (do the state/provinces have authority over taxing, spending, or legislating?), and STCONST (are the constituencies of the senators the states/provinces?). For further details, please see DPI 2017 codebook (pages 20-21).	DPI 2017
gdp_pc	GDP per capita (PPP)	GDP per capita, PPP (constant 2011 international $)	World Bank

Table S6: Codebook for: crossnational data *(continued)*

Variable Name	Variable Label	Definition	Source
gini	Income GINI	Index of disposable (after taxes and transfers) income inequality, based on SWIID 8.2 data. Ranges from 0 to 1, with higher values denoting more income inequality.	SWIID v8.2
gov_effect	Government effectiveness	Government Effectiveness captures perceptions of the quality of public services, the quality of the civil service and the degree of its independence from political pressures, the quality of policy formulation and implementation, and the credibility of the government's commitment to such policies. Estimate gives the country's score on the aggregate indicator, in units of a standard normal distribution, i.e. ranging from approximately -2.5 to 2.5.	World Bank Indicators
hdi	HDI	Human Development Index	UNDP via GHSI
health_equality	Health equality (VDem)	Extent to which high quality basic healthcare is guaranteed to all and sufficient to enable citizens to exercise their basic political rights. Scale: Ordinal and converted to interval by the measurement model. Range (0-4): Extreme inequality = 0: Because of poor-quality healthcare at least 75 percent (%) of citizens ability to exercise their political rights as adult citizens is undermined. Equality = 4: Basic health care is equal in quality and less than five percent (%) of citizens cannot exercise their basic political rights as adult citizens.	V-Dem Dataset - Version 10
health_exp_pc	Healthcare spending/capita	Healthcare spending/capita	GHSI
health_index	Health sector robustness (GHSI)	Index reporting on sufficient and robust health sector to treat the sick and protect health workers.	GHSI
healthcare_qual	Healthcare quality index (GHSI)	Healthcare Access and Quality Index based on mortality from causes amenable to personal health care (0-100)	GHSI
hosp_beds_pc	Hospital beds / capita (GHSI)	Hospital beds per capita.	GHSI
inequality	Income share top 10%	Income share held by highest 10%. Percentage share of income or consumption is the share that accrues to subgroups of population indicated by deciles or quintiles. (SI.DST.10TH.10)	World Bank
infection	Ebola/SARS/MERS exposure	The Ebola/SARS/MERS exposure measure captures a countries recent experience with SARS, Ebola, or MERS, draws on data from the World Health Organization and reports on whether a country displays at least 100 cases for either MERS, SARS, or Ebola.	WHO (HDX)
journal_harass	Harassment of journalists (VDem)	Are individual journalists harassed - i.e. threatened with libel or arrested or imprisoned or beaten or killed – by governmental or powerful nongovernmental actors while engaged in legitimate journalistic activities? Range (0-4): 0= 0: No journalists dare to engage in journalistic activities that would offend powerful actors because harassment or worse would be certain to occur. 4= Journalists are never harassed	V-Dem Dataset - Version 10
life_exp_2017	Life expectancy	Life expectancy at birth indicates the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life.	World Bank
med_age_2013	Median age (2013)	Median age of population in 2013.	World Bank
media_critical	Print/broadcast media criticalness (VDem)	Of the major print and broadcast outlets - how many routinely criticize the government? Range(0-3): 0 = None. 3 = All major media outlets criticize the government at least occasionally. Scale: Ordinal converted to interval by the measurement model.	V-Dem Dataset - Version 10
migration_share	Share foreign born	International migrant stock is the number of people born in a country other than that in which they live. It also includes refugees. The data used to estimate the international migrant stock at a particular time are obtained mainly from population censuses. The estimates are derived from the data on foreign-born population–people who have residence in one country but were born in another country. When data on the foreign-born population are not available, data on foreign population–that is, people who are citizens of a country other than the country in which they reside–are used as estimates. After the breakup of the Soviet Union in 1991 people living in one of the newly independent countries who were born in another were classified as international migrants. Estimates of migrant stock in the newly independent states from 1990 on are based on the 1989 census of the Soviet Union. For countries with information on the international migrant stock for at least two points in time, interpolation or extrapolation was used to estimate the international migrant stock on July 1 of the reference years. For countries with only one observation, estimates for the reference years were derived using rates of change in the migrant stock in the years preceding or following the single observation available. A model was used to estimate migrants for countries that had no data. 2015 data.	United Nations Population Division, Trends in Total Migrant Stock: 2008 Revision via World Bank

Table S6: Codebook for: crossnational data *(continued)*

Variable Name	Variable Label	Definition	Source
oil	Oil rents (% of GDP)	Measure of oil rents as a share of GDP. Estimates based on sources and methods described in "The Changing Wealth of Nations: Measuring Sustainable Development in the New Millennium" (World Bank, 2011). "NY.GDP.PETR.RT.ZS"	World Bank Indicators
pandemic_prep	Pandemic preparedness	A comprehensive assessment of countries' ability to prevent infectious disease outbreaks and to detect and report, and rapidly respond to mitigate the spread. It also accounts for health system capacities and compliance with international norms to improving national capacity, along with countries' overall risk environment and vulnerability to disease spread.	GHSI 2019
polar_rile	Party polarization (MARPOR)	Polarization on a Right-Left dimension, computed as the sum of the weighted squared deviations from the mean position on the Right-Left dimension in the party system, using the party vote shares in the election as weights (Taylor and Herman, 1971). Higher values denote a greater degree of political polarization in the party system. Measure includes only manifestos for elections that took place since 2015.	MARPOR, version 2019b
polariz_veto	Polarization between executive party and 4 main parties in legislature	Constructed from the ideological leaning (left, center, right) of the 5 main legislative parties. Computed as the maximum difference between the chief executive's party's value and the values of the 3 largest government parties and the largest opposition party	DPI2017
polity	Polity2	Autocracy-democracy index (polity2)ranging between -10 (total autocracy) and 10 (total democracy) from the Polity IV dataset	Polity IV data
pop_density_log	Population density (log)	Population density (people per sq. km of land area)	FAO and World Bank
pop_tot_log	Total population (logged)	Total population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship. The values shown are midyear estimates. (In '000,000)	World Bank
pos_gov_lr	Left-Right Government	The Left-Right position of the governing coalition (or party). Computed from the ideological placements of all the parties that form the government. The final score represents the weighted average of the governing parties' Left-Right position, using their legislative seats as weights. Higher values denote a more Right-leaning governing coalition.	ParlGov
pr	PR electoral system	Candidates are elected based on the percent of votes received by their party.	DPI 2017
precip	Precipitation (mm/month)	Average precipitation for Jan-Mar 2018, in mm per month	Climatic Research Unit, University of East Anglia
property_rights	Property rights (VDem)	Do citizens enjoy the right to private property? Range (0-5): 0 = Virtually no one enjoys private property rights of any kind. 5 = Virtually all citizens enjoy all or almost all property rights. Scale: Interval.	V-Dem Dataset - Version 10
resp_disease_prev	Respiratory disease prevalence	Combined prevalance of upper and lower respiratory disease as a % of the total population in 2017.	Institute for Health Metrics and Evaluation (IHME)
respond_index	Epidemic response capacity	Capacity to rapidly respond and mitigat the spread of an epidemic	GHSI
rq_polarization	Reynal-Querol ethnic polarization	Reynal-Querol (2002) index of ethnic polarization	EPR data
share_health_ins	Share with health insurance	Insurance coverage here includes affiliated members of health insurance, as well as the population having free access to healthcare services provided by the government.	ILO via Our World in Data
share_older	Share 65+	World Bank staff estimates using the World Bank's total population and age/sex distributions of the United Nations Population Division's World Population Prospects: 2019 Revision. Based on measures SP.POP.65UP.MA.IN and SP.POP.65UP.FE.IN	World Bank
share_powerless	Prop. of marginalized groups	Share of the population that is defined as "powerless" in the EPR. This refers to groups with political representatives that are excluded from national-level decision making, without being explicitly discriminated. It's important to highlight that groups that hold power at the subnational level, but not at the national level, are still defined as "powerless".	EPR data
soc_contrib	Social contributions as % of gov't revenue	Social contributions include social security contributions by employees, employers, and self-employed individuals, and other contributions whose source cannot be determined. They also include actual or imputed contributions to social insurance schemes operated by governments.	World Bank
soc_insur_cov	Social insurance program coverage (%)	Coverage of social insurance programs shows the percentage of population participating in programs that provide old age contributory pensions (including survivors and disability) and social security and health insurance benefits (including occupational injury benefits, paid sick leave, maternity and other social insurance). Estimates include both direct and indirect beneficiaries.	World Bank

Table S6: Codebook for: crossnational data *(continued)*

Variable Name	Variable Label	Definition	Source
soc_safety	Social safety net coverage (%)	Coverage of social safety net programs shows the percentage of population participating in cash transfers and last resort programs, noncontributory social pensions, other cash transfers programs (child, family and orphan allowances, birth and death grants, disability benefits, and other allowances), conditional cash transfers, in-kind food transfers (food stamps and vouchers, food rations, supplementary feeding, and emergency food distribution), school feeding, other social assistance programs (housing allowances, scholarships, fee waivers, health subsidies, and other social assistance) and public works programs (cash for work and food for work). Estimates include both direct and indirect beneficiaries.	World Bank
state_fragility	State fragility	The state fragility index combines scores measuring two essential qualities of state performance: effectiveness and legitimacy; these two quality indices combine scores on distinct measures of the key performance dimensions of security, governance, economics, and social development	State Fragility Index
temp_mean	Temperature (Celsius)	Average temperature for Jan-Mar 2018, in Celsius degrees	Climatic Research Unit, University of East Anglia
trade	Trade (share of GDP)	Trade is the sum of exports and imports of goods and services measured as a share of gross domestic product. (NE.TRD.GNFS.ZS)	World Bank
transparent_law	Transparent laws with predictable enforcement	Index indication to what extent are laws transparent and rigorously enforced and public administration impartial and to what extent citizens do enjoy access to justice and secure property rights and freedom from forced labor and freedom of movement and physical integrity rights and and freedom of religion. Range (0-1).	V-Dem Dataset - Version 10
trust_gov	Institutional trust	% of respondents who reported trusting the government "a great deal" or "quite a lot" obtained from waves 5 and 6 of the WVS (WVS 2018) and the 2018 wave of the Latin Barometer. Computed based on item E069_11, which asks respondents to self-rate their degree of confidence in the central government: 1=a great deal; 2=quite a lot; 3=not very much; 4=none at all. Measure includes only countries sampled since 2009.	WVS, LAPOP
trust_people	Interpersonal trust	% of respondents who believe that "most people can be trusted", when given the option between this and "you can't be too careful". Measures are based on waves 5 and 6 of the World Values Surveys (WVS), and obtained from Our World in Data (Ortiz-Ospina and Roser 2020), as well as from wave 5 of the Afrobarometer. Measures include only countries sampled since 2009. Where a country was included in both OWiD and Afrobarometer data, the most recent survey was given priority.	WVS, Afrobarometer
urban	Urban populario (percent)	Population in urban agglomerations of more than 1 million (% of total population) (EN.URB.MCTY.TL.ZS)	UB via World Bank
vdem_libdem	Liberal democracy	To what extent is the ideal of liberal democracy achieved? Clarifications: The liberal principle of democracy emphasizes the importance of protecting individual and minority rights against the tyranny of the state and the tyranny of the majority. The liberal model takes a "negative" view of political power insofar as it judges the quality of democracy by the limits placed on government. This is achieved by constitutionally protected civil liberties, strong rule of law, an independent judiciary, and effective checks and balances that, together, limit the exercise of executive power. To make this a measure of liberal democracy, the index also takes the level of electoral democracy into account.	V-Dem v10 (QoG)
vdem_mecorrpt	Media independence	Do journalists, publishers, or broadcasters accept payments in exchange for altering news coverage? Responses: 0: The media are so closely directed by the government that any such payments would be either unnecessary to ensure pro-government coverage or ineffective in producing anti-government coverage. 1: Journalists, publishers, and broadcasters routinely alter news coverage in exchange for payments. 2: It is common, but not routine, for journalists, publishers, and broadcasters to alter news coverage in exchange for payments. 3: It is not normal for journalists, publishers, and broadcasters to alter news coverage in exchange for payments, but it happens occasionally, without anyone being punished. 4: Journalists, publishers, and broadcasters rarely alter news coverage in exchange for payments, and if it becomes known, someone is punished for it.	V-Dem v10 (QoG)
woman_leader	Women leaders	Woman head of government on 1 Jan 2020. Does not include heads of state or joint heads.	Wikipedia list

Table S7: Codebook for: India

Variable Name	Variable Label	Definition	Source
percentage_of_women	share female legislators	Proportion of state assembly representatives who are female	Total number of state assembly seats from https://www.elections.in/upcoming-elections-in-india.html. Number of female legislators calculated from data provided by Lok Dhaba, The Trivedi Center for Political Data at Ashoka University, https://lokdhaba.ashoka.edu.in/browse-data?et=AE, accessed Sep 26, 2020
avg_margin_pc_per_state	margin of victory	Average margin of victory across all seats in most recent state assembly elections	Calculated from data provided by Lok Dhaba, The Trivedi Center for Political Data at Ashoka University, https://lokdhaba.ashoka.edu.in/browse-data?et=AE, accessed Sep 26, 2020
dif_btw_pop_seat_shares	measure of malapportionment	Difference between seat and population share	Lok Sabha seat data from https://www.elections.in/upcoming-elections-in-india.html. Population data from the 2011 population census of India, Primary Census Abstract data Highlights (India & States/UTs - District Level), https://censusindia.gov.in/2011census/population_enumeration.html, accessed 27 Sep 2020. Population for Telangana from http://www.populationu.com/in/telangana-population, accessed 5 Oct 2020.
reserve_proportion	proportion of reserved seats	Proportion of assembly seats reserved for representatives of scheduled castes and scheduled tribes	Total number of state assembly seats from https://www.elections.in/upcoming-elections-in-india.html. Number of reserved seats calculated from data provided by Lok Dhaba, The Trivedi Center for Political Data at Ashoka University, https://lokdhaba.ashoka.edu.in/browse-data?et=AE, accessed Sep 26, 2020.
average_events_per_state	average number of violent events	Average number of violent events per year 2015-2019, calculated from total number of violent events in 2015 through 2019. Events included are those classed by ACLED as Battles, Explosions/Remote violence, Protests, Riots, and Violence against civilians.	https://acleddata.com/data-export-tool/, accessed 26 Sep 2020
election_soon	electoral pressure	Election scheduled for 2020	https://www.elections.in/upcoming-elections-in-india.html
pan_prep	pandemic preparedness	Pandemic preparedness based on recent experience with dengue, coded 1 where state experienced an average of 10+ deaths from dengue per year between 2015 and 2019	Government of India, https://nvbdcp.gov.in/index4.php?lang=1&level=0&linkid=431&lid=3715, accessed 27 Sep 2020
women_executive	female leader	Female Chief Minister coded 1, as of 1 Jan 2020	Names of Chief Ministers from https://www.elections.in/government/chief-minister.html. Sex coded based on photographs available at https://www.jagranjosh.com/general-knowledge/list-of-current-chief-ministers-in-india-1492596487-1.
leader_experience	executive experience	Experience of Chief Minister, coded as the number of months in office prior to Jan 2020	Calculated based on election dates reported in https://www.elections.in/government/chief-minister.html
party_mis_fed	partisan misalignment between state and federal levels	Partisan misalignment between state Chief Minister and federal Prime Minister, coded 1 when the state Chief Minister is from any party other than the BJP	Coding based on partisanship of Chief Ministers reported in https://www.elections.in/government/chief-minister.html

Table S7: Codebook for: India *(continued)*

Variable Name	Variable Label	Definition	Source
hospital_beds_pc	hospital beds per 1,000 inhabitants	Hospital beds per 1,000 inhabitants	Estimated total number of hospital beds in private and public hospitals as of 20 Apr 2020 from Geetanjali Kapoor et al, "COVID-19 in India: State-wise estimates of current hospital beds, intensive care unit (ICU) beds and ventilators," Center for Disease Dynamics, Economics & Policy and Princeton University, 20 April 2020, downloaded from https://cddep.org/publications/covid-19-in-india-state-wise-estimates-of-current-hospital-beds-icu-beds-and-ventilators/, accessed 13 Nov 2020. To calculate total beds per 1,000 people, population figures from 2011 population census of India, Primary Census Abstract data Highlights (India & States/UTs - District Level), https://censusindia.gov.in/2011census/population_enumeration.html, accessed 27 Sep 2020. Population for Telangana from http://www.populationu.com/in/telangana-population, accessed 5 Oct 2020.
log_pop	log of population	Total population, logged.	2011 population census of India, Primary Census Abstract data Highlights (India & States/UTs - District Level), https://censusindia.gov.in/2011census/population_enumeration.html, accessed 27 Sep 2020. Population for Telangana from http://www.populationu.com/in/telangana-population, accessed 5 Oct 2020.
minority_pct	minority percent	Percent of total population classed as scheduled caste or scheduled tribe.	2011 population census of India, Primary Census Abstract data Highlights (India & States/UTs - District Level), https://censusindia.gov.in/2011census/population_enumeration.html, accessed 27 Sep 2020. Population for Telangana from http://www.populationu.com/in/telangana-population, accessed 5 Oct 2020.
urban_pct	percent of population in urban areas	Percent of total population classed as urban.	2011 population census of India, Primary Census Abstract data Highlights (India & States/UTs - District Level), https://censusindia.gov.in/2011census/population_enumeration.html, accessed 27 Sep 2020. Population for Telangana from http://www.populationu.com/in/telangana-population, accessed 5 Oct 2020.
pop_density	population density	Total population per square kilometer.	https://www.niti.gov.in/niti/content/population-density-sq-km, accessed 18 Oct 2020.
gdp_pc	GDP per capita	GDP per capita calculated as Gross State Domestic Product at current prices (base year 2011-12) in crore Rupees per total population as of 2018-2019.	Gross State Domestic Product from Directorate of Economics & Statistics of respective state governments, reported by Government of India, Ministry of Statistics and Programme Implementation, http://mospi.nic.in/data, accessed 27 Sep 2020. Total population from 2011 population census of India, Primary Census Abstract data Highlights (India & States/UTs - District Level), https://censusindia.gov.in/2011census/population_enumeration.html, accessed 27 Sep 2020. Population for Telangana from http://www.populationu.com/in/telangana-population, accessed 5 Oct 2020.
homicides_pc	per capita homicides	Homicides per capita. Homicides include murder (sec. 302), culpable homicide not amounting to murder (sec. 304), dowry deaths (sec. 304B), and infanticide (sec. 315).	Government of India, Ministry of Home Affairs, National Crime Records Bureau, extit{Crime in India 2018. Statistics}, vol. 1 (New Delhi 2019). Total population from 2011 population census of India, Primary Census Abstract Data Highlights (India & States/UTs - District Level), https://censusindia.gov.in/2011census/population_enumeration.html, accessed 27 Sep 2020. Population for Telangana from http://www.populationu.com/in/telangana-population, accessed 5 Oct 2020.

Table S7: Codebook for: India *(continued)*

Variable Name	Variable Label	Definition	Source
pct_over_65	percentage of population aged 65 and over	Percentage of population aged 65 and over	Number of residents 65 and older calculated using 2011 population census of India, C-13 Single Year Age Returns by Residence and Sex, httpscensusindia.gov.in/2011census/C-series/C-13.html, accessed 17 Oct 2020. Population for Telangana from http: //www.populationu.com/in/telangana-population, accessed 5 Oct 2020.
avg_rainfall	average rainfall	Average rainfall in milimeters, 2015-2019	For 2015, India Meteorological Department, Rainfall Statistics of India - 2015, table 14, p.38; for 2016, ibid. 2016, table 16, p. 42; for 2017,ibid. 2017, table 15, pp. 34-35; for 2018, ibid. 2018, table 15, p. 34; for 2019, table 15, p. 34; for 2019, India Meteorological Department, Office of Climate Research & Services, Annual Climate Summary - 2019, table 1, p. 26.
infant_mortality	infant mortality rate	Estimated infant mortality rate 2019 per 1,000 live births	https://censusindia.gov.in/vital_statistics/SRS_Bulletins/SRS%20Bulletin_2018.pdf, accessed 19 Oct 2020
pct_poor	poverty rate	Percentage of population below poverty line 2011-12	Reserve Bank of India, Table 162, https://www.rbi.org.in/scripts/PublicationsView.aspx?id=16603, accessed 19 Oct 2020
resp_disease_prev	acute respiratory disease infection deaths per 1k ppl	Average number of deaths due to acute respiratory infection per 1,000 people (2014-18)	Respiratory deaths data from Indiastat, available from 2014 to 2018 only, delivered by email on 28 Feb 2021; population data from 2011 India population census, Primary Census Abstract data Highlights (India & States/UTs - District Level), https://censusindia.gov.in/2011census/population_enumeration.html, accessed 27 Sep 2020; population data for Telangana from http: //www.populationu.com/in/telangana-population, accessed 5 Oct 2020.
covid_deaths	cumulative number of COVID-19 deaths	Cumulative number of COVID-19 deaths since 1 Jan 2020 as of today	Government of India, https://www.mygov.in/corona-data/covid19-statewise-status/, accessed 16 November 2020

Table S8: Codebook for: Mexico

Variable Name	Variable Label	Definition	Source
log_pop_mx	population, logged	Logarithm of state population in 2019 (forecast based on the 2010 census)	CONAPO: National population council. https://datos.gob.mx/herramientas/indicadores-demograficos-de-mexico-de-1950-a-2050-y-de-las-entidades-federativas-de-1970-a-2050?category=web&tag=economia, date accessed Oct 11, 2020
pct_over65_mx	percent of population aged 65 and over	Percentage of population aged 65 and over in 2019 (forecast based on the 2010 census)	CONAPO: National population council. https://datos.gob.mx/herramientas/indicadores-demograficos-de-mexico-de-1950-a-2050-y-de-las-entidades-federativas-de-1970-a-2050?category=web&tag=economia, date accessed Oct 11, 2020
pop_density_mx	population density	Total population per square kilometer in 2015	INEGI Banco de Indicadores: Población, Distribución de la Población. https://www.inegi.org.mx/app/indicadores/, date accessed Oct 12, 2020
pct_urban_mx	percent of population in urban areas	Percent of the population 18 and older residing in urban areas out of the total population in 2015	INEGI Banco de Indicadores: Población de 18 años y más que habita en áreas urbanas de cien mil habitantes y más (Personas). https://www.inegi.org.mx/app/indicadores/, date accessed Oct 14, 2020
pct_religious_mx	percent of religious population	Percent of religious population in 2010 (when the last population census was conducted), constructed as one minus percent of non-religious population.	INEGI: Población que no profesa ninguna religión (total). http://en.www.inegi.org.mx/app/tabulados/interactivos/?px=Religion_03&bd=Religion, date accessed Oct 12, 2020
pct_catholic_mx	percent of catholic population	Percent of Catholic population in 2010 (when the last population census was conducted)	INEGI: Población que profesa religión católica por entidad federativa, 1990 a 2010 (total). http://en.www.inegi.org.mx/app/tabulados/interactivos/?px=Religion_01&bd=Religion, date accessed Oct 10, 2020

Table S8: Codebook for: Mexico *(continued)*

Variable Name	Variable Label	Definition	Source
pct_tertiaryemp_mx	percent employed in the tertiary sector	Percent of employed in the tertiary sector in 2010	Censos y Conteos de Poblacion y Vivienda: Serie histórica censal e intercensal (1990-2010), Conjunto de datos: Indicadores sociodemográficos de la población. https://www.inegi.org.mx/programas/ccpv/cpvsh/default.html#Tabulados, date accessed Oct 13, 2020
pct_indiglang_mx	percent indigenous language-speaking population	Percentage of indigenous language-speaking population, 5 years and over, in 2010	INEGI: Pobliacion; Ethnicidad. http://en.www.inegi.org.mx/app/tabulados/interactivos/?px=Lengua_03&bd=LenguaIndigena, date accessed Oct 13, 2020
gdppc_mx	GDP per capita	GDP per capita in 2018 (last available year), in constant 2013 prices; mln pesos per capita	INEGI: Producto Interno Bruto por entidad federativa (PIBE), 2018 revisado, año base 2013 serie detallada. Source: http://en.www.inegi.org.mx/programas/pibent/2013/default.html#Datos_abiertos, date accessed Oct 10, 2020
pct_informal_mx	percent of informal employment	Percentage of employed in the informal sector in q12014-q42019 (aged 15 and over), calculated as the ratio between the number of the employed in the informal sector and the number employed overall	INEGI Banco de Indicadores: Empleo y Ocupación, Población Ocupada (total; en el sector informal). https://www.inegi.org.mx/app/indicadores/, date accessed Oct 11, 2020
gini_mx	Gini coefficient	Gini coefficient of economic (vertical) inequality, on average in 2014-2018	CONEVAL Indicadores de Cohesion Social nacional y entidad federativa 2008-2018. https://www.coneval.org.mx/Medicion/MP/Documents/Cohesion_social, date accessed Oct 12, 2020
pct_extreme_poverty_mx	percent living in extreme poverty	Percentage of people living in extreme poverty, on average in 2014-2018	Sistema Nacional de Informacion Estadistica y Geografica (collected by CONEVAL): Porcentaje de población en situación de pobreza extrema. https://www.snieg.mx/cni/escenario.aspx?idOrden=1.1&ind=6300000108&gen=218&d=n, date accessed Oct 20, 2020
pct_healthins_mx	percent of population with health insurance	Percentage of people covered by health insurance in 2015	INEGI Banco de Indicadores: Salud y Seguridad Social; Derechohabiencia; Porcentaje de la población derechohabiente en el Seguro popular, PEMEX, SDN, SM, IMSS, ISSSTE; https://www.inegi.org.mx/app/indicadores/, date accessed Oct 20, 2020
tuberc_cases_mx	tuberculosis cases per 100,000 people	Confirmed cases of tuberculosis per 100,000 people in 2018	Sistema Nacional de Informacion Estadistica y Geografica: Tasa de incidencia de tuberculosis pulmonar (por 100 mil habitantes); Número de casos confirmados de tuberculosis pulmonar por cada 100,000 habitantes (de todas las edades) en un año determinado. https://www.snieg.mx/cni/escenario.aspx?idOrden=1.1&ind=6300000002&gen=137&d=n, date accessed Oct 13, 2020
trust_people_mx	index of interpersonal trust	Index of interpersonal trust in 2019 (normalized score; higher values mean more trust), constructed from the individual-level data as the first principal component based on questions about trust in neighbors, colleagues/clasmates, relatives and friends	ENVIPE 2019: Base de Datos. https://www.inegi.org.mx/programas/envipe/2019/default.html#Microdatos. Section V on Institutional Trust (Question 5.2), date accessed Oct 13, 2020
trust_inst_mx	index of institutional trust	Index of institutional trust in 2019 (normalized score; higher values mean more trust), constructed from the individual-level data as the first principal component based on questions about trust in federal government institutions (Policía Federal; Policía Ministerial o Judicial; Ministerio Público (MP) y Procuradurías Estatales; Procuraduría General de la República (PGR); Ejército; Marina; Jueces)	ENVIPE 2019: Base de Datos. https://www.inegi.org.mx/programas/envipe/2019/default.html#Microdatos. Section V on Institutional Trust (Questions 4.6 to 4.10), date accessed Oct 13, 2020
health_expendpc_mx	healthcare expenditures per capita	Healthcare expenditures per capita in 2017, in constant 2017 prices (mln pesos per capita)	Gasto en Salud 1993-2017, miles de pesos constantes, 2017 = 100 (in constant 2017 prices): Gasto Público Total en Salud 2017 (miles de pesos). Source: http://www.dgis.salud.gob.mx/contenidos/sinais/gastoensalud_gobmx.html, date accessed Oct 13, 2020

Table S8: Codebook for: Mexico *(continued)*

Variable Name	Variable Label	Definition	Source
hospital_bedspc_mx	hospital beds per 1,000 people	Hospital beds per 1,000 people in 2018	INEGI Banco de Indicadores: Total camas area hospitalizacion. https://www.inegi.org.mx/app/indicadores/, date accessed Oct 13, 2020
av_precipit_mx	average precipitation	Average precipitation (mm/month) in 2019	https://smn.conagua.gob.mx/es/climatologia/pronostico-climatico/precipitacion-form, date accessed Oct 21, 2020
infant_mort_mx	infant mortality rate	The number of infant deaths for every 1,000 live births, on average in 2013-2018	Sistema Nacional de Informacion Estadistica y Geografica (collected by Dirección General de Información en Salud): Tasa de mortalidad infantil: Es el número de defunciones de niños menores de 1 año de edad por cada mil nacidos vivos, en el año de referencia. https://www.snieg.mx/cni/escenario.aspx?idOrden=1.1&ind=6300000011&gen=146&d=n, date accessed Oct 21, 2020
pand_exp_mx	pandemic preparedness	Data on dengue fever cases in 2019 as a proxy for pandemic preparedness: 1 if more than 10 cases, 0 otherwise	CDC WONDER site https://wonder.cdc.gov/nndss/static/2019/52/2019-52-table1j.html, date accessed Sept 21, 2020
elxn_margin_mx	margin of victory	Margin of Victory	National Electoral Institute (INE) website for the information on margin with: https://siceen.ine.mx:3000/#/primeros-tres-lugares
women_exec_mx	female leader	Female governor coded 1, as of 1 Jan 2020	Wikipedia page of governors in office https://es.wikipedia.org/wiki/Anexo:Gobernadores_de_M%C3%A9xico_en_funciones (consulted October 19, 2020)
women_leg_mx	percent of female legislators	Percentage of female legislators	https://igualdad.ine.mx/wp-content/uploads/2017/10/C%C3%A1mara-de-Diputados.as_.xlsx
party_exec_right_mx	right-party executive	State executive is member of right-wing party	Wikipedia page of governors in office https://en.wikipedia.org/wiki/List_of_current_state_governors_in_Mexico (consulted October 19, 2020)
party_mis_fed_mx	federal-state executive partisan mismatch	Partisan mismatch between federal head of state and state-level excecutive	https://es.wikipedia.org/wiki/Anexo:Legislaturas_de_los_estados_de_M%C3%A9xico; https://en.wikipedia.org/wiki/List_of_current_state_governors_in_Mexico
party_mis_state_mx	executive-legislative partisan mismatch	Partisan mismatch between governor and majority party in legislature	https://es.wikipedia.org/wiki/Anexo:Legislaturas_de_los_estados_de_M%C3%A9xico; https://en.wikipedia.org/wiki/List_of_current_state_governors_in_Mexico
leader_experience_mx	leader experience	Number of years state executive (governor) has been in office	National Electoral Institute (INE) Election Calendar https://www.ine.mx/voto-y-elecciones/calendario-electoral/ (consulted October 19, 2020)
elxn_soon_mx	local election in 2020	State election coded 1 in 2020 (mayors and local congress)	National Electoral Institute (INE) Election Calendar https://www.ine.mx/voto-y-elecciones/calendario-electoral/ (consulted October 19, 2020)
malapportion_mx	malapportionment	Over or underrepresentation in Federal Congress (Camara de Diputados), difference between weighted seat share and population share	Seat share calculated from Camara de Diputados Composicion por Entidad Federativa, http://sitllx.diputados.gob.mx/composicion_politicanp.php (consulted October 20, 2020). Population share calculated from CONAPO: National population council, https://datos.gob.mx/herramientas/indicadores-demograficos-de-mexico-de-1950-a-2050-y-de-las-entidades-federativas-de-1970-a-2050?category=web&tag=economia (accessed Oct 11, 2020).
avg_homicide_rate_mx	average homicide rate	Average Homicide Rate per 100,000 2015-2019	Secretariado Ejecutivo del Sistema Nacional de Segurida Publica https://www.gob.mx/sesnsp/acciones-y-programas/datos-abiertos-de-incidencia-delictiva?state=published (consulted October 6, 2020)

Table S8: Codebook for: Mexico *(continued)*

Variable Name	Variable Label	Definition	Source
irag_rate_mx	serious respiratory decease cases per 100,000 people	Serious Acute Respiratory infection rates averaged during the seasonal influenza from week 40 of prior year to week 20 of current year, from 2017 to 2020	Dirección General de Epidemiología, Secretaría de Salud, Informes Semanales para la Vigilancia Epidemiológica de Influenza (various years).https://www.gob.mx/salud/documentos/informes-semanales-para-la-vigilancia-epidemiologica-de-influenza-2020
inf_rate_mx	influenza cases per 100,000 people	Influenza rates averaged during the seasonal influenza from week 40 of prior year to week 20 of current year, from 2017 to 2020	Dirección General de Epidemiología, Secretaría de Salud, Informes Semanales para la Vigilancia Epidemiológica de Influenza (various years).https://www.gob.mx/salud/documentos/informes-semanales-para-la-vigilancia-epidemiologica-de-influenza-2020
infdeath_rate_mx	influenza deaths per 100,000 people	Influenza deaths averaged during the seasonal influenza from week 40 of prior year to week 20 of current year, from 2017 to 2020	Dirección General de Epidemiología, Secretaría de Salud, Informes Semanales para la Vigilancia Epidemiológica de Influenza (various years).https://www.gob.mx/salud/documentos/informes-semanales-para-la-vigilancia-epidemiologica-de-influenza-2020

Table S9: Codebook for: USA

Variable Name	Variable Label	Definition	Source
covid_deaths	cumulative covid deaths	Total deaths per state from the COVID Tracking Project as of 12pm CET Nov 16, 2020	COVID Tracking Project: https://covidtracking.com/data/download/all-states-history.csv
corrected score	legislative professionalism	The measure is drawn from the Squire Index of each US state legislature's professionalism. The legislative professionalism score ranges from 0 to 1 (where 1 is the most professional), and incorporates information on legislator pay, number of days in session, and staff per legislator, all compared to those characteristics in Congress during the same year. We use the latest index that is based on data from the year 2015.	SAGE Journals provides the study that presents the Squire Index: https://journals.sagepub.com/doi/10.1177/1532440017713314
leader_experience	leader experience	Experience of the Governor, coded as the number of months the state executive had been in office prior to January 2020	Calculated based on election dates reported by BallotPedia: https://ballotpedia.org/Partisan_composition_of_governors
pan_prep	pandemic preparedness	Coded as 1 if the state reported any dengue fever cases in 2019	Data on Dengue Fever cases from the CDC WONDER site: https://wonder.cdc.gov/nndss/static/2019/52/2019-52-table1j.html
avg_homicide_rate	homicide rate	Average crude homicide rates are created based on data of the years 2014-2018	We obtain data on both homicide rates and numbers from the CDC site: https://www.cdc.gov/nchs/pressroom/sosmap/homicide_mortality/homicide.htm
trust_instit	institutional trust	Percentage of 2016 VSG respondents in each state who report that they trust the federal government 'just about always' or 'most of the time' in response to the question: 'How much of the time do you think you can trust the government in Washington to do what is right?'	Data set obtained from the Voter Study Group: https://www.voterstudygroup.org/publication/2016-voter-survey
trust_people	interpersonal trust	Percentage of 2016 VSG respondents in each state who report believing that 'most people can be trusted,' in response to the question: 'Generally speaking, would you say that most people can be trusted or that you can't be too careful in dealing with people?'	Data set obtained from the Voter Study Group: https://www.voterstudygroup.org/publication/2016-voter-survey

Table S9: Codebook for: USA *(continued)*

Variable Name	Variable Label	Definition	Source
civil_society	percent of people who participate in civic life	Percentage of people in 2016 in each state who say they have participated in civic life. This is a sum of the proportion of people in each state who responded affirmatively to the following question: 'In general, how often, if at all, do you participate in a nonreligious activity group, such as sports team, book club, PTA or neighborhood association?' There were multiple frequency options (e.g., more than once a week, once a week, etc.). To be clear, excluded from the percent are those people who said 'Never' in response to the question	Data set obtained from the Voter Study Group: https://www.voterstudygroup.org/publication/2016-voter-survey
women_executive	female executive	Coded as 1 if a female governor was in power in the state as of January 1, 2020	Center for American Women and Politics: https://cawp.rutgers.edu/women-statewide-elective-executive-office-2020
percentage_of_women	women in state assembly (percent)	Percentage of female legislators in state assembly in 2019.	Center for American Women and Politics: https://cawp.rutgers.edu/women-state-legislature-2019
party_executive_right	right-party executive	Coded as 1 if the state executive is a member of the Republican party as of January 15, 2020.	Party affiliation obtained from BallotPedia: https://ballotpedia.org/Partisan_composition_of_governors
party_executive_right	right-party majority in legislature	Coded as 1 if the Republicans hold the majority in the state senate in 2019	We use the data on pre-election 2020 majority in the senate from BallotPedia: https://ballotpedia.org/Partisan_composition_of_state_legislatures#2019_Elections
h_diffs	legislative polarization (assembly)	Legislative polarization within the state assembly is measured as the difference in party medians within the house of assembly in 2018, following the logic of Shor and McCarty (2020)	The dataset was obtained from the Harvard Dataverse: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/AP54NE
s_diffs	legislative polarization (senate)	Legislative polarization within the state senate is measured as the difference in party medians within the senate in 2018, following the logic of Shor and McCarty (2020)	The dataset was obtained from the Harvard Dataverse: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/AP54NE
party_mis_state	executive-legislative partisan mismatch	Coded as 1 if there is a partisan mismatch between governor and majority party in legislature as of January 15, 2020.	Party affiliation of Governor from BallotPedia: https://ballotpedia.org/Partisan_composition_of_governors and Majority in State Senate from: https://ballotpedia.org/Partisan_composition_of_state_legislatures#2019_Elections
party_mis_fed	federal-state executive partisan mismatch	Coded as 1 if there is a partisan mismatch between the federal head of state and the state-level executive, i.e. if the Governor was not affiliated with the Republican party as of January 15, 2020	Data on Party Affiliation of the Governor was obtained from BallotPedia: https://ballotpedia.org/Partisan_composition_of_governors
election_soon	executive election in 2020	The variable is coded as 1 if an executive state-level election took place or is scheduled for 2020	Election dates from BallotPedia: https://ballotpedia.org/Gubernatorial_elections,_2020.
avg_margin_pc_per_state	legislative margin of victory	Measured as the difference between the share of votes cast for the winning candidate and the second-place candidate in an election. We obtain data on the average margin of victory within the lower house from the year 2018	Data on the average margins in the lower house can be obtained from BallotPedia: https://ballotpedia.org/State_legislative_elections,_2020
ethnic_frac_score	panethnic diversity	Panethnic Diversity in 2017 by state as introduced in the paper by Lee et al. in 2017	Lee et al. (2017). State-level changes in US racial and ethnic diversity. Demographic Research. Source: https://www.demographic-research.org/volumes/vol37/33/37-33.pdf
pct_religious	population very religious (percent)	Percentage of the state's population who report themselves to be 'very religious', as opposed to 'moderately religious' or 'nonreligious' in 2019	Data set obtained from Statista: https://www.statista.com/statistics/221454/share-of-religious-americans-by-state/
pct_foreign	percent foreign born	Percentage of the state population born outside of United States in 2019	Data set obtained from Statista: https://www.statista.com/statistics/312701/percentage-of-population-foreign-born-in-the-us-by-state/
pct_minority	percent minority population	Percentage of the state's population identifying as nonwhite in 2017	Data set obtained from the Governing site: https://www.governing.com/gov-data/census/state-minority-population-data-estimates.html
living_in_poverty	percent population living in poverty	The average percentage from the years 2018 and 2019 of the state's population living in poverty. We draw upon a data set by U.S. Census Bureau where the percentage of people in poverty is determined using a set of dollar value thresholds that vary by family size and composition, following the Office of Management and Budget's (OMB) Statistical Policy Directive 14	Data set obtained from the Census site: https://www.census.gov/library/publications/2020/demo/p60-270.html

Table S9: Codebook for: USA *(continued)*

Variable Name	Variable Label	Definition	Source
gdppc	GDP per capita	State-level GDP per capita in 2019 measured in USD	Data set obtained from the Bureau of Economic Analysis: https://apps.bea.gov/iTable/iTable.cfm?reqid=70&step=1#reqid=70&step=1
gini_coef	Gini coefficient	Gini coefficient as a measure for household income distribution inequality for U.S. states in 2019	Data set obained from Statista: https://www.statista.com/statistics/227249/greatest-gap-between-rich-and-poor-by-us-state/
pct_urban	percent of urban population	Percentage of the state's population living in urban areas in 2010	Data set obtained from Statista: https://www.statista.com/statistics/301961/us-urban-population-by-state/
log_pop	population (logged)	Logged state population in 2010	Data set on population estimates obtained from the US Census Bureau: https://www.census.gov/search-results.html?q=population+per+state&page=1&stateGeo=none&searchtype=web&cssp=SERP&_charset_=UTF-8
pop_density	population density (logged)	State's population per square mile in 2019	Data set obtained from Statista: https://www.statista.com/statistics/183588/population-density-in-the-federal-states-of-the-us/
avg_rainfall	Average yearly precipitation in 2017	Average yearly precipitation (rain and snow), centimeters	Yearly average from Netstate: https://www.netstate.com/states/tables/state_precipitation.htm
beds_sum	hospital beds per 1,000 people	Number of hospital beds per 1,000 people in a given state in 2018	Data set obtained from KFF State Health Facts: https://www.kff.org/other/state-indicator/beds-by-ownership/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D
health_spending_per_capita	health care expenditures per capita	Health care expenditures per capita by state of residence in 2014	Data set obtained from KFF State Health Facts: https://www.kff.org/other/state-indicator/health-spending-per-capita/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D
pct_health_ins	percent of population with health insurance	Percentage of the state's population with health insurance coverage in 2019. The variable created as the column 'uninsured' subtracted from the total of 1 (100%)	Data set obtained from KFF State Health Facts: https://www.kff.org/other/state-indicator/total-population/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D
pct_pop_over_65	percent of population aged 65 and over	Percentage of population aged 65 and over in 2018	Census Bureau's 2018 population estimates: https://www.prb.org/which-us-states-are-the-oldest/
resp_disease_prev	respiratory disease prevalence	Number of deaths due to chronic lower respiratory disease in 2018	Data set obtained from the CDC site: https://www.cdc.gov/nchs/pressroom/sosmap/lung_disease_mortality/lung_disease.htm
malapportionment_congress	malapportionment_congress	Difference between the share of seats in Congress (House of Reps) and the share of people in 2019 for each state	The data for Congressional seats per state are obtained from Britannica: https://www.britannica.com/topic/United-States-House-of-Representatives-Seats-by-State-1787120, date accessed: 19 Nov 2020. The data on population can be obtained from the Cencus site: Data set on population estimates obtained from the US Census Bureau: https://www.census.gov/search-results.html?q=population+per+state&page=1&stateGeo=none&searchtype=web&cssp=SERP&_charset_=UTF-8, date accessed: 19 Nov 2020

S2.6 Forecasting Details

In the forecasting portion of this project we used the Social Science Prediction Platform to elicit expert assessments of the models. The 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 first was 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 another challenge.

Figure S4 reports the number of respondents who entered and completed each forecasting exercise. There is substantial dropoff for the initial, crossnational forecasts. Indeed, just 42.6% and 30% of individuals who entered the system completed the horse race and stacking forecasts, respectively. As such, the difference in completion rates for the initial forecast was 12.6 percentage points ($p = 0.06$). We note that the ratio of additional forecasts to initial forecast

completion is 43/35 and 49/48 for the horse race and stacking exercises, respectively. Collectively these completion rates suggest that the stacking forecasts may have been more challenging or taxing than the horse race forecasts. We detail the procedures for each elicitation in the next subsections.

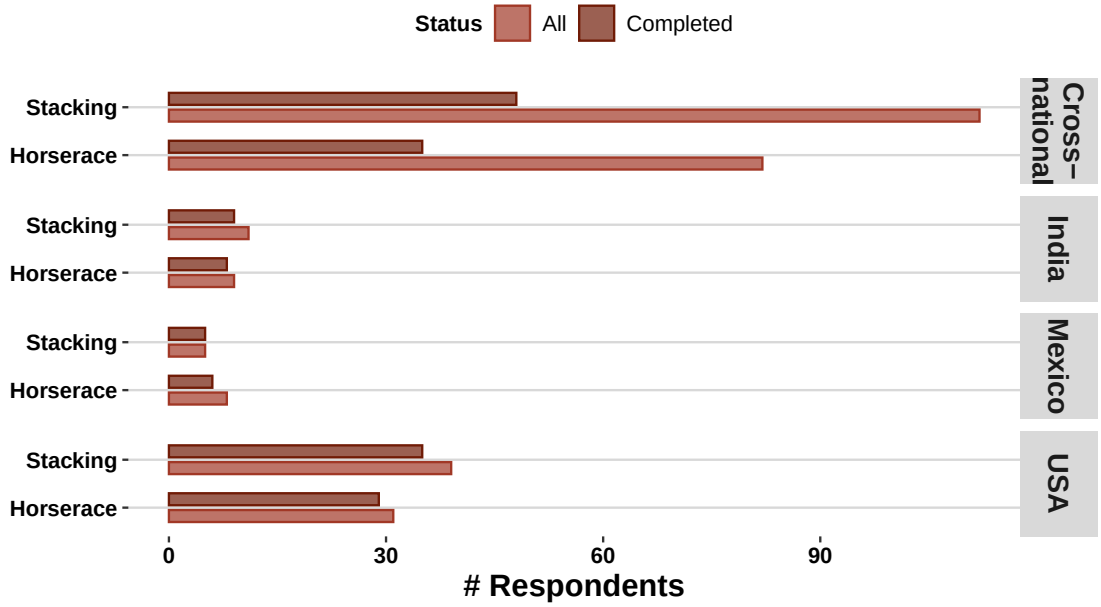


Figure S4: Selection into the forecasting activty.

S2.7 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 there varied across respondents since 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.

S2.8 Stacking Forecasts

The goal of the stacking forecasts was to elicit the **stacking weights**, analogous to those that we estimate using (10) 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 there varied across respondents since different respondents saw different subsets of the general models.

"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.

S2.9 Relationship between models

In Figure 6, we plot two measures of the distance between the models submitted in the crossnational general challenge. In this section, we provide information on the definition of both distance measures and report analogous graphs from the India, Mexico, and US general challenges.

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

(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).

regression setting (Jones 2019). For two models with predictor sets $\mathbf{x}_1, \mathbf{x}_2$, we measure their multivariate R^2 by (4).

$$\begin{aligned} \text{MV-}R^2_{2\to1} &= 1 - \frac{\sum_{i=1}^{N} d(\mathbf{x}_{1,i}, \hat{\mathbf{x}}_{1,i})}{\sum_{i=1}^{N} d(\mathbf{x}_{1,i}, \bar{\mathbf{x}}_1)} \\ \text{MV-}R^2_{1\to2} &= 1 - \frac{\sum_{i=1}^{N} d(\mathbf{x}_{2,i}, \hat{\mathbf{x}}_{2,i})}{\sum_{i=1}^{N} d(\mathbf{x}_{2,i}, \bar{\mathbf{x}}_2)} \\ \text{Avg. MV-}R^2 &= \frac{1}{2}\left(R^2_{2\to1} + R^2_{1\to2}\right) \end{aligned} \tag{4}$$

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 (5).

$$\text{Adj. Correlation} = \frac{1}{2}\left(\frac{\sum_{i=1}^{N}(y_{1,i} - \bar{y}_1)(y_{2,i} - \bar{y}_2)}{\sqrt{\sum_{i=1}^{N}(y_{1,i} - \bar{y}_1)^2 \sum_{i=1}^{N}(y_{2,i} - \bar{y}_2)^2}} + 1\right) \tag{5}$$

For ease of interpretation of the network graphs, Table S10 gives the directory of general models in the four challenges with their indices in the network plots of figures 6 and S6. The indexing is given by the rank of a model by its pseudo-R^2 statistic in descending order.

Figure S6 replicates the analysis from Figure 6 for the other three national challenges.

S3 Evaluating the Models

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

S3.1 Fitting the Predictive Models

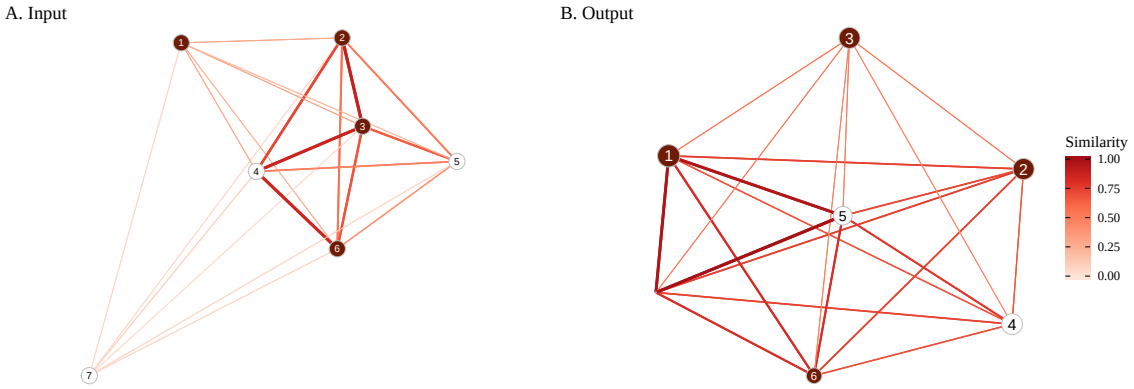
First, we denote models submitted as part of the COVID-19 Models Challenge as:

$$y_i = f_k(\mathbf{x}_i|\boldsymbol{\theta}_k) \tag{6}$$

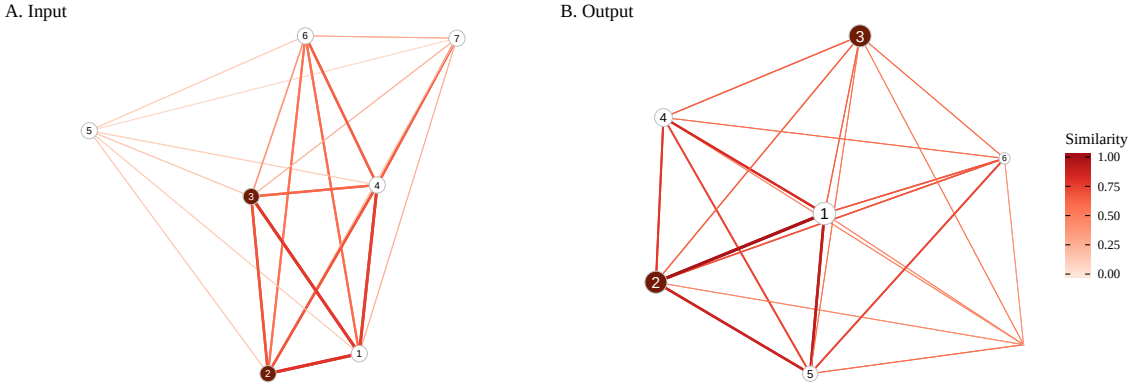
We index observations by i and models by k. The outcome, logged cumulative COVID-19 deaths per million, is denoted

Table S10: Directory of general models in the model network for all challenges, ranked by their pseudo-R^2 statistics

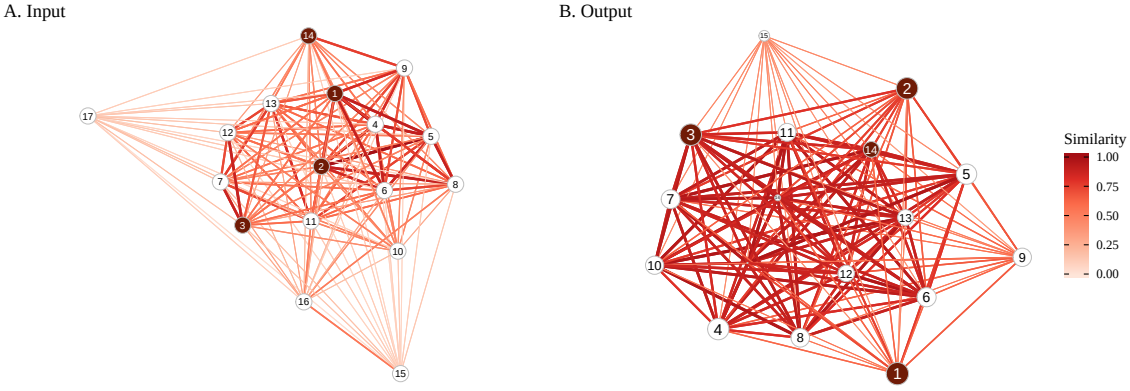
ID	Crossnational	India	Mexico	USA
1	Trust in Authoritarian Government with Strong Age Effect	Health Sector Capacity	Socio-Political	Inequality and Polarization
2	Government Capacity and Social Inequality	Interactions and Political Pressures	Trust and Catholicism	Population Differences
3	Perverse Development	Urbanisation and Health Care	Trust Poverty and TB	Inequality and Capacity
4	Health Sector	Business and Density Model	Poverty, Election, and Public Goods	Right Party Power and Income Inequality Model
5	Pandemic Experience + Inequity	GDP TB and Othering	Government Experience	Religiosity Model
6	Effective Developing Country	Minority Representation and Urbanization	Interactions and Political Pressures	Women in Leadership Model
7	Development	Government Capacity	Investment Inequality	Ethnicity, Inequality, and Capacity
8	Populism and Trust			Social Contact
9	Trust and Nature			Institutional Trust Model
10	Democracy			Community Equality and Trust
11	State, Society, and Elderly			Religion, Economic Inequality, and Minority Status
12	Trust and Institutions			Inequality and Urbanity
13	Government Capacity			Poverty
14	Dictatorships and Misinformation			Institutional Trust and Race
15	Pandemic Readiness			Vaccination Coverage
16	Long-Run Factors			Health
17	Shackled Leviathan			Government Experience
18	Institutional Trust			
19	Trust, Development, and State			
20	Trust and Air Travel			
21	Liberalism, Capitalism and Media Independence			
22	Cultural Tightness			
23	Trust and Social Safety			
24	Language and Culture			
25	Polarization and Populism			
26	Competitiveness of Executive Recruitment Model			



(a) Model networks for India.



(b) Model networks for Mexico.



(c) Model networks for the USA.

Figure S6: 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.

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 specific models.

For **general** models, note that 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:

$$\widehat{y}_{ik}^{\text{loo}} = f_k(\mathbf{x}_i|\widehat{\boldsymbol{\theta}_k^{-i}}) \tag{7}$$

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:

$$\widehat{y}_{ik} = f_k(\mathbf{x}_i|\boldsymbol{\theta}_k) \tag{8}$$

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

S3.2 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:

$$v_k = 1 - \alpha \frac{\sum_i (\widehat{y}_{ik} - y_i)^2}{\sum_i (\overline{y}_i - y_i)^2} \tag{9}$$

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

For the *level* approach, we evaluate (9) 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. Note that 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 (9) 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.

S3.3 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:

$$w = \arg\min_{w} \sum_{i=1}^{n} \left(y_i - \sum_{k} w_k \widehat{y}_{ik} \right)^2 \text{ s.t. } w_k \geq 0 \; \forall k, \sum_{k=1}^{K} w_k = 1 \tag{10}$$

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. Predictive ability of each model is increasing in w.

S3.4 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. Note that 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:

$$\hat{k}^c = \text{argmax}_k \left\{ \sum_j w_k^j \right\} \tag{11}$$

   This yields model predictions given by:

$$\widehat{y}_i^c = \widehat{y}_{i\widehat{k}^c} \tag{12}$$

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:

$$\widehat{y}_i^j = \sum_k \widehat{w}_k^j \widehat{y}_{ik}^{\text{loo}} \tag{13}$$

   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 (9) 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:

$$w^r = \{w^j | v^j = \text{median}(v^h)_{h \in H}\} \tag{14}$$

   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:

$$w_k^c = \frac{\sum_j \widehat{w}_k^j}{\sum_k \sum_j \widehat{w}_k^j} \tag{15}$$

This yields model predictions given by:

$$\widehat{y}_i^c = \sum_k w_k^c \widehat{y}_{ik}^{\text{loo}} \tag{16}$$

S4 Gathering: Supplementary results

In this section, we provide complementary results for the "gathering" stage of our analysis. Table S11 provides an overview of the "general" models submitted cross all four challenges, including information about the (i) functional form of the models; (ii) the number of predictors used and use of predictors outside our dataset; (iii) the theoretical justification for the models; and (iv) the number of modelers that submitted each model. Table S12 provides information about the model challenge participants ("modelers"). Finally, Figures S7 to S9 depict the co-occurrence of predictors in model submissions in the country-specific challenges.

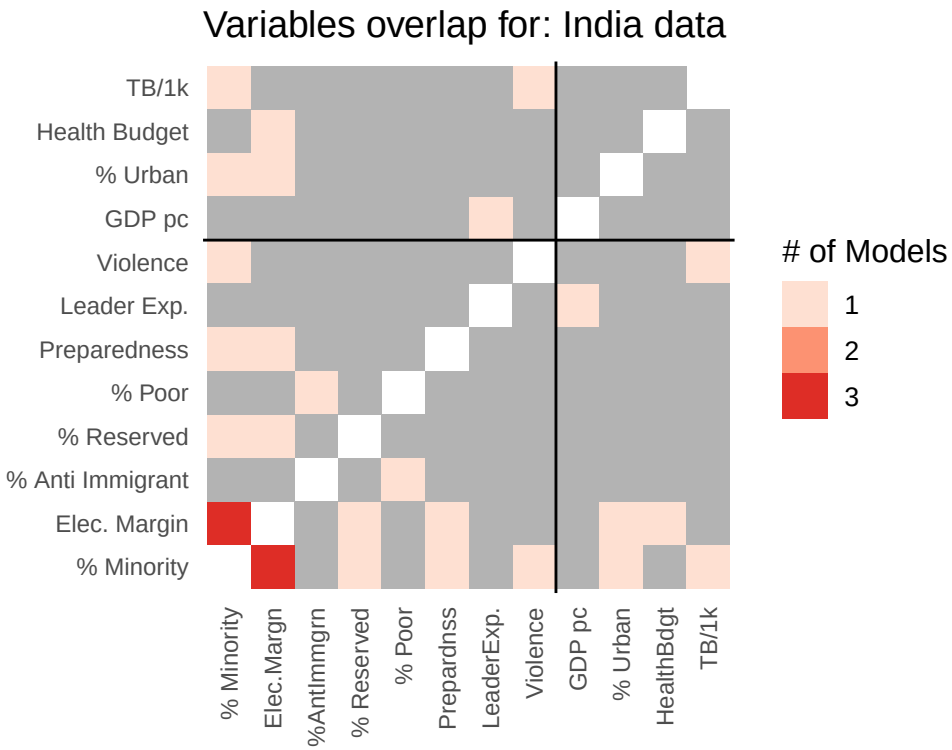


Figure S7: Country-specific challenge, India data

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 S11: 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. Note that three submissions did not identify all modelers by name.

Table S12: Tally of model challenge participants.

Challenge	# Modeler	# Institution	# Country
Crossnational	42	21	9
India	18	6	5
Mexico	15	6	3
US	29	15	6
All	60	32	10

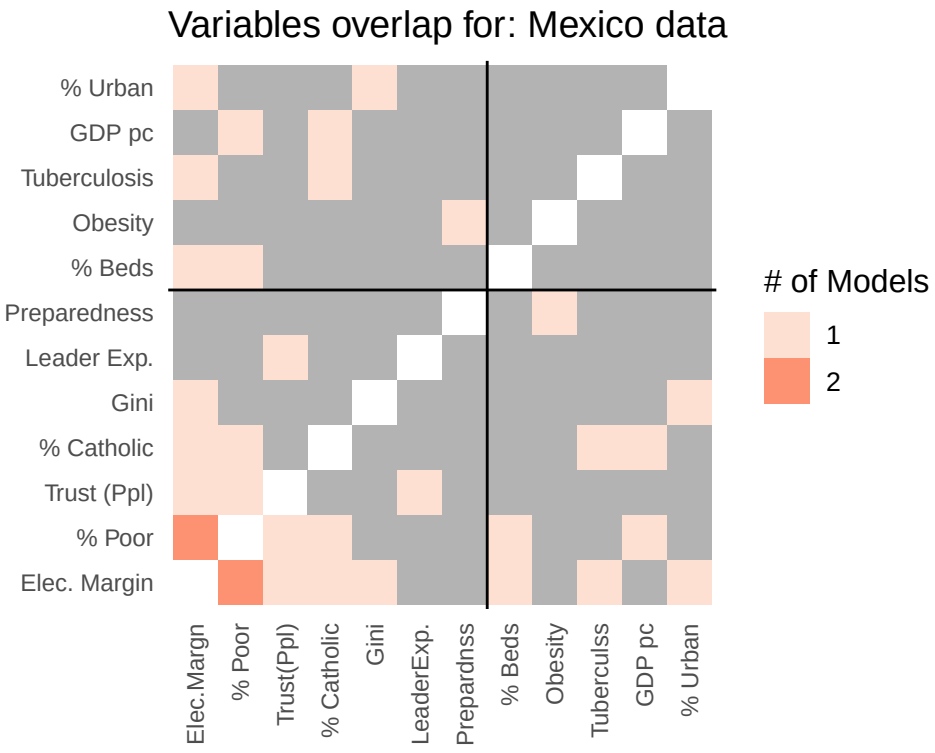


Figure S8: Country-specific challenge, Mexico data

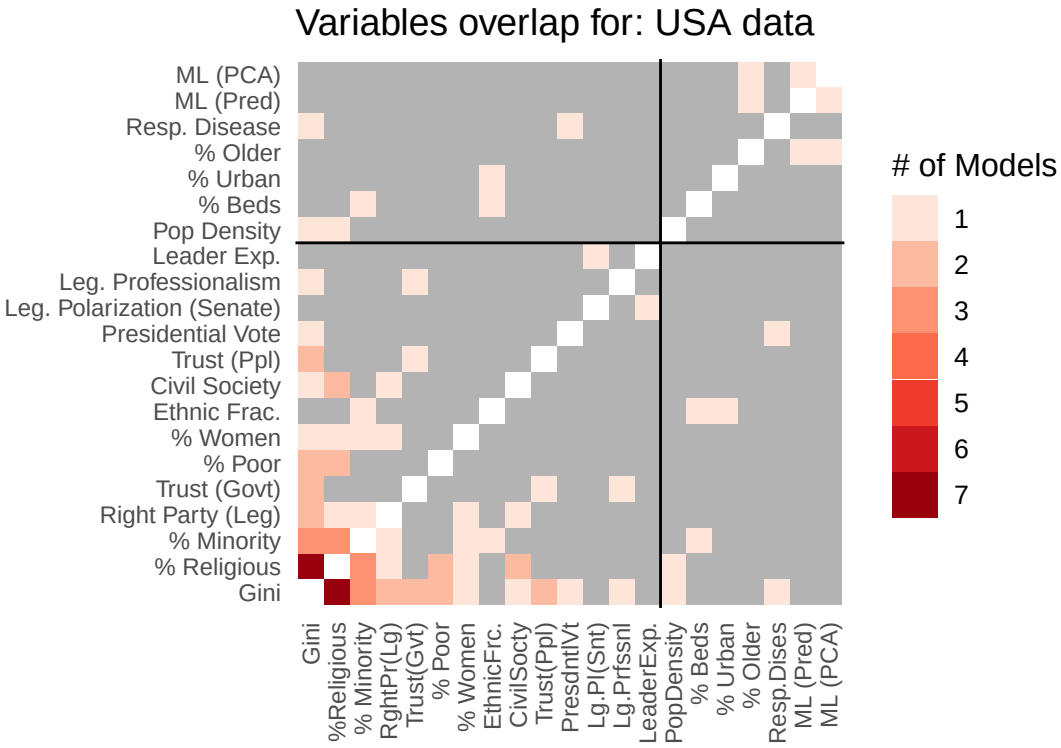


Figure S9: Country-specific challenge, USA data

S5 Evaluating: Supplementary results

In this section, we report the results from the "evaluating" stage for the additional seven challenges. We first report results on individual model performances. Figure S13 summarizes the performance of the best and median-performing model in each challenge.

			Pseudo-R^2	
Challenge	# of models	Figure	Best	Median
Crossnational, general	28	2	0.530	0.171
Crossnational, parameterized	16	S10	0.141	-0.379
India, general	9	S11	0.422	0.295
India, parameterized	7	S12	-1.67	-3.12
Mexico, general	9	S13	0.455	0.04
Mexico, parameterized	6	S14	0.619	-4.72
USA, general	19	S15	0.549	0.325
USA, parameterized	7	S16	-0.102	-5.56

Table S13: Summary of results from the "evaluating" analysis of pseudo-R^2 for each challenge.

We also examine the robustness of our model evaluation to alternative treatments of missing predictors. In the model challenge, the instructions indicated:

> "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 S12-S16 when 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 S14.

Note that each model is a regression model. Our fit statistics (pseudo-R^2 and correlation) summarize the results of each model. For illustrative purposes, we show the top three performing general models (ranked by pseudo-$\hat{R}^2$) in the

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

Table S14: Percent of observations remaining after listwise deletion for any missing predictor.

Table S15: Pseudo-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

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

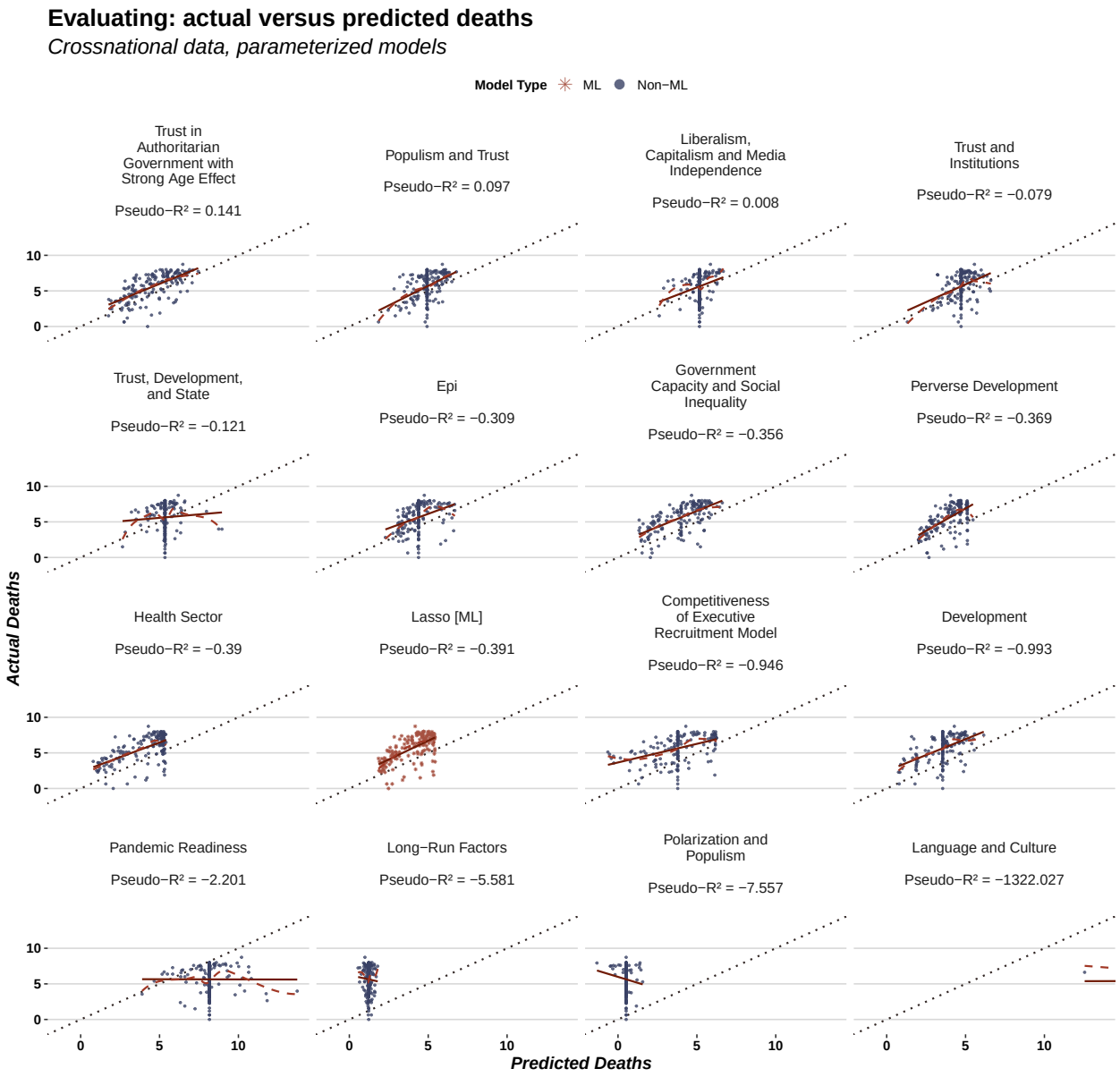
crossnational challenge in Table S15.

Tables S10-S16 show the gathering analysis in Table 2 for the remaining seven challenges.

Figure S17 shows the (linear) correlation in model performances (pseudo-R^2) between general and parameterized forms of the same models across all four challenges. Notice that for display reasons, we have left-censored pseudo-R^2 at -1. The "raw" correlation reflects the uncensored result; the "adjusted" correlation reflects the censored result.

Now we report the results of our model selection exercise. Table S16 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.

In an effort to assess the robustness of predictions to the specific date at which they are evaluated (August 31, 2021), Figure S22 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 with strong age



Note: "Language and Culture" model mostly not shown due to extreme values.

Figure S10: Summary of model predictions and observed COVID-19 mortality from crossnational specific models.

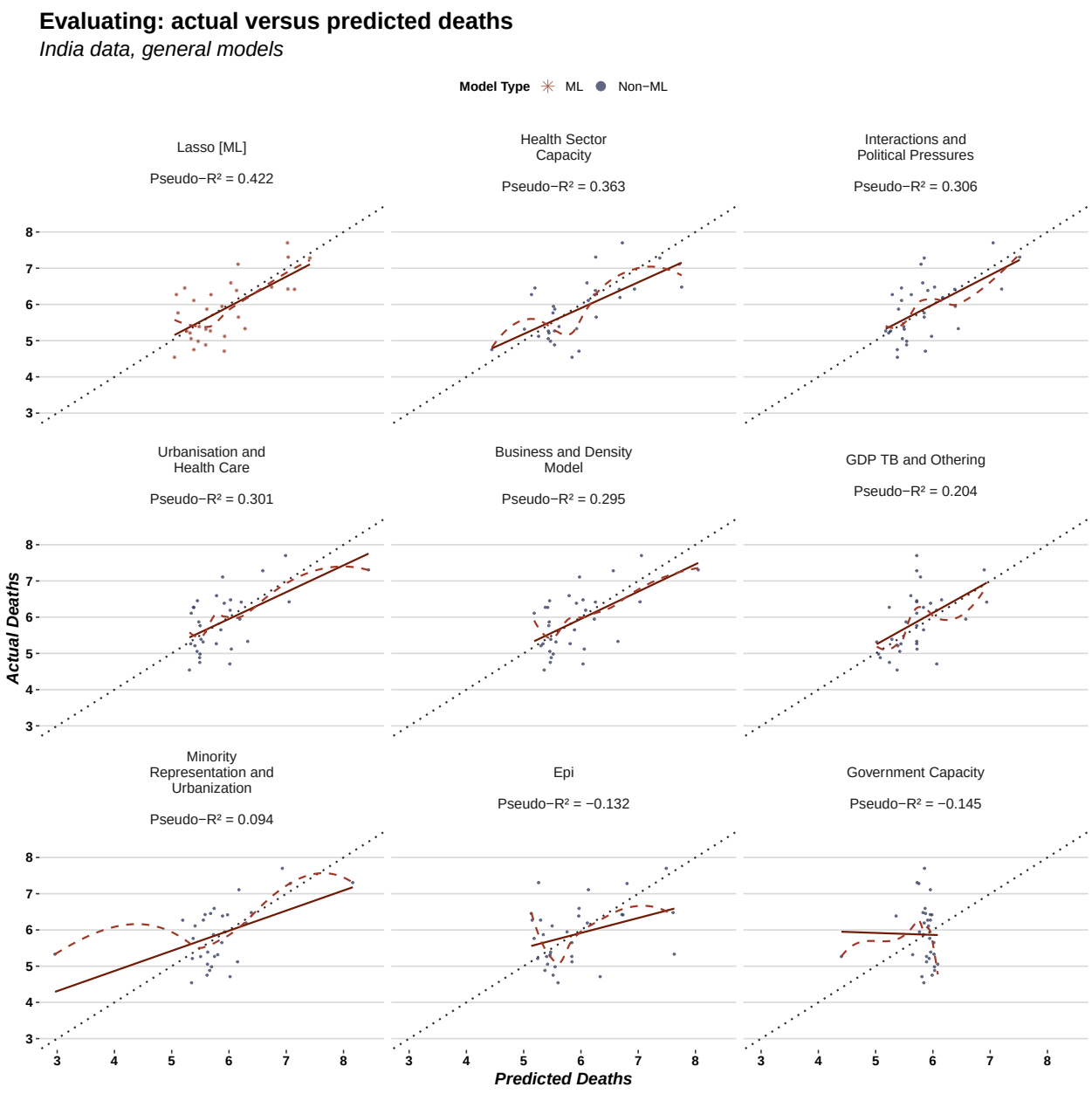


Figure S11: Summary of model predictions and observed COVID-19 mortality from general models for the India data.

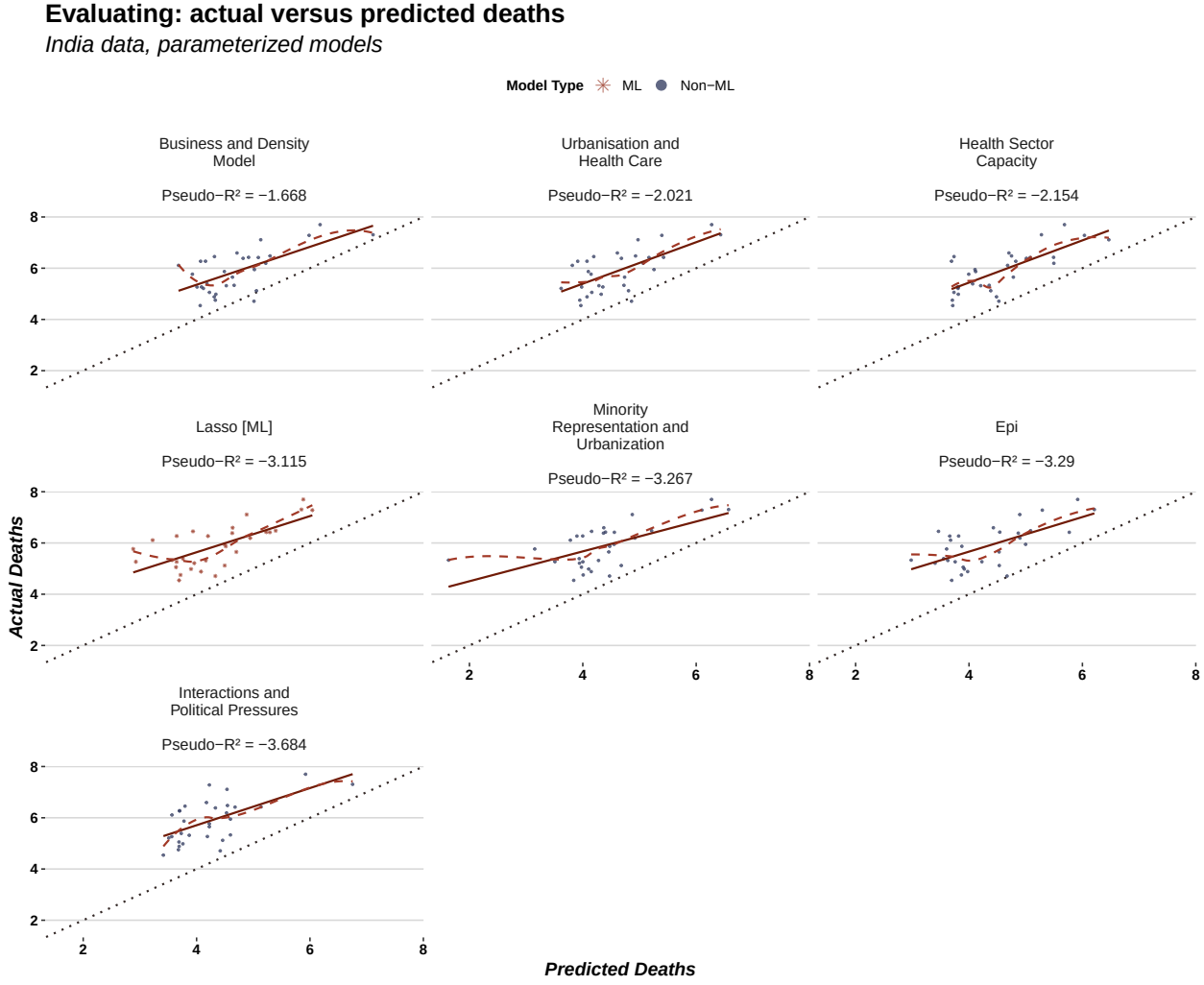


Figure S12: Summary of model predictions and observed COVID-19 mortality from parameterized models for the India data.

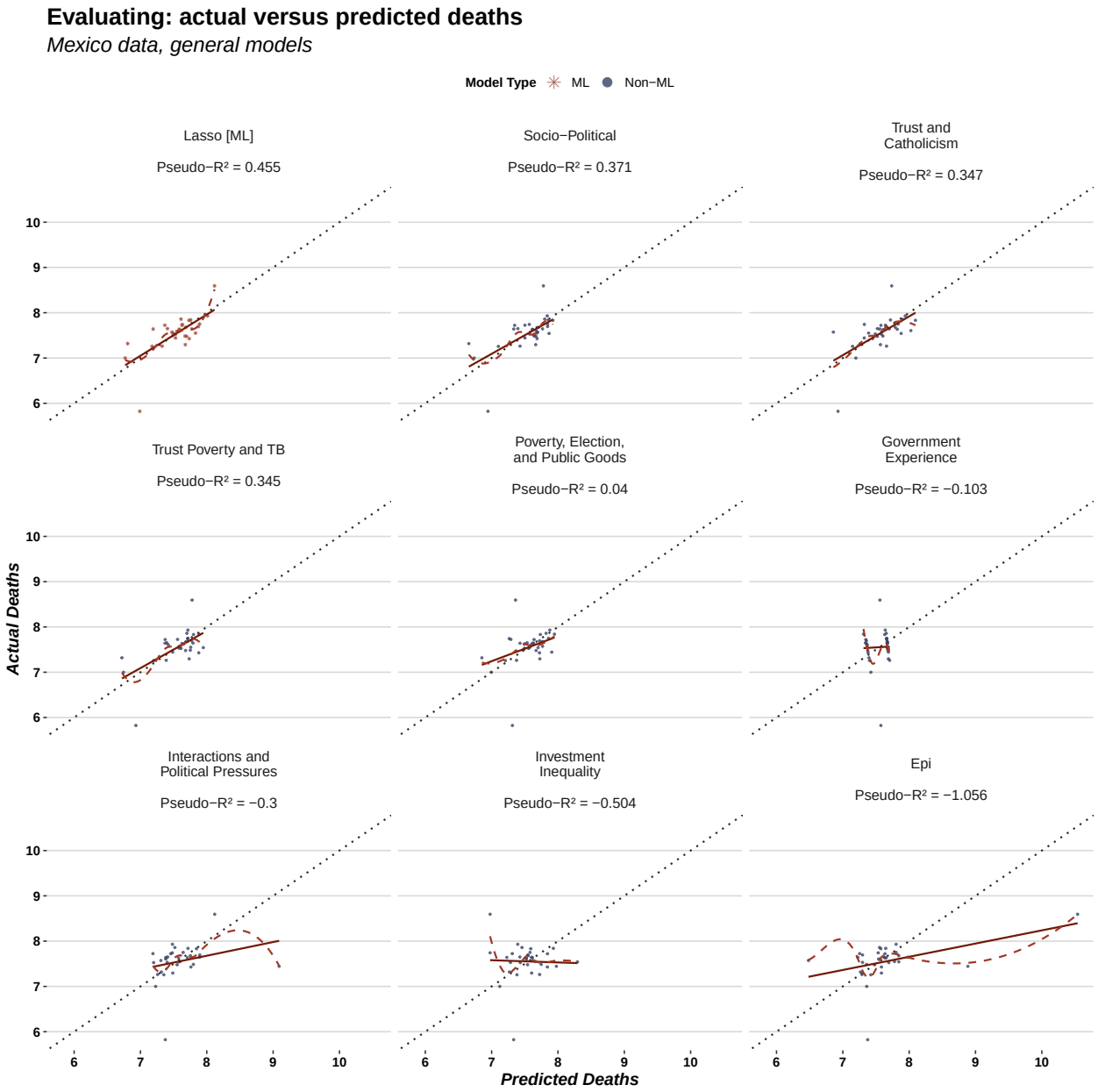


Figure S13: Summary of model predictions and observed COVID-19 mortality from general models for the Mexico data.

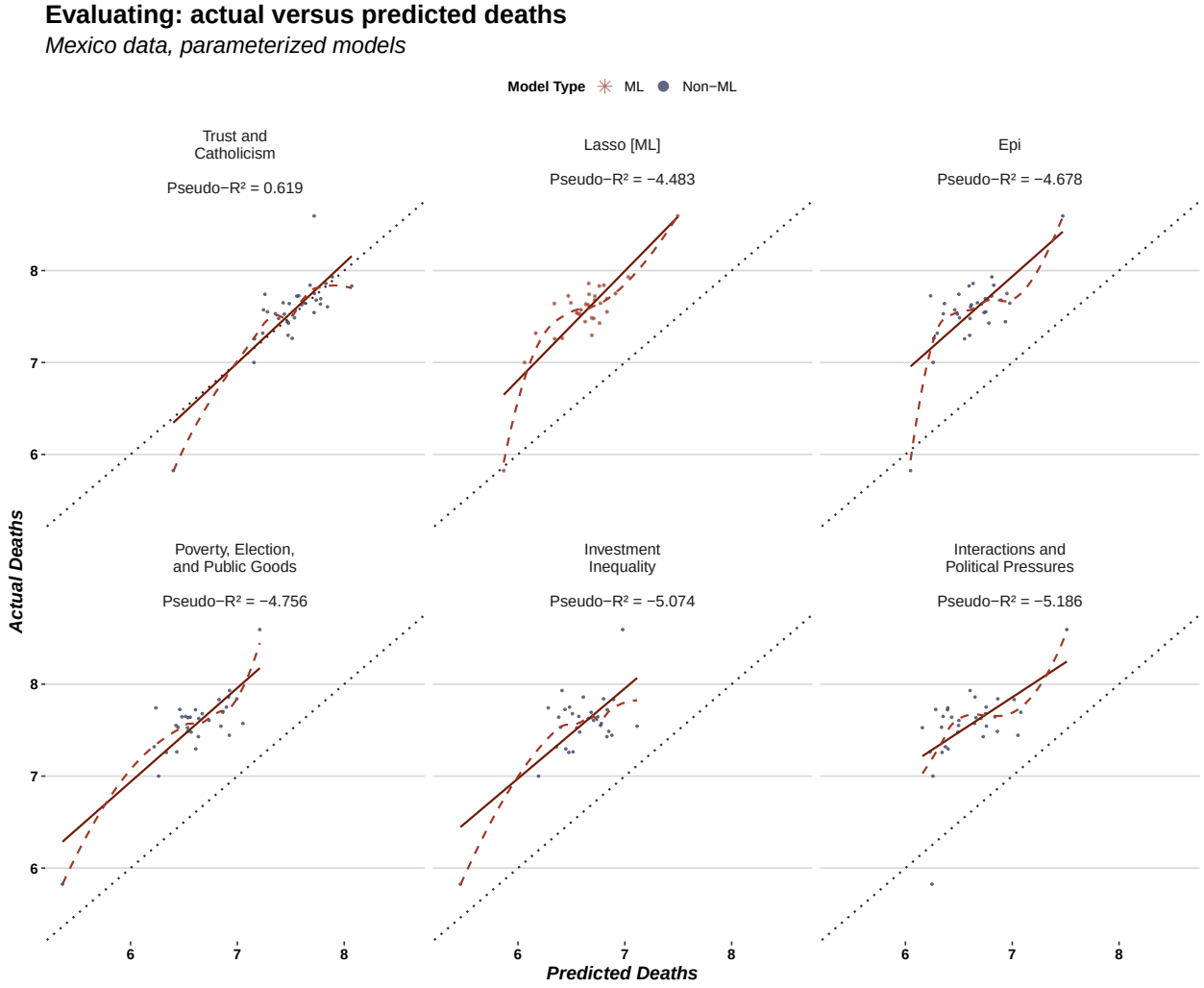


Figure S14: Summary of model predictions and observed COVID-19 mortality from parameterized models for the Mexico data.

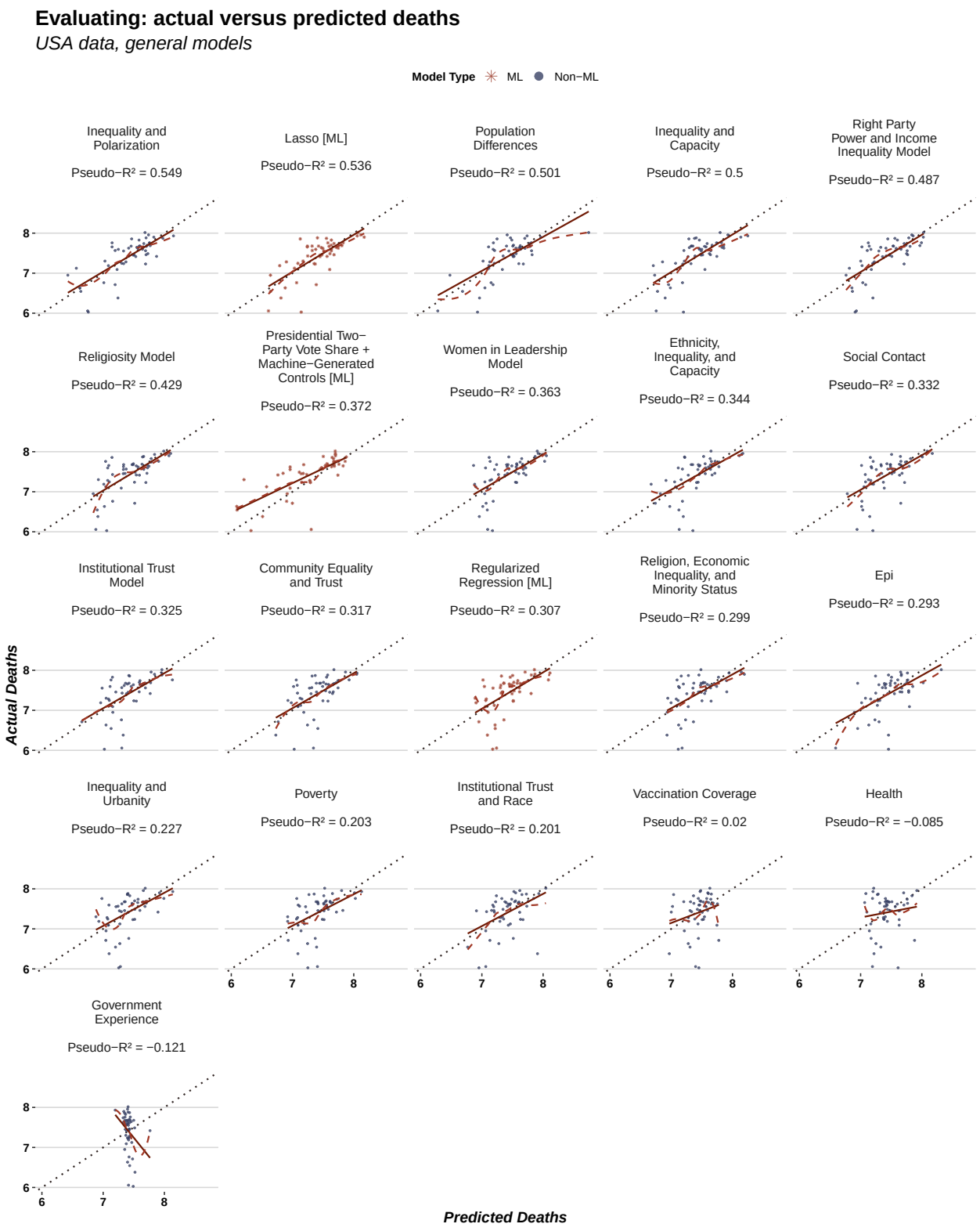


Figure S15: Summary of model predictions and observed COVID-19 mortality from general models for the US data.

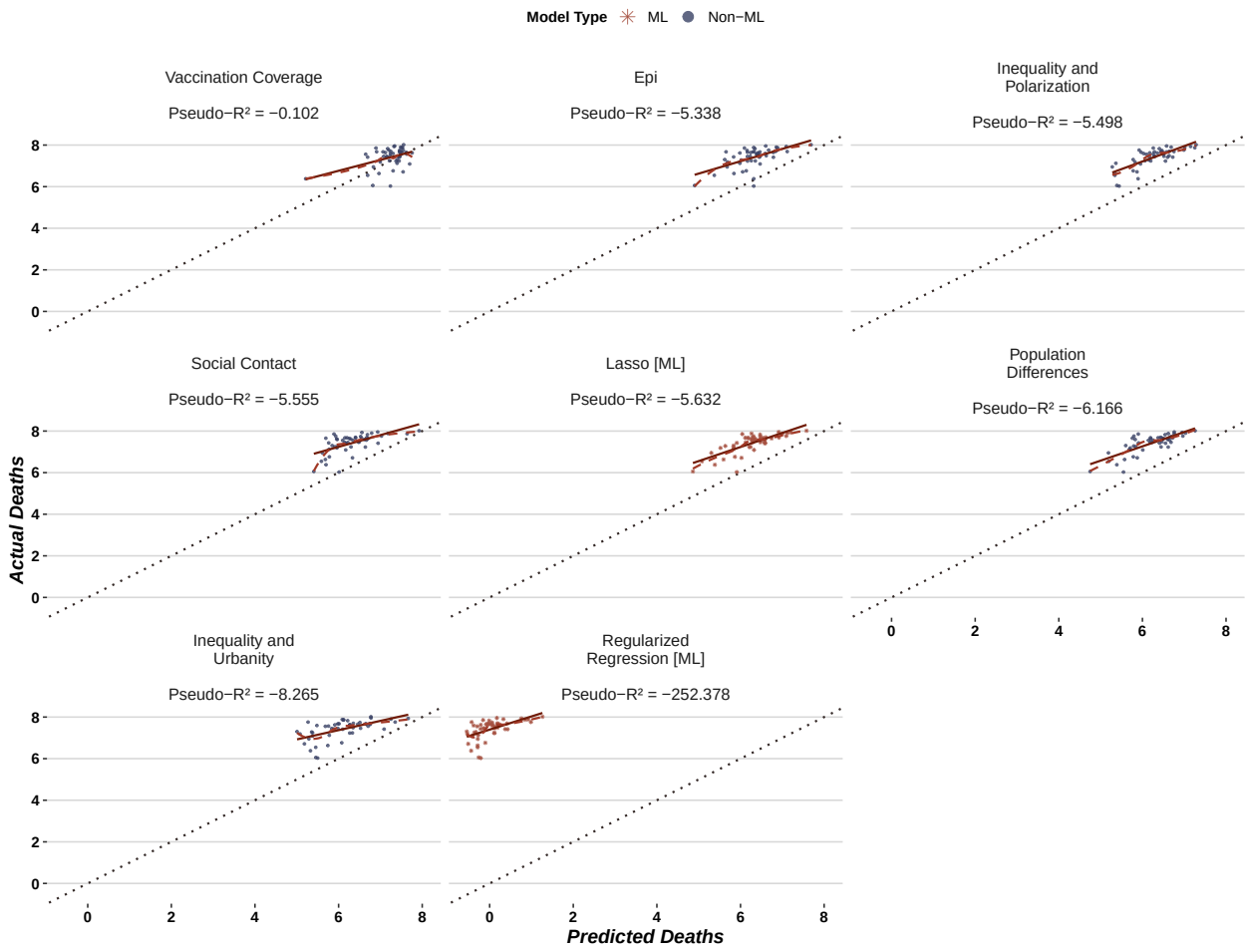


Figure S16: Summary of model predictions and observed COVID-19 mortality from parameterized models for the US data.

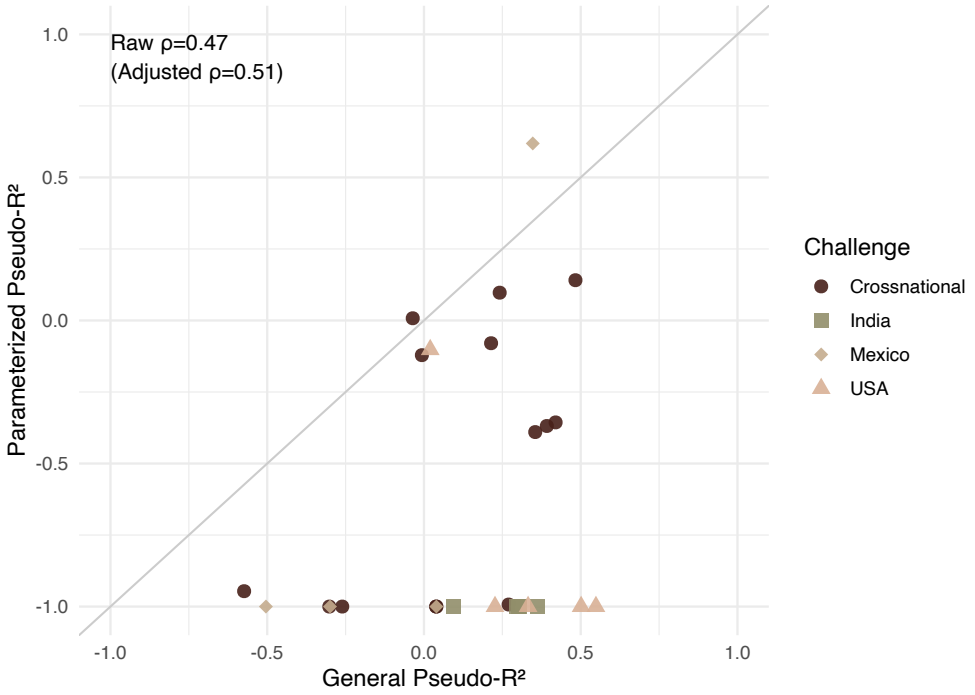


Figure S17: Correlation of pseudo-R^2 between general and parameterized forms of the same model. Values for the latter are left-censored at -1.

effect" and "government capacity and social inequality") are the top two models throughout the post-prediction period. Second, after August 31, however, we see a decrease in the weight afforded to the "government capacity and social inequality" model and a marked increase in the weight afforded to the "development" model.

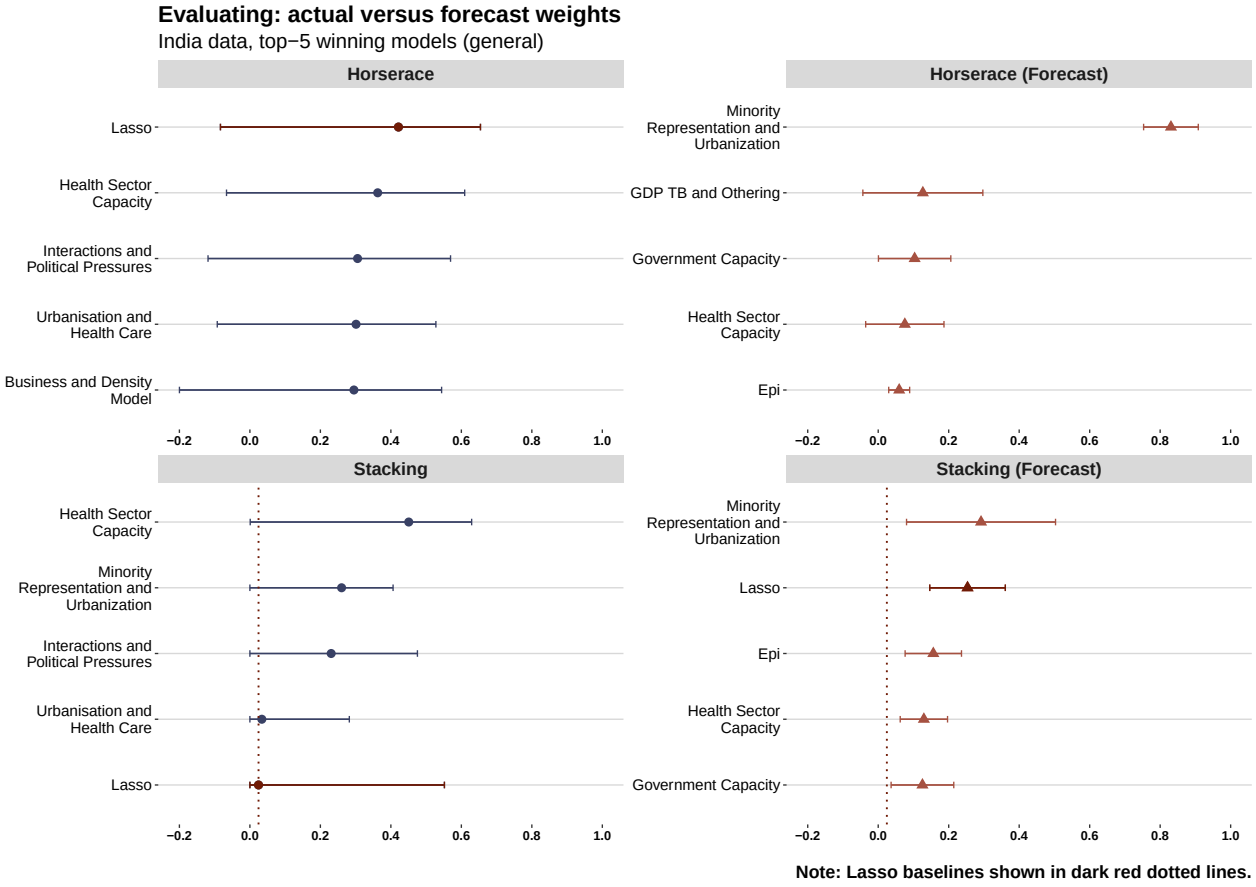


Figure S18: Model selection using four methods for the general models from India.

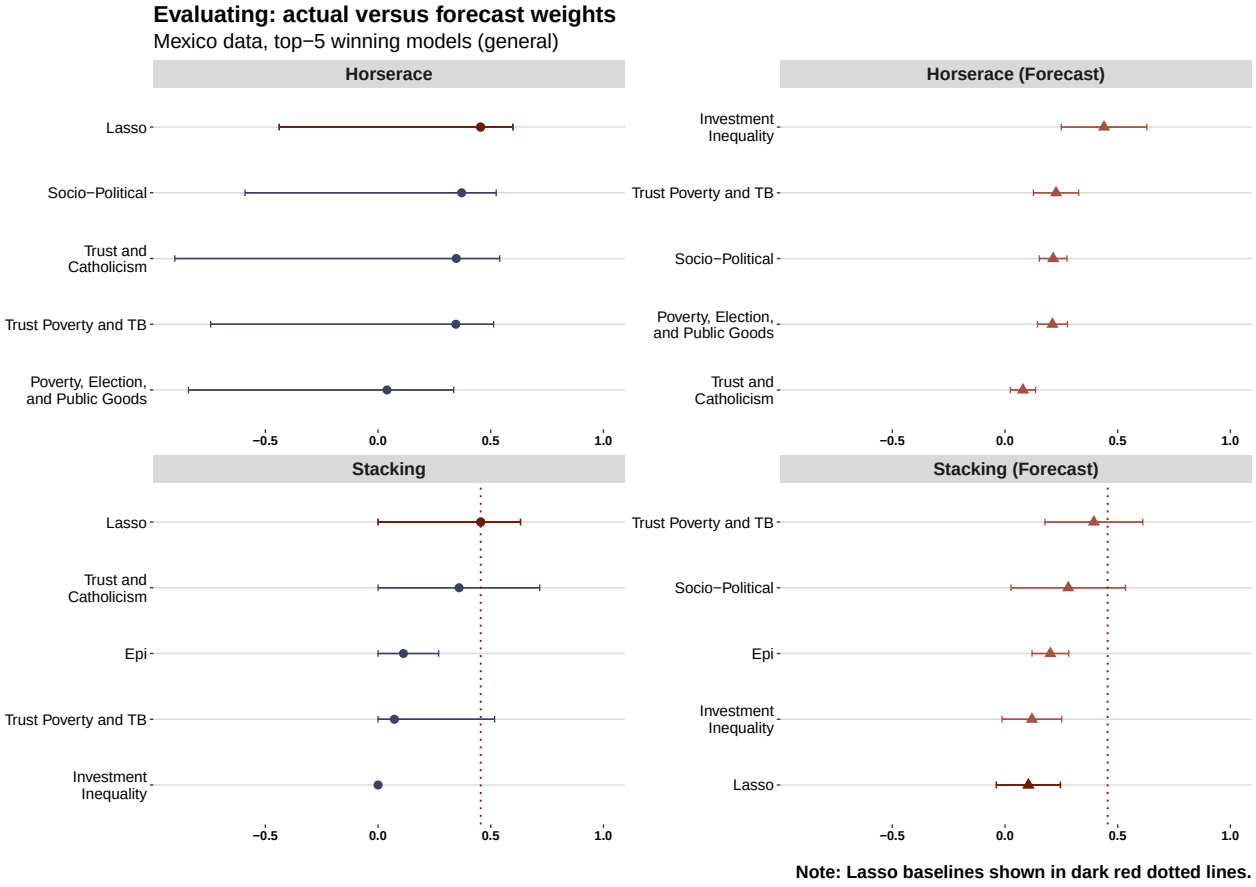


Figure S19: Model selection using four methods for the general models from Mexico

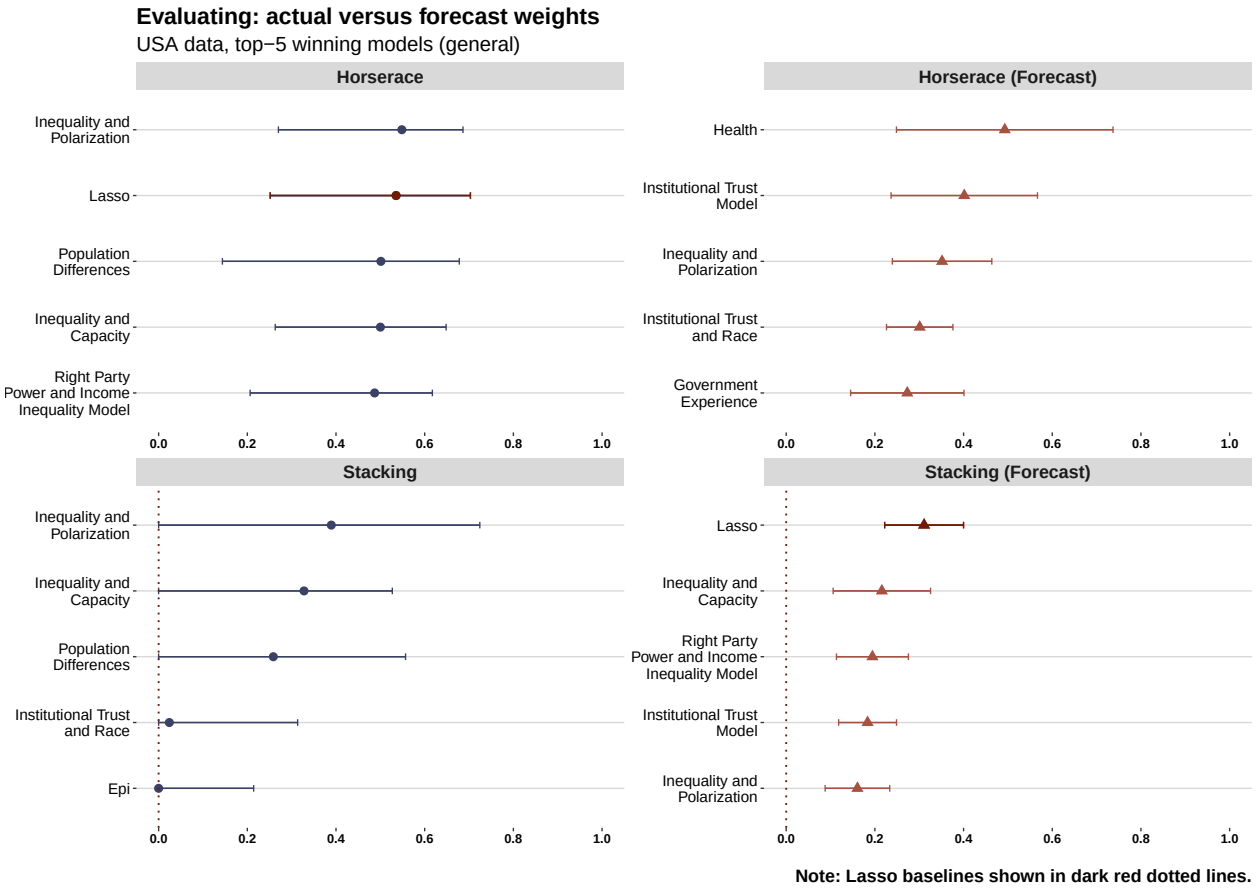
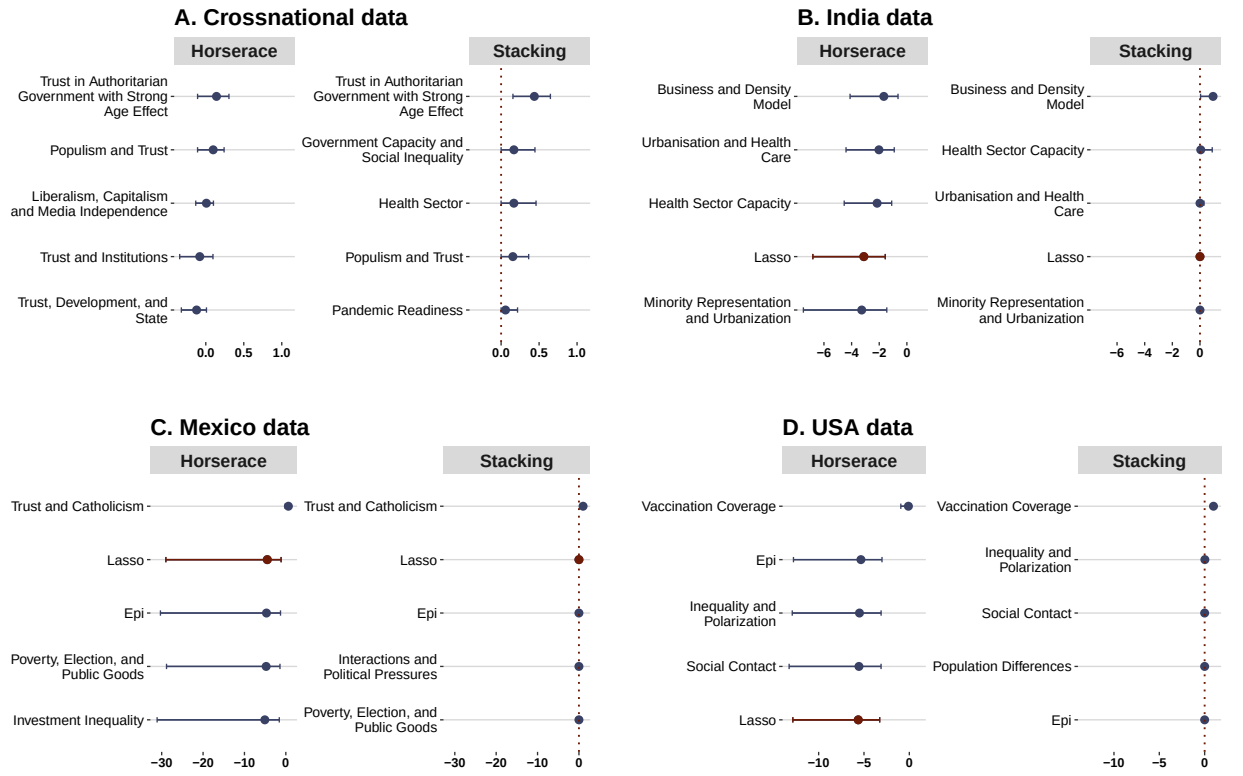


Figure S20: Model selection using four methods for the general models from the US

Evaluating: actual vs forecast weights

All four challenges, top−5 winning models (parameterized)

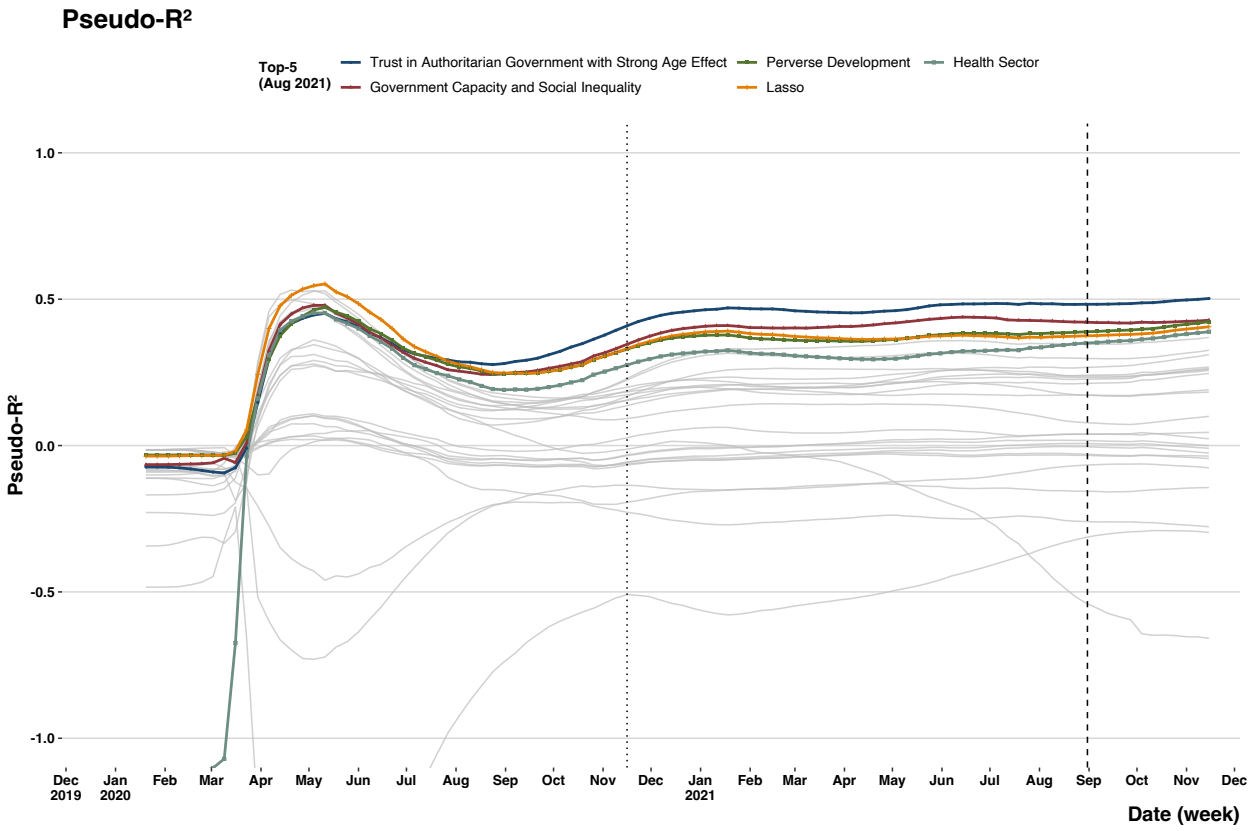


Note: Lasso baselines shown in dark red.

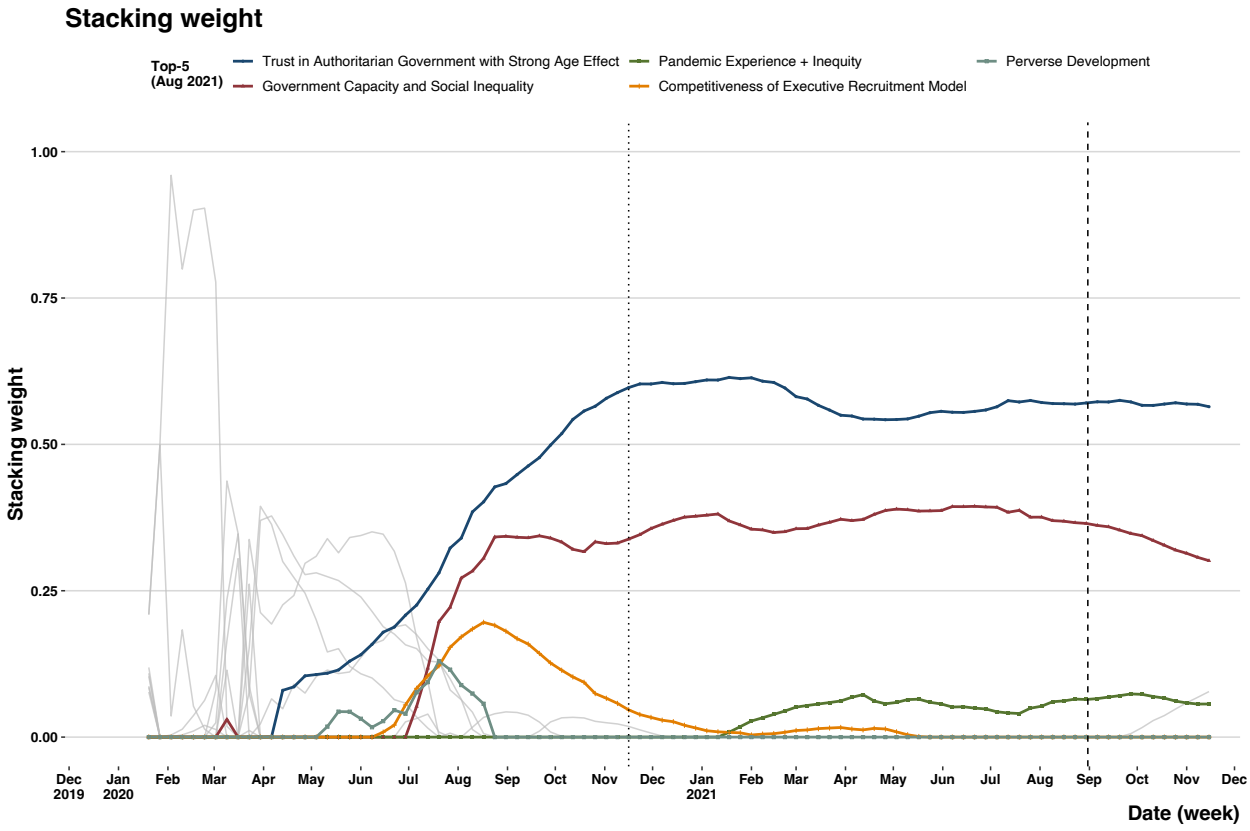
Figure S21: Summary of model predictions using two algorithmic methods for parameterized models from all challenges.

	# of models		Figure	Range (Top-5)			
Challenge	# Algorithmic	# Elicited		Horserace	Stacking	Horserace exp.	Stacking exp.
Crossnational, general	7	9	4	[0.34, 0.478]	[0.015, 0.524]	[0.273, 0.513]	[0.157, 0.347]
Crossnational, parameterized	8	-	S21	[-0.132, 0.127]	[0.076, 0.416]	*Not applicable*	
India, general	7	6	S18	[0.176, 0.334]	[0.085, 0.386]	[0.06, 0.831]	[0.125, 0.292]
India, parameterized	6	-	S21	[-3.652, -1.88]	[0, 0.794]	*Not applicable*	
Mexico, general	7	7	S18	[-0.201, 0.254]	[0.1, 0.363]	[0.08, 0.439]	[0.104, 0.394]
Mexico, parameterized	6	-	S21	[-8.569, 0.512]	[0, 0.985]	*Not applicable*	
USA, general	7	8	S20	[0.401, 0.527]	[0.069, 0.349]	[0.273, 0.493]	[0.161, 0.311]
USA, parameterized	6	-	S21	[-6.928, -0.176]	[0.00, 0.964]	*Not applicable*	

Table S16: Summary of model selection results. Note that 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.



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



(b) Evolution of stacking weights over time.

Figure S22: Evolution of measures of model performance over time.

S6 Aggregating: Supplementary results

In this section, we report the results from the aggregation analysis of the other challenges to complement Figure 5. We first summarize the results of aggregating across the four general and four parameterized challenges in Table S17. Figures S23-S29 report the full results for each of the challenges.

	# of models in		Figure	Estimate					
Challenge	Algorithmic metrics	Elicited metrics		Best single	Median single	Stacking	Expert favored	Rep. expert	Wisdom of crowds
Crossnational, general	28	26	5	0.483	0.155	0.536	0.417	0.340	0.361
Crossnational, parameterized	16	-	S23	0.142	-0.404	0.490	*Not applicable*		
India, general	9	9	S24	0.437	0.234	0.592	0.033	0.419	0.466
India, parameterized	7	-	S25	-1.839	-3.228	-1.812	*Not applicable*		
Mexico, general	9	9	S26	0.414	-0.024	0.571	0.208	0.343	0.358
Mexico, parameterized	6	-	S27	0.512	-7.824	0.516	*Not applicable*		
USA, general	19	18	S28	0.582	0.309	0.655	0.513	0.524	0.537
USA, parameterized	7	-	S29	-0.176	-6.343	-0.160	*Not applicable*		

Table S17: Summary of model aggregation metrics across challenges. Note that because we did not elicit expert forecasts over the parameterized models, we do not include those metrics for the parameterized models.

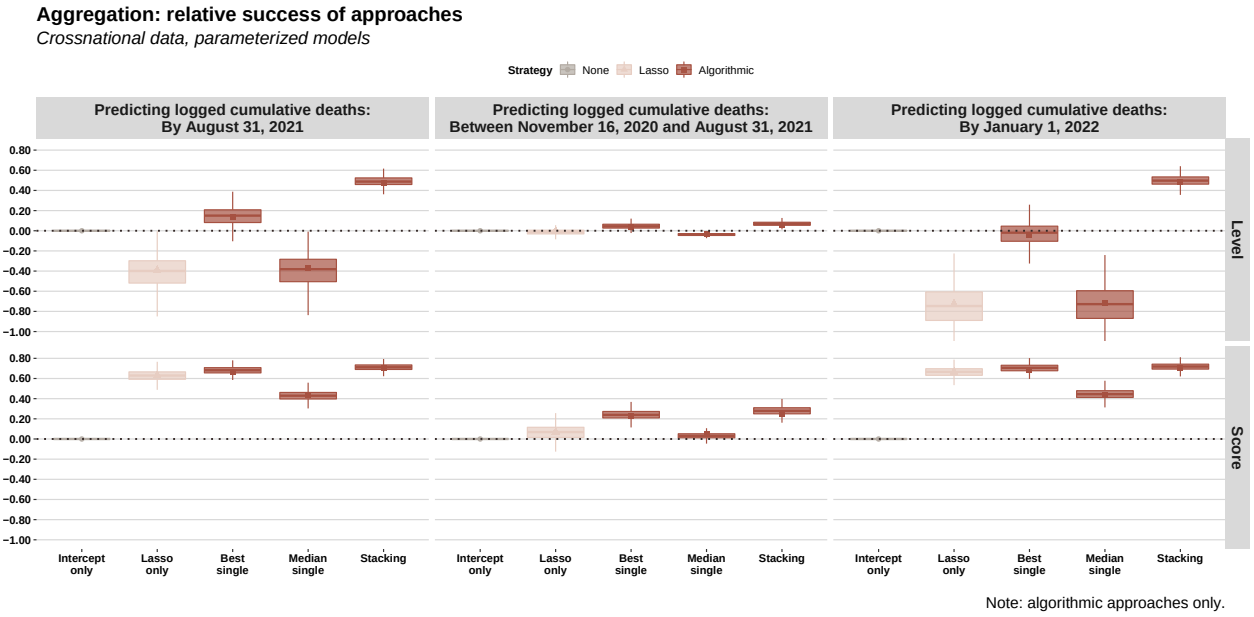


Figure S23: Prediction aggregation metrics for crossnational parameterized models. Note that because we did not elicit forecasts over the parameterized models, we do not include the 'expert favored', 'representative expert', or 'wisdom of the crowds' metrics.

For the "expert-favored" strategy in aggregating, 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 where the idea of combining models through stacking has been prevalent. Here we provide results using an alternative selection metric, namely the forecasters' horserace predictions (probability of each model of being the best-performing model).

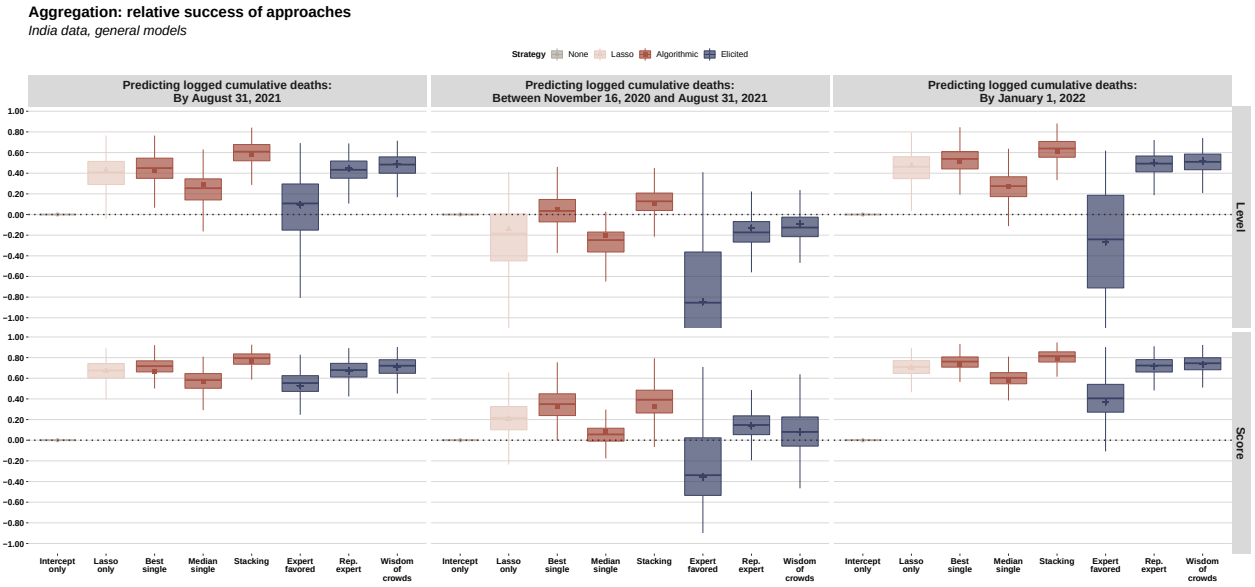


Figure S24: Prediction aggregation metrics for general models for India.

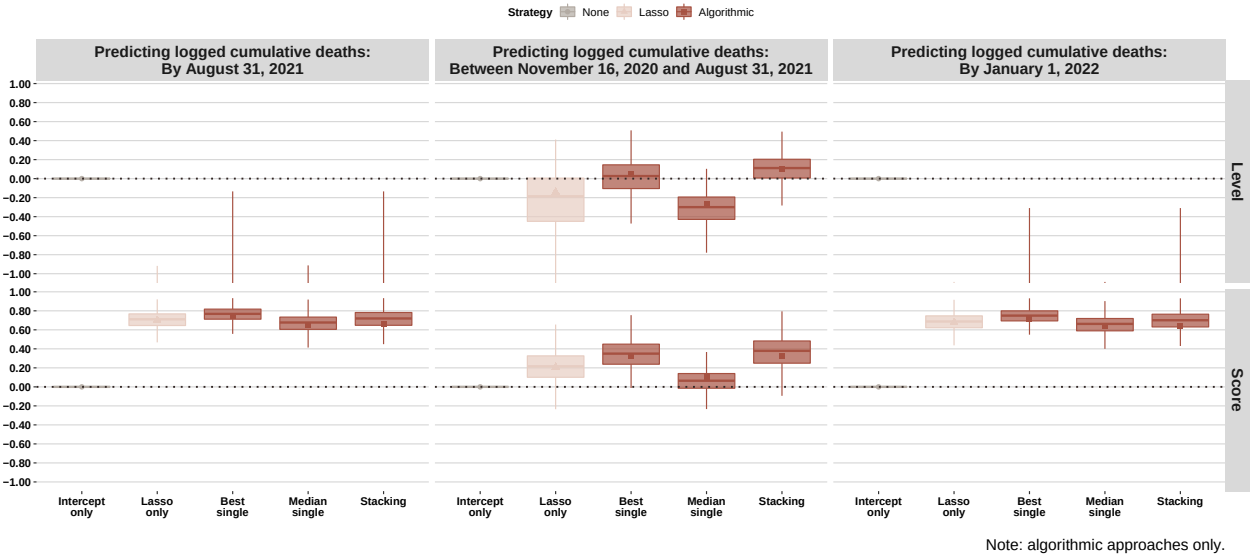


Figure S25: Prediction aggregation metrics for parameterized models for India. Note that because we did not elicit forecasts over the parameterized models, we do not include the 'expert favored', 'representative expert', or 'wisdom of the crowds' metrics. Note that for the level approach, some metrics fall outside the $[-1, 1]$ domain and are therefore omitted from the plot. This generally occured due to an inaccurate prediction of the intercept.

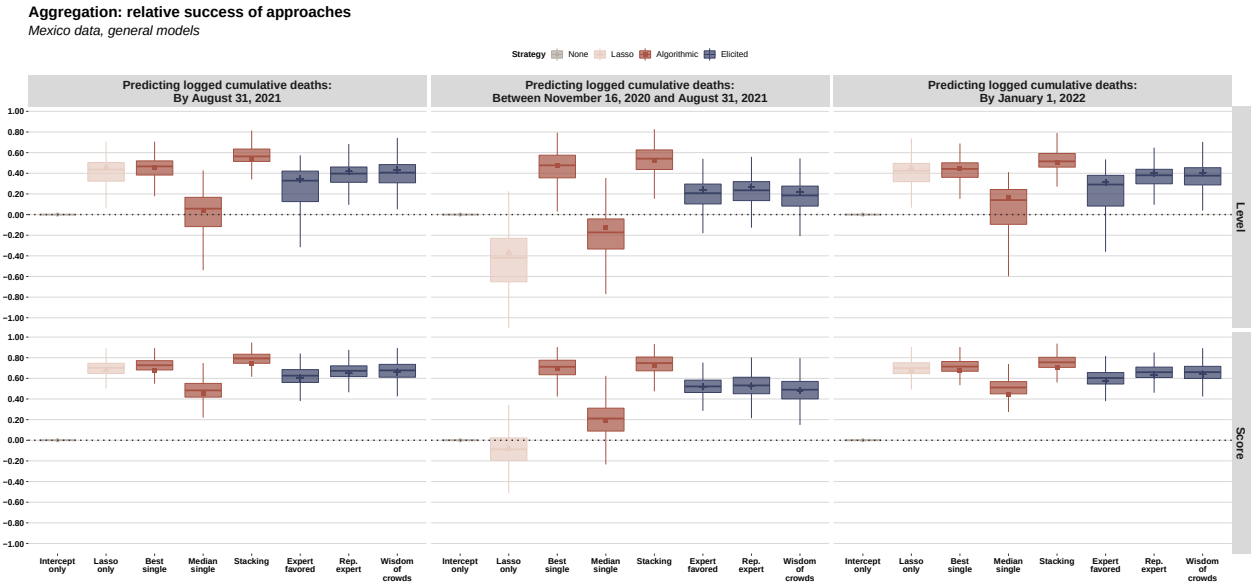


Figure S26: Prediction aggregation metrics for general models for Mexico.

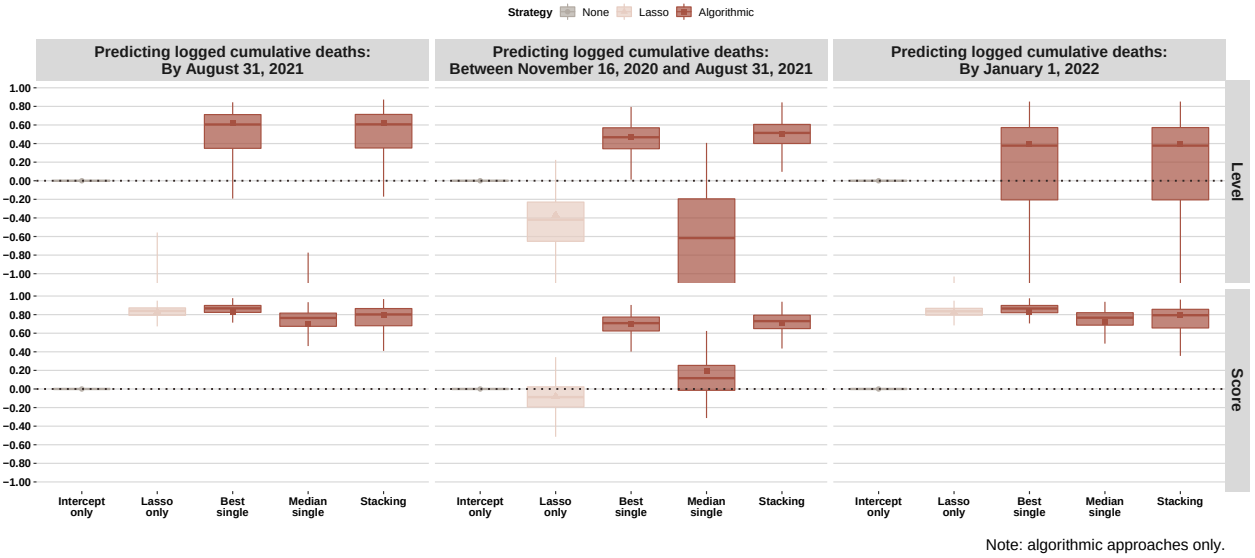


Figure S27: Prediction aggregation metrics for parameterized models for Mexico. Note that because we did not elicit forecasts over the parameterized models, we do not include the 'expert favored', 'representative expert', or 'wisdom of the crowds' metrics. Note that for the level approach, some metrics fall outside the $[-1, 1]$ domain and are therefore omitted from the plot. This generally occured due to an inaccurate prediction of the intercept.

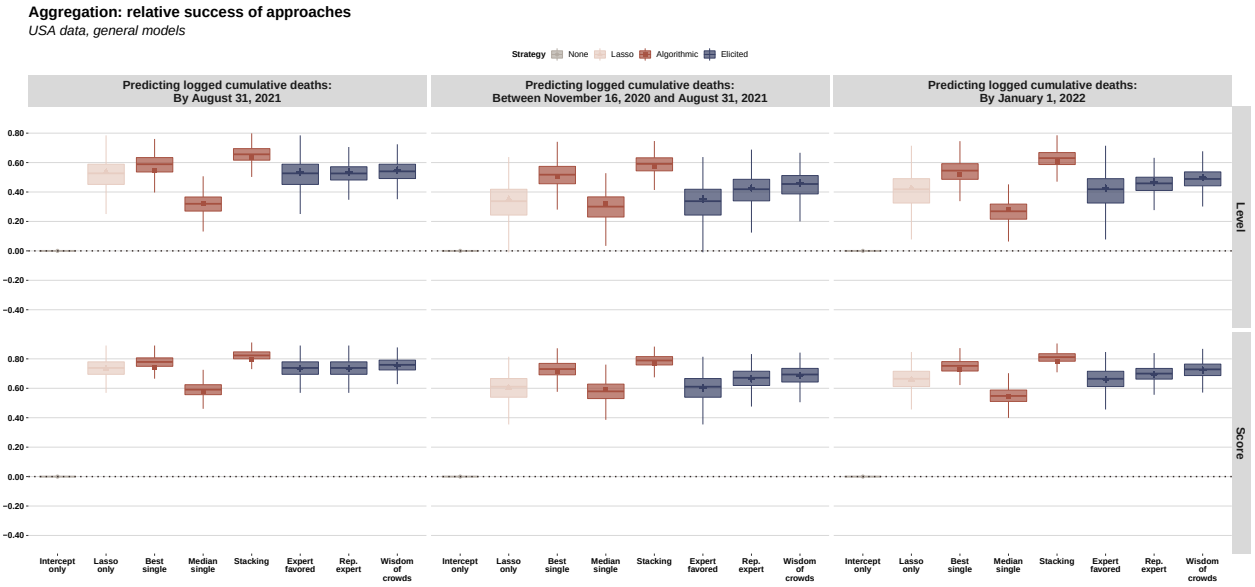


Figure S28: Prediction aggregation metrics for general models for the US.

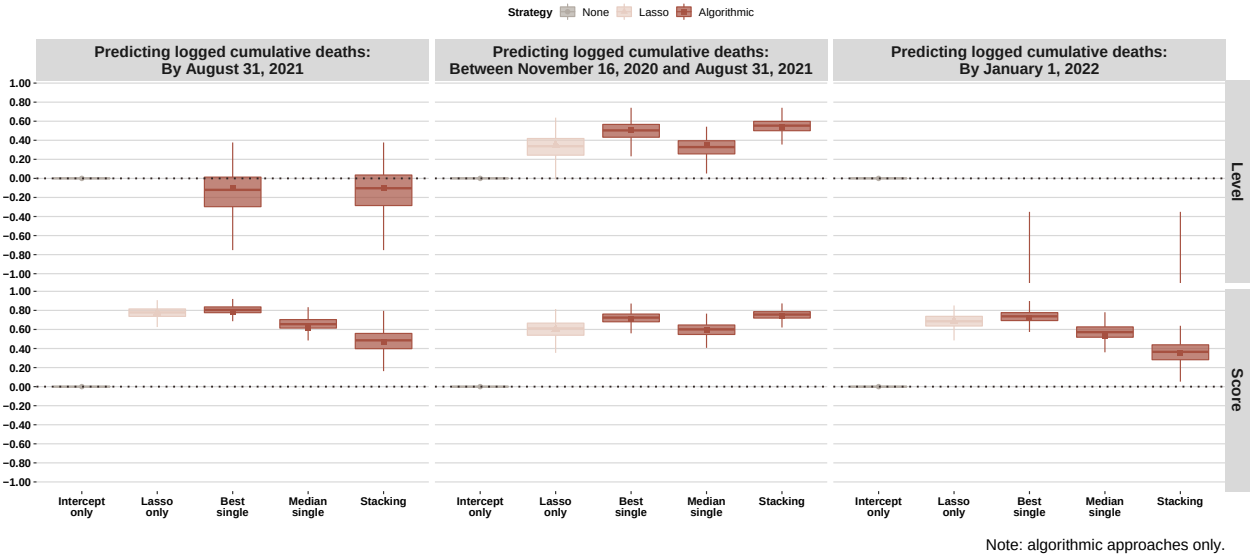


Figure S29: Prediction aggregation metrics for parameterized models for the US. Note that because we did not elicit forecasts over the parameterized models, we do not include the 'expert favored', 'representative expert', or 'wisdom of the crowds' metrics. Note that for the level approach, some metrics fall outside the $[-1, 1]$ domain and are therefore omitted from the plot. This generally occured due to an inaccurate prediction of the intercept.

Forecast	Model	Weight
Crossnational		
Horserace	Trust, Development, and State	0.513
Stacking	Government Capacity and Social Inequality	0.347
India		
Horserace	Minority Representation and Urbanization	0.831
Stacking	Minority Representation and Urbanization	0.292
Mexico		
Horserace	Investment Inequality	0.439
Stacking	Trust Poverty and TB	0.394
USA		
Horserace	Health	0.493
Stacking	Lasso	0.311

Table S18: Expert-favored models using two different forecast weights.

The most preferred models selected under the horserace and the stacking predictions are shown in Table S18 alongside their average weights, for each model challenge. With the exception of India, experts selected different models in each forecast. Figure S30 shows the best models' performances in the actual aggregating analysis using each forecasting metric. Since the elicited stacking model consistently outperforms the horserace best model, we consider the former as an upper bound for aggregating performance using the expert-favored strategy.

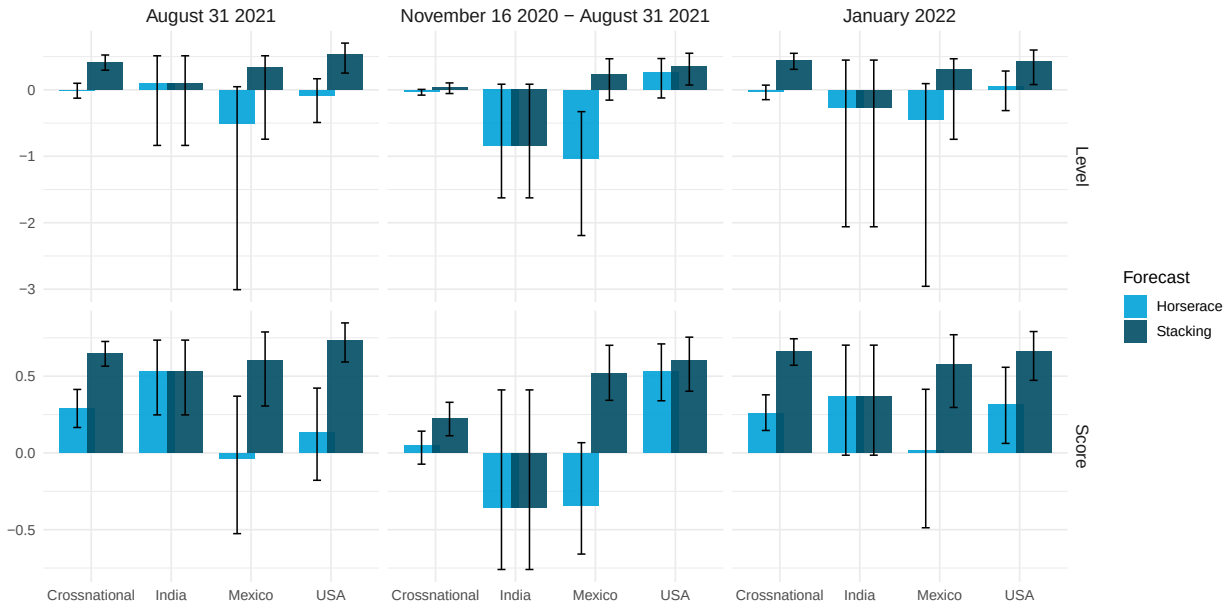


Figure S30: Comparison of performances for the "expert-favored" models selected using two types of forecast weights.

S7 Simulation Results

In order to ascertain the probability that randomly-generated models would perform as well as the model challenge submissions, we conduct two simulations, described below.

In the *model-level* simulation, we construct all permutations of three-variable models from the Models Challenge datasets that use linear functional forms. We calculate the pseudo-R^2 of each model. The distribution of pseudo-R^2 statistics for this simulation forms null distribution. We plot the null distribution (as a density) for the each challenge in Figure S31 in blue. The overlaid histogram shows the observed distribution of models submitted in the Models Challenge. We see that in general, several models in each challenge offer excellent performance relative to the null distribution. However, the median model does not perform particularly well.

In the *challenge-level* simulation, we conduct a simulated stacking exercise. Suppose that there are J forecasts in one of the challenges (J varies across each of the challenges). The simulation proceeds by:

1. Randomly select J three-predictor models with a linear functional form from the universe of permutations in the model-level simulation.
2. Fit a stacking model on the J models sampled in step #1.
3. Calculate the pseudo-R^2 of the stacking model in step #2.
4. Repeat steps #1-3 n times to generate a null distribution of the performance of stacking models.
5. Compare the pseudo-R^2 of the stacking models of each general challenge to this simulated distribution.

Figure S32 shows the results of our challenge-level simulation. It shows that in 3 of 4 challenges, our estimated stacking models outperform every simulated model. In the Mexico challenge, 4% of the simulated models outperform our estimated stacking models ($p = 0.04$).

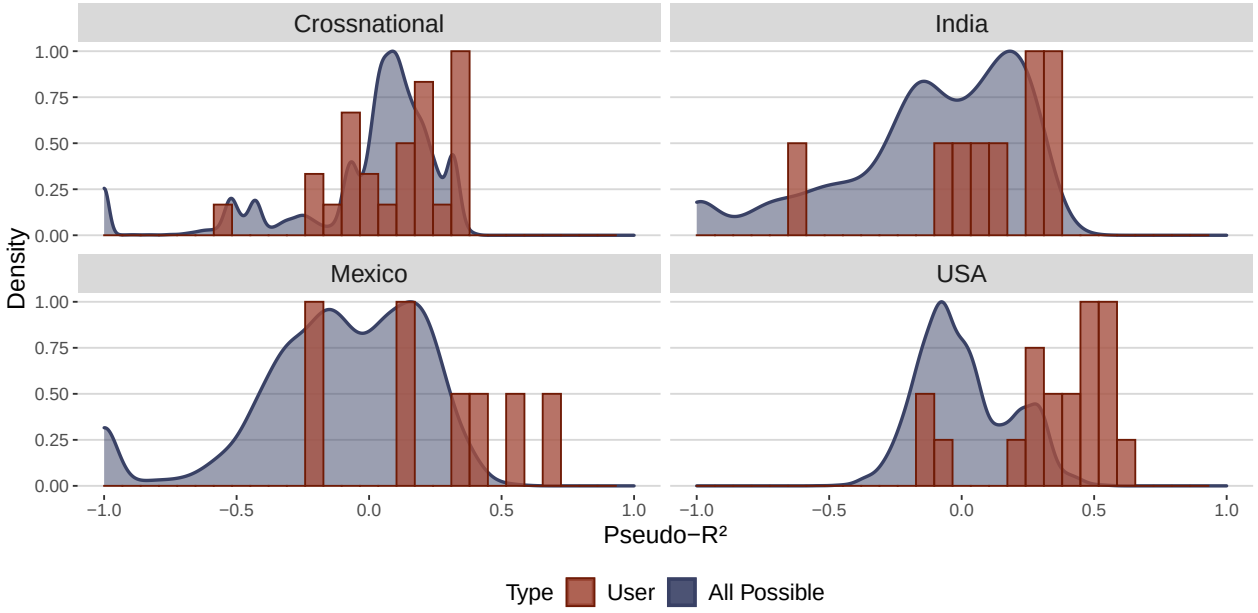


Figure S31: Horserace simulation.

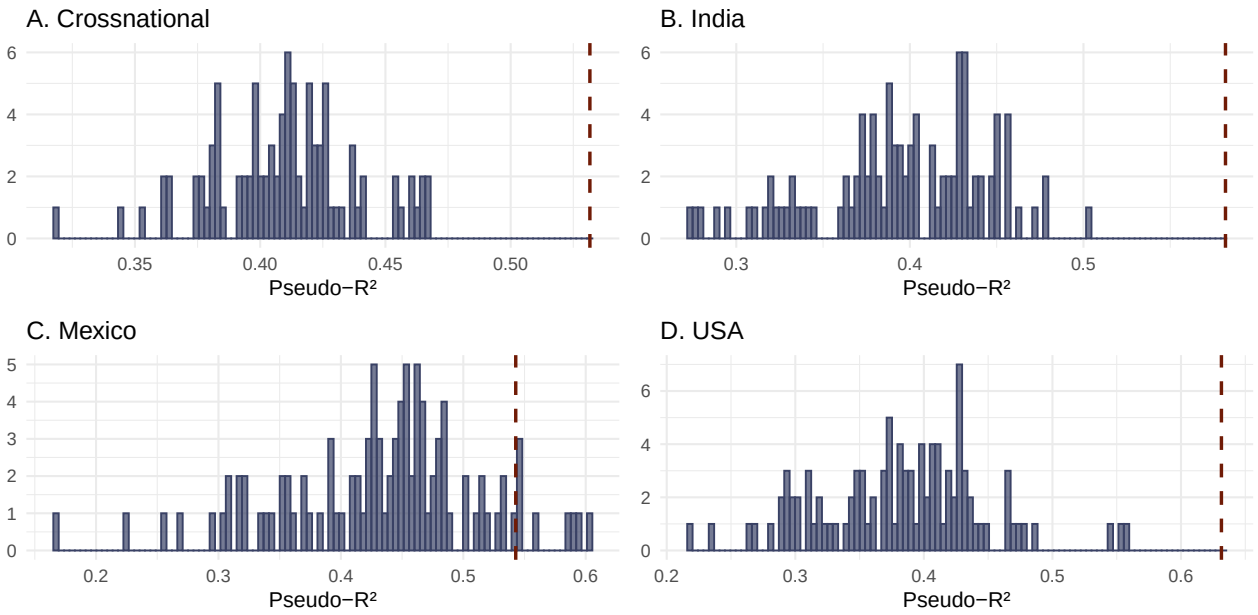


Figure S32: Stacking simulation.

S8 Imputation results

Due to the excessive level of listwise missingess in the crossnational data (see S14) we have adopted a multiple imputations procedure to deal with missingess 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 S33 to S56 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 that have been most affected by missingness. In particular, we notice that pre-prediction imputation has led to a reshuffling of mode performance rankings in the evaluating analyses, where previously worst-affected models now achieve usually higher pseudo-R^2 and stacking weights than before, compared to their less-affected competitors. In contrast, the aggregating analysis experiences a less significant improvement in terms of aggregate model performances.

Note that for 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 that challenge.

S8.1 Evaluating

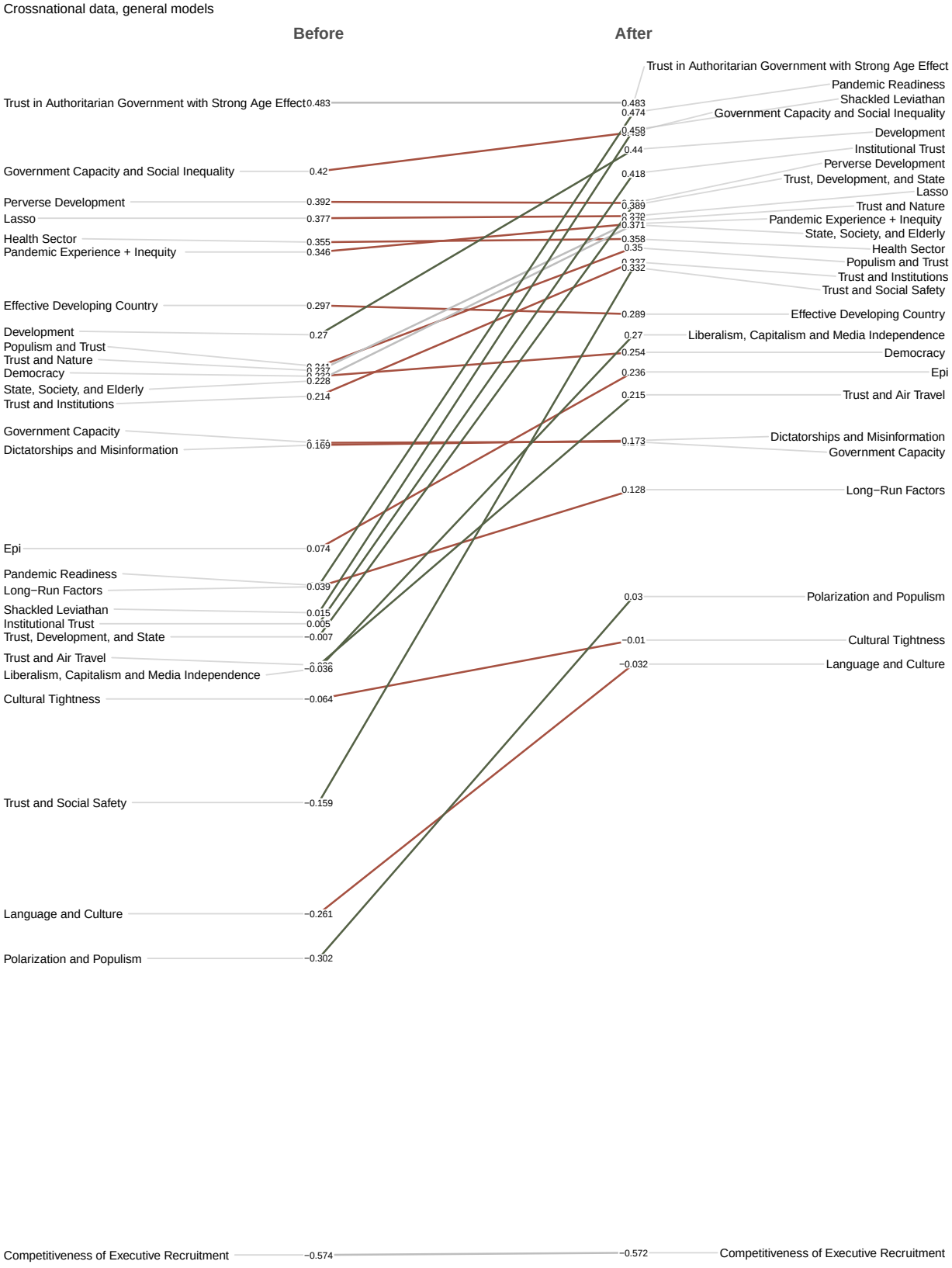


Figure S33: Pre- and post-imputation performances of the general models for crossnational data.

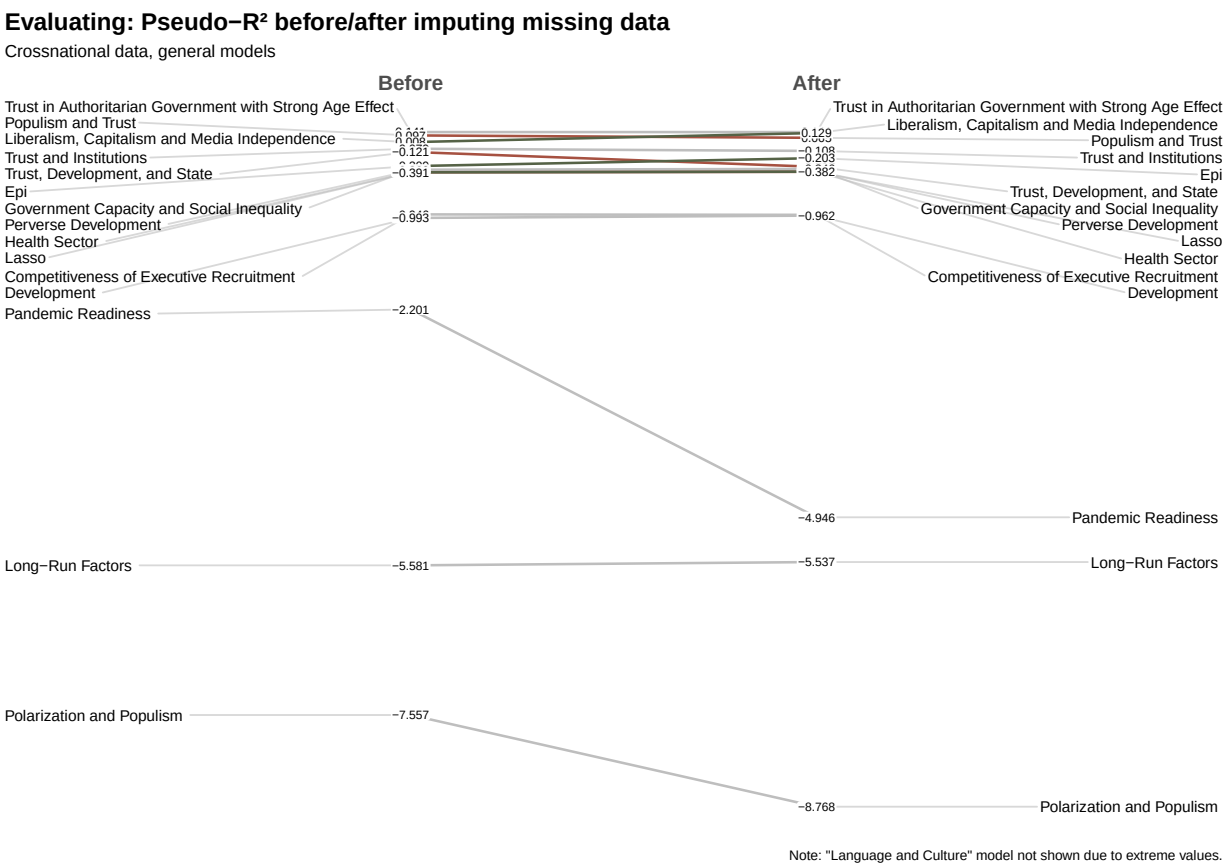


Figure S34: Pre- and post-imputation performances of the parameterized models for crossnational data.

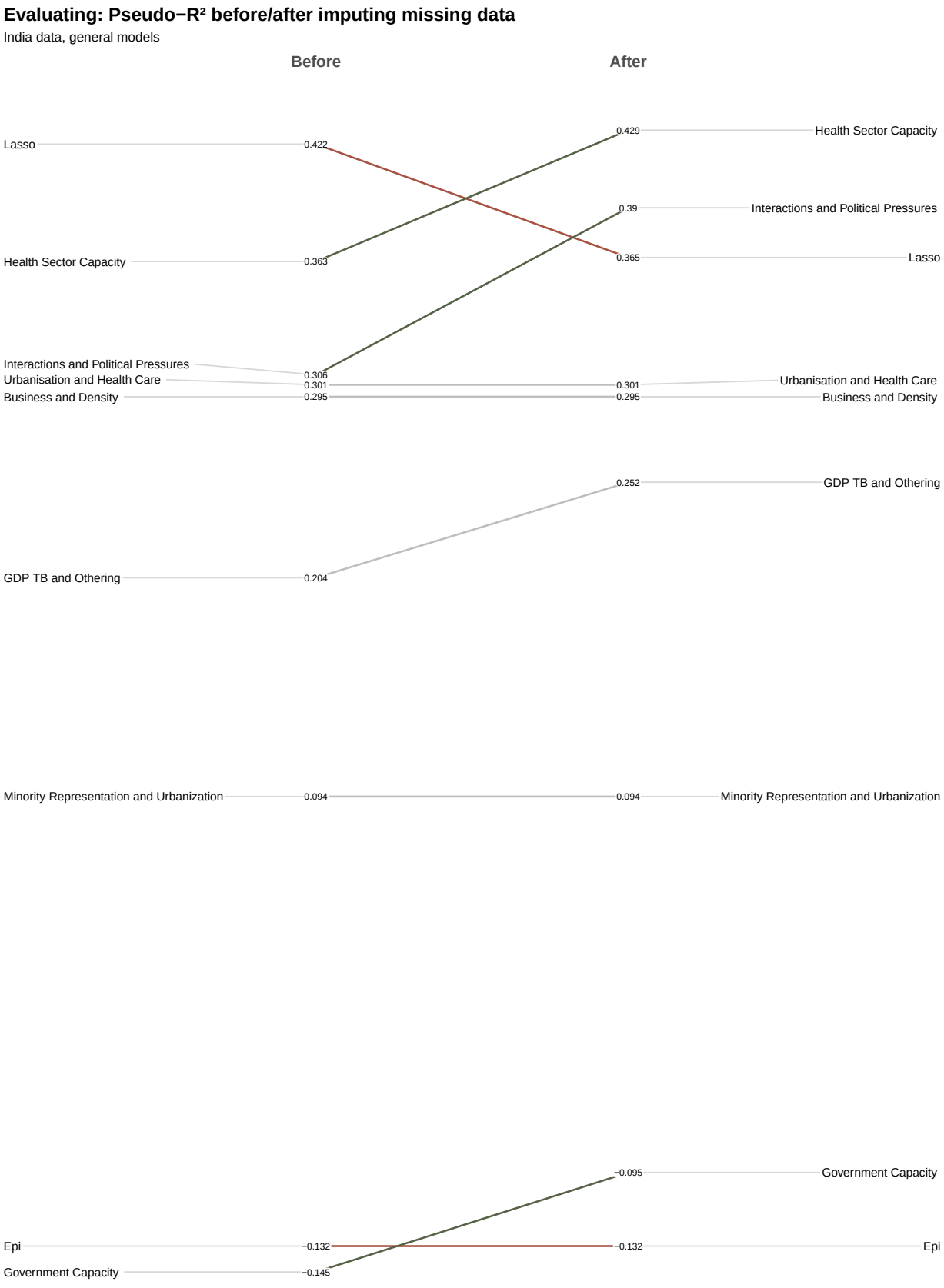


Figure S35: Pre- and post-imputation performances of the general models for India.

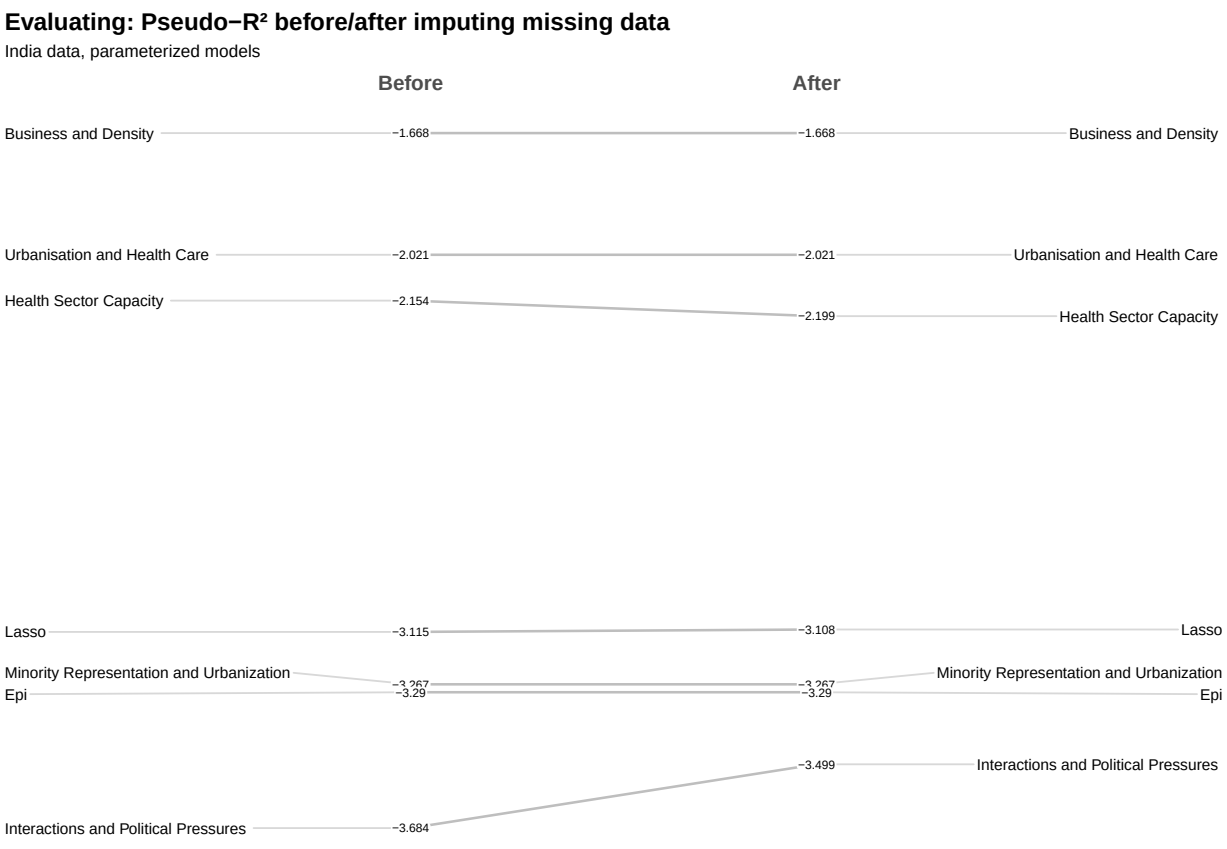


Figure S36: Pre- and post-imputation performances of the parameterized models for India.

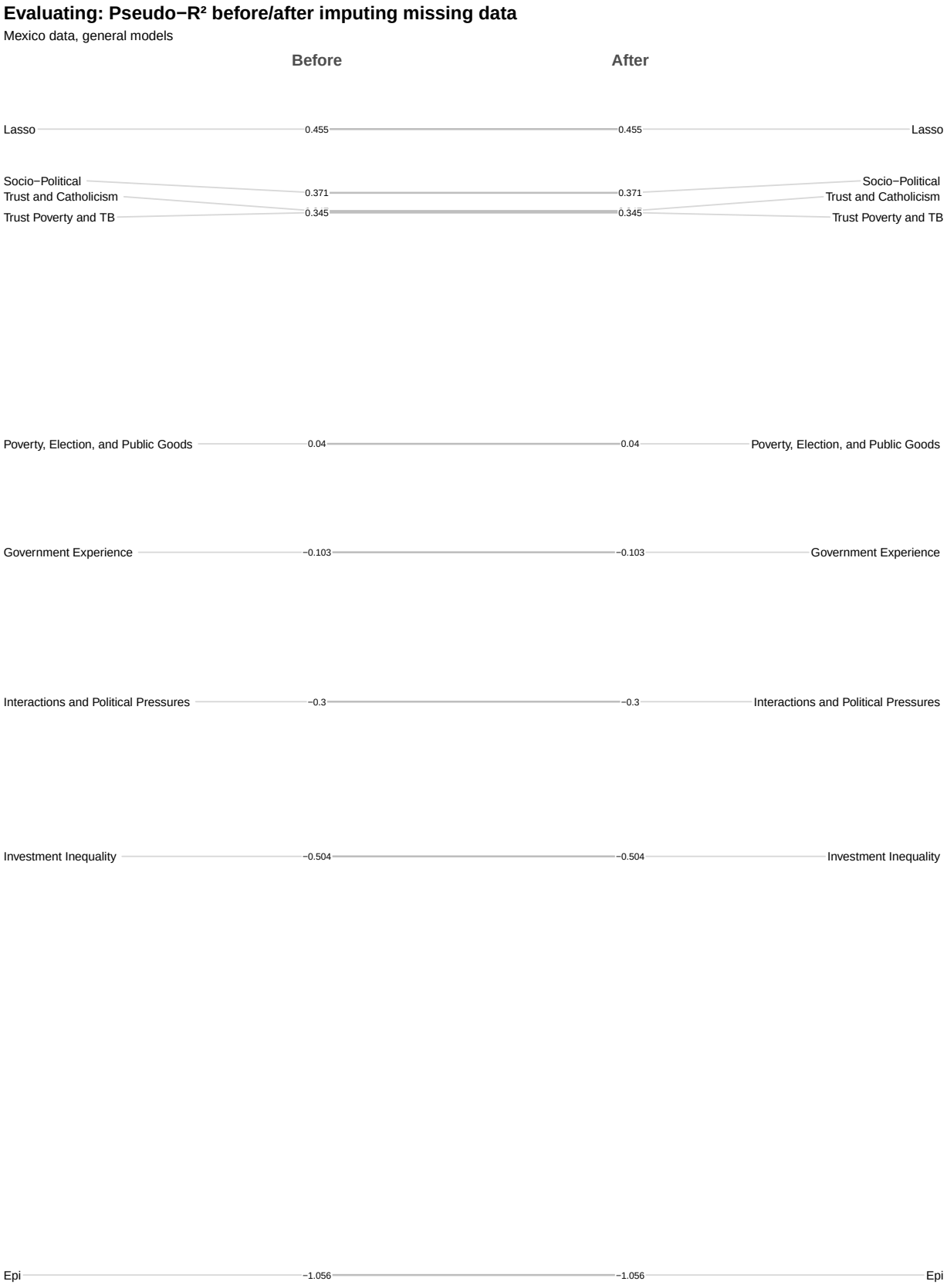


Figure S37: Pre- and post-imputation performances of the general models for Mexico.

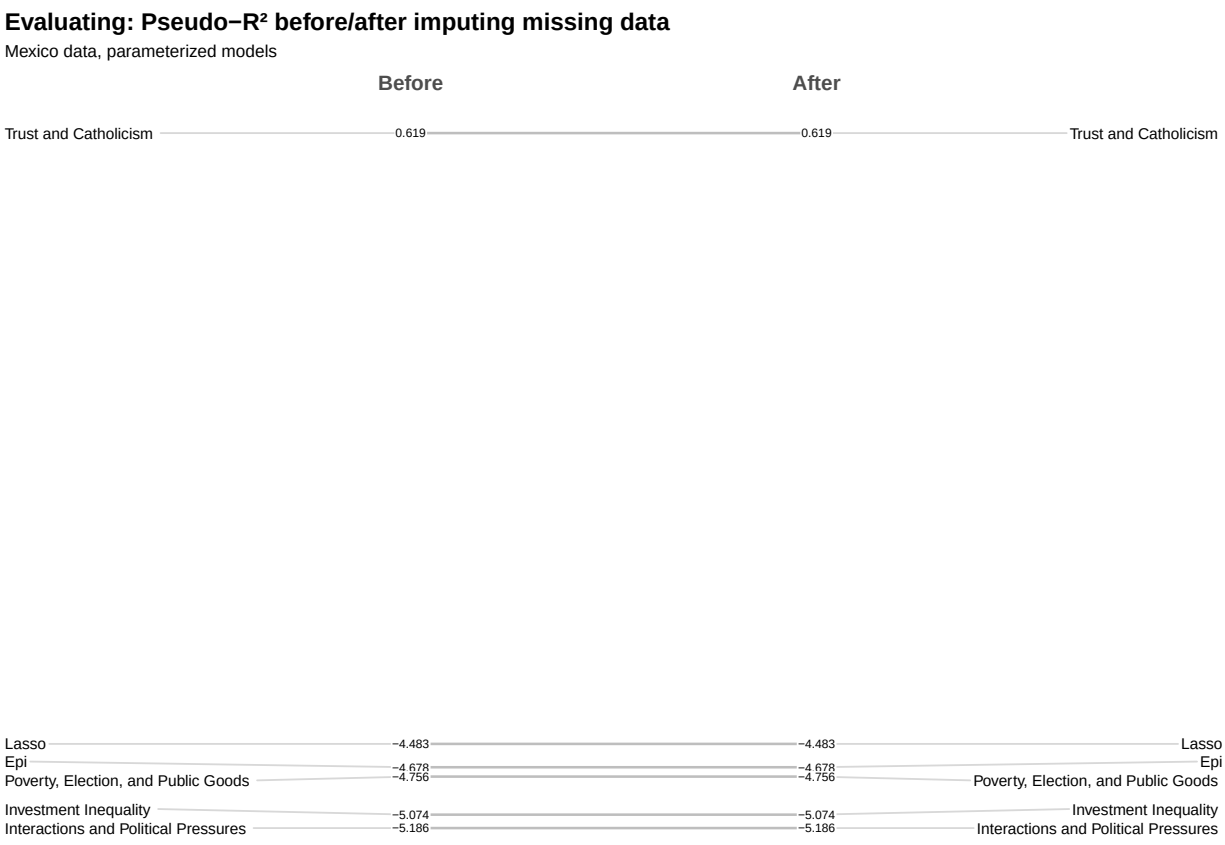


Figure S38: Pre- and post-imputation performances of the parameterized models for Mexico.

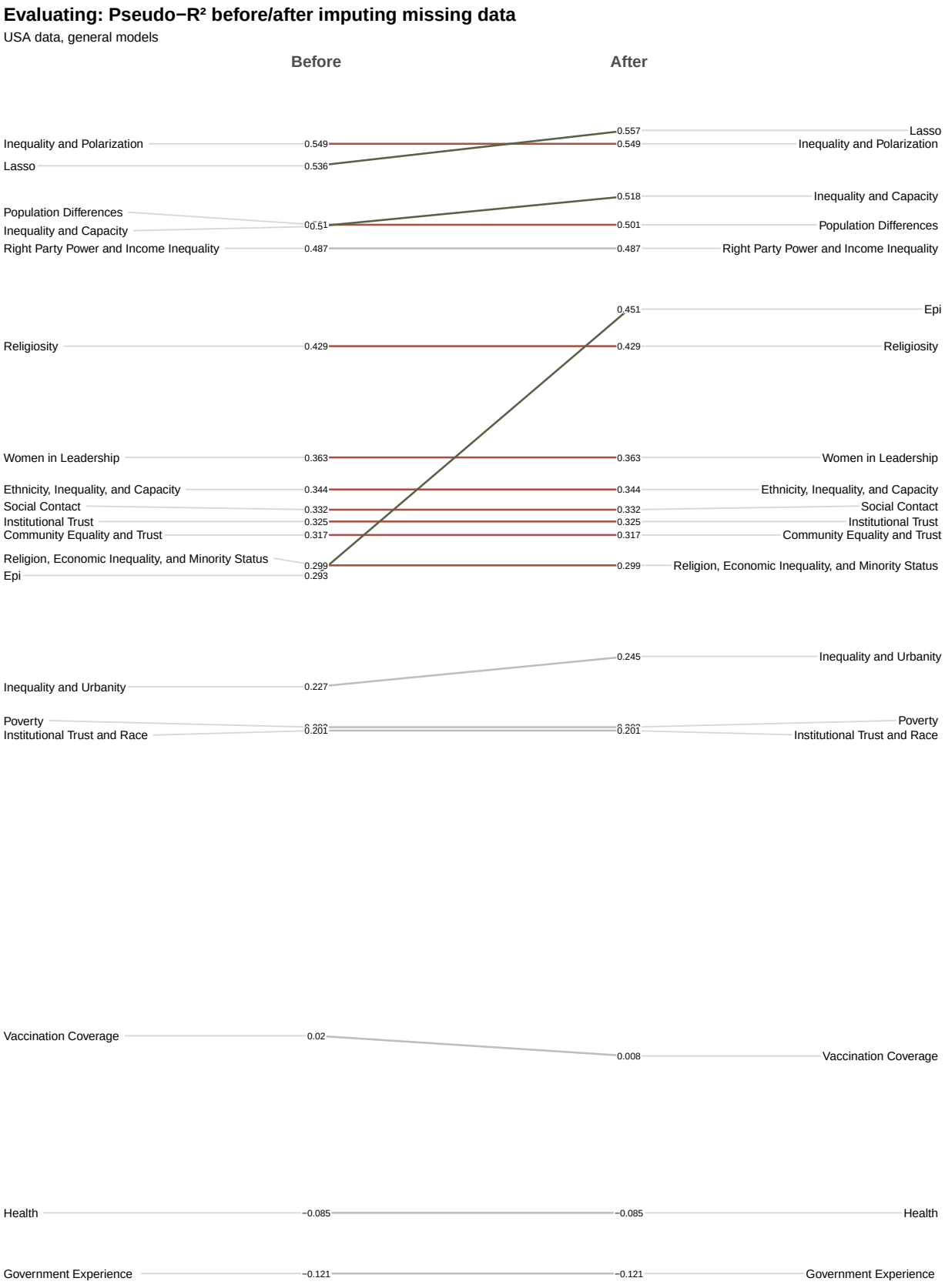


Figure S39: Pre- and post-imputation performances of the general models for the US.

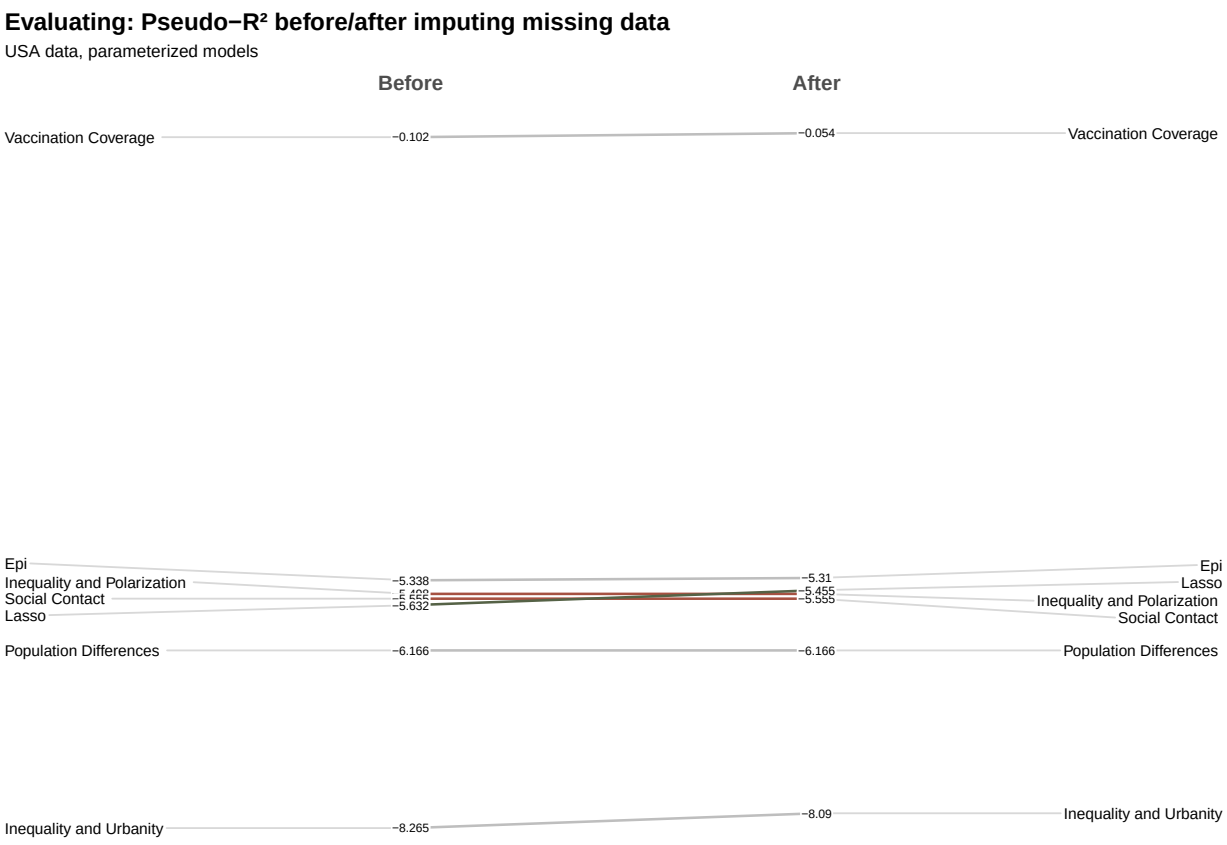


Figure S40: Pre- and post-imputation performances of the parameterized models for the US.

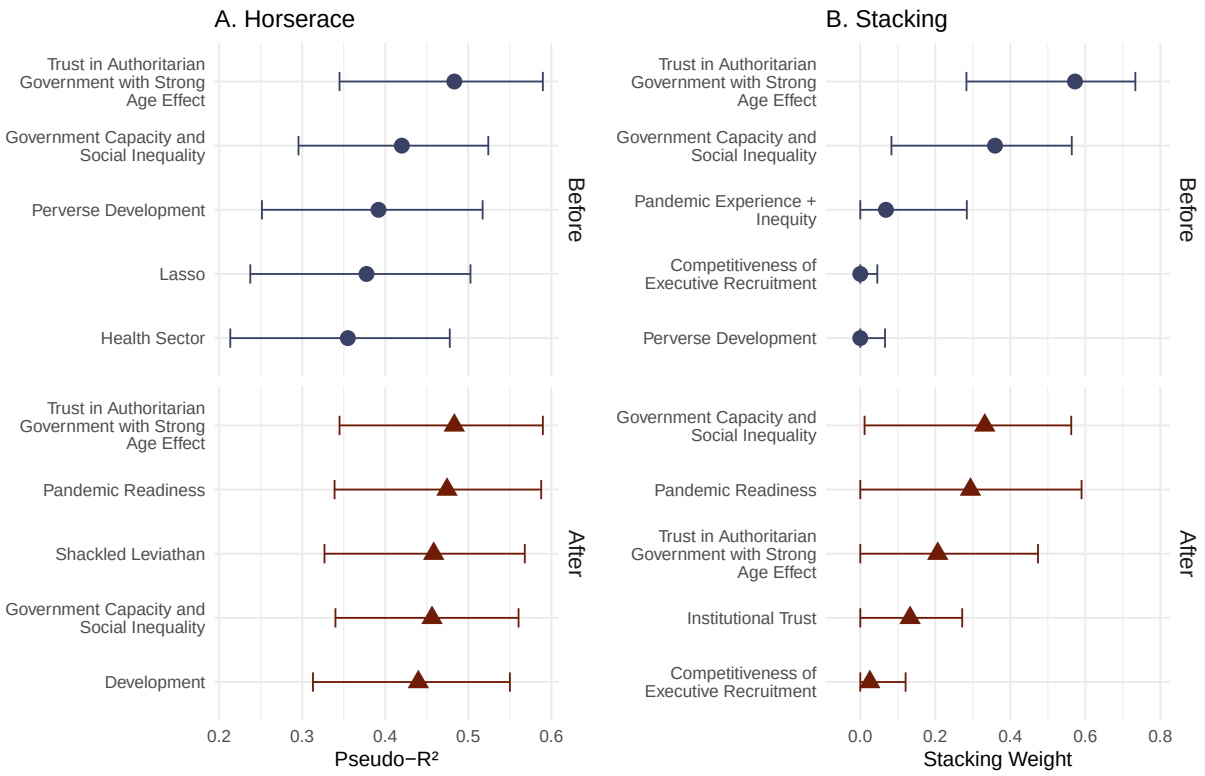


Figure S41: Model selection before and after imputation for crossnational general models.

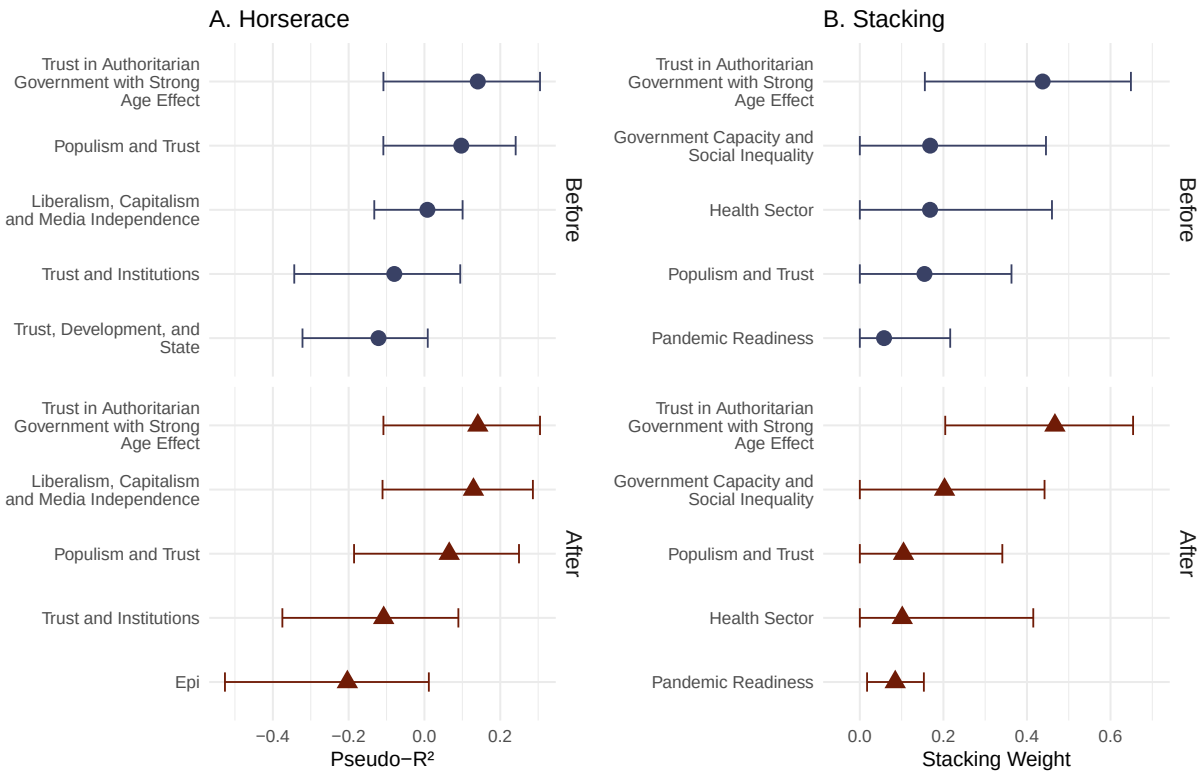


Figure S42: Model selection before and after imputation for crossnational parameterized models.

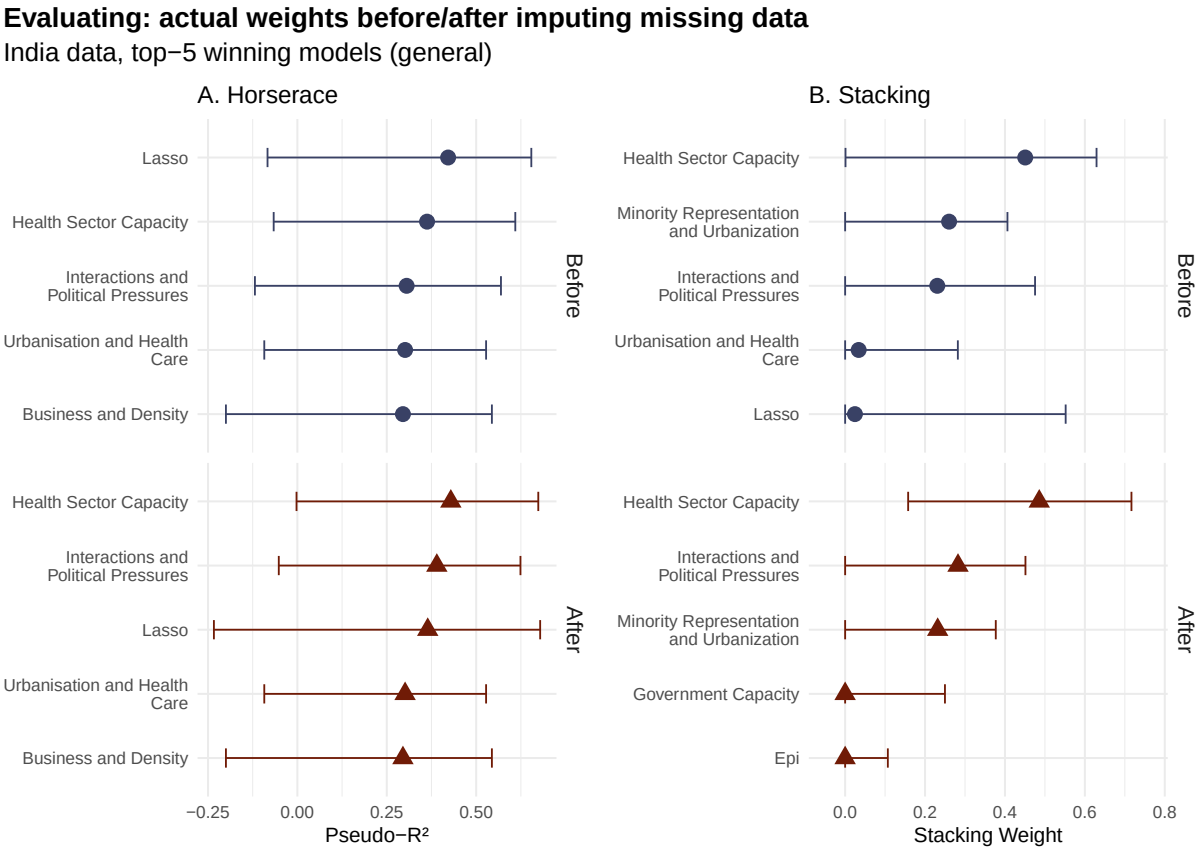


Figure S43: Model selection before and after imputation for the general models for India.

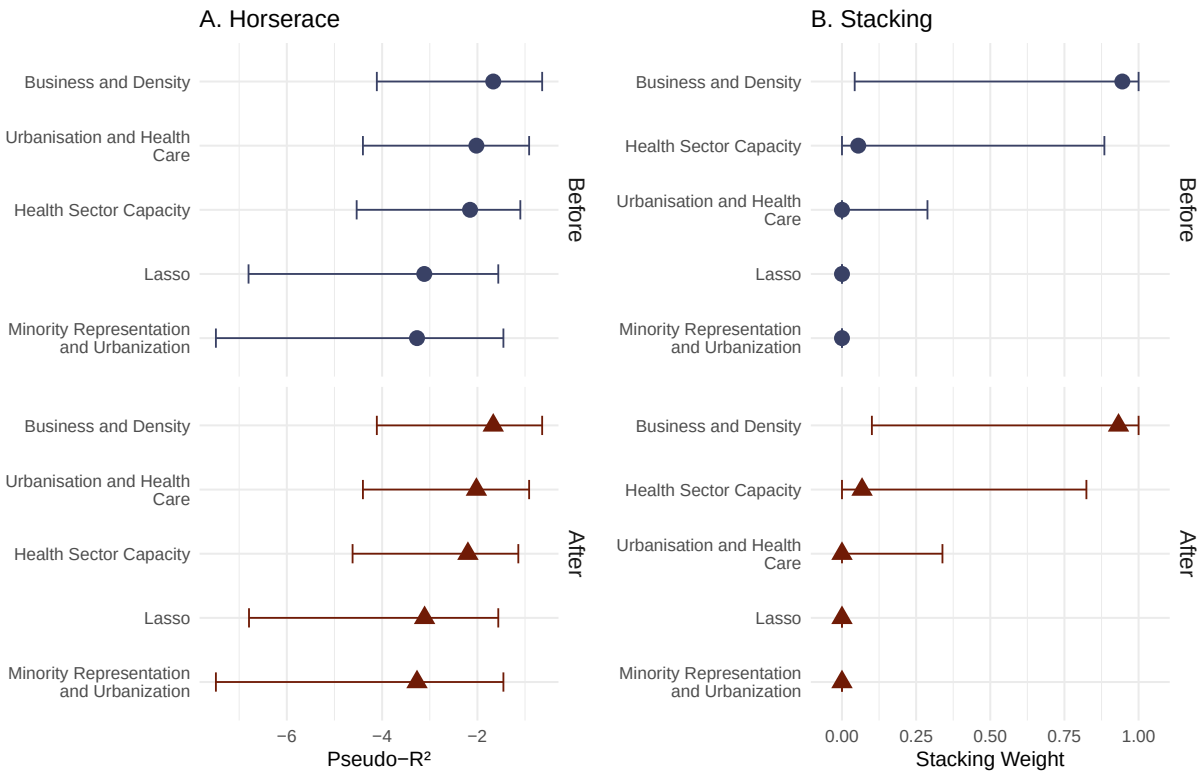


Figure S44: Model selection before and after imputation for the parameterized models for India.

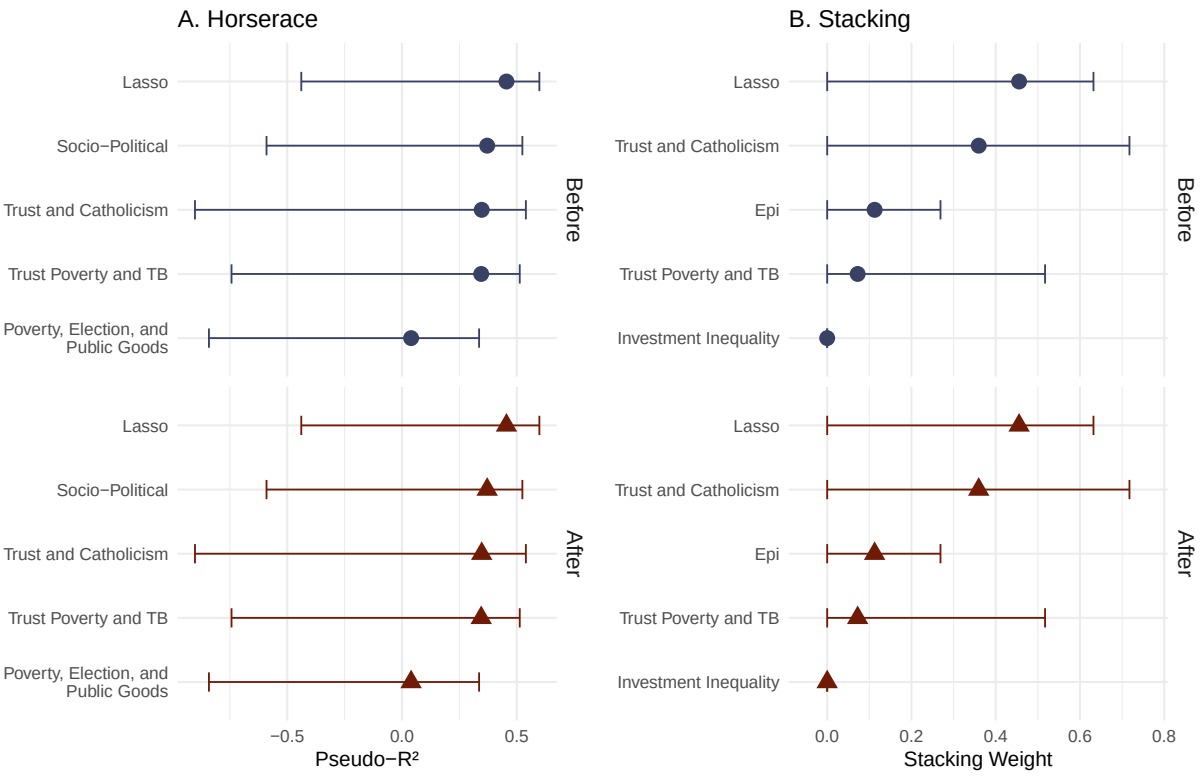


Figure S45: Model selection before and after imputation for the general models for Mexico.

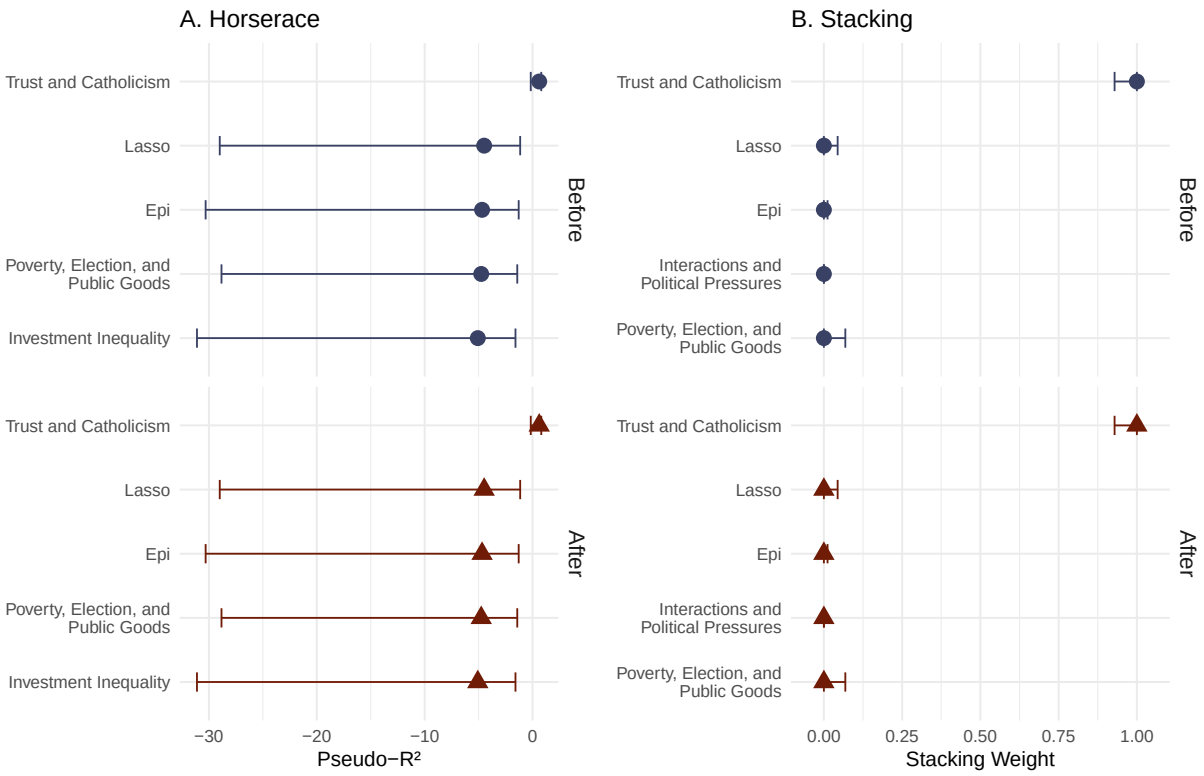


Figure S46: Model selection before and after imputation for the parameterized models for Mexico.

Figure S47: Model selection before and after imputation for the general models for the US.

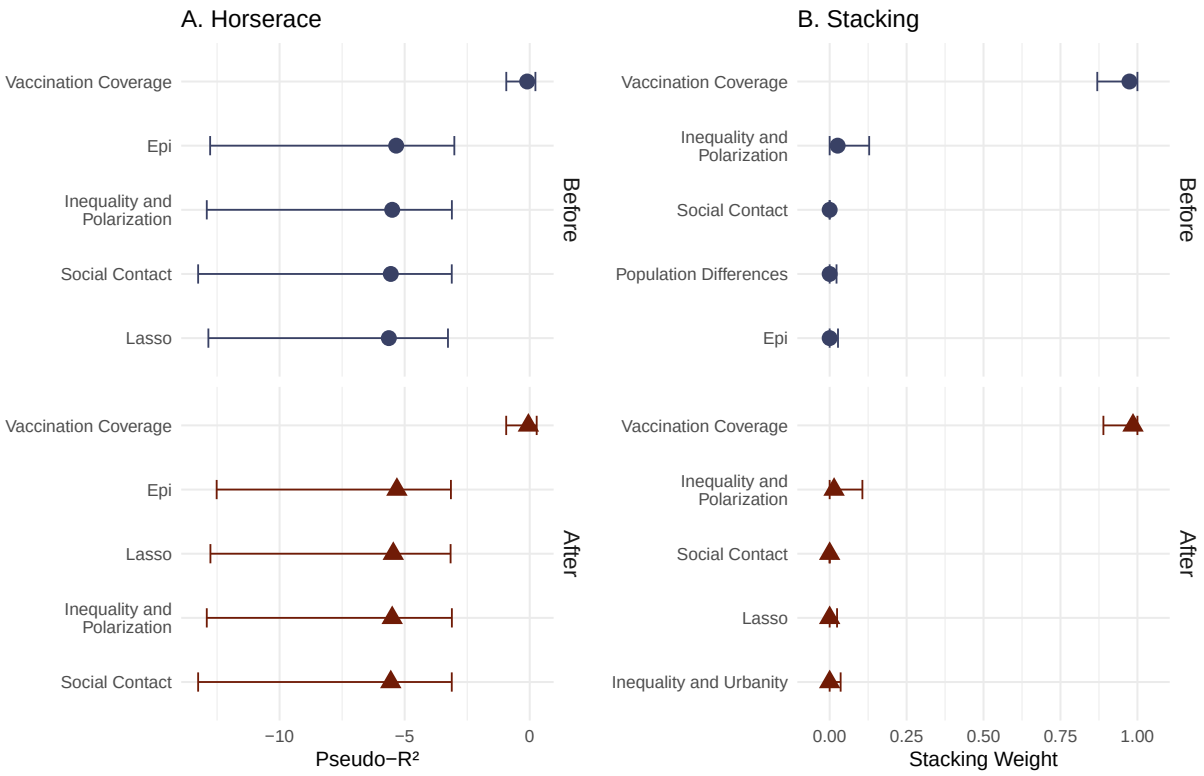


Figure S48: Model selection before and after imputation for the parameterized models for the US.

S8.2 Aggregating

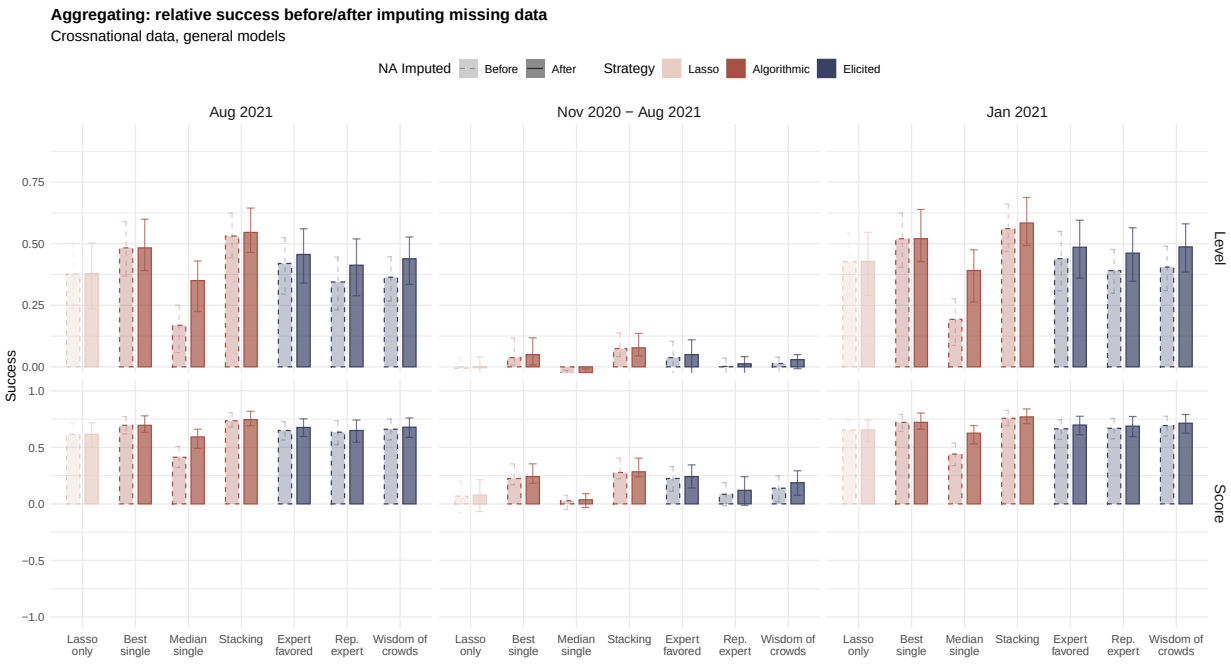


Figure S49: Prediction aggregation metrics for crossnational general models before and after imputation.

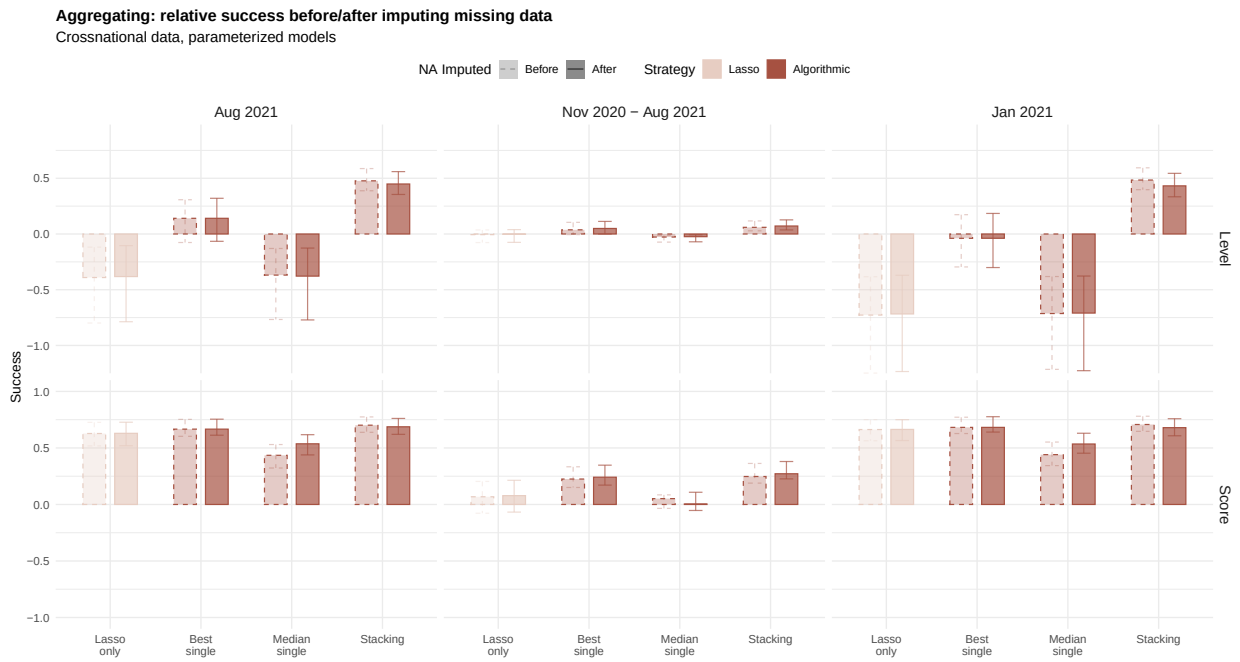


Figure S50: Prediction aggregation metrics for crossnational parameterized models before and after imputation. Note that because we did not elicit forecasts over the parameterized models, we do not include the 'expert favored', 'representative expert', or 'wisdom of the crowds' metrics.

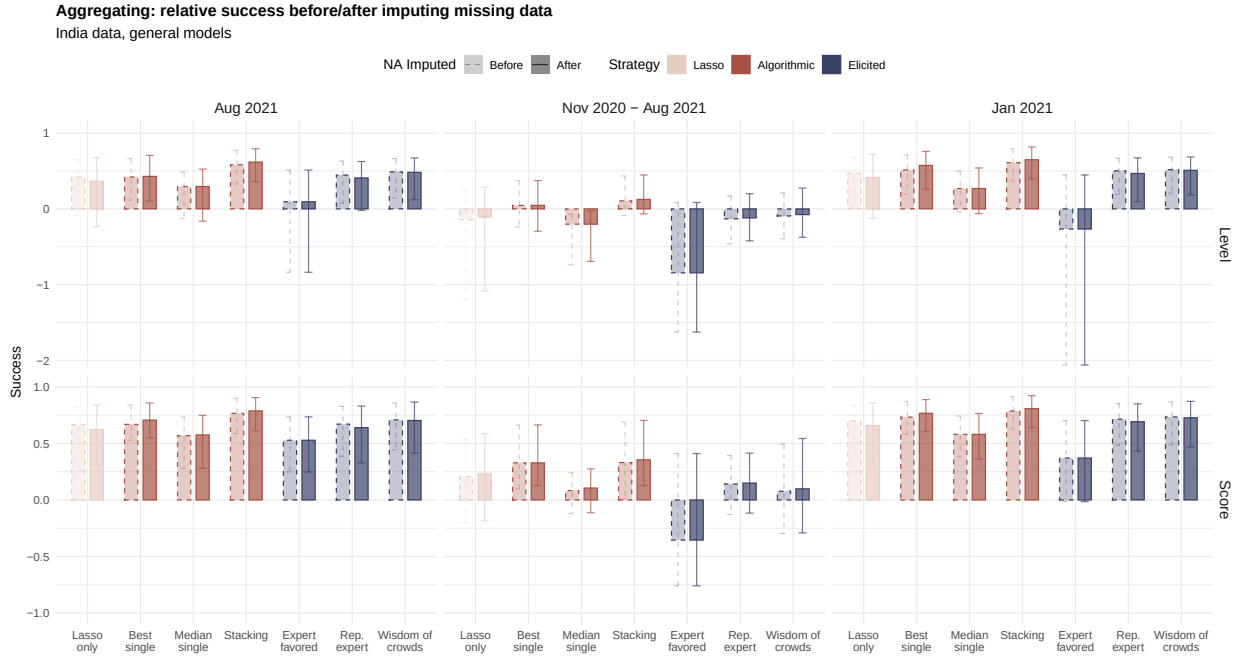


Figure S51: Prediction aggregation metrics for general models before and after imputation for India.

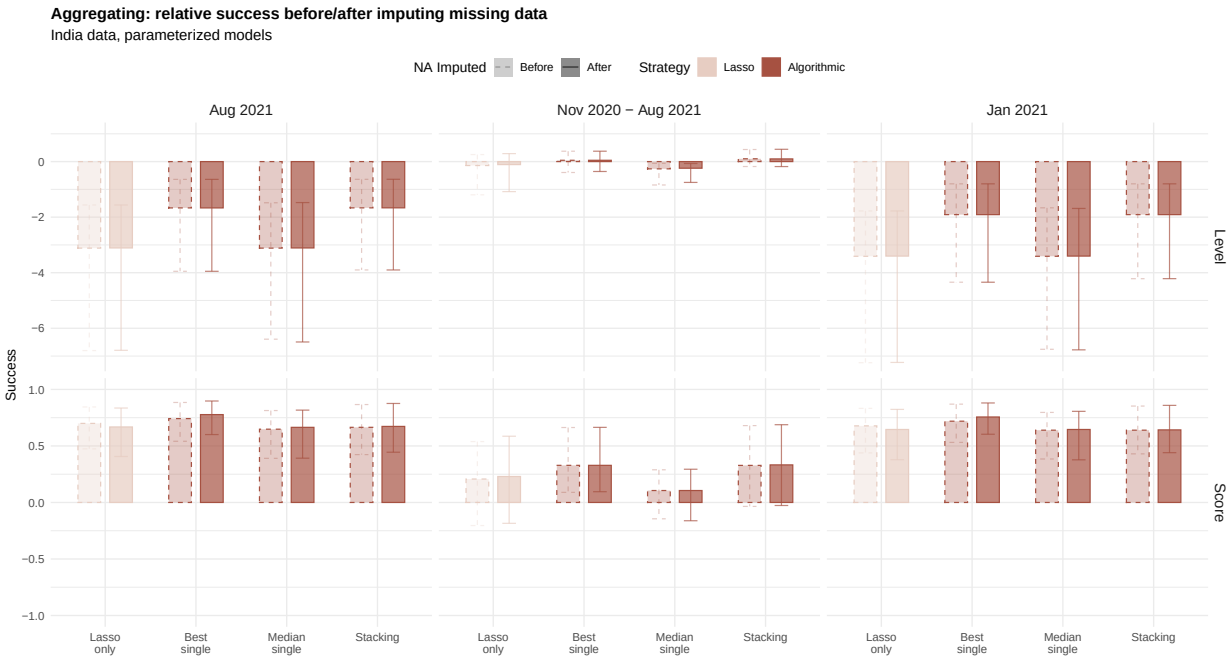


Figure S52: Prediction aggregation metrics for parameterized models before and after imputation for India. Note that because we did not elicit forecasts over the parameterized models, we do not include the 'expert favored', 'representative expert', or 'wisdom of the crowds' metrics.

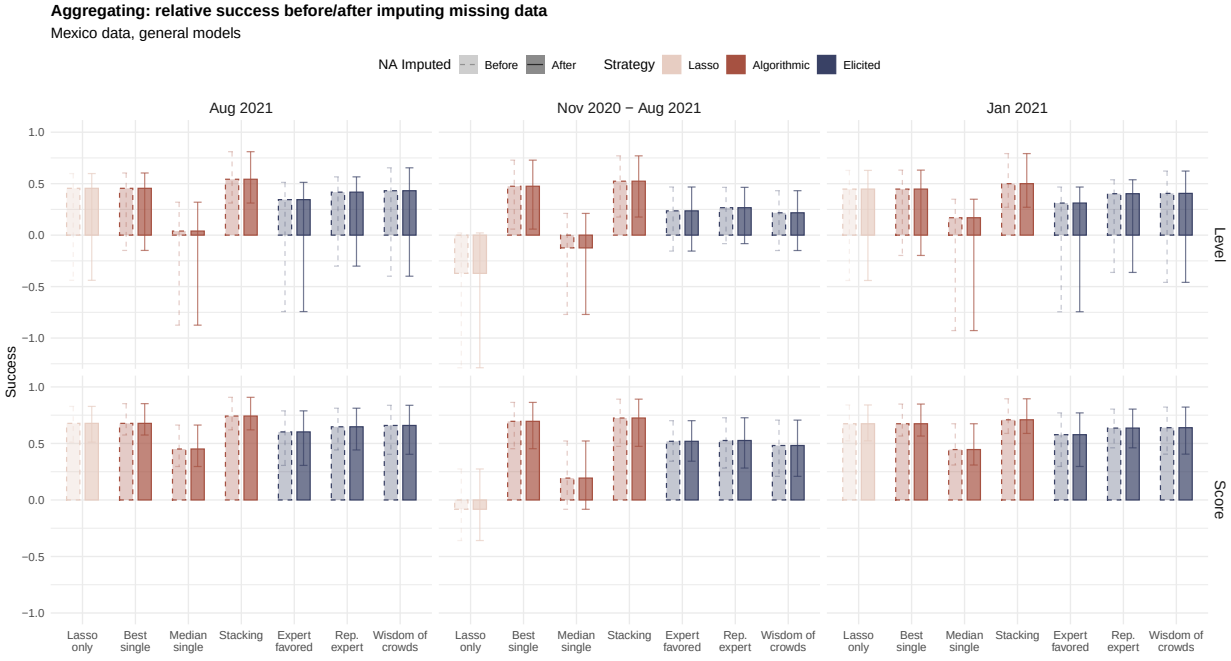


Figure S53: Prediction aggregation metrics for general models before and after imputation for Mexico.

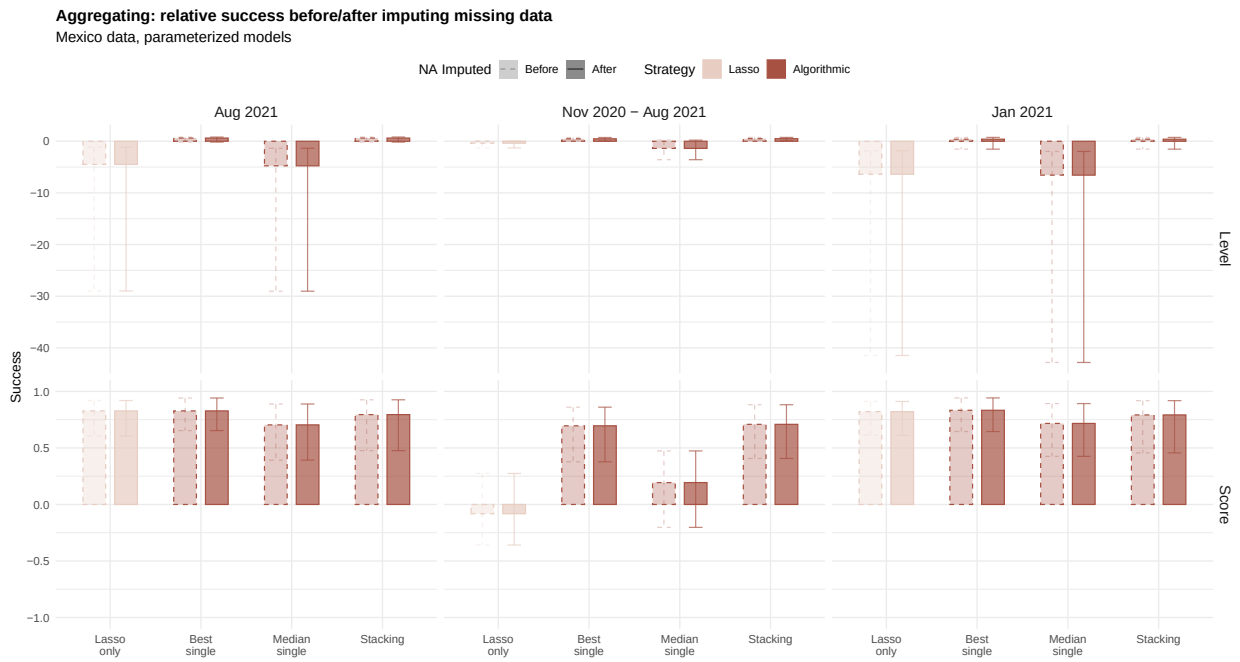


Figure S54: Prediction aggregation metrics for parameterized models before and after imputation for Mexico. Note that because we did not elicit forecasts over the parameterized models, we do not include the 'expert favored', 'representative expert', or 'wisdom of the crowds' metrics.

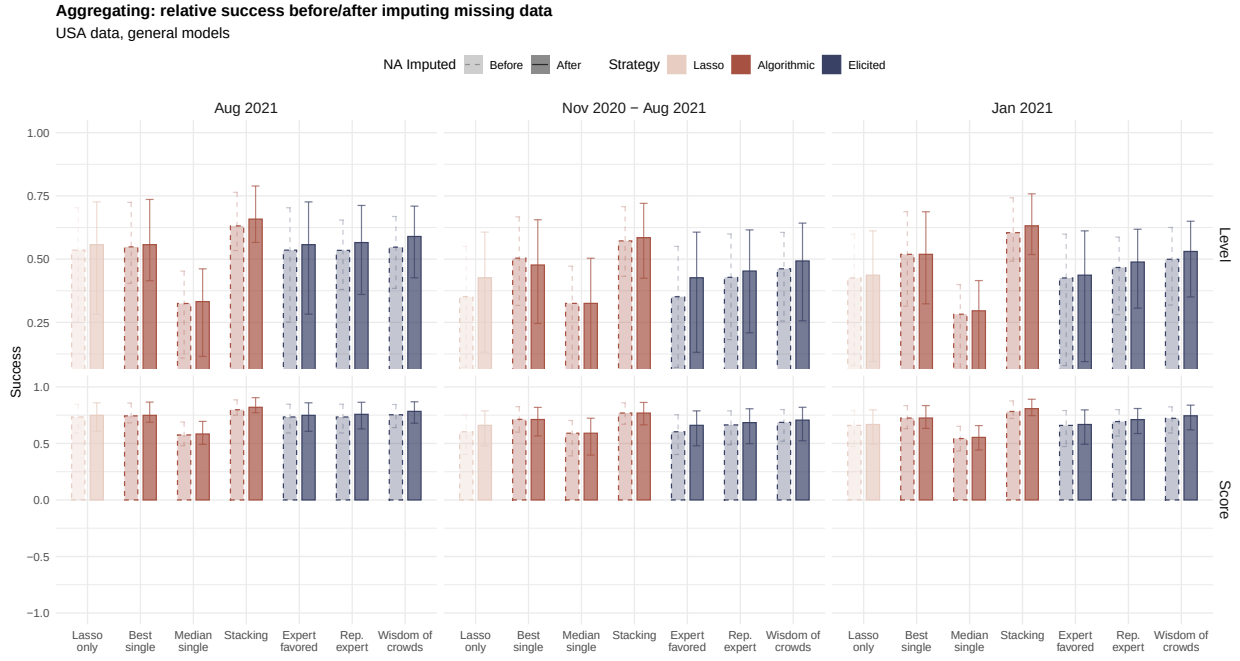


Figure S55: Prediction aggregation metrics for general models before and after imputation for the US.

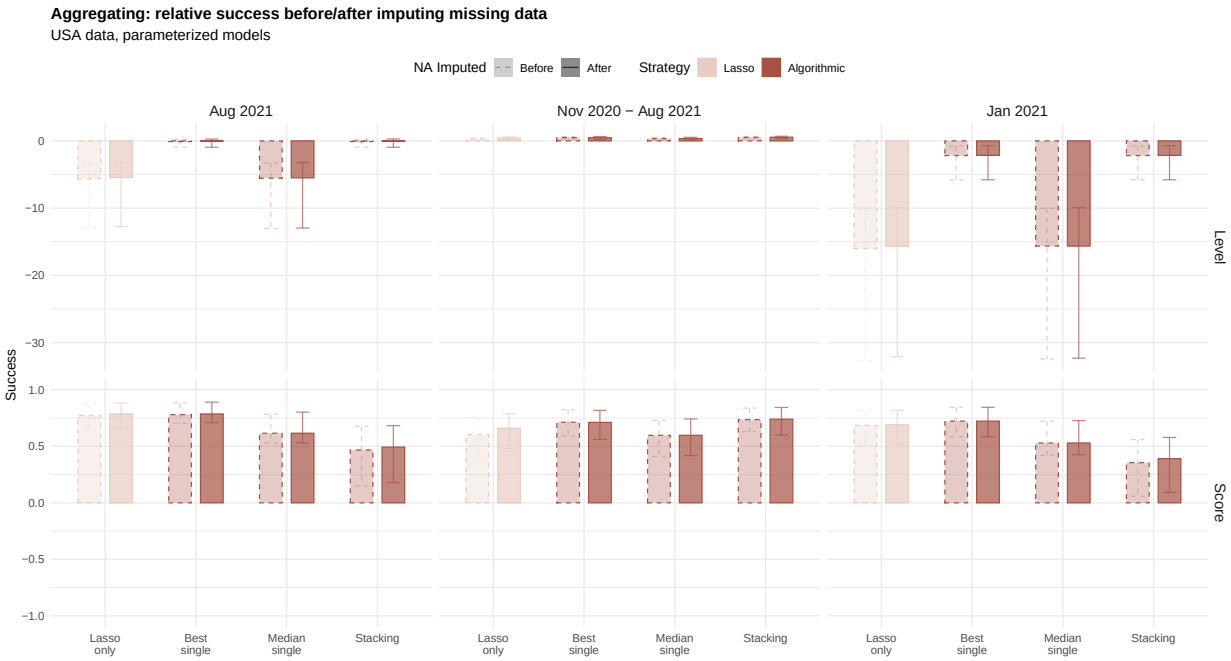


Figure S56: Prediction aggregation metrics for parameterized models before and after imputation for the US. Note that because we did not elicit forecasts over the parameterized models, we do not include the 'expert favored', 'representative expert', or 'wisdom of the crowds' metrics.

S9 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 (9).

2. The pre-analysis plan suggests three steps: gathering, selecting, and aggregating. We have re-conceptualized the gathering stage in this paper to focus on the *content* of the submitted models, not their predictive performance. The 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 when pursuing listwise deletion of observations with missing covariates (see Table S14), 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 88 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 5 uses predictions from the stacking forecasts only. We include Figure S30 to show that this indicates that our approach (using stacking forecasts), if anything, overstates expert abilities to aggregate.