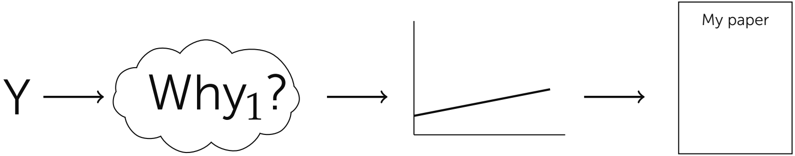
Gathering, evaluating, and aggregating social scientific models of COVID-19 mortality

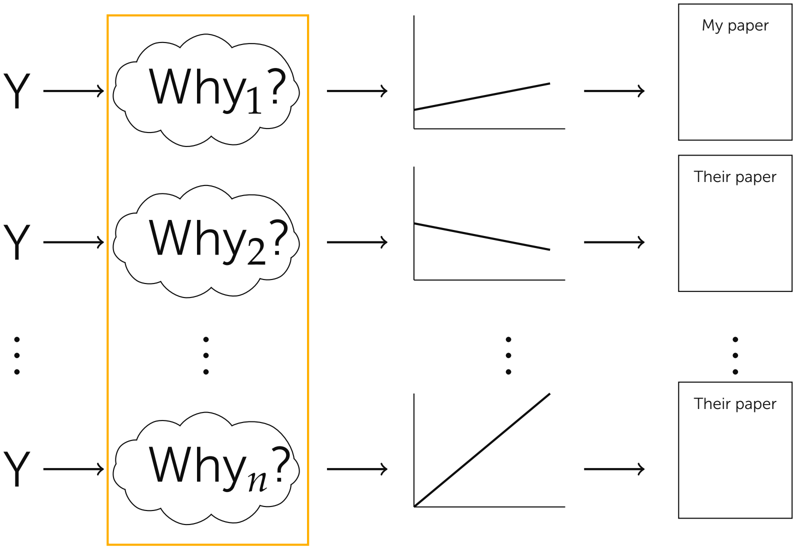
Tara Slough with Miriam Golden, Alexandra Scacco, Macartan Humphreys, Haoyu Zhai, Alberto Díaz-Cayeros, Kim Yi Dionne, Sampada KC, Eugenia Nazrullaeva, and Eva Vivalt et al.

September 14, 2022

The production of empirical social science



The production of empirical social science, ctd.



The problem: the proliferation of explanations

- Social science is filled with discrete rival explanations for outcomes that many people care about.
  - In many literatures we confront many disjointed explanations for outcomes of interest.
  - When do many discrete explanations allow us to understand a phenomenon?
  - How might we aggregate or discriminate between explanations?

- Why does this problem emerge?
  - Incentives for novelty.
  - Norms encouraging simultaneous theoretical and empirical contributions.
  - Focus on "effects of causes," not explanatory questions.

- We lack a framework for filtering or combining multiple explanations → a cumulative learning problem.

A framework for filtering and combining explanations

	Analytic review essays	Meta-analysis	This paper
# of treatments or predictors	Any	1	**Many**
# of outcomes	Any	1	**1**
Sample	Any	Multiple	**Same**
Quantities of interest	–	Common structural parameters across studies or samples.	**Metrics of predictive accuracy.**

Comparison of research designs for aggregating evidence in social science.

- Fundamentally not about external validity → a single sample.
- "Dimension reduction" plays an important role in efforts to cumulate.

The Models Challenge

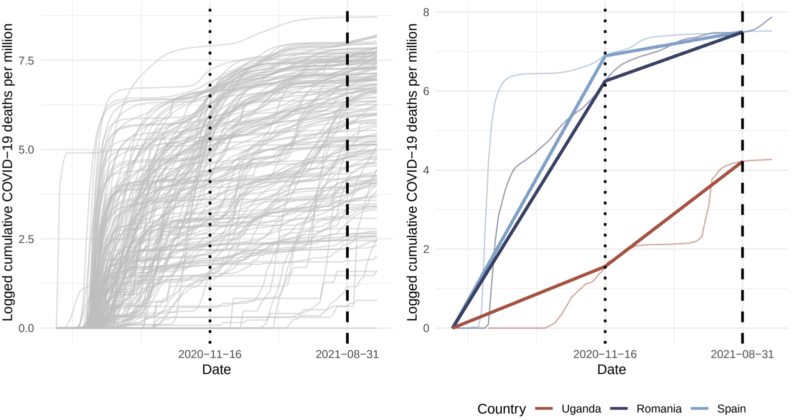
The COVID-19 Model Challenges

- We crowdsourced predictive statistical models of COVID-19 mortality from social scientists.
    - Predictions collected: **December 2020-January 2021** using a Shiny interactive web platform.
    - We asked for predictions of future mortality on **August 31, 2021**.

- To simulate human aggregation of multiple explanations, we elicited expert forecasts about the anticipated predictive performance of the crowdsourced models.
    - Forecasts elicited in **May 2021**.
    - Forecasters reported anticipated predictive preformance on **August 31, 2021**.

The outcome

- Logged COVID-19 mortality per million as of August 31, 2021.



Eight challenges

- Four samples: crossnational, Indian states, Mexican states, US states
- Two types of models:
  - General: only functional form → parameters are estimated from data, e.g., $y_i = \beta_0 + \beta_1 x_i$.
  - Parameterized: functional form and parameters, e.g., $y_i = 4 + 2x_i$.
- We added Lasso and epidemiological benchmarks for each challenge.

	Substantively motivated		Machine Learning		Total
Challenge	General	Parameterized	General	Parameterized	
Crossnational	27	15	1	1	44
India	8	6	1	1	16
Mexico	8	5	1	1	15
US	17	6	3	2	29
Total	61	32	6	5	104

Framework for model evaluation, aggregation

Evaluating model performance

- We use out-of-sample predictions to evaluate each set of model predictions:
  - For general models → **leave-one-out** (LOO) predictions.
  - For parameterized models → **model** predictions.

- Our two metrics of predictive performance are variants of:

$$M = 1 - \alpha \frac{\sum_{i=1}^{N} (\widehat{y}_{ik} - y_{ik})^2}{\sum_{i=1}^{N} (\overline{y}_{ik} - y_{ik})^2}$$

  - **Pseudo-R^2**: $\alpha = 1$
  - **Correlation**: $\alpha = 1/2$, standardize $\widehat{y}_{ik}$ and y_{ik}.

A meta-model

- We use model stacking to combine estimates from different predictive models into a single model. (e.g., Yao et al., 2018).
  - The stacking model places **weights** on some subset of individual models.
  - To generate the stacking model, we estimate these weights by optimizing:

$$\underset{w_k}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \sum_k w_k \widehat{y}_{ik} \right)^2 \text{ s.t. } w_k \geq 0 \forall k, \sum_{k=1}^{K} w_k = 1$$

- The stacking model generates the prediction:

$$\hat{y}_i = \sum_k w_k^* \widehat{y}_{ik}$$

Eliciting expected performance

- Are social scientists able to accurately discriminate between or aggregate multiple predictive models?
- We answer this question by using our forecasting exercise using the Social Science Prediction Platform.
- Experts were randomly assigned to give one of two types of forecasts:
    - Horserace: Which model is most likely to predict the most variation in the outcome?
    - Stacking: Assign larger weights to models if you would pay relatively more attention to the predictions of those models when forming an overall prediction.

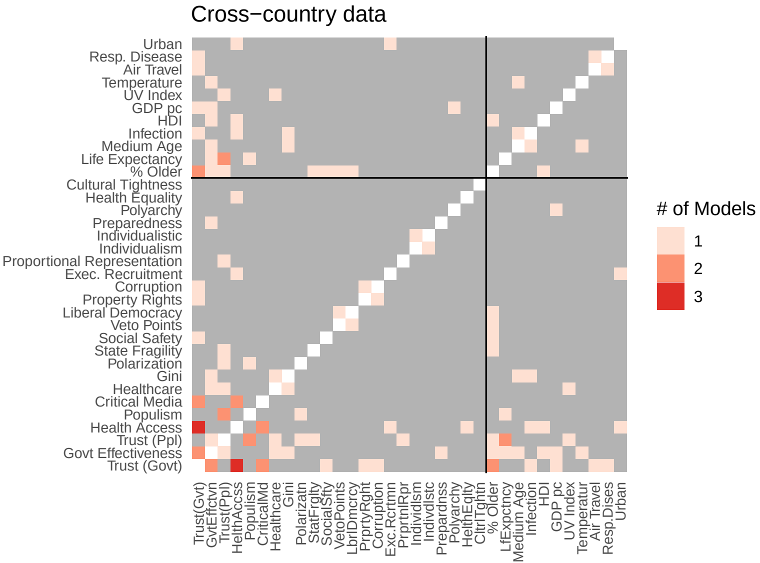
Results

Gathering models: results

1. Descriptively, there was a fairly high degree of overlap in the content of predictive models.

2. Most common predictors: measures of trust and government effectiveness; few models used measures of regime type or political institutions.

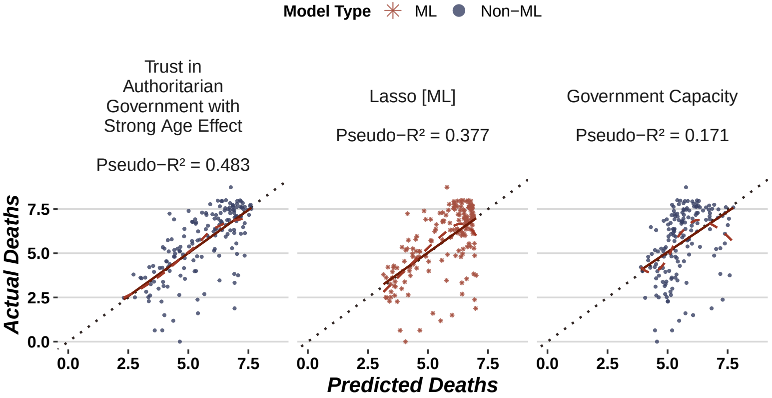
Gathering models: overlap in predictors



Evaluating models: results

1. High variability in predictive performance: the strongest substantively-justified models outperform Lasso in 3/4 general challenges. But the median models in each challenge perform poorly.

Evaluating models: cross-national general challenge

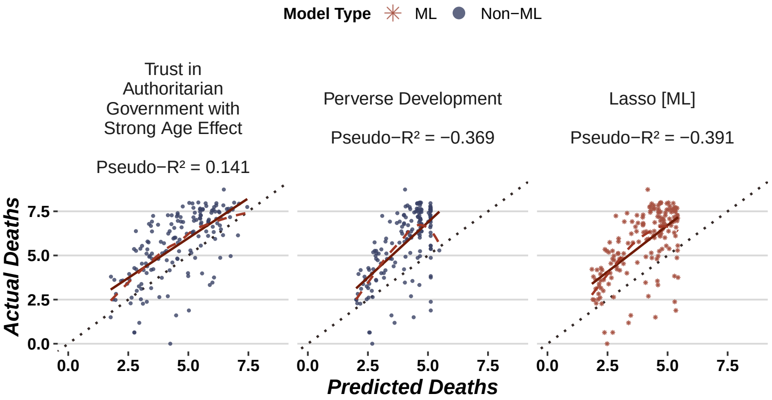


Evaluating models: results

1. High variability in predictive performance: the strongest substantively-justified models outperform Lasso in 3/4 general challenges. But the median models in each challenge perform poorly.

2. Parameterized models make substantially less accurate predictions.
    - Predictions correlate strongly with outcomes, but almost all models substantially underpredict mortality.

Gathering models: cross-national parameterized challenge



Evaluating models: results

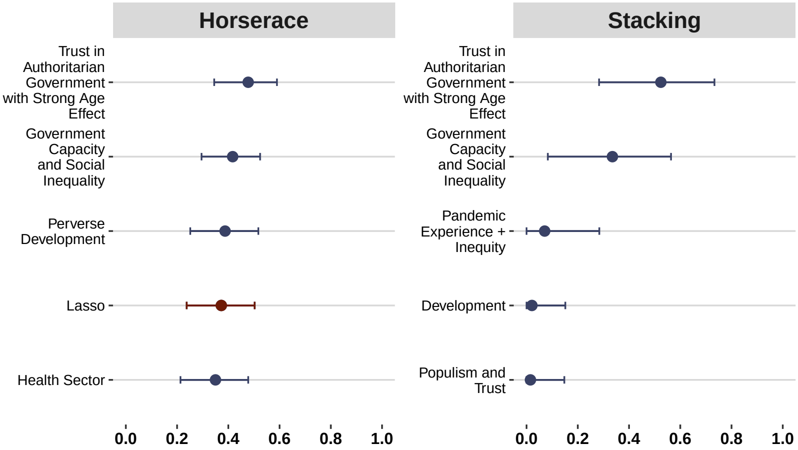
1. High variability in predictive performance: the strongest substantively-justified models outperform Lasso in 3/4 general challenges. But the median models in each challenge perform poorly.

2. Parameterized models make substantially less accurate predictions.

3. Stacking and horserace contests converge on top two models, but stacking weights are heavily skewed toward top two models.
    - Stacking weights reward models that provide different information.

Evaluating models: horserace vs. stacking

- Top five models in the algorithmic horserace and stacking contests.



Evaluating models: results

1. High variability in predictive performance: the strongest substantively-justified models outperform Lasso in 3/4 general challenges. But the median models in each challenge perform poorly.

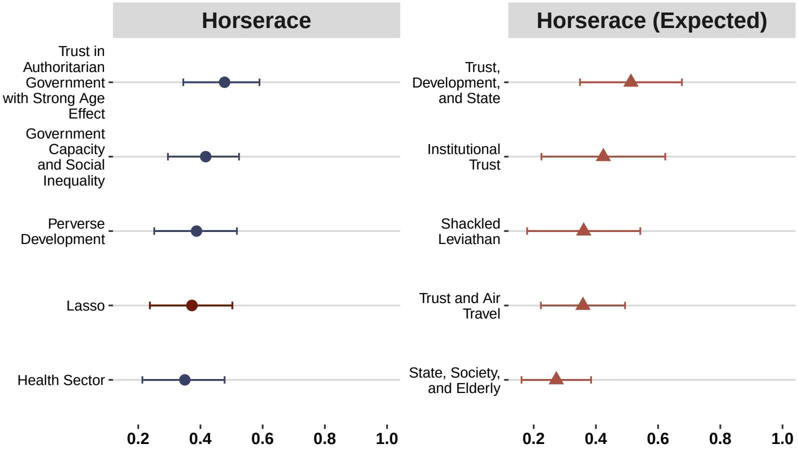
2. Parameterized models make substantially less accurate predictions.

3. Stacking and horserace contests converge on top two models, but stacking weights are heavily skewed toward top two models.

4. Algorithmic horserace and elicited horserace rankings do not overlap at top of the distribution.
    - Note slightly different measure of model performance (psuedo-R^2 vs. Pr(best model in set)).

Evaluating models: algorithmic vs. elicited horserace

- Top five models in the algorithmic horserace and elicited horserace contest.



Evaluating models: results

1. High variability in predictive performance: the strongest substantively-justified models outperform Lasso in 3/4 general challenges. But the median models in each challenge perform poorly.

2. Parameterized models make substantially less accurate predictions.

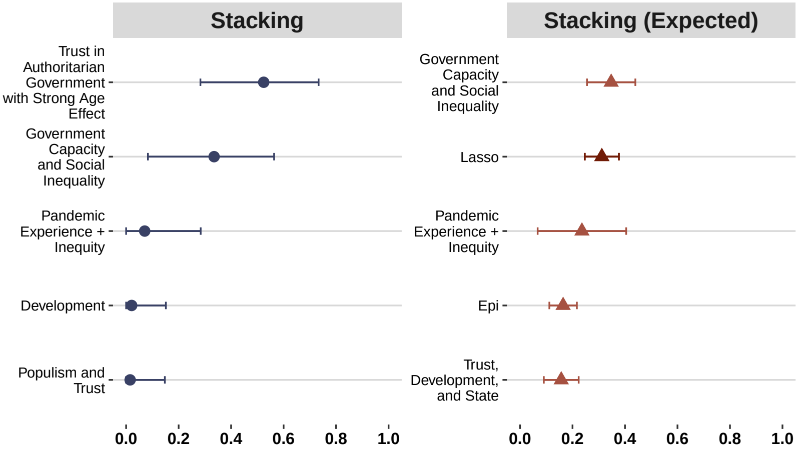
3. Stacking and horserace contests converge on top two models, but stacking weights are heavily skewed toward top two models.

4. Algorithmic horserace and elicited horserace rankings do not overlap at top of the distribution.

5. Algorithmic and elicited stacking rankings overlap somewhat, but elicited weights are much less skewed.

Evaluating models: algorithmic vs. elicited stacking

- Top five models in the algorithmic horserace and elicited stacking contest.



Aggregating models: results

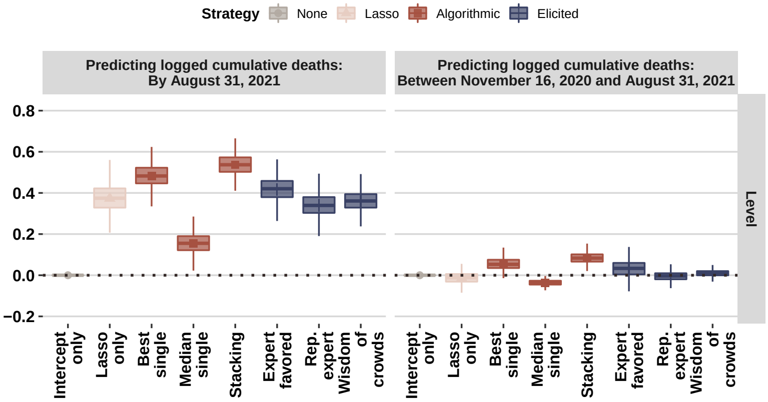
1. Aggregate stacking-based prediction outperforms best model (as it should!), but modestly.
    - But stacking models with the submitted models vastly outperform stacking models generated from an equal number of randomly-generated predictive models.

2. Metrics based on expert stacking perform surprisingly poorly.

3. Models performed substantially worse at predicting only future deaths.

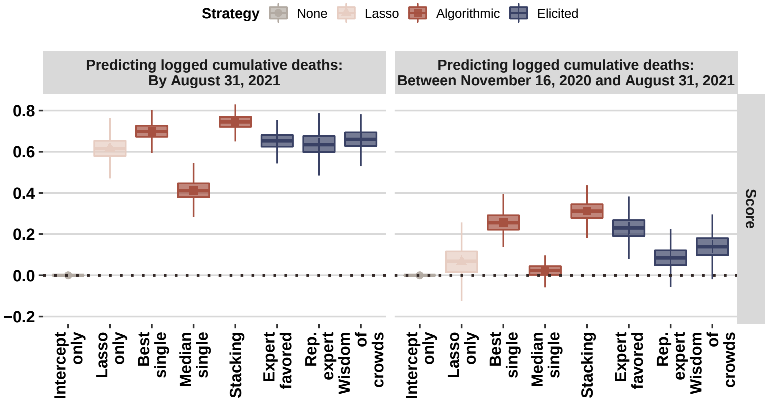
Aggregating models: levels approach

- Comparison of approaches to aggregation using pseudo R^2 measure:



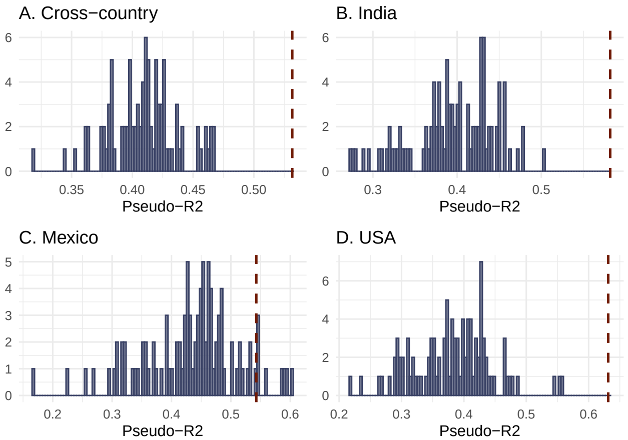
Aggregating models: correlation approach

- Comparison of approaches to aggregation using correlation measure:



Aggregating models: is stacking's edge mechanical?

- Stacking on the submitted models outperforms:
  - The **best model** → mechanical as this is equivalent to setting $w_{best} = 1$.
  - Stacking on a **randomly generated** sets of K models.



Discussion and conclusion

- We highlight the problem of accumulating explanations and provide a framework to measure the severity of problem and provide a path forward.

- A framework for aggregation of multiple explanations:
  - Applicable to many other literatures.
  - Many potential applications in a secondary meta-study.
    - But beware of publication bias, **selection** into the published literature.

- Forecasting results researchers perform poorly at filtering or combining explanations.
  - At least when abstracting from usual heuristics etc.
  - A reason to use our framework!