S8. SUPPLEMENTAL MATERIALS

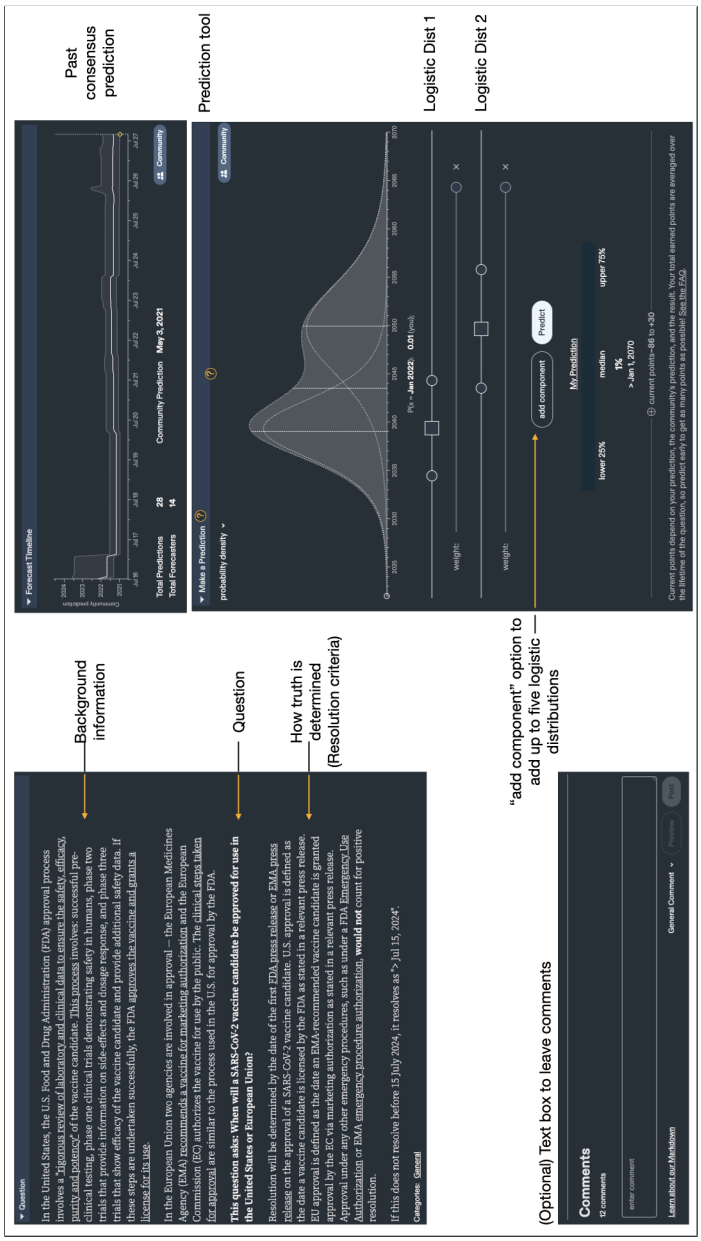

FIG. S6: The Metaculus prediction interface which presents background information, the question, and resolution criteria to the forecaster. Forecasters can use a mixture of five logistic distributions to form their prediction, and optionally leave comments.

Dear <Expert>,

We read your work on <insert work> with great interest, and invite you to join a collaborative group of select experts.

We are building expert consensus predictions about the development of SARS-CoV-2 vaccines and COVID-19 therapeutics each month with a small group of researchers involved in the study of novel therapeutics/vaccines. Experts are surveyed about their predictions about the future development of SARS-CoV-2 vaccines and COVID-19 therapeutics and the results are bei aggregated into a consensus.

We feel your skill set defines you as an expert in this field. Would you participate, alongside other experts, in our survey project to forecast the development of SARS-CoV-2 vaccines and COVID-19 therapeutics?

Your anonymized predictions will contribute to an expert consensus made available to the public, and sent to the CDC to provide support for public health decision making. Our main goal is to provide public health officials probabilistic predictions from experts on the research and development of vaccines and therapeutics. We have made the results from our first survey available here

Our past work, in collaboration with Thomas McAndrew at the University of Massachusetts at Amherst, focused on forecasting the early trajectory of COVID-19 in the United States. Forecasts and predictions generated by 41 experts in the modeling of infectious disease were featured in outlets such as Science, FiveThirtyEight, and The Economist, and were also sent to the CDC to support the US COVID-19 response.

If you are available to participate in at least one survey, please respond to this email, preferably by <MMDD>. We expect to administer the next monthly survey on <MMDD>. We would be thrilled to welcome you—if only just for one round. An expert consensus can produce forecasts on a diverse range of vaccine and therapeutic solutions that computational and statistical methods cannot, and we feel your expertise will make impactful and meaningful contributions.
Please reach out to us with any questions.

Sincerely,

FIG. S7: A template email used to solicit forecasting participation from subject matter experts in molecular and cellular biology, microbiology, virology, biochemistry, and infectious disease who have had several years of experience studying vaccine, antiviral, or biological related to infectious agents.

Dear X,

The fourth COVID-19 Countermeasures session has just opened! This month's survey focuses on the differences between a SARS-CoV-2 vaccine approved in the US via a normal approval process and an emergency approval process, as well as on the recent contradictory statements in the US between the White House, FDA, and CDC on vaccine timelines and mask wearing. Finally, for the first time we are asking experts and trained forecasters to provide us with purely text-based responses (Q6 and Q7) — a unique aspect to this work that computational models cannot provide.

The goal of these set of predictions is to support public health decision making, provide best estimates that allow the public to make informed decisions, and address current controversies between the White House, FDA, and CDC.

Here are all of the questions:

1. When will a SARS-CoV-2 vaccine candidate be approved for use in the US through a normal approval process?
2. When will a SARS-CoV-2 vaccine candidate be approved for use in the US through an emergency approval process?
3. What will be the efficacy ratio of the first SARS-CoV-2 vaccine candidate approved on an emergency basis (numerator) compared to the first approved through a normal process (denominator)?
4. What is the probability of at least ten serious adverse events (SAEs) being attributed within one year to the first SARS-CoV-2 vaccine approved in the US through a normal approval process?
5. What is the probability of at least ten serious adverse events (SAEs) being attributed within one year to the first SARS-CoV-2 vaccine approved in the US through an emergency approval process?
6. What will be the most common adverse event, serious or not serious, of the first US-approved SARS-CoV-2 vaccine?
7. How long after the approval of a SARS-CoV-2 vaccine of at least 50% efficacy would you continue to recommend the general public wear masks? What percentage of the US population would have to be vaccinated for your view to change?

We very much encourage you to share your reasoning and analyses in the comments with other experts, especially for questions 6 and 7 since these are entirely comments-based.

From now until 25 September, the community prediction will be hidden. Subsequently from 25 September until 30 September 11:59 PM EST, the community prediction will be viewable. Feel free to revise your predictions at any time.

All questions will close on 30 September at 11:59 PM EST. We will subsequently collect the survey results and compile a report to be sent to the Centers for Disease Control and Prevention (CDC). In addition, the report and aggregate de-identified raw data will be made available to the public at the following site in October.

As always, if you have any questions or feedback feel free to reply to this email. Thank you for your participation. Those of you who have participated in previous sessions can find our first three reports here.

Regards,

FIG. S8: An example email sent to subject matter experts and trained forecasters signaling new questions were available to forecast. Forecasters were given a summary of the questions and line list of each question. Questions contained hyperlinks that, when clicked, directed the forecaster to the forecasting platform and the corresponding question. Forecasters were told up until what time predictions could be submitted and the dates when the linear pool prediction would could be observed. The email ended by reiterating forecasters could direct questions and feedback to the authors.

Question
Survey 1 (June, 2020)
What will be the efficacy of the Oxford/AstraZeneca ChAdOx1 nCoV-19 vaccine candidate according to the results of Phase II/III testing?
When will the first SARS-CoV-2 vaccine to be approved in the US or EU be administered to >100K people?
How many SARS-CoV-2 vaccine candidates will be in human trials as of 1 August 2020?
When will a SARS-CoV-2 vaccine candidate be approved for use in the United States or European Union?
When will a SARS-CoV-2 vaccine candidate demonstrate $\geq$ 70 % efficiacy?
When will a COVID-19 therapeutic or therapeutics cocktail show a statistically significant survival benefit for the treatment group in a n>200 RCT?
Survey 2 (July, 2020)
When will a SARS-CoV-2 antiviral show a statistically significant survival benefit for the treatment group in an n>200 RCT? What will be the efficacy of the first US- or EU- approved SARS-CoV-2 vaccine based on a non-replicating viral vector platform?
When will a SARS-CoV-2 monoclonal antibody or antibody cocktail show a statistically significant survival benefit for the treatment group in an n>200 RCT?
What will be the efficacy of the first US- or EU- approved SARS-CoV-2 vaccine based on a protein subunit platform?
What will be the efficacy of the first US- or EU- approved SARS-CoV-2 vaccine based on an inactivated virus platform?
What will be the efficacy of the first US- or EU- approved SARS-CoV-2 vaccine based on a DNA or RNA platform?
When will a SARS-CoV-2 vaccine candidate be approved for use in the United States or European Union?
Survey 3 (August, 2020)
When will a SARS-CoV-2 vaccine candidate be approved for use in the US or EU through a normal approval process?
When will a SARS-CoV-2 vaccine candidate be approved for use in the US or EU through an emergency approval process?
What will be the efficacy of the first US- or EU- approved SARS-CoV-2 vaccine candidate approved through a normal approval process?
What will be the efficacy of the first US- or EU- approved SARS-CoV-2 vaccine candidate approved through an emergency approval process?
How many weeks after approval will the first 100 million doses of the first US- or EU- approved SARS-CoV-2 vaccine candidate based on a DNA or RNA platform be manufactured?
How soon after approval will the first 100 million doses of the first US- or EU- approved SARS-CoV-2 vaccine candidate based on a non-replicating viral vector platform be manufactured?
When will an orally administered SARS-CoV-2 antiviral show a statistically significant survival benefit for the treatment group in an n>200 RCT?
What will be the SARS-CoV-2 infectivity of children relative to adults when schools are open?
Survey 4 (September, 2020)
When will a SARS-CoV-2 vaccine candidate be approved for use in the US through a normal approval process?
When will a SARS-CoV-2 vaccine candidate be approved for use in the US through an emergency approval process?
What will be the efficacy ratio of the first SARS-CoV-2 vaccine candidate approved on an emergency basis (numerator) compared to the first approved through a normal process (denominator)?
What is the probability of at least ten serious adverse events (SAEs) being attributed within one year to the first SARS-CoV-2 vaccine approved in the US through a normal approval process?
What is the probability of at least ten serious adverse events (SAEs) being attributed within one year to the first SARS-CoV-2 vaccine approved in the US through an emergency approval process?

TABLE S2: List of all questions stratified by survey.

Question	True value
When will the first SARS-CoV-2 vaccine to be approved in the US or EU be administered to >100K people?	Jan. 19, 2021
How many SARS-CoV-2 vaccine candidates will be in human trials as of 1 August 2020?	26
When will a SARS-CoV-2 vaccine candidate be approved for use in the United States or European Union?	Dec. 21, 2020
When will a SARS-CoV-2 vaccine candidate demonstrate >70% efficacy?	Dec. 10, 2020
When will a COVID-19 therapeutic or therapeutics cocktail show a statistically significant survival benefit for the treatment group in a n>200 RCT?	July 17th, 2020
What will be the efficacy of the first US- or EU- approved SARS-CoV-2 vaccine based on a non-replicating viral vector platform?	66.9%
What will be the efficacy of the first US- or EU- approved SARS-CoV-2 vaccine based on a DNA or RNA platform?	95%
When will a SARS-CoV-2 vaccine candidate be approved for use in the United States or European Union?	Dec. 21, 2020
When will a SARS-CoV-2 vaccine candidate be approved for use in the US or EU through a normal approval process?	Dec. 21, 2020
When will a SARS-CoV-2 vaccine candidate be approved for use in the US or EU through an emergency approval process?	Dec. 10th, 2020
What will be the efficacy of the first US- or EU- approved SARS-CoV-2 vaccine candidate approved through a normal approval process?	95%
What will be the efficacy of the first US- or EU- approved SARS-CoV-2 vaccine candidate approved through an emergency approval process?	95%
When will a SARS-CoV-2 vaccine candidate be approved for use in the US through a normal approval process?	Aug. 23, 2021
When will a SARS-CoV-2 vaccine candidate be approved for use in the US through an emergency approval process?	Dec. 10th, 2020
What will be the efficacy ratio of the first SARS-CoV-2 vaccine candidate approved on an emergency basis (numerator) compared to the first approved through a normal process (denominator)?	1.0

TABLE S3: A list of all fifteen questions to date to have ground truth data available.

Forecaster	Total predictions with ground truth data	Number of forecasters
Expert	79	10
Not Expert	110	11
Linear pool	45	3

TABLE S4: A summary of the number of predictions with ground truth data stratified by expert, non-expert, and linear pool model.

S8.1. An analysis of logarithmic scores

The accuracy of a linear pool of trained forecasters plus experts was in between the accuracy of a linear pool generated from trained forecasters and the accuracy of a linear pool generated from experts except for a single question where a subset of individual experts' accuracy was very poor (Fig. S9 and Fig. S10). Across all fifteen questions where the truth could be determined, the log scores for a linear pool of trained forecasters plus experts had a smaller interquartile range when compared to individual forecasters, though median scores were similar between individuals and all three linear pool distributions (Fig. S11).

The mean log score across all fifteen questions was for individual trained forecasters 2.12 (80CI: [0.32, 4.66]) and for individual subject matter experts was 1.45 (80CI: [-0.45, 3.65]). The standard deviation of the log score was 1.75 for trained forecasters and 1.87 for experts.

The linear pool mean log score generated from both trained forecasters and experts was 2.06 (80CI: [0.49, 4.17]) compared to a mean log score of a trained forecasters only linear pool of 2.07 (80CI: [0.73, 4.03]), and for an experts only linear pool of 2.15 (80CI: [0.19, 4.72]). The linear pool of trained forecasters and experts generated a higher average log score than both the linear pool of trained forecasters and experts for 1/6 questions (proportion: 0.16, 80CI: [0, 0.33]). Compared to a linear pool of trained forecasters, a linear pool of experts produced a higher average log score for 4/15 questions (proportion: 0.27, 80CI: [0.14, 0.43]). A trained forecaster linear pool produced a higher average log score than a linear pool of experts for 0/5 efficacy questions (Fig. S9 top) and had a lower average log score when asked to predict the number of SARS-CoV-2 vaccine candidates that will be in human trials by Aug 1st, 2020 (Fig. S10 middle). The average log score for an expert linear pool was higher compared to a linear pool of trained forecasters when asked for the date a vaccine will be approved in the US/EU, the date a vaccine will be approved through an emergency authorization (asked in Aug and Sept.), and the date when a COVID-19 therapy will show a survival benefit (Fig. S9 and Fig. S10)

The 25^{th} and 75^{th} percentiles for log scores, from the largest interval to smallest interval, was [0.42, 2.98] for all individuals, [0.65, 3.07] for a linear pool of experts, [0.98, 2.96] for a linear pool of trained forecasters plus experts, and [1.24, 2.90] for a linear pool of trained forecasters (Fig. S11).

Trained forecasters had the highest log scores on average, followed by linear pool models, and then subject matter experts (Table I) however these results did not meet statistical significance. Ninety five percent confidence intervals around the difference in log scores between subject matter experts and trained forecasters, and between linear pool models and trained forecasters were large. We do not have enough data on forecast accuracy to conclude statistical significance at a type I error of 5% that trained forecasters made more accurate predictions than subject matter experts or linear pool models.

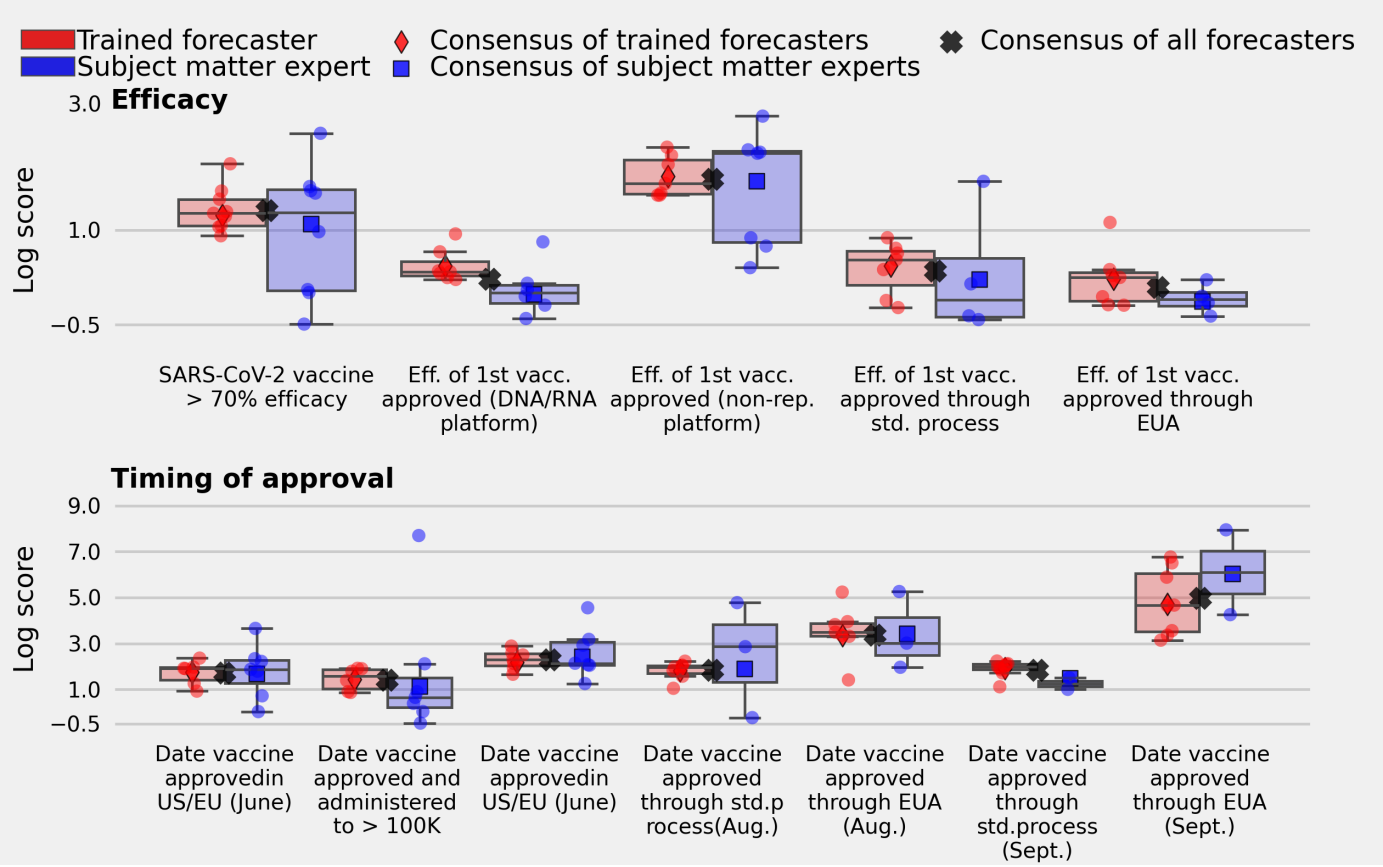

FIG. S9: Log scores for individual trained forecasters (red circles) and a linear pool of trained forecasters (red diamond), individual subject matter experts (blue circles) and an expert linear pool (blue diamond), and a linear pool of both trained forecasters and experts related (black X) for questions related to the efficacy and timing of approval of a vaccine. Individuals, and so linear pools, received higher average log scores for questions related to whether a vaccine would demonstrate an efficacy above 70% and the efficacy of a vaccine that uses a non-replicating platform when compared to questions related to the efficacy of the first vaccine approved. Individuals and linear pools received log scores for questions related to the timing of approval of a vaccine that were similar to questions related to vaccine efficacy. None of the linear pool predictions receive the highest log score for any one question, however the trained forecaster plus expert linear pool also never receives the lowest log score. Aggregating trained forecasters and subject matter experts has the potential to guard against an individual forecast with poor accuracy.

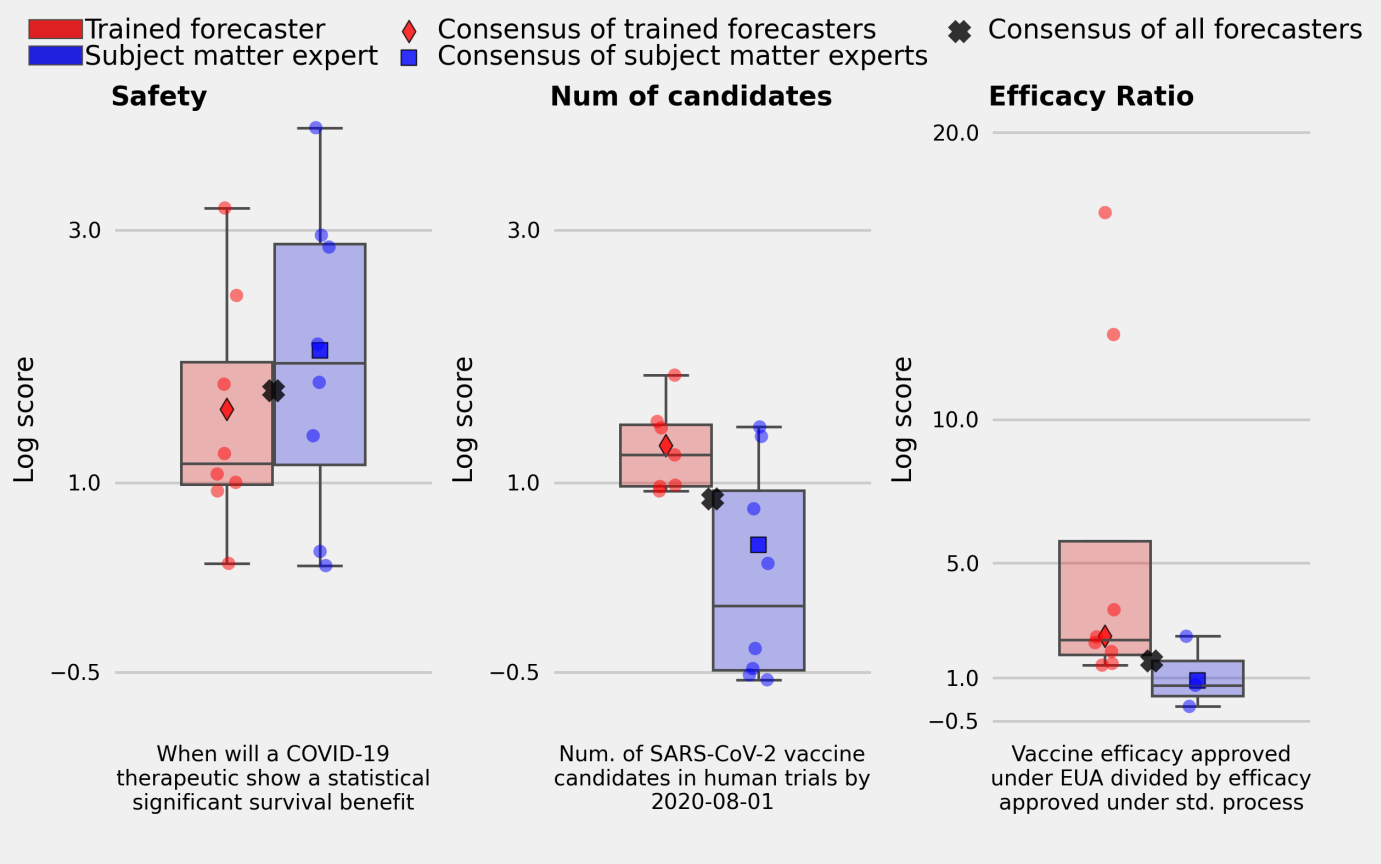

FIG. S10: Log scores for individual trained forecasters (red circles) and a linear pool of trained forecasters (red diamond), individual subject matter experts (blue circles) and an expert linear pool (blue diamond), and a linear pool of both trained forecasters and experts related (black X) for questions related to safety, the number of vaccine candidates, and the estimate efficacy of a vaccine approved under emergency authorization divided by a vaccine approved under a standard process. Linear pool predictions again never attain the highest nor the lowest log scores among any other forecaster.

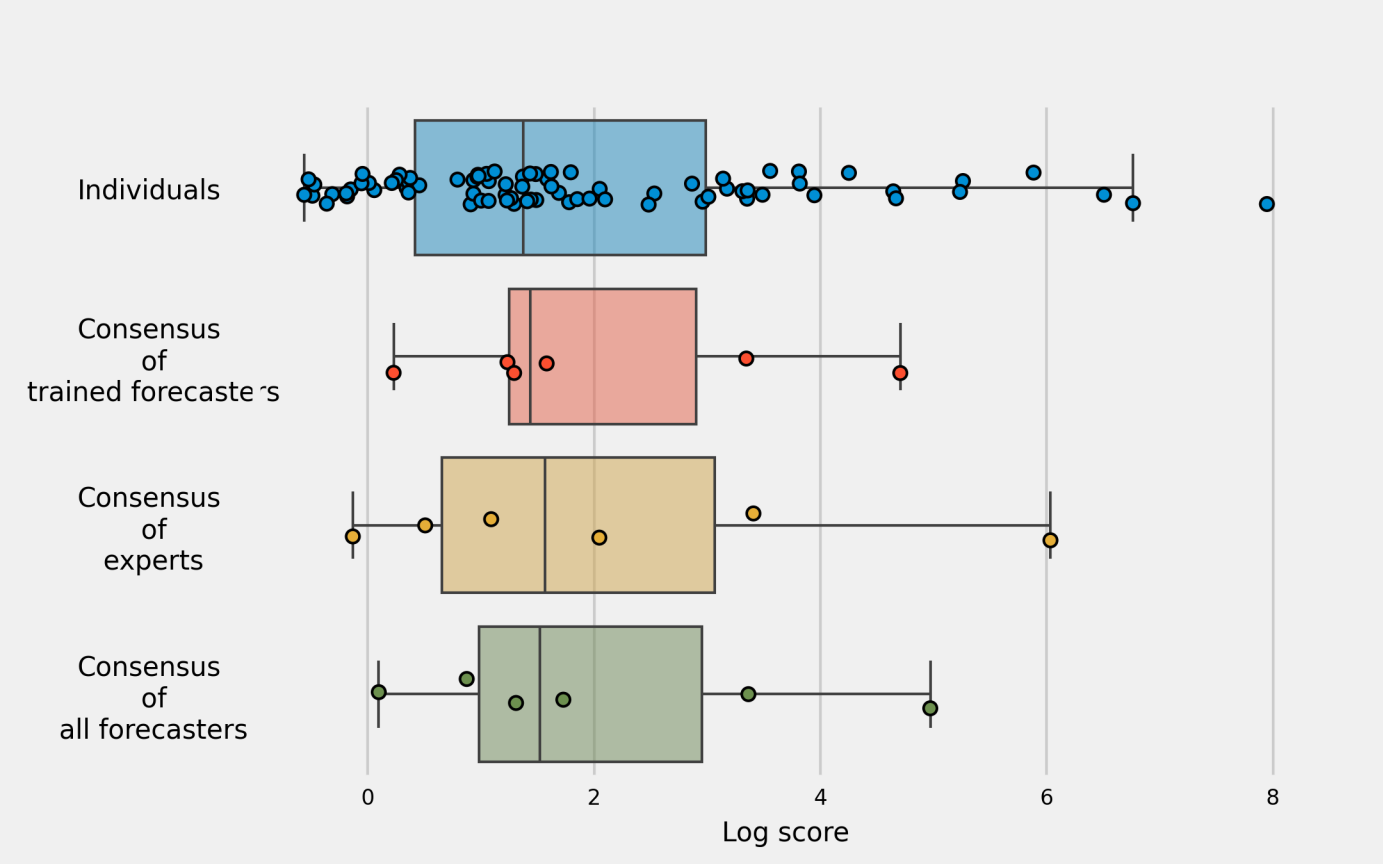

FIG. S11: The log score across all questions for individuals and the same three linear pool predictions. Though linear pool predictions were able to guard against poor performing individual predictions, there was little difference in predictive performance between a linear pool of trained forecasters, linear pool of experts, or both.

S8.2. An analysis of scaled ranks

Scaled ranks showed trained forecasters made more accurate predictions about vaccine efficacy, the number of potential vaccine candidates, and the difference in efficacies between a vaccine approved through an emergency process versus standard process while subject matter experts outperformed trained forecasters on questions about the timing of approval and safety of a vaccine (Fig. S12 and Fig. S13). Linear pool predictions were never the least accurate nor the most accurate forecasts for all questions.

A linear pool of trained forecasters plus subject matter experts scored above the 50th percentile for 9/15 questions while a trained forecaster linear pool scored above the 50th percentile on 10/15 questions with ground truth and an expert only linear pool scored above the 50th percentile on 5/15 questions with ground truth (See supplemental figure S12 and figure S13).

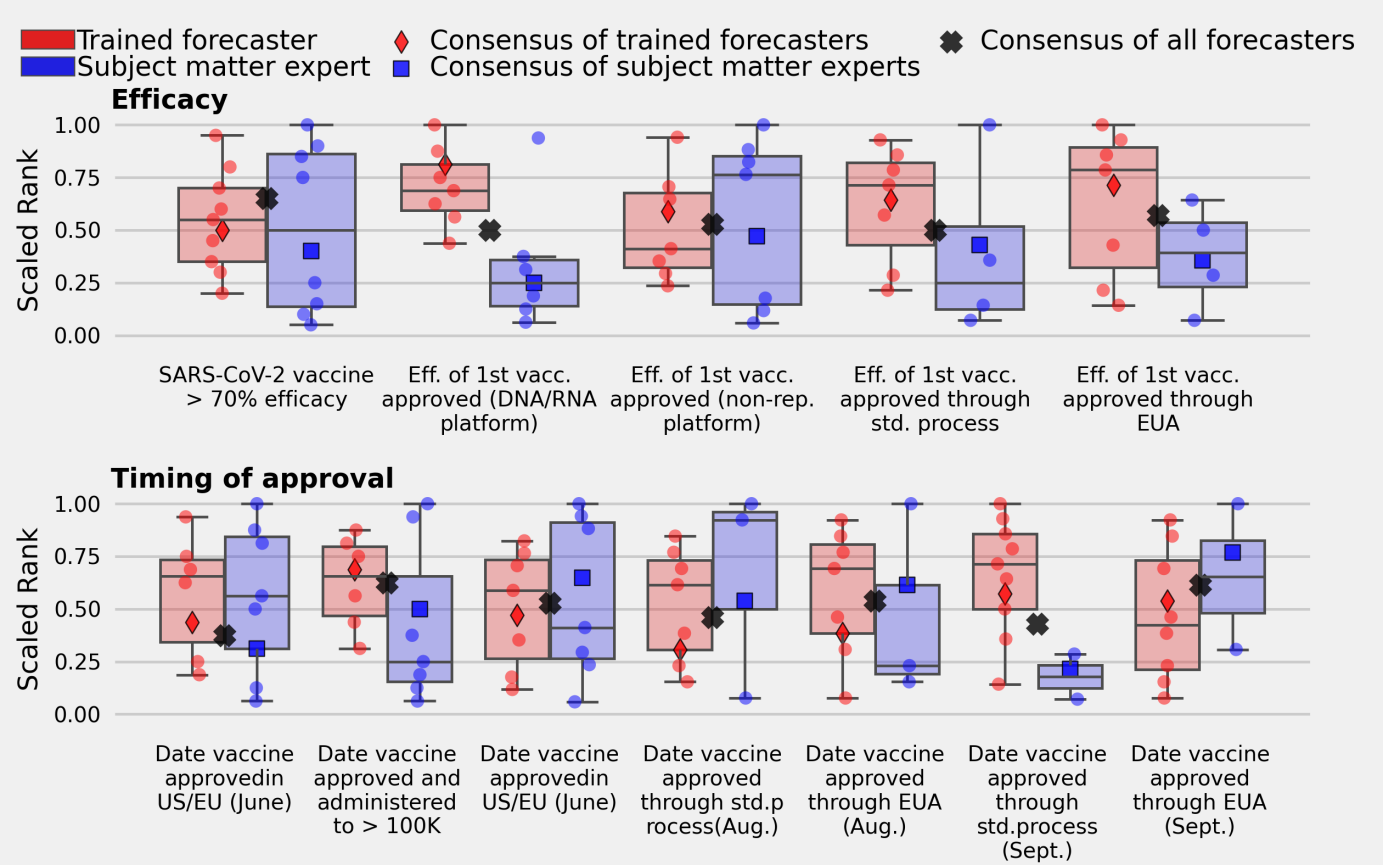

FIG. S12: Scaled ranks for individual trained forecasters (red circles), subject matter experts (blue circles), and three linear pool distributions: (i) a linear pool of trained forecasters (red diamond), of subject matter experts (blue square), and the linear pool of both trained forecasters and experts (black X) for 5 questions related to efficacy and 7 questions related to the timing of vaccine approval with ground truth.

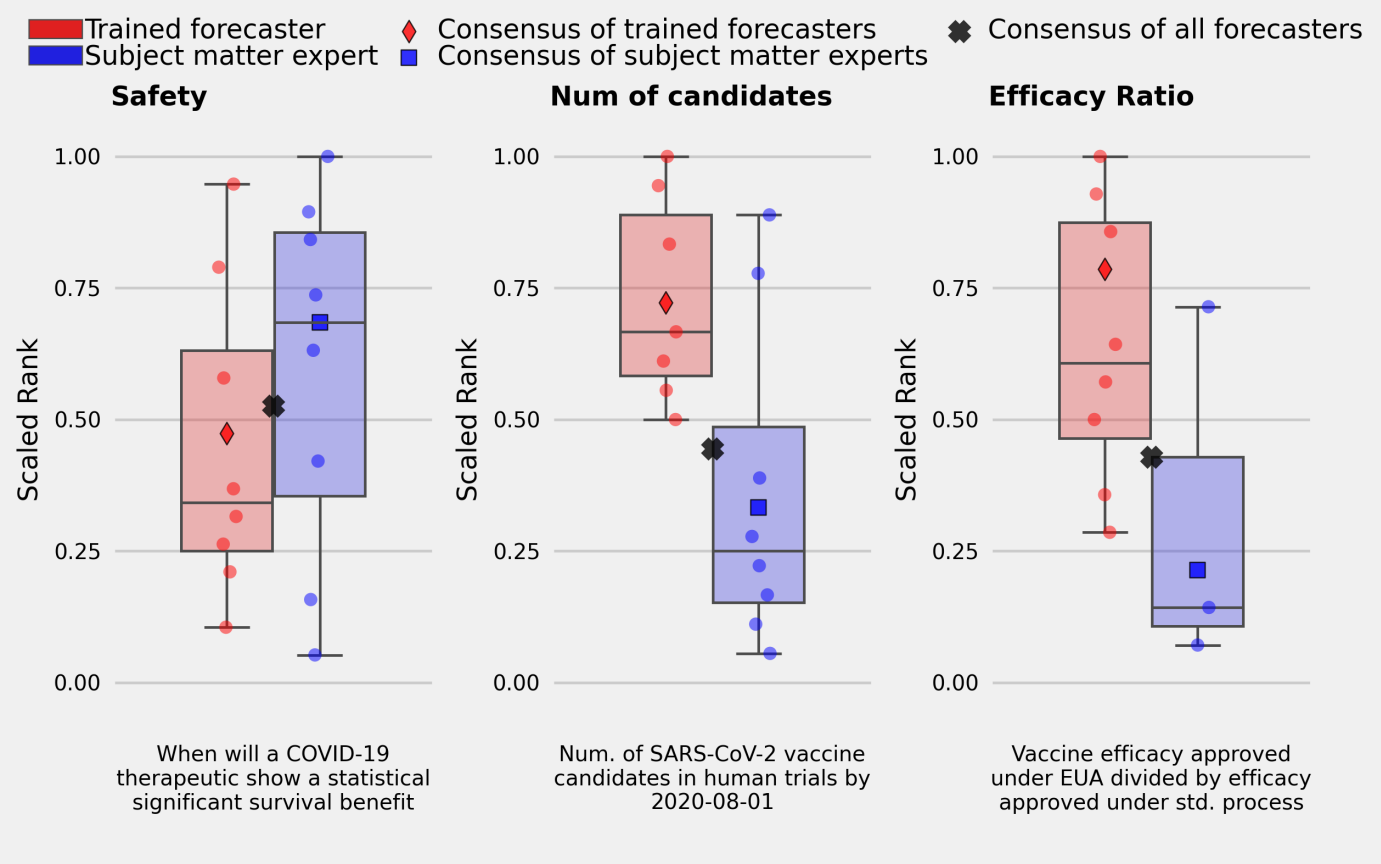

FIG. S13: Scaled ranks for individual trained forecasters (red circles), subject matter experts (blue circles), and three linear pool distributions: (i) a linear pool of trained forecasters (red diamond), of subject matter experts (blue square), and the linear pool of both trained forecasters and experts (black X) for questions related to the safety of a vaccine, number of vaccine candidates, and the difference in efficacy between a vaccine approved under an emergency process versus standard process with ground truth.