The Model Challenges: Gathering, Aggregating, and Evaluating Covid Mortality Models

Miriam A. Golden[1,6,*], Tara Slough[2], Haoyu Zhai[3], Alexandra Scacco[4], Macartan Humphreys[4], Eva Vivalt[5], Alberto Diaz-Cayeros[6], Kim Yi Dionne[7], Sampada KC[8], Eugenia Nazrullaeva[9], P. M. Aronow[10], Jan-Tino Brethouwer[11], Anne Buijsrogge[11], John Burnett[7], Stephanie L. DeMora[12], José Ramón Enríquez[6], Robbert Fokkink[11], Chengyu Fu[13], Nicholas Haas[14], Sarah Virginia Hayes[15], Hanno Hilbig[16], William Hobbs[17], Dan Honig[18], Matthew Kavanagh[15], Roy Lindelauf[19,20], Nina McMurry[4], Jennifer Merolla[7], Amanda Lea Robinson[21], Julio S. Solís Arce[13], Marijn ten Thij[22], Fulya Felicity Türkmen[7], Stephen M. Utych[23]

[1]European University Institute, [2]New York University, [3]New York University Abu Dhabi, [4]WZB Berlin, [5]University of Toronto, [6]Stanford University, [7]UC Riverside, [8]University of British Columbia, [9]LSE, [10]Yale University, [11]Delft University of Technology, [12]Stony Brook University, [13]Harvard University, [14]Aarhus University, [15]Georgetown University, [16]Princeton University, [17]Cornell University, [18]University College London, [19]Netherlands Defense Academy, [20]Tilburg University, [21]Ohio State University, [22]Maastricht University, [23]Independent Researcher, [24]UCLA

[*]Corresponding author: golden@ucla.edu

September 2, 2025

Version 5.0

Word count 9,480 excluding title page, abstract, bibliography, and supporting material

Key words: Meta-science; knowledge aggregation; ensemble techniques; COVID-19; comparative politics

Abstract

In 2020, shortly after the onset of the global Covid-19 pandemic, we devised a set of Model Challenges to explore what our discipline's inherited knowledge had to offer to accurately model (and understand) how governments would respond to the pandemic. Set up as a crowdsourced tournament, we first gathered models of political drivers of Covid mortality and then asked experts to assess which models would perform well or poorly. In this paper, we describe and take stock of this large scale collective effort. Our analysis suggests three main conclusions. First, the ability of political scientists to predict or explain which polities will react effectively to the pandemic and which not appears very limited. Second, even when our models are relatively successful, as a discipline we are not good at telling apart effective and ineffective models. Third, the best results are generated when models are combined. On the basis of these findings, we suggest that our discipline would benefit from the establishment of continued structured interactions that encourage multiple perspectives on phenomena of common interest and that aggregate ideas across researchers. As a discipline, we should sharpen our theories by seeking to predict future events instead of predicting the past. [199 words]

1 Introduction

Like social scientists across the world, when Covid-19 unexpectedly locked us all inside, we were eager to respond using our professional expertise. But what did we have to offer as political scientists? We were not medically-trained professionals who could save individual lives; we were not public health experts who could model disease flows and give advance warning; and we were not specialists in building or delivering vaccines. Could all our years of accumulated knowledge provide socially useful guidance in the face of a daunting global public health emergency? Having studied governments and political processes for years, could we say anything definitive about which pre-existing features would make some countries more successful than others in protecting the public from death by Covid, for instance?

When we started working on the Model Challenges (MCs) in summer and fall 2020, newspaper front pages were covered with hypotheses about the potential importance of political and social variables for Covid-19 outcomes. "The world needs more women leaders ..." proclaimed *The Conversation – Canada* in December 2020; "Poll: Most Republicans Say Covid Threat Overblown ..." *Forbes* reported in October of the same year; "Will COVID-19 kill democracy?" asked *Foreign Policy* in September 2020 (Adkins & Smith, 2020; Champoux-Paillé, 2020; Walsh, 2020). In the United States, Republicans were accused of allowing Covid to proceed unchecked due to their refusal to support policies of social distancing; around the world, women leaders were lauded for their superior communications during and responsiveness to the pandemic; and many worried that democracies might compare badly to authoritarian regimes in their capacity to manage the pandemic. Perhaps partisan polarization, the gender of political leadership, and regime type would prove consequential for how many people died from Covid.

Political scientists do not have an off-the shelf answer to this question. Asking about the political determinants of Covid-19 fatalities is asking about a new event, but also an event in a small class of events that are not typically at the center of research in the discipline (see Lynch, 2020 for a timely discussion of research on the politics of health). But at a more abstract level a

large body of political science research does focus on government effectiveness, government responsiveness, government engagement with crises.

Our goal was then to see how this broader knowledge could be marshaled to address a new question of global concern.

First we sought to bring together as many social scientists as possible and aggregate our combined expertise to understand the dynamics of the pandemic. There are multiple ways to gather and aggregate research on a topic. At one end there is a compilation approach, where excellent work by different researchers are brought together and published together. The *Perspectives on Politics* special issue on pandemic preparedness is a good example of this approach (Lynch et al., 2022). At another extreme are more fully deliberative processes that seek to produce a consensus on a question. The Intergovernmental Panel on Climate Change (IPCC) process is an example of this (for criticism see Oppenheimer et al., 2007). Meta-analysis occupies a middle ground where multiple studies, sharing a common treatment and outcome, are combined to reach an aggregate conclusion. Our approach is similar though importantly our interest was in eliciting and aggregating multiple accounts of a single event rather than single accounts of multiple events.

To do so, we built on previous crowdsourced competitions (such as Bennett and Lanning, 2007) and challenges (such as the Fragile Families Challenge (Salganik et al., 2020)). Our Steering Committee, drawn from six research institutions around the world, assembled datasets built from accessible and potentially theoretically relevant measures from 166 countries around the world as well as from first-tier subnational units (states and territories) in Mexico, the United States, and India. We then created an interactive website to crowdsource statistical models predicting future Covid mortality numbers. As an incentive for participants we offered co-authorship (reflected in the co-author list for this article) to those who submitted the most successful predictions of logged deaths per million—a decision that we return to below. Providing harmonized covariates and Covid-19 mortality data through November 16, 2020 we gave participants a six-week window (December 1, 2020 through Jan-

uary 20, 2021) to submit statistical models predicting future cumulative mortality as of August 31, 2021 with substantive justifications for their models. Modelers were permitted to include up to three covariates from the data we supplied or to provide data of their own on other variables.

Second, we wanted to understand whether political scientists agree on what constitutes a good explanation. Even if they do not generate models on a given topic, do they recognize a good model when they see it? To address this question, we put the submitted models on the Social Sciences Prediction Platform (SSPP) and asked social scientists to forecast the performance of the models that we had received in the MCs.

Following model submission and forecasting, we aggregated the models and the forecasts in a number of ways and then evaluated individual models, individual forecasts, and various aggregations of models and forecasts against the realized outcome data. How good were modelers who submitted to the MCs in identifying variables that accurately predicted Covid-19 mortality? How good were experts in identifying which models would perform well? If we pitted modelers and forecasters against a simple machine learning (ML) algorithm, would humans do as well as the algorithm? If we aggregated models and drew out their best features across submissions, would we find that our community as a whole out-performed individual modelers? Or were there exceptionally talented modelers in our crowd who could beat a machine? And finally, if we could accurately predict mortality rates from a global pandemic, were we also able to provide an understanding of why variations in those rates existed that was deeper than what we read in the newspapers — at a time when the best analytic news coverage of the pandemic achieved remarkable insight, especially through innovative data visualization (illustrated by work of *The Financial Times* (Burn-Murdoch, 2022))?

In this paper, we report on our experiences with the Model Challenges as a strategy for gathering and accessing knowledge. Our results are sobering but not depressing. The R^2_{loo} (the main measure of predictive accuracy that we use, consisting of leave-one-out predictive

success, detailed below) of the best single submitted model is 0.483 while that of the median model is only 0.171, indicating wide variation in model quality. A benchmark ML model, consisting of predictors selected by Lasso (least absolute shrinkage and selection operator), has an associated R^2_{loo} of 0.377. This tells us that most models that were contributed were much less accurate than a standard ML model, although a small number of contestants submitted highly predictive models that bested Lasso. The forecasting exercise found that forecasters were not very good at predicting which models would perform well, casting doubt on the ability of human experts to discern among competing theories of a common outcome. We also used an off-the-shelf statistical ensemble procedure (called stacking) as one way to aggregate models, and find that the results outperform the median model by a large margin, outperform the best model by 4 percent, and outperform aggregate predictions made by expert forecasters. Thus, *collectively* social scientists —– at least those that self-selected into our challenge —– possess substantive knowledge that can be aggregated using existing and well-validated statistical procedures, and doing this produces results that greatly outperform what any of us, however expert, can do alone. The challenge going forward is to devise methods and sites of intellectual exchange that bring together existing individual expertise and allow machines to filter and combine what we already know.

We focus on three specific arenas in which to distill lessons we learn from the Model Challenges. First, we discuss the *substantive findings* of the MCs, which we interpret as showing that the institutional structures and social processes that we might expect to predict which governments would be better positioned to tackle a challenge like the Covid-19 pandemic are in fact not very predictive of success. This is a negative but important substantive finding. Second, we discuss the promise — and also the logistical and ethical difficulties — of *crowdsourcing knowledge and massive coauthorship*. Finally, based on the statistical results of our analysis, we make the case for *combining our collective expertise with statistical algorithms*. We believe that the final point is the most general one to draw from the MCs: collectively, we know more than almost any of us individually. Aggregating competing statistical mod-

els and refining the results using algorithmic methods may therefore be the surest route to predictive success.

Substantively, the procedures built into the MCs sought to integrate explanation and prediction. We required each submission include a text explanation to justify the choice of statistical predictors, and we explicitly encouraged modelers to draw on the literature in writing those explanations. In essence, we were asking how transferable the understandings and explanations that our intellectual community had developed about other outcomes were to the outcome of Covid-19 mortality. In this way, we provide a kind of stress test of our existing theories for a newly-emerged domain of interest.

Our paper proceeds as follows. Section 2 describes the Model Challenges and the forecasting exercise as well as how we aggregate and evaluate models and forecasts. Section 3 provides the substantive results, explaining how models and forecasts perform and what we learn from the MCs about how governments affect Covid-19 mortality outcomes. In Section 4, we draw lessons about the underlying question itslef and the types of explanations that emerged, about largescale collaborations of this form, and about strategies for aggregating explanations in political science research.

2 What We Did

2.1 How the Model Challenges Operated

The Model Challenges sough to gather, aggregate, and evaluate rival accounts for a common phenomenon. With a focus on a single outcome – here Covid mortality — the substantive interest is in explanation: what *accounts for* variation in outcomes, or more specifically, what political logics account for diverging outcomes. Given multiple explanations, can we combine these into an aggregate account? And how can we evaluate these individual and aggregate accounts. Can we, as a discipline, recognize a good account when we see one?

	Tournament stage	COVID-19 mortality data provided as of:	Dates of stage	Models/forecasts evaluated against data as of:
1	Model generation (Model Challenges)	Nov. 16, 2020	Dec. 1, 2020 - Jan. 20, 2021	Aug. 31, 2021
2	Model assessment (SSPP forecasting)	Feb. 28, 2021	May 1-31, 2021	Aug. 31, 2021 Aug. 31, 2022

Table 1: Summary of the MC research timeline.

To answer these questions we implemented Model Challenges with two phases.

1. **Model generation**: In the first phase we invited researchers to develop models that use social and political variables to predict cumulative COVID-19 deaths, measured by logged deaths per million, as of a specific future date (August 31, 2021). We asked researchers to include written explanations for why selected socio-political variables would predict COVID-19 mortality. We created four challenges: (1) across countries; (2) across states in the USA; (3) across Mexican states; and (4) across Indian states. We selected these three countries because of data availability and team expertise, as well as a general desire to maximize geographic variation.

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

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

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

The sequence of activities in each stage of the tournament is shown in Table 1.

Challenge	no. obs.	Logged cumulative COVID-19 deaths per million, as of August 31, 2021				
		Median	Mean	Minimum	Maximum	Std. Dev.
Crossnational	166	6.12	5.64	0	8.73	1.87
India	31	5.87	5.87	4.5	7.7	0.84
Mexico	32	7.62	6.55	5.82	8.59	0.41
USA	50	7.56	7.41	6.03	8.02	0.47

Table 2: Summary statistics for the outcome measure by challenge. We add 1 to our cases per million prior to logging, such that 0 is interpretable as no deaths. (There were no reported COVID-19 deaths in the Solomon Islands as of August 31, 2021.)

Participants had access to mortality data as of November 16, 2020 (thus, as of a few weeks prior to the launch of the MCs) and were asked to predict cumulative mortality as of August 31, 2021, or about seven months into the future. Table 2 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. See Appendix A on data sources.

To provide context, the period when the Model Challenges were open to submission was one when questions about vaccine availability (Basta & Moodie, 2021; Bokemper et al., 2021; WHO, 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 (de Figueiredo et al., 2021; Lazarus et al., 2021; Solís Arce et al., 2021) were particularly salient. New variants (including Delta) emerged only after predictions had been made. Thus, uncertainty over the trajectory of the COVID-19 pandemic at the time of the challenges complicated the task for participants of making out-of-sample predictions of mortality.

We also provided rich data to modelers that could be used as determinants of future mortality. Substantively, we included candidates identified by broad literatures on possible determinants of how effective governments and societies may be in handling unexpected crises, including public health crises, and more broadly in how effectively they perform over-

all. We also limited measures to those that pre-dated the onset of Covid-19, which in practice meant we systematically coded variables as of no later than the start of 2019. We viewed policy responses to Covid-19 as mediators (or mechanisms) that may have been conditioned by preexisting social structures and political institutions. This constrained our modelers to work with a common set of stable *pre-existing* determinants measured prior to the outbreak of the pandemic.

We assembled approximately two dozen measures for each challenge, which we can categorize into three theoretically-relevant buckets of possible determinants of effectiveness in responding to Covid-19: (1) state capacity and general institutional effectiveness, including the extent of corruption and whether the government is unitary or federal; (2) societal capacity, including ethnic fractionalization and the extent of social trust; and (3) policy priorities, including the presence of female leaders and the extent of polarization. In addition, we provided data on the health system (access to sanitation, for instance), data on the level of development (GDP per capita), and data measuring what we called "epidemiological" variables, which include GDP per capita, share of population over 65, respiratory disease prevalence, hospital beds per capita, share of population living in urban areas, and population density. For a discussion of the underlying logic of the variables we included, see Bosancianu et al., 2020. As already noted, we also allowed modelers to submit their own variables and data (although in practice, few chose to do so).

The MC platform became publicly available on December 1, 2020. Prior to and after the launch of the MCs, we advertised to other social scientists — mainly political scientists and to a lesser extent economists — via an array of solicitations using social media (Twitter at the time) and professional listservs (the American Political Science Association, the European Political Science Association, the Society for Political Methodology, Evidence in Governance and Politics, and others). We also sent individual emails directed at political science researchers at the top 100 research institutions globally and specifically in the USA, Mexico, and India. To make the Model Challenges accessible to a wide range of scholars

— including social scientists who do not typically engage quantitative methods — steering committee members hosted a virtual hackathon-style workshop that demonstrated how to use the platform and participate in the Model Challenges.

The public MC interface, depicted in Figure 1, was programmed in Shiny to be an interactive website. It allowed researchers to:

1. Choose a model challenge to enter (see Figure 1b).
2. Select up to three predictors and see graphics showing the performance of linear bivariate models for each predictor on COVID-19 mortality data as of November 16, 2020 (see Figure 1b).
3. Optionally upload new regressors (see Figure 1b).
4. Optionally change the functional form of the models to allow interaction, polynomial, or custom model submissions (see Figure 1c).
5. Optionally predict parameter values for models, enabling submission of "parameterized models" (see Figure 1d). (We refer to models that do not have parameters provided as "general models.")
6. Provide a logic to explain the model (required). We encouraged researchers to describe why the set of predictors they chose mattered for the outcome, with references to relevant literatures (see Figure 1e).
7. Submit models (see Figure 1f).

Prior to the launch, the Model Challenges were deemed exempt or received approvals from the Institutional Review Boards (or equivalent bodies) at the six institutions with which the eight members of the Steering Committee were affiliated. Modelers provided informed consent as well as identifying information as part of the submission process. Separate IRB approvals or exemptions were received for the forecasting portion of the research, which did not collect any personally identifying information.

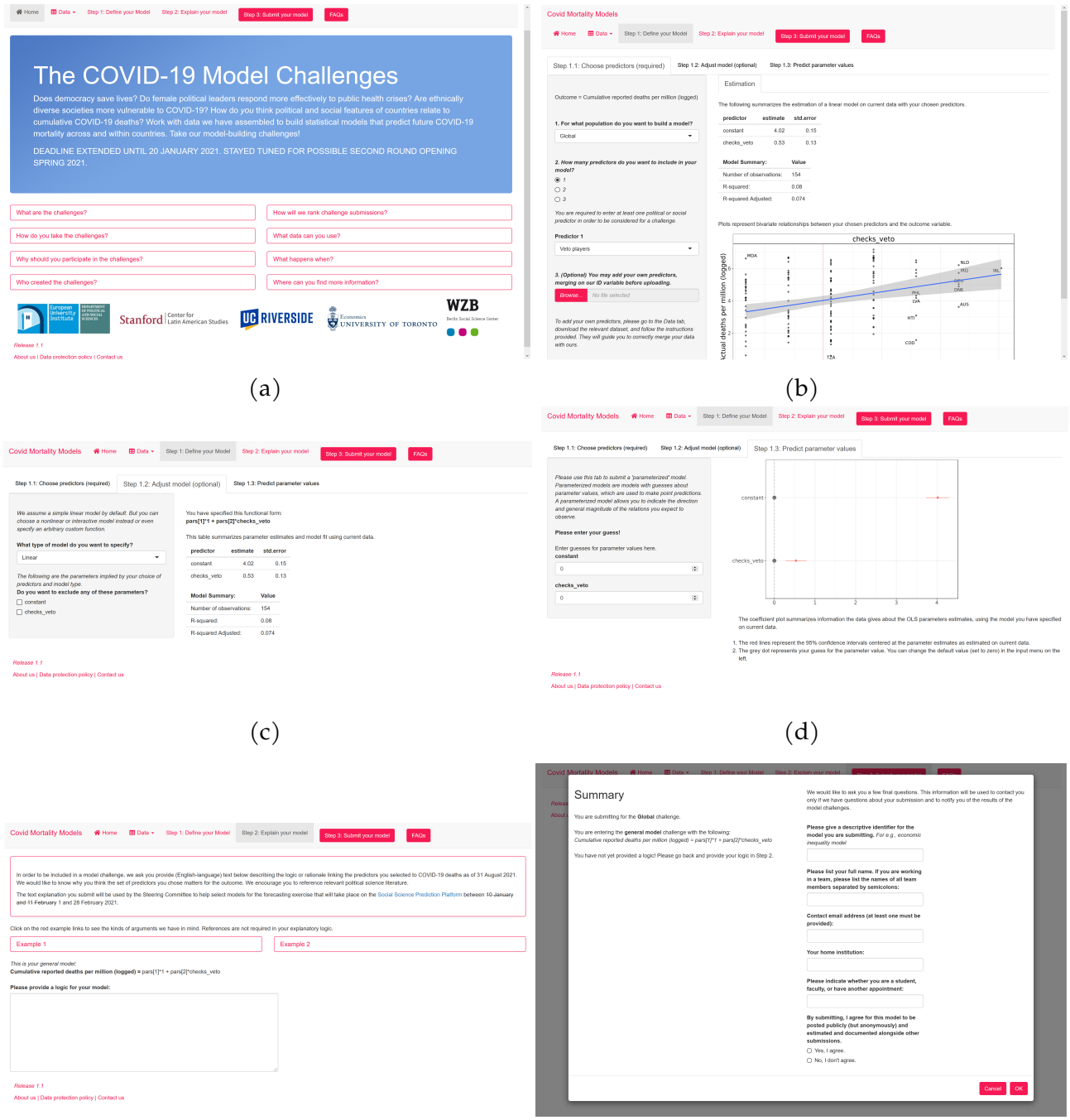


Figure 1: Screenshots from the MC interface. Plots and reported statistics were presented dynamically.

2.2 Submissions

In all, 116 general models were submitted across the challenges. Figure 2 provides an overview, including information about the functional form of the models, the number of predictors used and the addition of predictors from outside the MC datasets, whether the models were theoretically motivated — i.e., whether they included a theoretical argument for why their selected variables should predict COVID-19 mortality or whether the submission stated that the model had instead been generated using machine learning methods, whether the modeler included references to existing literature to justify inclusion of selected variables ("has model justification"), and the number of persons comprising each team of modelers. Fuller descriptions of all models are provided in Appendix Table C.1.

The average team submitting consisted of between two and three persons, although some were as large as eight. Although nearly two-thirds of the models had an accompanying theoretical motivation (meaning they were not generated using ML methods), we deemed only about half of the entries to have included a model justification (meaning they referenced existing literature to justify inclusion of determinants). Most modelers used the data that we had already assembled rather than providing their own. We received more crossnational than country-level submissions and among countries, more for the USA.

Typically participating individuals participated in multiple challenges. In all, the four MCs received 88 submissions from 60 different individuals based at 32 institutions in 10 countries (see Table B.1 for details). How should we think about these numbers? The Fragile Families Challenge, which is the nearest academic equivalent we know of, received submissions from 160 teams (Salganik et al., 2020), about double the MC number. In the case of the Fragile Families challenge, however, modelers were given more than four months (from March 21 to August 1, 2017), compared to the six weeks we provided for submission. This comparison suggests that our outreach was relatively successful given the compressed timeframe of the project.

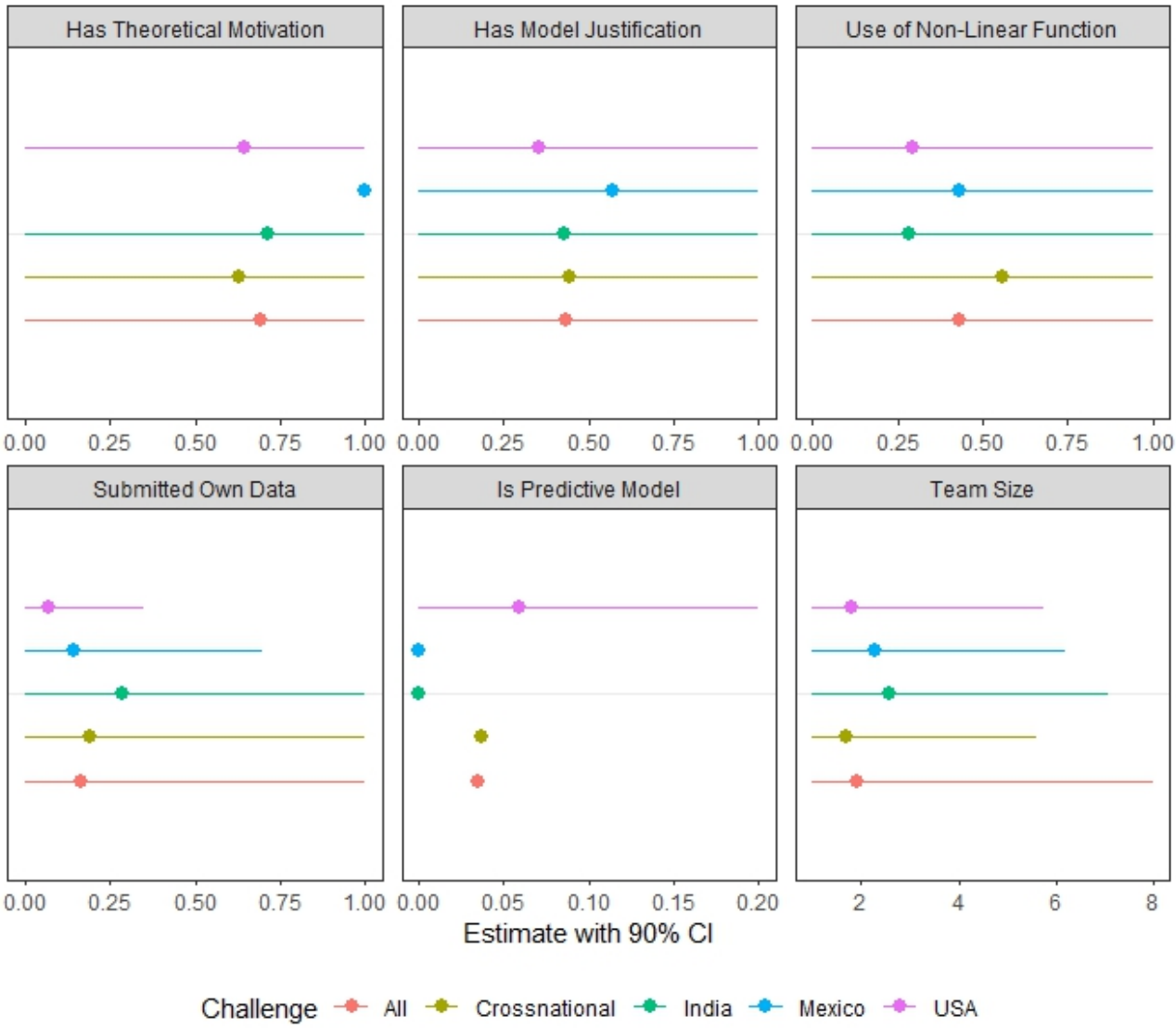


Figure 2: Overview of features of general models from all four challenges, organized by challenge. The top panel (*"Total"*) reports the sum of the subsequent challenge-specific panels. "Theoretical motivation" means that a substantive rationale for the model was provided with the model. "Model justification" means referred to existing literature to explain reasoning. "Is a predictive model" means that the model was selected by a (disclosed) machine-learning algorithm.

Given the heterogeneity of the teams, we have no way to relate modeller characteristics to the quality of the models submitted. Some teams comprised multiple ranks — ranging from undergraduate students to full professors — and some were cross-institution — including large R1 research universities and small liberal arts colleges. The composite nature of the submitting teams is itself an interesting aspect of the Model Challenges, but means we can say little about whether seniority or institutional ranking related to model performance.

2.3 How the Forecasting Exercise Operated

The second stage of the tournament elicited expert forecasts. Forecasting has become increasingly common across the social sciences (DellaVigna & Pope, 2018; DellaVigna et al., 2019), offering a way to access expertise about social processes. For the MCs, we sought expert evaluations of statistical models, which is, informally, how they are routinely evaluated, for instance via the peer review processes. In February and March 2021, we used the Social Science Prediction Platform (https://socialscienceprediction.org/) to elicit expert forecasts to evaluate the performance of the models earlier submitted to the MCs. We received 175 expert forecasts, 83 focused on how the crossnational models would perform on future data and 92 on country-specific models.

We randomly assigned experts into two groups to elicit two sets of forecasts, which we label "horserace" and "stacking" forecasts. In the horserace, experts saw a subset of six randomly-selected general models that had been submitted to a given challenge. Experts were asked to guess the probability that a model would be the most predictive in the set. In the stacking exercise, forecasters were asked to allocate weights across models over a subset of seven randomly-selected models.

The horserace forecast was designed 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 from among the theoretical (non-ML) general models submitted to a given challenge and (2) the epidemiological model. The set of models that forecasters viewed

varied across respondents; different respondents saw different subsets of the general models.

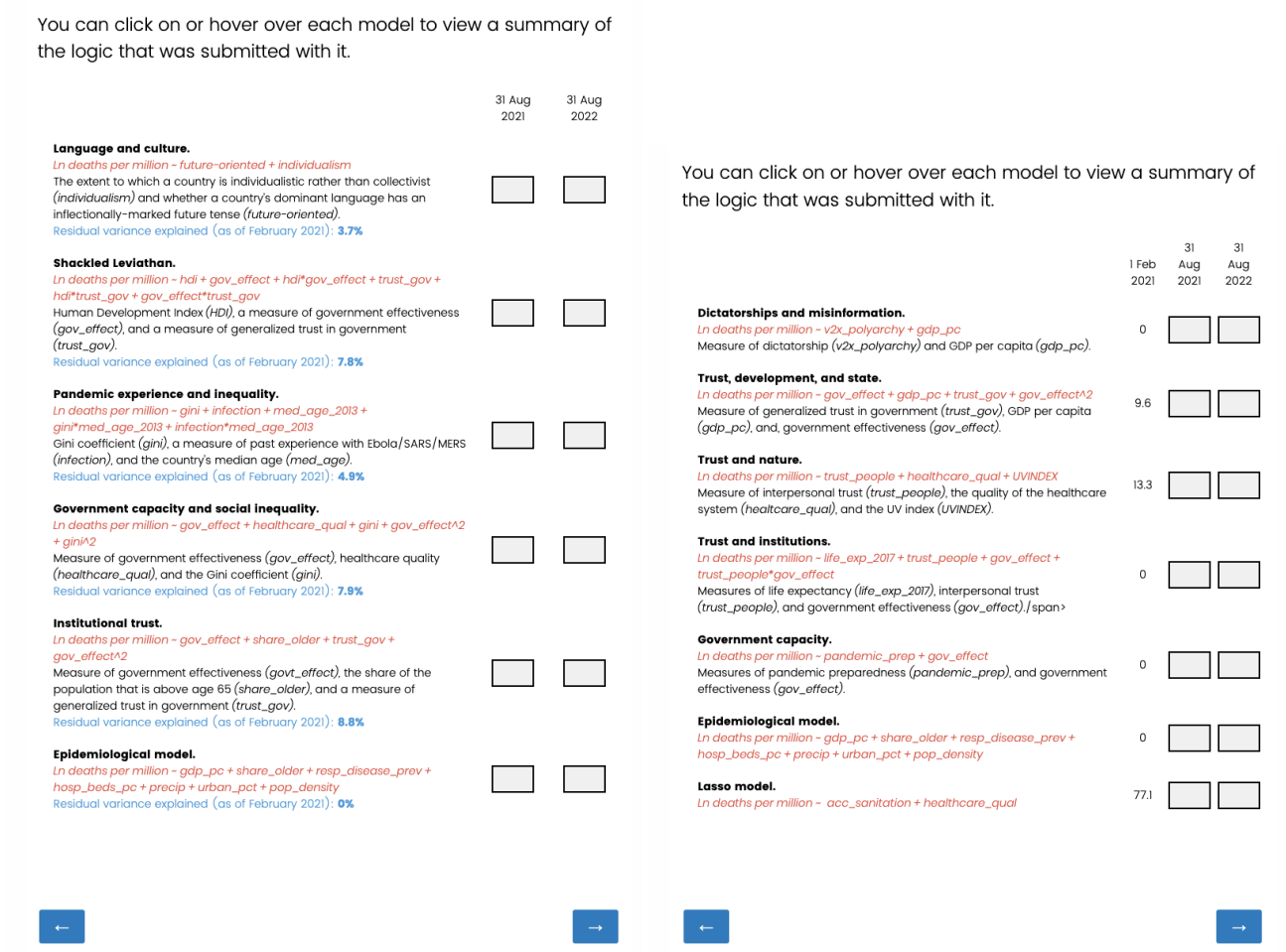
Forecasters read the following instructions for the “horserace" elicitation for the cross-national challenge:

> 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.

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

The instructions that stacking forecasters read were similar to those for the horserace but with appropriate adjustments in language. For details, see the text in Appendix K.

Figures 3a and 3b provide representative screenshots of the interface for a horserace forecast and a stacking forecast.



(a) Horeserace forecast interface

(b) Stacking forecast interface

Figure 3: Forecasting interface for two representative forecasts. Forecasters could hover over the models to read a description of the logic behind each model (using the text submitted by the modelers).

2.4 Analysis strategy: Aggregating and Evaluating Models and Forecasts

Having elicited models as well as forecasts for how well the models would perform against future data, we then use statistical methods to combine the models into a meta-model and

also to combine forecasts. Doing this generates material to assess how well we do collectively in producing and evaluating theories and to compare our collective, aggregate expertise with the expertise of single individuals.

We generate an *aggregate* prediction for each challenge, using the submissions to evaluate how much (if at all) predictive accuracy improves when we combine what were created as individual, competing theories. Aggregation is implemented using a stacking procedure that generates a meta-model based on all submitted models (Yao et al., 2021). The stacking estimator allocates weights to the predictions of each constituent model to maximize the meta-model's predictive accuracy. For technical details, see Appendix H.[1] By construction, the aggregated meta-model is designed to match or exceed the performance of its constituent models. To assess the performance of individual constituent models, we examine the weights that each contributes to the meta-model. We find that most models contribute zero weight to the meta-model, meaning that they effectively make no contribution to the aggregate model. We interpret this to suggest that the models submitted by most modelers have no value in our collective understanding of Covid-19 mortality.

We aggregate the forecasting results using two separate procedures to create: (1) a "representative expert" model, based on the performance of the median forecast; and (2) a "wisdom of the crowds" model, based on the normalized average weight placed on a model by experts. For technical details on both procedures, see Appendx I. Although these are not exactly akin to the meta-model that is generated from the MC submissions, they similarly use all the available data to bring together the collective expertise of forecasters.

With these various data and data transformations at hand, we are able to evaluate the predictive accuracy of all the submitted models, all the forecasts, the aggregated meta-model

[1] To clarify, we fed early results of the stacking estimates of models into the "stacking" version of the forecasting exercise, described above; in that version of forecasting, we asked experts to assign weights to submitted models and we allowed them to see an initial stacking estimate as of February 2021 (see the text in the instructions to the "stacking" elicitation, reprinted above). The stacking results that we analyze in this paper are based on the performance of the meta-model against mortality outcome data as of August 31, 2021 and then June 20, 2022. The June 20, 2022 cutoff was used instead of the initially-planned August 31, 2022 because some governments stopped collecting COVID-19 mortality data before the latter date.

of the submitted models, and the two aggregations of the forecasts. In what follows, we focus on the crossnational challenge and present the results of the country-specific challenges in Appendix F.

3 How Well Did Models and Forecasts Perform?

We can compare submitted models against each other as regards their predictive accuracy. But we also want to know how well human modelers do relative to what an algorithm produces. We thus benchmark the relative performance of individual models against two entirely atheoretical statistical models predicting COVID-19 mortality: (1) one with prespecified "epidemiological" covariates, described above, and (2) a model selected by a Lasso algorithm on the full set of assembled predictor variables. For details on the Lasso procedure, see Appendix L. We use Lasso because it is a widely-used algorithm that selects predictors to generate interpretable models (Tibshirani, 1996). These two benchmarks compare the specific theoretically-motivated expertise of social scientists with off-the-shelf atheoretical and algorithmic predictions.

For the forecasts, we proceed analogously. We compare rankings of models that are produced by the "horserace" forecasters to the "representative expert" and the "wisdom of the crowds" models that emerge from the stacking forecasts. We proxy the former with the median-performing stacking forecast and the latter with the average stacking weights made by all forecasters.

3.1 Models, Forecasts, and Aggregated Models and Forecasts

The procedures we implemented produce six sets of outcomes. The comparisons are: (1) how individual models perform in predictive accuracy against each other and against the epidemiological and Lasso benchmarks; (2) how forecasts perform in predictive accuracy against each other and against the representative expert and wisdom of the crowds bench-

marks. Before getting to these main results, we walk the reader through results of the predictive performance of the individual submissions to the MC.

3.1.1 Predictive assessment

We begin with an evaluation of the performance of individual models that were submitted to the general crossnational MC. Figure 4 depicts (on the x-axis) the leave-one-out predictions arising from each crossnational general model that was 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. The measure of predictive accuracy is calculated using Equation (G.1). A perfectly predictive model would have an R^2_{loo} of 1, where higher values indicate greater predictive power. In Figure 4, models are ordered from best to worst performing according to this metric.

Inspection of the data displayed in Figure 4 reveals that models vary substantially in their predictive power. The R^2_{loo} of the best model is 0.483 but only 0.171 for the median model. Interpreting these metrics on an absolute scale is more challenging than making relative comparisons. Because the Lasso model is fit on all common predictors, it provides one possible benchmark. On the one hand, the comparison between the Lasso-selected model and the MC submissions should favor Lasso: Lasso targets predictive accuracy rather than explanation, and we asked modelers to provide substantive explanations in their submitted models. Providing substantive explanations might have affected how modelers thought about predictive accuracy. On the other hand, the stated goal of the challenge was to predict future (August 31, 2021) COVID mortality. The modelers had this information when formulating their models, but the Lasso was fit on data only through December 2020. In this respect, the Lasso-selected model is disadvantaged relative to predictions made by humans. Considering the advantage and disadvantage of the Lasso-selected benchmark relative to the submissions is important when interpreting the Lasso benchmark. The R^2_{loo} of the Lasso model is 0.377 and it ranks fourth (out of 28 models) in predictive power. So three out of 28

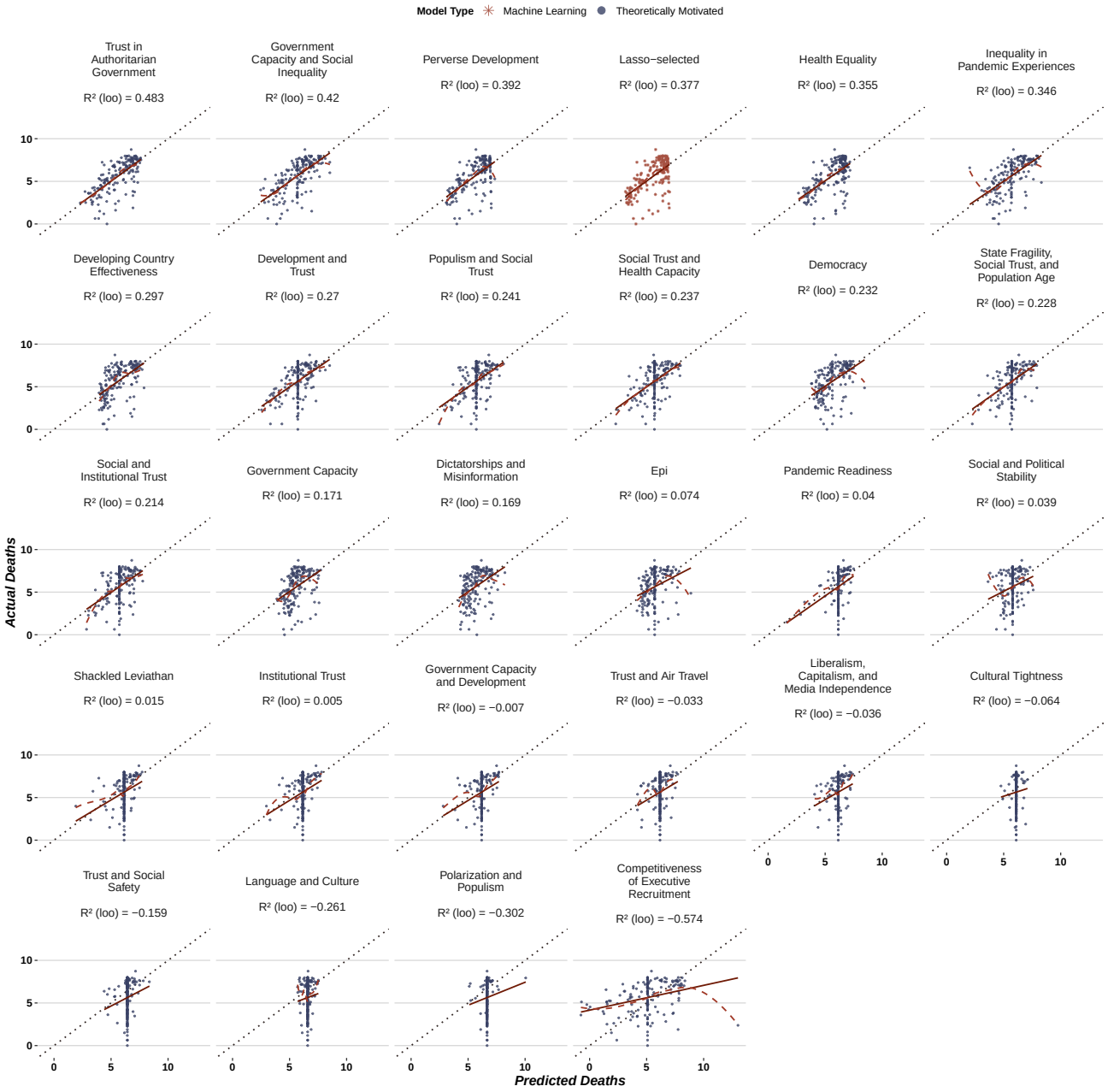


Figure 4: Evaluating: actual versus predicted deaths. Leave-one-out predictions of general models submitted to the crossnational MC and observed COVID-19 mortality as of August 31, 2021. Facets are ordered from highest to lowest R^2_{loo}. Dotted diagonal lines are 45 degree lines and fitted lines are estimated by OLS and LOESS. Non-machine learning models are theoretically-justified user submissions whereas machine learning models were generated using a known (or reported) machine learning algorithm.

modelers out-perform the Lasso algorithm; these three comprise 10 percent of submissions.[2]

[2]We might think of these three modelers as akin to the superforecasters studied by Tetlock and Gardner, 2016. Most of our modelers, however, resembled standard experts, in that they were not very good at forecasting outcomes (Tetlock, 2005).

Fully ninety percent of submissions did worse than an algorithm.

Table 3: Regression results for crossnational data (general models)

	Authoritarian Trust	**Govt. Capacity and Inequality**	**Perverse Dev.**
(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***		
	(0.16)		
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 $<$ 0.05; **p $<$ 0.01; ***p $<$ 0.001

3.1.2 Contribution assessment

We next assess models by examining their contributions to the aggregate meta-model that is created using stacking. The meta-model combines the most useful predictive features of individual models to generate a model that is by definition better than its single components. To evaluate the value of each single model's contribution to the meta-model, we consider

its stacking weight. In Figure 5, we plot results of these assessments. Comparing Model Challenge submissions of the horserace and stacking contests, we see that only a few models receive non-zero stacking weights. The stacking meta-model draws on only three constituent models ("Trust in Authoritarian Government," "Government Capacity and Social Inequality," and "Inequality in Pandemic Experiences") despite minimal differences in their predictive accuracy, as measured by R^2_{loo}.[3] Despite these minimal differences, the stacking estimates suggest that much of the collective predictive power of the models that we assess is concentrated in only a few of the best performing. The skew of estimated stacking weights towards the two top-performing models is striking.

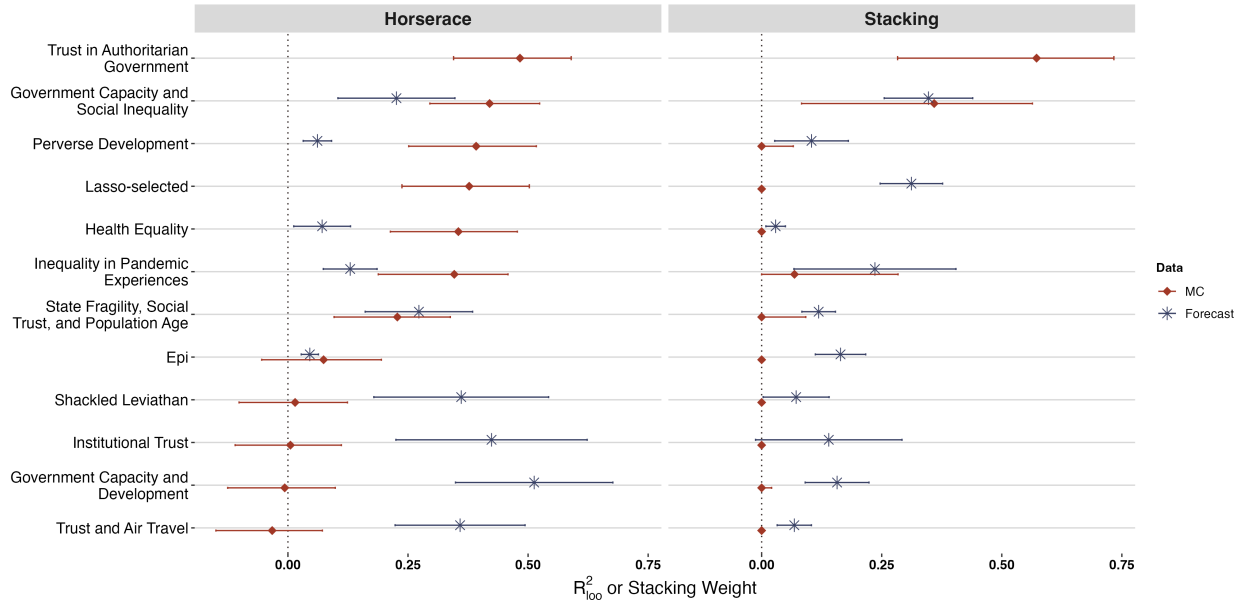


Figure 5: Comparison of models selected by horserace and/or by stacking weights. Each panel compares results using either individual MC submissions or forecasts. The models included are among the top-five performers in either. The 95% confidence intervals are generated by bootstrapping. Models with build-in procedures for improving fit (e.g. "Trust in Authoritarian Government") that were not standard across models were not included in the forecasting exercise.

This initial analysis of model selection yields two central findings. First, comparison (horserace) and aggregation (stacking) prioritize different sets of models. Stacking heavily favors very few models, putting much lower weights on the others. This occurs even when

[3]The weights estimated for each model are relative to the specific set of models evaluated.

differences in predictive accuracy of individual models are minimal. Application of these metrics in other contexts is necessary to establish the generality of these patterns, though they hold across all four MCs.[4] Second, we show that the average expert has very limited ability to accurately identify the most predictive models. Even when experts are provided baseline performance metrics, as in the forecasting exercise, they do not do very well. Our educated guess is that the forecasting exercise that we designed required novel cognitive tasks. Indeed, completion rates of 42.6 percent (horserace) and 30 percent (stacking) suggest that forecasting the models' out-of-sample predictive accuracy is challenging even for experts, most of whom gave up before completing a forecast. Perhaps with practice, forecasting performance would improve.

The results of the forecasting exercise show that, to the extent that traditional methods for organizing and synthesizing knowledge produced by an existing literature ask researchers to identify the strongest arguments, there are grounds for skepticism about their abilities to do so, at least if predictive accuracy serves as the primary evaluative criterion. Expert forecasters were not very good at predicting which submissions to the MCs would perform well (cf. Tetlock, 2005).

Finally, we turn to the results of six different aggregation methods, depicted in Figure 6. We describe our metrics of model success in Appendix J. The rows depict two different ways of assessing predictive performance. The top row evaluates predictions of observed outcomes. The second normalizes both model predictions and outcomes, providing information about the correlations between them. The two columns show predictions for different time periods. The left column presents estimates of predictions of cumulative COVID-19 mortality as of August 31, 2021, which is the prediction date that MC participants were asked to use. The second column presents out-of-sample predictions, which are evaluated as of June 20, 2022.

We provide two benchmarks for each method, benchmarks which we have already dis-

[4]For the country-specific results, see Appendix E.

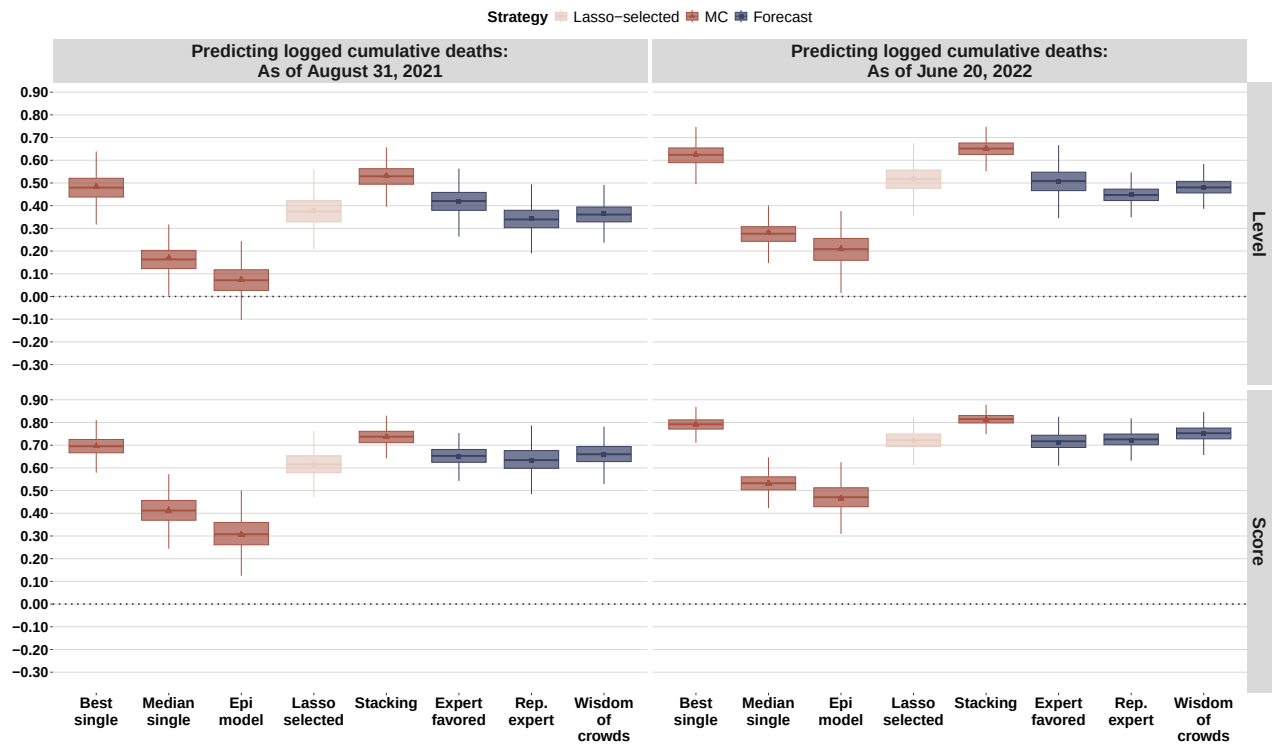


Figure 6: Comparison of predictive accuracy using different metrics. The top row of boxplots shows predictive performances (R^2_{loo}) of cumulative COVID-19 deaths per million and the bottom row shows correlations between predictions and actual mortality. The left column of boxplots assesses predictive accuracy on cumulative mortality as of August 31, 2021. The right column evaluates out-of-sample predictions of cumulative mortality through June 2022 using the models selected on the basis of the August 31, 2021 data. The boxplots show the interquartile range; whiskers are two standard deviations above and below the mean R^2_{loo}. Interquartile ranges and 95% confidence intervals are generated by bootstrapping.

cussed: the Lasso model selected on the basis of 2021 data to make out-of-sample 2022 predictions and the epidemiological model that uses what we consider standard predictors of disease mortality.

After the benchmarks, the third and fourth measures of predictive performance — the best- and median-performing models in the MC — follow directly from the discussion in Section 3. Point estimates in the top row report the R^2_{loo} of each model. The fifth prediction examines the outcomes using the stacking meta-model. For purposes of out-of-sample predictions for 2022 in the righthand panel, we use the best, median, Lasso, epidemiological and stacking models that were selected on the basis of the 2021 data. For all these, we construct

a sampling distribution of model performance by bootstrapping the data (resampling 166 countries with replacement).

The remaining three methods, illustrated in Figure 6, aggregate expert forecasts. The first metric examines the predictive power of the expert-favored model. As we already documented in Figure 5, the model that experts deem most likely to be the most predictive does not align with the model that is objectively found to be most predictive. The next two methods aggregate expert stacking forecasts. The "representative expert" forecast depicts the median aggregate stacking forecast. The final metric presents a "wisdom of the crowds" stacking model that aggregates over forecasters' stacking weights.

The main result that we highlight from Figure 6 is the consistent ability of the stacking method to outperform all other metrics for measuring the success of either models or forecasts. Indeed, it does better against both the top model from the MC or from the forecasts. While stacking by construction will always match the best constituent model, it will not necessarily do so by a large margin, which is what we observe here.

When we evaluate models submitted to the Model Challenges, we find that the most successful theoretically-motivated models outperform a Lasso-selected model. However, the performance of the "typical" median-performing model is far worse than the Lasso-selected benchmark. In Panel (a) of Figure 7, we compare the user-submitted models to a random sample of 130,000 $(5,000 \times 26)$ models generated from the MC-provided data and Shiny application. Specifically, we randomly sample permutations of three-predictor models from the MC-provided data and then randomly select the functional form of the model (linear, quadratic, or with interactions). For each model form, we randomly select the parameters to include in the model (see Appendix M for the sampling algorithm). Results show that the strongest of the submitted models clearly falls in the top percentiles of all possible models; thus, eliciting models from experts provides an advantage over any algorithmic production of models. However, many weaker submitted models do not perform well relative to the distribution of all possible models. In Panel (b), we compare the stacking prediction to stacking

predictions generated from the identical 5000 random samples of 26 random three-predictor models generated from the crossnational MC data. The stacking prediction outperforms all of the predictions from a "null" distribution of stacking models ($p \approx 0$). By aggregating expert models via stacking, we can substantially enhance the predictive performance of a set of models. Implementing an ensemble algorithm to combine features of multiple models adds predictive value, documenting the utility of aggregation, and specifically of algorithmic aggregation.

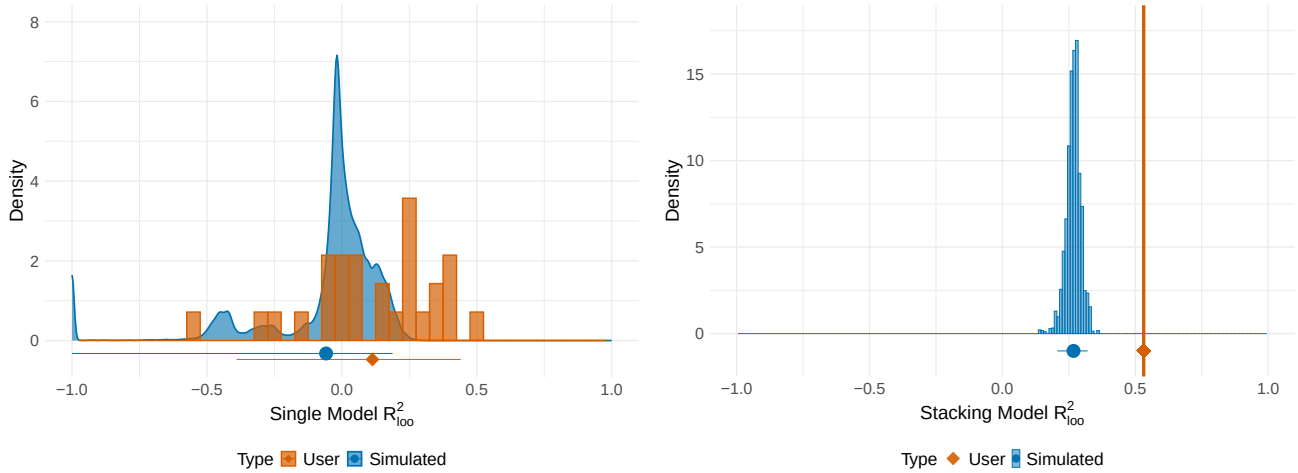


Figure 7: Panel (a) depicts the observed distribution of R^2_{loo}'s and the density plot depicts the distribution of R^2_{loo}'s from our sample of 130,000 linear, quadratic, and interactive three-predictor models using the common MC dataset. Panel (b) shows how the predictive performance of the stacking model compares to the predictive performance of randomly generated stacking models. All models are crossnational general models.

To conclude our presentation of analytical results, we stress the superior performance of the ensemble stacking method in predictive performance over the performance of models produced by humans. But we caveat that humans do a better job than algorithms at producing models in the first place, a point we substantiate in the next subsection.

3.1.3 Explanatory assessment

We now consider the explanatory success of the best performing models. The models that outperform Lasso are all theoretically motivated in the sense that their authors provided reasons justifying the inclusion of each variable.

To better understand the extent to which good performance reflects strong logics we show the performance and discuss the logic of the top three performers, shown in Table 3.

The best performer, “Trust in Authoritarian Government," is a simple linear model that uses three variables: a measure of trust in government, the presence or absence of a critical mass media, and access to sanitation. The text explaining the logic for this model states that health access is a control variable and that a critical media is included (based on observation of the Chinese experience) to proxy data manipulation by government. In fact, its explanatory power appears to derive from the control rather than from the media variable, which does not enter into the model at a level that is statistically significant. While the model appears motivated by logics in authoritarian settings, regime type does not enter in the statistical model.

The second best performing model, “Government Capacity and Social Inequality," includes measures of government effectiveness, the quality of healthcare, and economic inequality; it uses two quadratic terms. The text predicted non monotonic relations for each of these which are seen, in the expected directions, in both cases. The logic for non monotonicity of effectiveness drew from a combination of a response logic (at the high end) and a reporting logic (at the low end). The logic for inequality however drew more from early Covid experiences than from general theory.

The third theoretically-motivated model, “Perverse Development," include two measures of general levels of development: access to sanitation and the human development index (HDI). Both of these were intended, according to the logic supplied by the modeler, to capture a country’s level of exposure to Covid risks and not to capture government responses. The model predicted higher levels of COVID mortality in more developed contexts, as is borne out in these models.

Strikingly these three best performing models do not in general rely heavily on well defined political theories. Two draw more from observations of Covid responses and one fo-

cused on demographic risks (age distribution, mobility, density) rather than political logics. In each case the logics provided were variable centered, focusing on why different variables might matter, rather then outcome centered, focusing on how multiple variables combine to form an explanation.

4 Lessons

We draw lessons regarding the pandemic itself, how to organize collective crowdsourced aggregation processes of this form, and how to aggregate knowledge.

4.1 The Substantive Lessons

Did the Model Challenges produce substantive knowledge about COVID-19?

Our findings suggest that, broadly, institutional and government characteristics are not strongly predictive of COVID-19 mortality rates: high and low mortality rates can be found among both democracies and autocracies, for instance. Like other studies, the MCs also provide some support for the importance of social and political trust.

Evidence for the claim that many institutions were potentially compatible with more successful policy responses to the unexpected COVID-19 crisis comes from the highly heterogeneous nature of the variables that modelers thought would be predictive of mortality outcomes. In Appendix Figure D.1, we show a visual summary of the most common predictors found in crossnational models.

Overall the most commonly-used variables are trust in others, trust in government, and government effectiveness, and the single most common pairing of variables is trust in government and healthcare access — a coupling used, however, in only three of 26 submissions. The frequency with which distinct submissions used common predictors is distinguishable from random selection of predictors drawn from the MC-provided dataset (p-value $= 0.004$), which suggests that participants considered common arguments from the literature

or shared intuitions when constructing their models — but also wide variability in how those arguments were interpreted or which intuitions were drawn on.

The exception to this wide variability in possible determinants of mortality is trust. Across the four challenges, 30 percent of models include measures of trust in society or trust in government in predicting COVID-19 mortality. (For the country-specific data, see Appendix **??**). If modelers agreed on anything, it was that trust could be an important predictor of policy success. Even if variables measuring trust proved to be important, however, the MC did not elicit very predictive *theories* of mortality outcomes.

The fact that institutional variables were not very predictive of mortality corrorborates other research investigating contextual factors explaining COVID-19 outcomes. A study by the COVID-19 National Preparedness Collaborators (2022) analyzes data from 177 countries and sublocations and reports that 44 percent of the variation in COVID-19 mortality remains unexplained by demographic, health, economic, social, or political variables, making COVID-19 "an epidemiological mystery" (COVID-19 National Preparedness Collaborators, 2022, p. 1505). The single most important variables predicting mortality were, first, the age-structure of the population and, second, an obesity index. Countries which had more elderly and more obese experienced higher fatality rates among those infected.

Regarding social and political determinants, the COVID-19 National Preparedness Collaborators (2022) reports that trust in government and social trust appear to be especially important, in particular because they appear to promote vaccine uptake. Other studies also find that trust affects vaccine uptake (Adhikari et al., 2022; Sapienza & Falcone, 2022).

From the Model Challenge results, we find confirmation of these broader findings. We also find that these features not only appear to have predictive power but that they stand out when pitted against and combined with a wide variety of other measures.

More problematically, we do not see a distinctive *theory* emerging that accounts for why some states perform well and others poorly. "Explanation" remains very much at the variable

level. Modelers appear to have used the opportunity to include three variables as a way to add multiple possible independent explanatory variables rather than as a way to propose a theoretically integrated set of predictors.

4.2 Lessons for Crowdsourcing Social Scientific Research

As proof-of-concept, this exercise showed that a small group of unfunded but dedicated researchers could rapidly assemble systematic data for many observations at multiple levels and build a platform for users to interact with the data and submit comparable statistical models.

What lessons can we draw for selection procedures and incentive strategies?

Our selection process demonstrated that a robust community of political scientists around the world would voluntarily engage in the kind of mixed competition/collaboration that we offered, and that some of them — although admittedly not many — would even contribute additional data that they sourced themselves. We infer that members of our discipline are willing and even eager to engage in large-scale collaborative enterprises, perhaps building on the growth of co-authorship in political science in general (Grossman et al., 2025).

With that said, we cannot claim that the respondents represent the discipline. Given the size of the discipline the numbers responding were relatively small, and they are also self selected. Alternative procedures might include selection via a professional association.

The incentives we provided to take part were modest. Researchers would be contributing their insights to understanding a major societal challenge, which for some is reward. In addition, we incentivized effort by offering coauthorship to those who submitted the most predictive models, defined as models receiving non-zero weight (≥ 0.001) in the stacking exercise. As a result, 25 of the 60 MC participants were offered coauthorship on the basis of the merits of their models. Of those 25, 22 elected to serve as coauthors of the current manuscript.

The question of authorship-based incentivization raises ethical concerns. Our offer of coauthorship on the basis of a specific output is consistent with other mega-studies in the social sciences, including Salganik et al. (2020). But is it ethical for coauthorship to be granted on the basis of submission of a statistical model only? We believe that it is, provided contributions are communicated transaparently. Fundamentally, contributors contributed the models that were used in the analysis. In addition we did offer coauthors the opportunity to review a completed draft of the paper prior to submission for publication, and we received some feedback; this, we sought to incorporate. But the 22 coauthors had no other opportunities to participate in the Model Challenges. We do see risks however. We see potential ethical issues arising when coauthorship is used to incentivize participants if (1) coauthorship is provided when it is not merited or (2) coauthorship is given to some but not to others who contribute equally (as might be the case if coauthorship were awarded by lottery). In the MC case, the first concern is addressed by the fact that coauthorship depends on a clear criterion for minimal but nevertheless a substantive contribution — producing a model that generates a non-zero stacking weight — and the nature of each coauthor's contribution is communicated clearly via this article's use of the Contributor Roles Taxonomy (CRediT) (from https://credit.niso.org/). A subtle form of the second concern arises because, although in our case there is no lottery component, there is a form of interdependency in that whether a contribution generates a positive weight depends on which other models happen to be provided. That introduces a measure of uncertainty that does not depend solely on the intrinsic properties of the statistical model submitted. But this is not very different than any other research setting, where every theory and every model is in implicit or even explicit competition with every other theory and every other model, and where we always evaluate the performance of each relative to the others in the field. The random element in the MCs is not much different than the randomness that characterizes all our professional engagements.

4.3 Lessons for Aggregating our Knowledge

The best models submitted to the MCs outperform a Lasso-selected benchmark and are over-represented among top performers, as shown in the results of the simulation depicted in Figure 7. To a skeptic of social science, it may not be obvious that social scientists are capable of generating highly predictive models — that, in other words, they possess the ability to accurately explain the social world. However, the sharp drop-off in performance between the best and median models, combined with experts' limited abilities to accurately identify the best-performing models (Figure 5), is cause for concern. If the development of knowledge depends on the abilities of experts to assess the merits of multiple empirically-supported claims, social scientists should address issues of aggregation more systematically.

That experts do far less well than algorithms in model evaluation and aggregation is perhaps surprising. Success in combining intuitions generated by many scholars to explain a common outcome is often viewed as a subtle art requiring deep expertise and insight. Our analysis finds that a statistical algorithm in fact performs better at this in our context. One possible implication of our findings is that scholars could profitably devote more resources to systematizing the models characterizing their explanations so that these can be aggregated using statistical methods. Ensemble procedures seem likely to produce more credible meta-models than informal reasoning, and social scientists do better when their expertise is combined than almost any of them do alone. Of course, predictive accuracy is only one evaluative criterion for social scientific explanations. Yet, it is one for which explanations can be evaluated systematically and, as we demonstrate, is easily amenable to different approaches to comparison and aggregation.

Several aspects of our findings should, we believe, be developed in future work. First, the MCs focused on a public health outcome. Disease mortality may well have been a relatively unfamiliar outcome to many political scientists. Studying the quality of explanations and social scientists' evaluations of explanations for outcomes of longer-standing disciplinary interest will be important to assess the robustness of our findings. Second, lower response

rates in our forecasting task relative to more common forecasting tasks around treatment effects (DellaVigna & Pope, 2018) suggest that social scientists are unaccustomed to evaluating the predictive success for explanations of outcomes. To the extent that this is a useful evaluative criterion for explanations in the social sciences, probing these challenges — or cultivating this ability — seems important. Finally, one may worry that by limiting the number of variables to three or by requiring provision of a verbal explanation of a model, we disadvantaged participants relative to the Lasso-selected benchmark or to the stacking aggregation method. While our discussion has highlighted other advantages of modelers relative to algorithmic approaches, evaluating the sensitivity of our findings to these specific constraints will be important moving forward.

Beyond the MC setting, several features of the procedures we implemented may be worth replicating in more established political science topical domains. In particular, there is a need to evaluate competing theories on common samples using common measures of an outcome. The algorithmic tools that we employ — model comparison based on predictive power and model stacking to generate an aggregate prediction — can easily be implemented in such settings. These forms of model assessment and combination harness the aggregate inputs of social scientists, documenting the strength of collective over individual knowledge. We end with the conclusion that we would all benefit from settings that systematically combined our ideas.

References

Adhikari, B., Cheah, P. Y., & von Seidlein, L. (2022). Trust is the common denominator for COVID-19 vaccine acceptance: A literature review. *Vaccine: X*, *12*, 100213.

Adkins, T. L., & Smith, J. (2020). Will covid-19 kill democracy? *Foreign Policy*, *September 18*.

Baden, L. R., El Sahly, H. M., Essink, B., Kotloff, K., Frey, S., Novak, R., Diemert, D., Spector, S. A., Rouphael, N., Creech, C. B., McGettigan, J., Khetan, S., Segall, N., Solis, J., Brosz, A., Fierro, C., Schwartz, H., Neuzil, K., Corey, L., … Zaks, T. (2021). Efficacy and safety of the mrna-1273 sars-cov-2 vaccine. *New England Journal of Medicine*, *384*, 403–416.

Basta, N., & Moodie, E. (2021, April 28). *Covid-19 vaccine tracker* (tech. rep.).

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

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

Bosancianu, M., Yii, D. K., Hilbig, H., Humphreys, M., KC, S., Lieber, N., & Scacco, A. (2020). *Political and social correlates of covid-19 mortality* [SocArXiv].

Burn-Murdoch, J. (2022, December). *Coronavirus tracked: See how your country compares*. https://ig.ft.com/coronavirus-chart/?areas=gbr&areas=usa&areas=bra&areas=ita&areas=mex&areas=swe&areasRegional=usny&areasRegional=usnj&cumulative=0&logScale=1&per100K=1&startDate=2022-01-01&values=deaths

Champoux-Paillé, L. (2020). *The world needs more women leaders - during covid-19 and beyond*. dvance-lexis-com.stanford.idm.oclc.org/api/document?collection=news&id=urn%3acontentItem%3a61DS-TRK1-JB75-939V-00000-00&context=1519360&identityprofileid=7N5PFF56124

COVID-19 National Preparedness Collaborators. (2022). Pandemic preparedness and covid-19: An exploratory analysis of infection and fatality rates, and contextual factors associated with preparedness in 177 countries, from Jan 1, 2020 to Sept 30, 2021. *Lancet*, *399*, 1489–512.

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

DellaVigna, S., & Pope, D. (2018). Predicting experimental results: Who knows what? *Journal of Political Economy*, *126*(6), 2410–2456.

DellaVigna, S., Pope, D., & Vivalt, E. (2019). Predict science to improve science. *Science*, *366*(6464), 428–429.

Folegatti, P. M., Ewer, K. J., Aley, P. K., Angus, B., Becker, S., Belij-Rammerstorfer, S., Bellamy, D., Bibi, S., Bittaye, M., Clutterbuck, E. A., Dold, C., Faust, S. N., Finn, A., Flaxman, A. L., Hallis, B., Heath, P., Jenkin, D., Lazarus, R., Makinson, R., … Pollard, A. J. (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*, *396*, 1979–1993.

Grossman, G., Dinneen, W., & Torreblanca, C. (2025). *The evolving landscape of political science: Two decades of scholarship in a growing discipline* [Unpublished paper.].

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

Logunov, D. Y., Dolzhikova, I. V., Shcheblyakov, D. V., Tukhvatulin, A. I., Zubkova, O. V., Dzharullaeva, A. S., Kovyrshina, A. V., Lubenets, N. L., Grousova, D. M., Erokhova, A. S., Botikov, A. G., Izhaeva, F. M., Popova, O., Ozharovskaya, T. A., Esmagambetov, I. B., Favorskaya, I. A., Zrelkin, D. I., Voronina, D. V., Shcherbinin, D. N., … Gintsburg, A. L. (2021). Safety and efficacy of an rad26 and rad5 vector-based heterologous prime-boost COVID-19 vaccine: An interim analysis of a randomised controlled phase 3 trial in Russia. *The Lancet*, *397*, 671–681.

Lynch, J. (2020). *Regimes of inequality: The political economy of health and wealth*. Cambridge University Press.

Lynch, J., Bernhard, M., & O'Neill, D. (2022). Pandemic politics. *Perspectives on Politics*, *20*(2), 389–394.

Mulligan, M. J., Lyke, K. E., Kitchin, N., Absalon, J., Gurtman, A., Lockhart, S., Neuzil, K., Raabe, V., Bailey, R., Swanson, K. A., Li, P., Koury, K., Kalina, W., Cooper, D., Fontes-Garfias, C., Shi, P.-Y., Türeci, Ö., Tompkins, K. R., Walsh, E. E., … Jansen, K. U. (2021). Phase i/ii study of covid-19 rna vaccine bnt162b1 in adults. *Nature*, *586*, 589–593.

Oppenheimer, M., O'Neill, B. C., Webster, M., & Agrawala, S. (2007). The limits of consensus. *Science*, *317*(5844), 1505–1506.

Polack, F. P., Thomas, S. J., Kitchin, N., Absalon, J., Gurtman, A., Lockhart, S., Perez, J. L., Pérez Marc, G., Moreira, E. D., Zerbini, C., Bailey, R., Swanson, K. A., Roychoudhury, S., Koury, K., Li, P., Kalina, W. V., Cooper, D., Frenck, R. W. J., Hammitt, L. L., … Gruber, W. C. (2021). Safety and efficacy of the bnt162b2 mrna covid-19 vaccine. *New England Journal of Medicine*, *385*, 1761–1773.

Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., Altschul, D. M., Brand, J. E., Carnegie, N. B., Compton, R. J., Datta, D., Davidson, T., Filippova, A., Gilroy, C., Goode, B. J., Jahani, E., Kashyap, R., Kirchner, A., McKay, S., … McLanahan, S. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, *117*(15), 8398–8403. https://doi.org/10.1073/pnas.1915006117

Sapienza, A., & Falcone, R. (2022). The role of trust in COVID-19 vaccine acceptance: Considerations from a systematic review. *International Journal of Environmental Research and Public Health*, *20*, 665.

Solís Arce, J. S., Warren, S. S., Meriggi, N. F., Scacco, A., McMurry, N., Voors, M., Syunyaev, G., Malik, A. A., Aboutajdine, S., Adeojo, O., Anigo, D., Armand, A., Asad, S., Atyera, M., Augsburg, B., Awasthi, M., Ayesiga, G. E., Bancalari, A., Nyqvist, M. B., … Omer, S. B. (2021). Covid-19 vaccine acceptance and hesitancy in low- and middle-income countries. *Nature medicine*, *27*, 1385–1394.

Tetlock, P. E. (2005). *Expert political judgement: How good is it? How can we know?* Princeton University Press.

Tetlock, P. E., & Gardner, D. (2016). *Superforecasting: The art and science of prediction*. Crown.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, *58*(1), 267–288.

Voysey, M., Clemens, S. A. C., Madhi, S. A., Weckx, L. Y., Folegatti, P. M., Aley, P. K., Angus, B., Baillie, V. L., Barnabas, S. L., Bhorat, Q. E., Bibi, S., Briner, C., Cicconi, P., Collins, A. M., Colin-Jones, R., Cutland, C. L., Darton, T. C., Dheda, K., Duncan, C. J. A., … Pollard, A. J. (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*, *397*, 99–111.

Walsh, J. (2020). Poll: Most Republicans say covid threat overblown, U.S.handled outbreak well. *Forbes*, *October 19*.

WHO. (2021, April 23). *Draft landscape and tracker of covid-19 candidate vaccines* (tech. rep.).

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

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

Appendices

A Sources of Data on Covid Mortality

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

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

- **India challenge**: Government of India (https://www.mygov.in/corona-data/covid19-statewise-status/), accessed November 16, 2020; March 23, 2021; September 7, 2021; March 2, 2022; and October 18, 2022.

- **Mexico challenge**: Government of Mexico (https://coronavirus.gob.mx/datos/#DownZCSV), accessed November 16, 2020; March 23, 2021; September 8, 2021; March 4, 2022; and October 2, 2022.

- **United States challenge**: The COVID Tracking Project at *The Atlantic* (https:/covidtracking.com/data/download/all-states-history.csv), accessed November 16, 2020 and March 23, 2021; The COVID-19 Response at the Centers for Disease Control and Prevention (CDC) (https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36), accessed September 14, 2021; March 3, 2022; and October 18, 2022.

Figure A.1 shows the outcome measure for the crossnational challenge. The left panel shows the evolution of logged deaths per million over the relevant period. The two vertical lines denote (1) the data shown to modelers during the Model Challenge and (2) the date when we evaluate the predictions made by the models they submitted (August 31, 2021). Each horizontal line represents a country. To illustrate the changes in COVID-19 mortality that participants predicted, we depict the three countries at the 10^{th}, 50^{th}, and 90^{th} percentiles in (percent) change in COVID-19 mortality between November 16, 2020 and August 30, 2021.

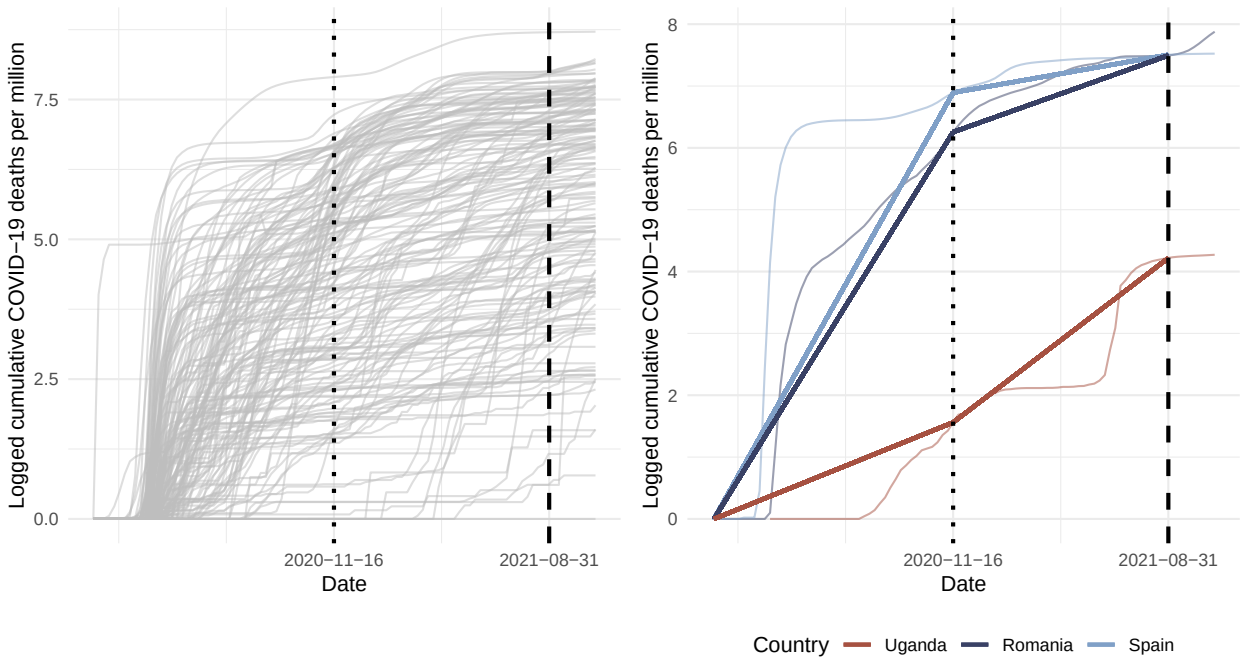


Figure A.1: Outcome data for the crossnational challenge. The left panel depicts the growth of logged cumulative COVID-19 deaths per million. Each line represents a country. Vertical lines reflect (1) the data provided in the MC and (2) the main outcome, mortality as of August 31, 2021. The right panel shows changes in the outcome between November 16, 2020 and August 31, 2021 for countries at the first decile (Spain), median (Romania), and top decile (Uganda).

The countries are Spain, Romania, and Uganda, respectively.

B Distribution of participants across challenges

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

Table B.1: Tally of model challenge submissions and participants (raw numbers). "All" may be less than the sum of the four challenges because some modellers submitted to multiple challenges.

C Submitted models

Table C.1 provides summaries of all submitted models. They are ranked, within challenge, by the R^2_{loo} metric (evaluated for general models using leave-one-out predictions; see Appendix G).

Table C.1: Model composition and performance by challenge

	Model Name	Variables included and functional form specified	R^2_{loo}	Stacking weight
Crossnational, parameterized				
1	Trust in Authoritarian Government	deaths_per_mio_log = 4.75 + 0.9 * acc_sanitation - 1.25 * trust_gov	0.141	0.437
2	Populism and Social Trust	deaths_per_mio_log = 4.8 + 0.9*electoral_pop + 0.9*electoral_pop*trust_people + 0.75*life_exp_2017 -0.9*trust_people	0.097	0.155
3	Liberalism, Capitalism, and Media Independence	deaths_per_mio_log = 5 + 0.7*property_rights -0.7*trust_gov -0.1*vdem_mecorrpt	0.008	0.000
4	Social and Institutional Trust	deaths_per_mio_log = 4.5 -0.5*gov_effect + life_exp_2017 -0.5*trust_people + 0.5*trust_people*gov_effect	-0.079	0.000
5	Government Capacity and Development	deaths_per_mio_log = 5 + 0.1*gdp_pc -0.5*gov_effect + 0.5*gov_effect^2 -0.9*trust_gov	-0.121	0.000
6	Government Capacity and Social Inequality	deaths_per_mio_log = 4 + 0.2*gini + 0.15*gini^2 -0.6*gov_effect -0.2*gov_effect^2 + 2*healthcare_qual	-0.356	0.168
7	Perverse Development	deaths_per_mio_log = 4 + 0.4*acc_sanitation + 0.6*hdi	-0.369	0.000
8	Health Equality	deaths_per_mio_log = 4 + 1.5*acc_sanitation + 0*acc_sanitation*respond_index -0.1*health_equality + 0*health_equality*acc_sanitation + 0.01*health_equality*acc_sanitation*respond_index + 0*health_equality*respond_index + 0*respond_index	-0.390	0.168
9	Competitiveness of Executive Recruitment	deaths_per_mio_log = 1.46*acc_sanitation + 0.85*acc_sanitation^2 -0.14*urban -0.04*urban^2 + 1.37*xrcomp_2018	-0.946	0.000
10	Development and Trust	deaths_per_mio_log = 4 + 1*hdi + 0.5*hdi*trust_people + 0.25*share_older -0.5*share_older*hdi + 0.25*share_older*hdi*trust_people -1*share_older*trust_people -0.55*trust_people	-0.993	0.000
11	Pandemic Readiness	deaths_per_mio_log = 9 -2*acc_sanitation + 1.4*infection -1.2*trust_gov	-2.201	0.058
12	Social and Political Stability	deaths_per_mio_log = 1.5 -0.2*gini -0.3*pr -0.1*trust_people + 0.1*trust_people*gini + 0.1*trust_people*pr	-5.581	0.000
13	Polarization and Populism	deaths_per_mio_log = 1*electoral_pop -0.5*polar_rile + 0.5*trust_people	-7.557	0.000

Table C.1: Model composition and performance by challenge (*continued*)

ID	Model	Functional Form	R^2_{loo}	Stacking Weight
14	Language and Culture	deaths_per_mio_log = 3.5 + 1.5*idv + 0.05*inflectional_ftr	-1322.027	0.013
India, general				
1	Health Sector Capacity	deaths_per_mio_log ~ β_0 + β_1*pandemic_prep + β_2*pct_poor + β_3*pandemic_prep*pct_poor + β_4*hosp_beds_pc + β_5*pandemic_prep*hosp_beds_pc + β_6*pct_poor*hosp_beds_pc + β_7*pandemic_prep*pct_poor*hosp_beds_pc	0.363	0.451
2	Interactions and Political Pressures	deaths_per_mio_log ~ β_0 + β_1*gdp_pc + β_2*urban_pct + β_3*election_margin	0.306	0.231
3	Urbanisation and Healthcare	deaths_per_mio_log ~ β_0 + β_1*gdp_pc + β_2*public.health.total.budget.2015 + β_3*urban_pct	0.301	0.034
4	Business and Density	deaths_per_mio_log ~ β_0 + β_1*minority_pct + β_2*gdp_pc + β_3*urban_pct	0.295	0.000
5	GDP, TB Prevalence, and Anti-immigration Attitudes	deaths_per_mio_log ~ β_0 + β_1*pct_anti_immig + β_2*tb_per_100k + β_3*gdp_pc	0.204	0.000
6	Minority Representation and Urbanization	deaths_per_mio_log ~ β_0 + β_1*reserve_proportion + β_2*urban_pct + β_3*reserve_proportion*urban_pct	0.094	0.260
7	Government Capacity	deaths_per_mio_log ~ β_0 + β_1*average_events_per_state + β_2*leader_experience	-0.145	0.000
India, parameterized				
1	Business and Density	deaths_per_mio_log = 4.7 + 0.2*gdp_pc -0.2*minority_pct + 0.5*urban_pct	-1.668	0.945
2	Urbanisation and Health Care	deaths_per_mio_log = 5.35 + 0.41*gdp_pc -13*public.health.total.budget.2015 + 0.44*urban_pct	-2.021	0.000
3	Health Sector Capacity	deaths_per_mio_log = 4.3 + 0.4*hosp_beds_pc + 1.8*pandemic_prep + 2*pandemic_prep*hosp_beds_pc + 1.5*pandemic_prep*pct_poor + 2.5*pandemic_prep*pct_poor*hosp_beds_pc -0.35*pct_poor -0.25*pct_poor*hosp_beds_pc	-2.154	0.055
4	Minority Representation and Urbanization	deaths_per_mio_log = 4.2 -0.6*reserve_proportion -0.6*reserve_proportion*urban_pct + 0.2*urban_pct	-3.267	0.000
5	Interactions and Political Pressures	deaths_per_mio_log = 4.25 + 0.05*election_margin + 0.4*gdp_pc + 0.4*urban_pct	-3.684	0.000
Mexico, general				
1	Political Leadership, Poverty, and Obesity	deaths_per_mio_log ~ β_0 + β_1*election_margin + β_2*pct_poor + β_3*obesity	0.371	0.000
2	Social Trust and Catholicism	deaths_per_mio_log ~ β_0 + β_1*pct_catholic + β_2*election_margin + β_3*trust_people + β_4*pct_catholic*trust_people	0.347	0.360

Table C.1: Model composition and performance by challenge (*continued*)

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

Table C.1: Model composition and performance by challenge (*continued*)

ID	Model	Functional Form	R^2_{loo}	Stacking Weight
7	Ethnicity, Inequality, and Healthcare Capacity	deaths_per_mio_log ~ β_0 + β_1*gini + β_2*hosp_beds_pc + β_3*ethnic_frac_score	0.344	0.000
8	Social Contact	deaths_per_mio_log ~ β_0 + β_1*pct_religious + β_2*pct_poor + β_3*pop_density	0.332	0.000
9	Institutional and Social Trust	deaths_per_mio_log ~ β_0 + β_1*pop_density + β_2*trust_gov + β_3*gini	0.325	0.000
10	Community Equality and Trust	deaths_per_mio_log ~ β_0 + β_1*gini + β_2*civil_society + β_3*trust_people	0.317	0.000
11	Religion, Economic Inequality, and Minority Status	deaths_per_mio_log ~ β_0 + β_1*pct_religious + β_2*minority_pct + β_3*gini	0.299	0.000
12	Inequality and Urbanity	deaths_per_mio_log ~ β_0 + β_1*urban_pct + β_2*gini	0.227	0.000
13	Poverty and Social Exclusion	deaths_per_mio_log ~ β_0 + β_1*minority_pct + β_2*pct_poor + β_2*gini	0.203	0.000
14	Institutional Trust and Race	deaths_per_mio_log ~ β_0 + β_1*trust_gov + β_2*pop_density + β_3*minority_pct + β_4*minority_pct^2	0.201	0.024
15	Vaccination Coverage	deaths_per_mio_log ~ β_0 + β_1*s_diffs + β_2*share_older + β_3*Influenza_vaccination_age_65*share_older	0.020	0.000
16	Population Health, Religiosity, and Compliance	deaths_per_mio_log ~ β_0 + β_1*pct_religious + β_2*resp_disease_prev + β_3*share_older	-0.085	0.000
17	Government Experience	deaths_per_mio_log ~ β_0 + β_1*leader_experience + β_2*corrected_score	-0.121	0.000
USA, parameterized				
1	Vaccination Coverage	deaths_per_mio_log = 7.3 -0.31*Influenza_vaccination_age_65*share_older -0.23*s_diffs -0.17*share_older	-0.102	0.974
2	Inequality and Polarization	deaths_per_mio_log = 6 + 0.5*gini + 0.5*party_leg_right -0.4*party_leg_right*gini + 0.4*pop_density -0.2*pop_density*gini	-5.498	0.026
3	Social Contact	deaths_per_mio_log = 6.3 + 0.1*pct_poor + 0.25*pct_religious + 0.5*pop_density	-5.555	0.000
4	Population Differences	deaths_per_mio_log = 6.3 + 0.3*gini + 0*gini^2 + 0.3*pct_religious -0.1*pct_religious^2 + 0.2*pop_density + 0.01*pop_density^2	-6.166	0.000
5	Inequality and Urbanity	deaths_per_mio_log = 6.1 + 0.6*gini + 0.1*urban_pct	-8.265	0.000

Notes: Lasso and Epidemiological models not shown in this table; displayed in Tables L.1 and L.2, respectively.

The Lasso model receives positive stacking weight in the India and Mexico general challenges.

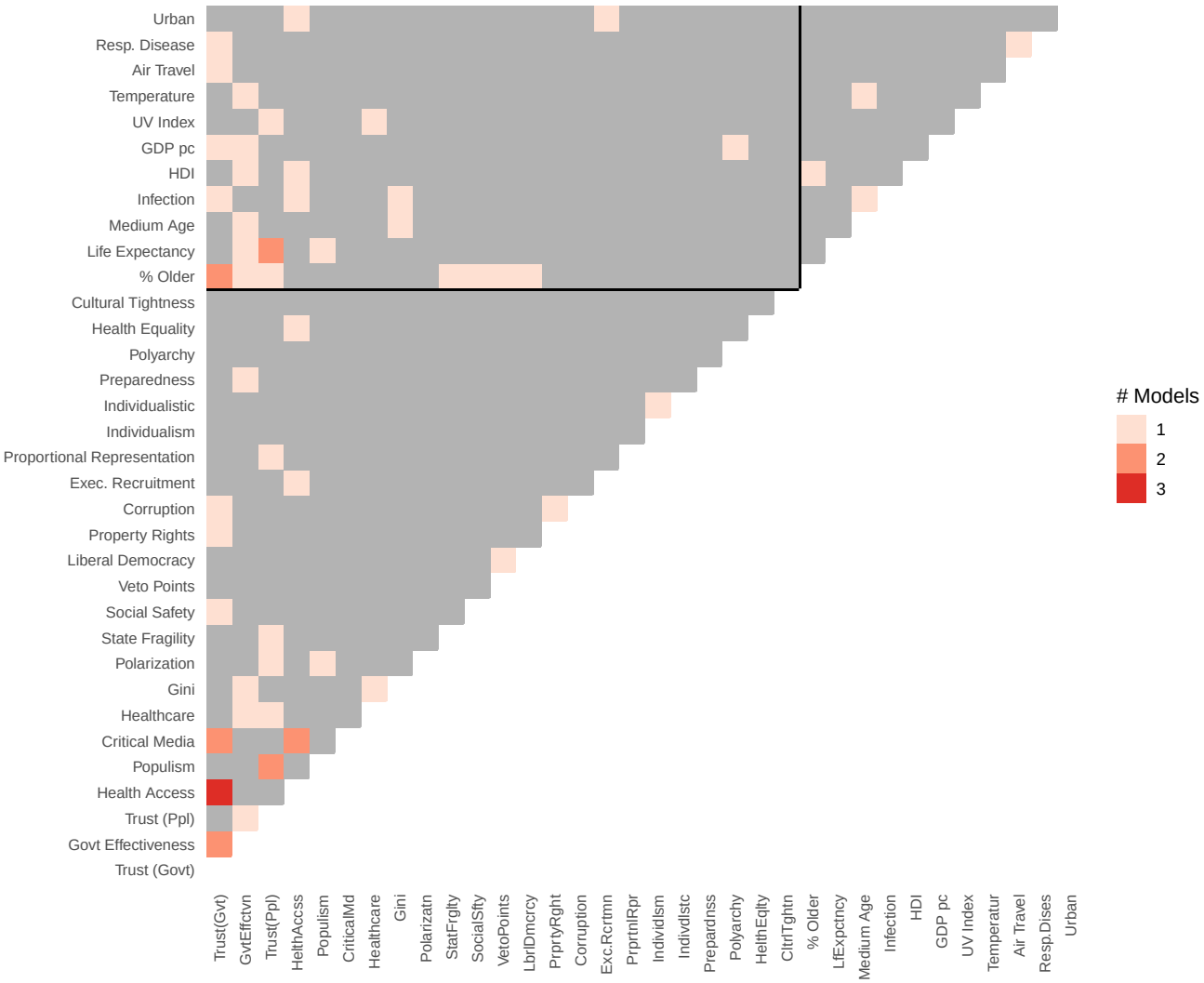


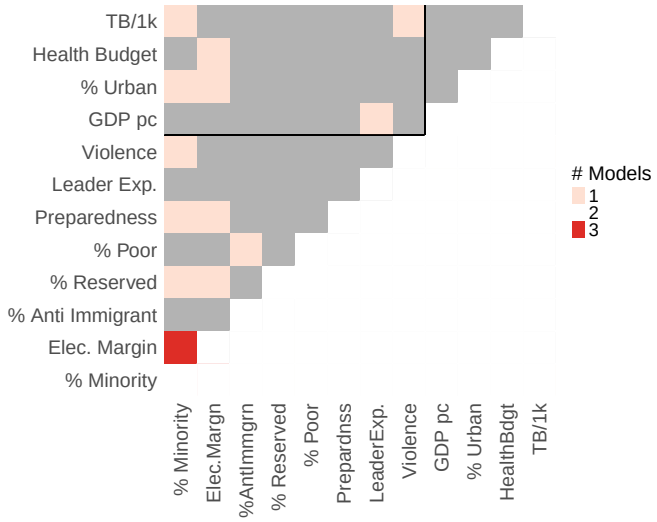
Figure D.1: Pairwise combinations of variables submitted to the crossnational MC. The lower left quadrant shows social and political variables provided in the MC. The upper right quadrant shows other variables provided in the MC. Variable definitions available at https://osf.io/pgydn.

D Pairwise Combinations of Predictors

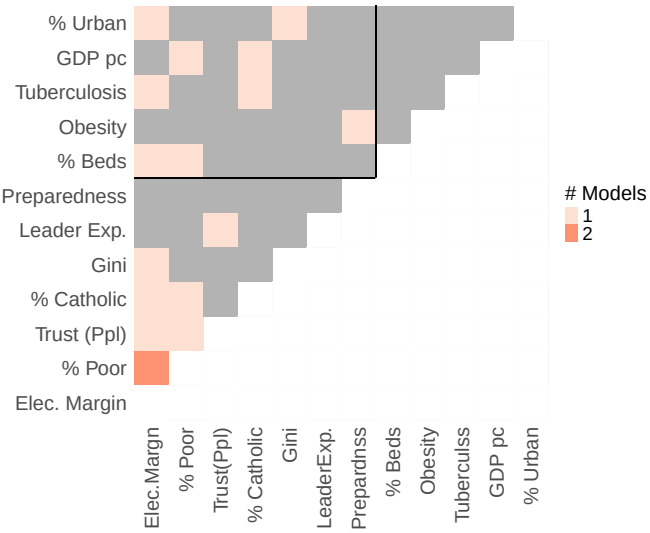
Figure D.1 shows a visual summary of the most common predictors found in crossnational models.

It first orders political and social variables and then orders other — mostly health and demographic — variables, all by how often they appear in submissions. Color coding indicates how frequently pairs of variables were entered together. The data depicted in the figure shows that

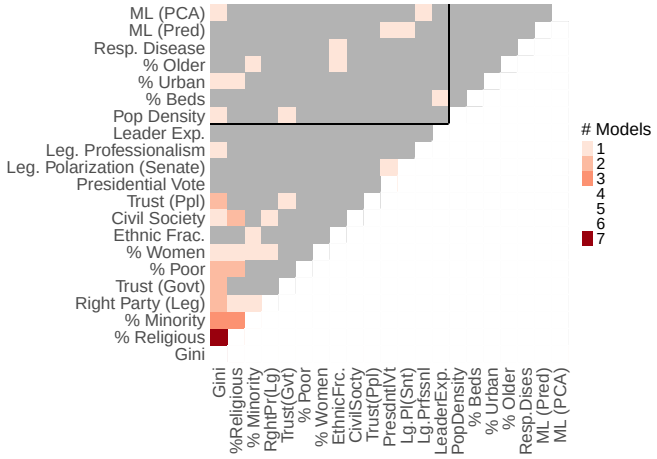
Figure D.2 depicts the pairwise combinations of predictors in the country-specific challenges analogous to Figure D.1. See the discussion of Figure D.1 in the main text for information on the interpretation of these plots.



(a) Variable combination for India.



(b) Variable combination for Mexico.



(c) Variable combination for the USA.

Figure D.2: Pairwise combinations of variables submitted to the country-specific challenges.

E Stacking and Model Selection Results by Country

We now report the country-specific results of our model selection exercise. Figures E.1-E.3 are analogous to Figure 5 in the main text. Note that we only show general model results for the country-specific challenges, as before.

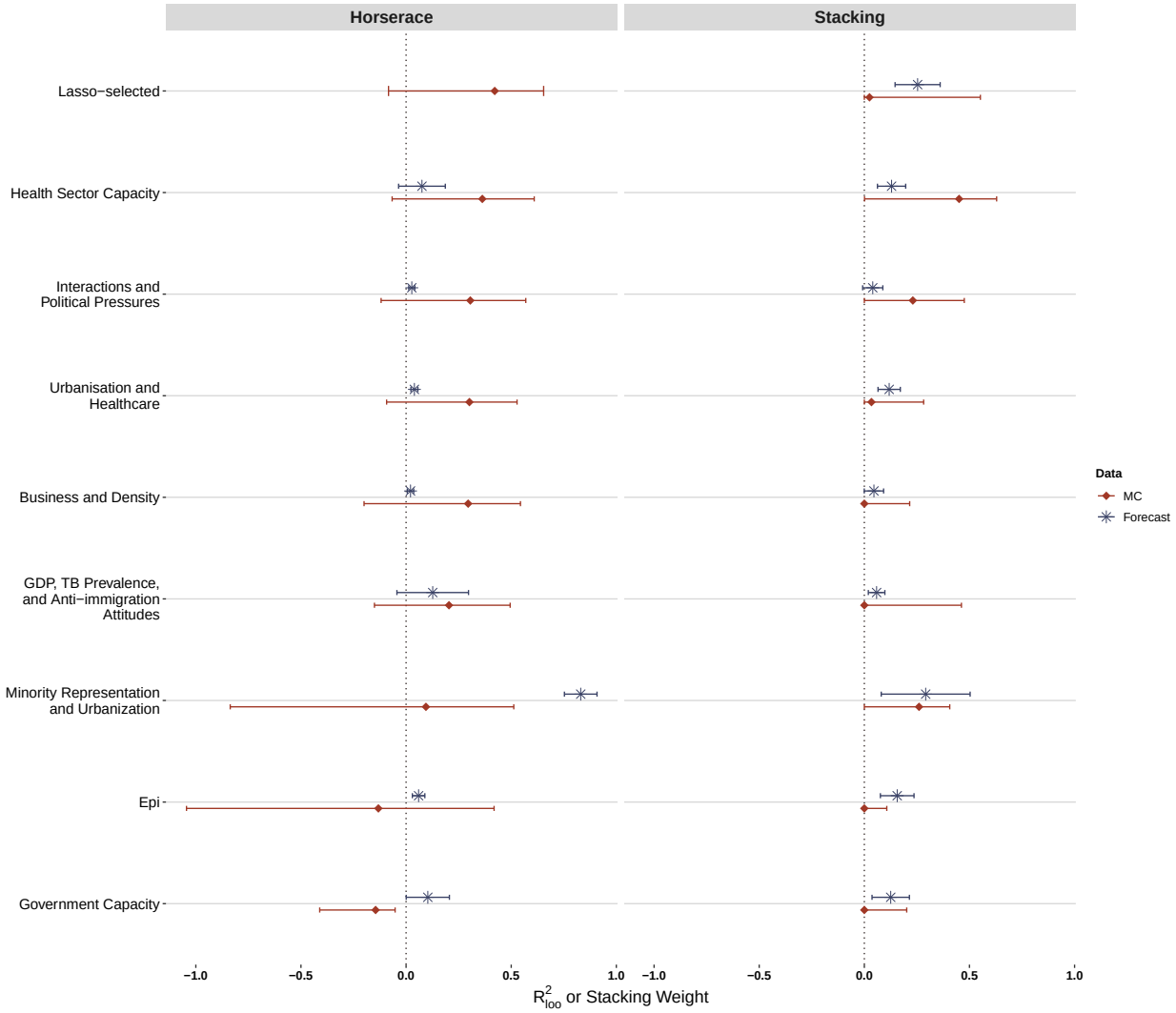


Figure E.1: Model selection using four methods for the general models from India.

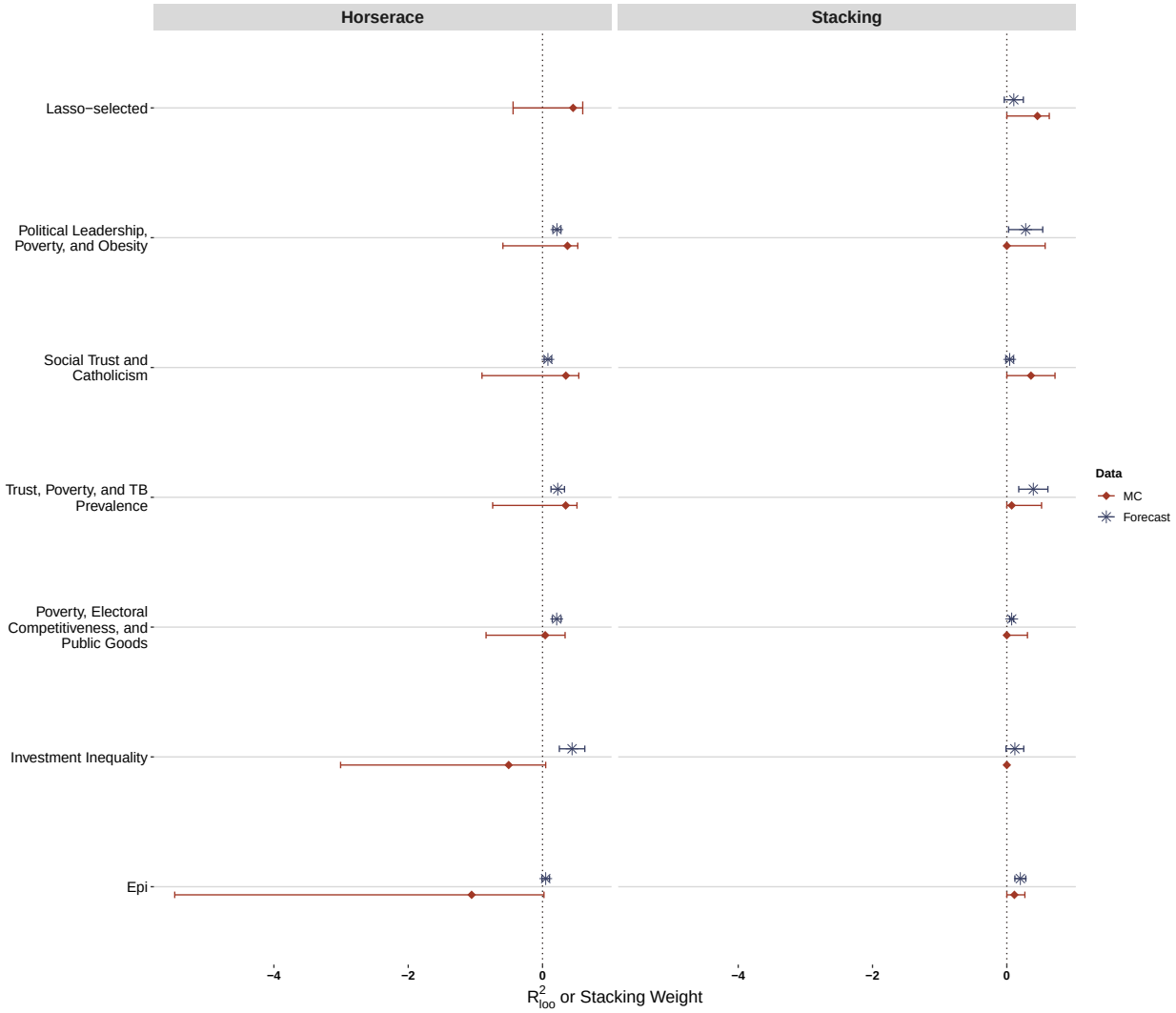


Figure E.2: Model selection using four methods for the general models from Mexico.

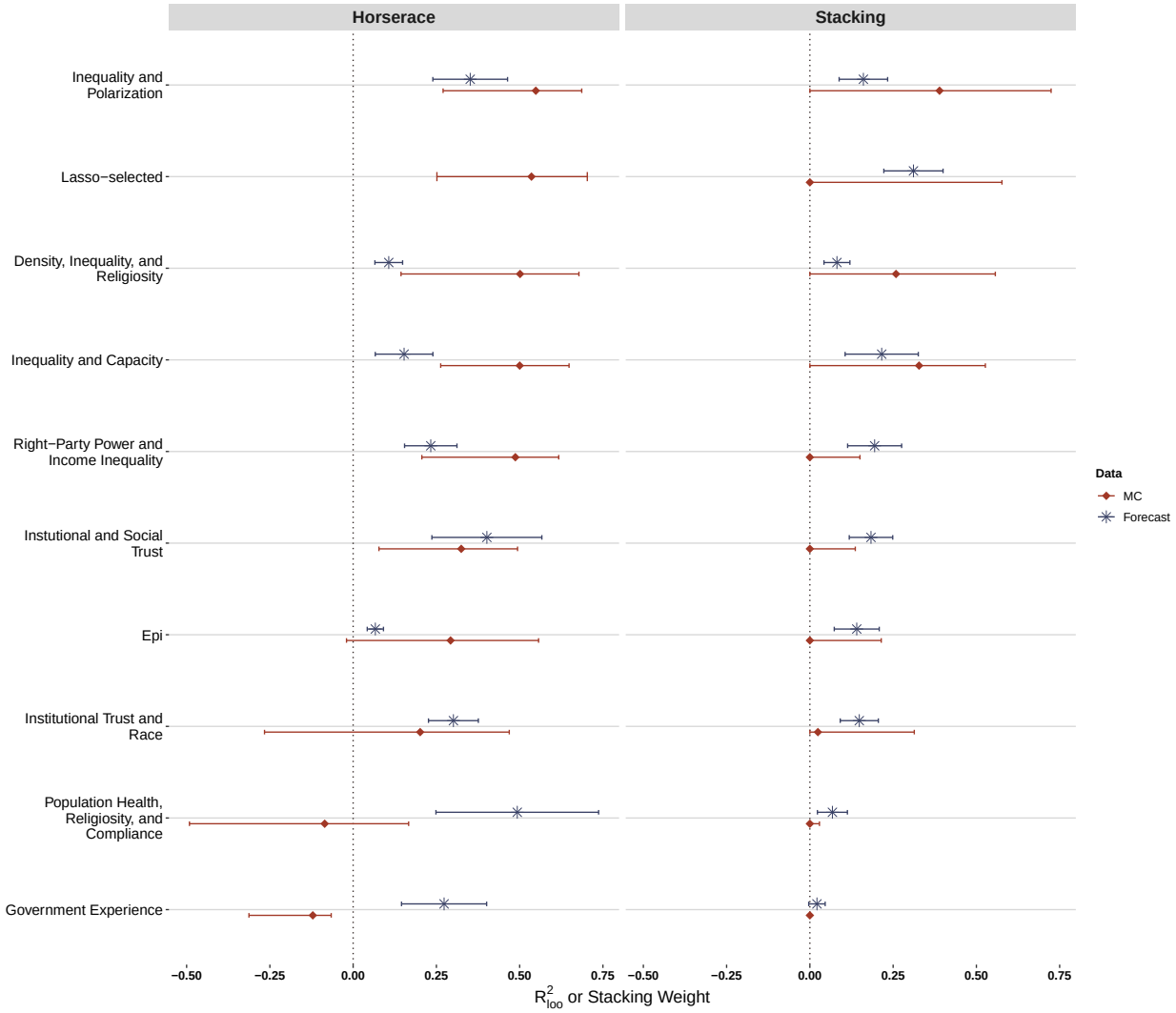


Figure E.3: Model selection using four methods for the general models from the US.

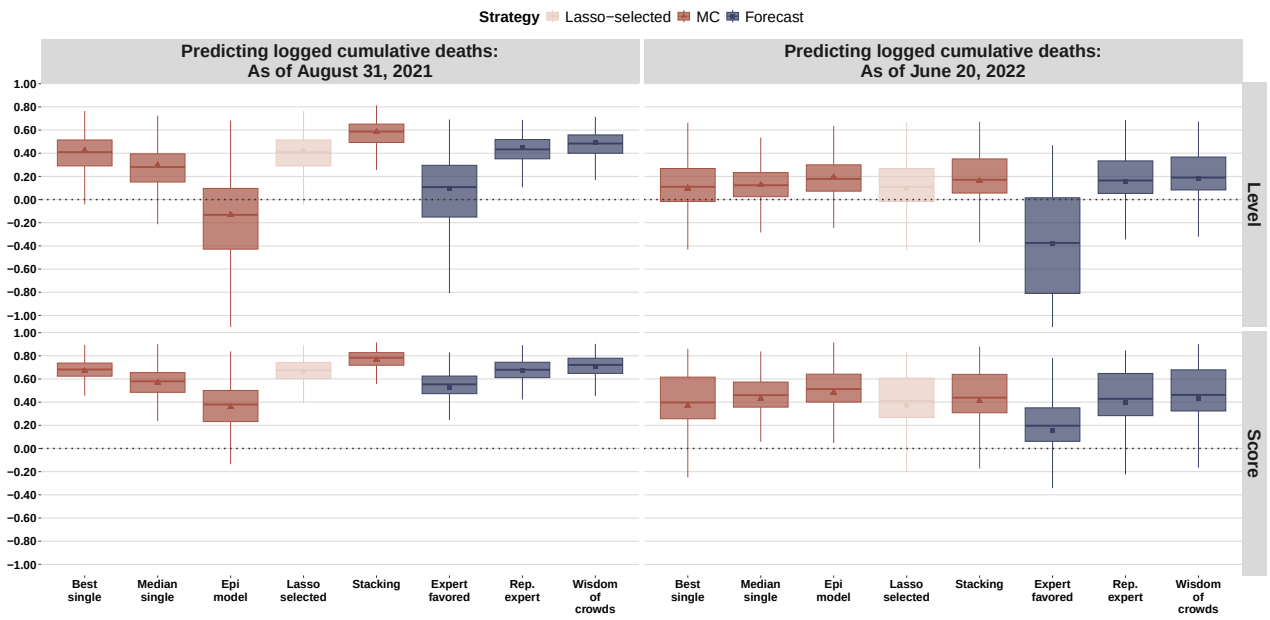


Figure F.1: Prediction aggregation metrics for general models for India.

F Aggregation Results for Other Challenges

We now report the results of aggregating across the three country-specific challenges. Figures F.1-F.3 report the full results for each challenge analogous to 6 in the main text. Note that we only show results for the general models from the country-specific challenges as before.

G Measure of Predictive Accuracy

We use a metric of the predictive accuracy that resembles the R^2 but is evaluated (for general models) using leave-one-out predictions. We do this even though general models make predictions about future (out-of-sample) COVID-19 mortality; the reason is that parameters of general models are estimated using the (out-of-sample) August 2021 outcome data. The use of out-of-sample predictions is unnecessary for the parameterized models, where modellers also predicted the values of these parameters.

We label the metric that we use R^2_{loo}. It is generated via Equation G.1. It is a rescaling of mean squared error (MSE) but offers a more easily interpretable scale. A R^2_{loo} of 1 constitutes

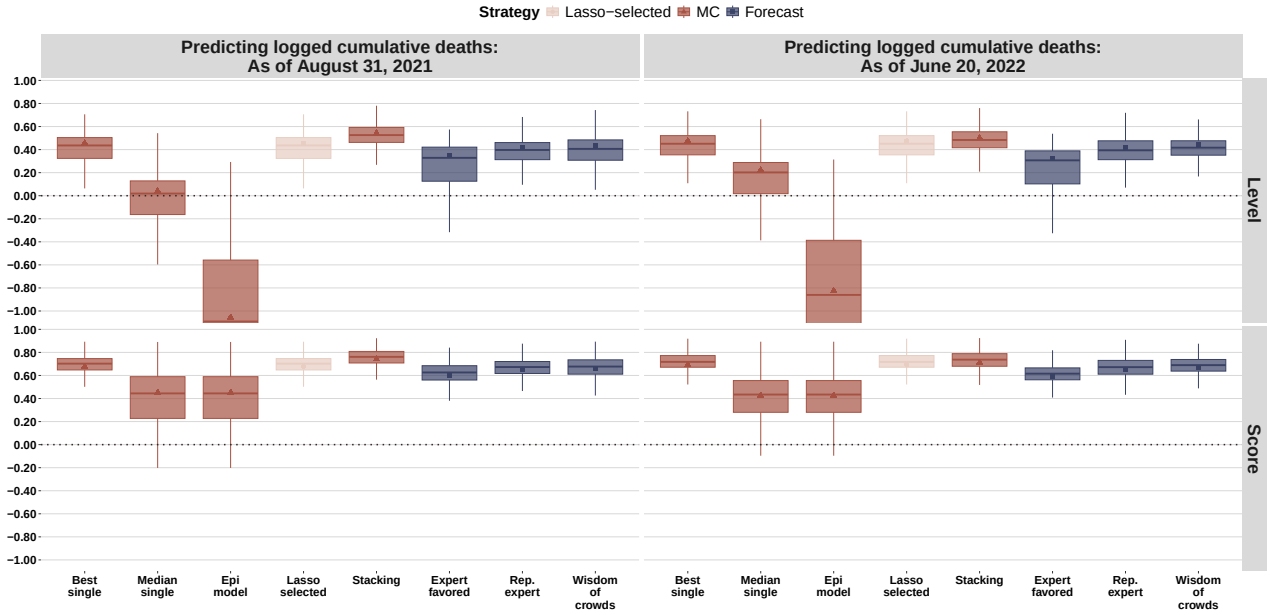


Figure F.2: Prediction aggregation metrics for general models for Mexico.

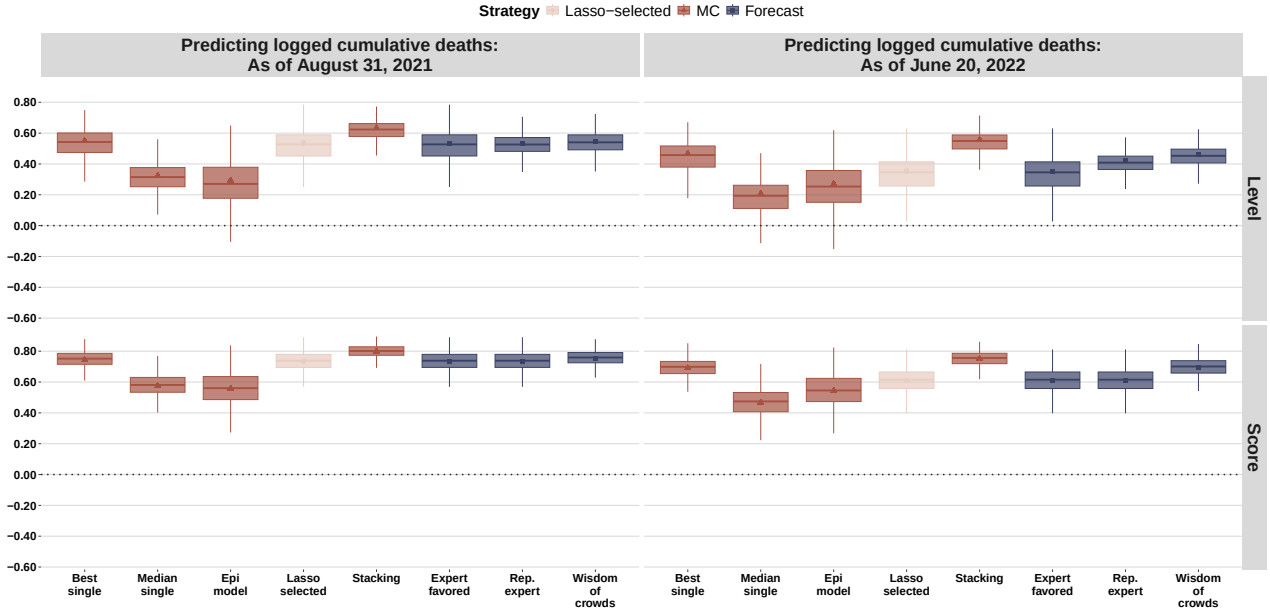


Figure F.3: Prediction aggregation metrics for general models for the USA.

a perfect prediction and a R^2_{loo} of 0 means a model does no better than a prediction of the mean outcome for all units.[5] We also compare the correlation between leave-one-out predictions and observed outcomes, which abstracts from the levels (or intercepts) of the predictions. This is particularly useful for evaluating parameterized models when predicted intercepts depart substantially from realized COVID-19 mortality.

In the following expressions, let $\boldsymbol{x_i}$ denote a vector of explanatory variables for unit i and $\hat{f}^k_{-i}$ the predictive model k trained on data that excludes unit i. Then the leave-one-out prediction for unit i under model k is $\widehat{y}_{ik} = \hat{f}^k_{-i}(\boldsymbol{x_i})$. The (squared) error for unit i produced by outcome k is $(\overline{y}_i - y_i)^2$. Our measure of accuracy combines these errors across units according to:

$$R^2_{loo,k} = 1 - \frac{\sum_{i=1}^{N}(\widehat{y}_{ik} - y_i)^2}{\sum_{i=1}^{N}(\overline{y}_i - y_i)^2} \tag{G.1}$$

H Model Stacking

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

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

Intuitively, a vector of weights w placed on models, results in an aggregated prediction for the unit $y_i^{\text{stacking}}(w) = \sum_k w_k \widehat{y}_{ik}$ and loss is assessed by how far the vector $y^{\text{stacking}}(w)$ is from the observed outcomes y. We estimate the stacking weights employed in Figures 5 and 6 using Equation (H.1).

[5] Values below 0 are possible if models perform worse than this. The measure has no lower bound.

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

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.

I Aggregating Forecasts

I.1 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:

$$\widehat{y}_i^j = \sum_k \widehat{w}_k^j \widehat{y}_{ik}^{\text{loo}} \tag{I.1}$$

We use the leave-one-out designation here to remind readers that forecasts were only elicited over general models where we employ the leave-one-out predictions in all metrics of prediction accuracy. We can plug this into Equation (J.1) 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:

$$w^r = \{w^j | v^j = \text{median}(v^h)_{h \in H}\} \tag{I.2}$$

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

I.2 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:

$$w_k^c = \frac{\sum_j \widehat{w}_k^j}{\sum_k \sum_j \widehat{w}_k^j} \tag{I.3}$$

This yields model predictions given by:

$$\widehat{y}_i^c = \sum_k w_k^c \widehat{y}_{ik}^{\text{loo}} \tag{I.4}$$

J Defining Model Success for 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:

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

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

For the level approach, we evaluate (J.1) 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 R^2_{loo}. 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 R^2_{loo}, $v_k \in (-\infty, 1]$. Higher values of v_k indicate more accurate predictions.

For the score approach, we evaluate (J.1) 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.

K Information on the Stacking Forecast

Experts randomized into the stacking forecast read the following instructions:

> *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 com-*

bined 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.

L Information on the Creation of the Lasso Benchmark, Epidemiological Models

The following variables were selected by the Lasso procedure for each challenge:

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

Table L.1: Lasso models for each challenge. The parameterized form was fit on outcome data as of November 16, 2020.

Table L.2 reports the epidemiological models used as a benchmark in each challenge.

M Simulating Model Selection by Machine

We extend the analysis from Figure 7 to the other challenges in this section. First, we outline our algorithm for sampling of models. The sampling strategy parallels the format of the MCs and the Shiny app that was provided to modelers. For each challenge we:

1. Randomly sample three predictors from the MC predictors.
2. Randomly select one type of model: polynomial (quadratic), interaction, or neither, each with probability 1/3.
3. For the selected type of model, we follow the Shiny menu of options to select terms to be excluded from the statistical model. We do so by generating a Bernoulli random

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

Table L.2: The epidemiological models used as a benchmark in each challenge. The parameterized form was fit on outcome data as of November 16, 2020.

variable (with $p = 0.5$) for each term and including the term if the draw takes the value 1 and omitting the term if the draw takes the value 0.

Following this algorithm, we sample $5000 \times M_c$ (M_c is the total number of user-submitted models in each challenge c) models per challenge. Figure M.1 shows the performance of the randomly generated models relative to the user submitted models in each MC. Table M.1 presents the corresponding summary statistics for the two types of models per MC. Figure M.2 compares the performance of the stacking model estimated on the user submitted relative to the stacking model estimated on equivalent-sized sets of randomly generated models. In three of four challenges, our estimated stacking models outperform every simulated model. In the Mexico challenge, 2.1 percent percent of the simulated models outperform our estimated stacking models ($p = 0.021$). Consistent with our interpretation of Figure 7, this suggests that the best user-submitted models outperform machine-selected models. These highly-predictive models yield performance gains of the stacking meta-model.

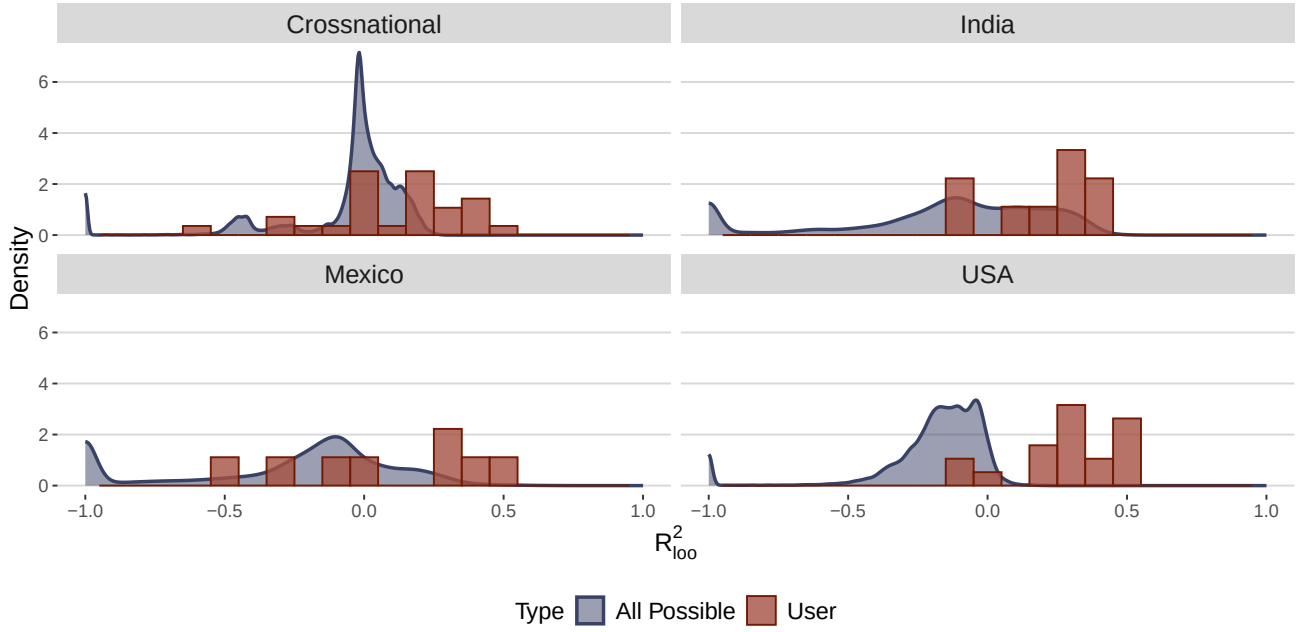


Figure M.1: Horserace simulation. The density plots represent the distribution of R^2_{loo}'s from 5,000 sets of simulated three-predictor models in the common MC datasets for each challenge.

Table M.1: Comparing model performances between user-submitted and simulated three-predictor models.

	User-submitted						Simulated					
Challenge	Mean	SD	95% CI (L)	95% CI (U)	Skew	Kurtosis	Mean	SD	95% CI (L)	95% CI (U)	Skew	Kurtosis
Crossnational	0.11	0.24	-0.39	0.44	-0.83	3.55	-0.06	0.24	-1	0.19	-2.37	9.18
India	0.19	0.21	-0.14	0.41	-0.74	2.08	-0.17	0.40	-1	0.35	-0.88	2.83
Mexico	-0.04	0.49	-0.90	0.44	-0.79	2.51	-0.29	0.40	-1	0.29	-0.68	2.37
USA	0.30	0.20	-0.11	0.54	-0.80	2.90	-0.19	0.20	-1	0.02	-2.49	10.24

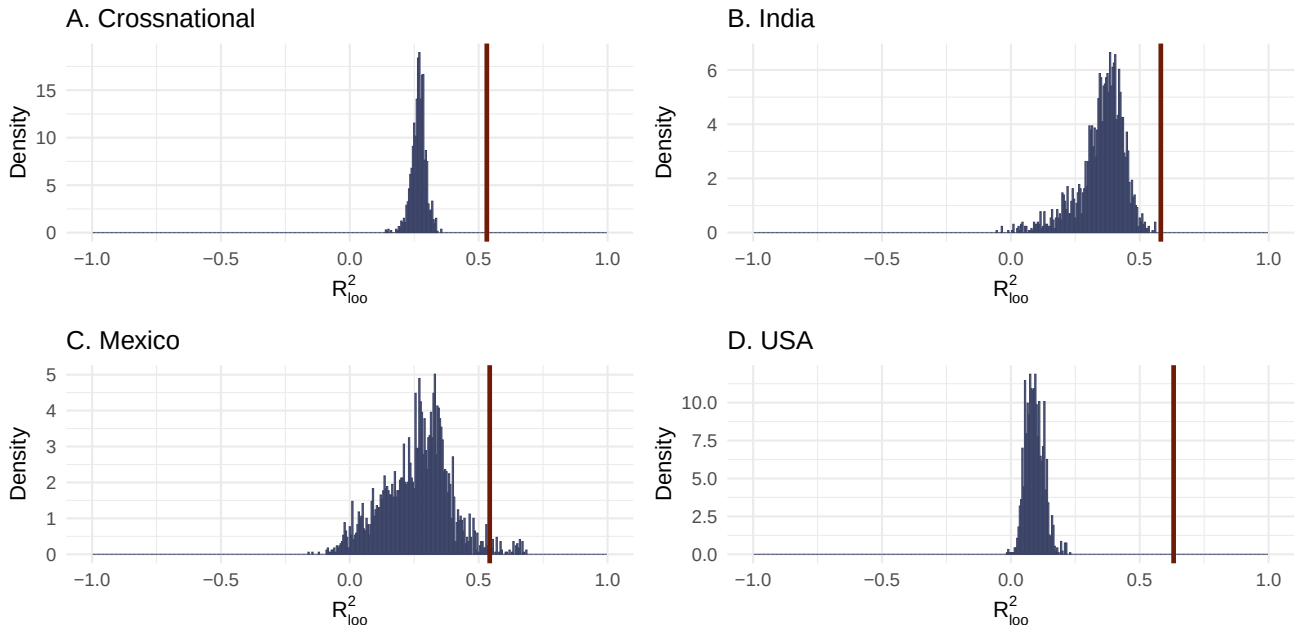


Figure M.2: Stacking simulation. The histograms represent the distribution of R^2_{loo}'s from 5,000 simulated stacking models in the common MC datasets for each challenge.