Informing pandemic response in the face of uncertainty. *An evaluation of the U.S. COVID-19 Scenario Modeling Hub*

Emily Howerton (The Pennsylvania State University (PSU)), Lucie Contamin (University of Pittsburgh), Luke C Mullany (Johns Hopkins University Applied Physics Lab (JHU/APL)), Michelle Qin (Harvard University), Nicholas G. Reich (University of Massachusetts Amherst), Samantha Bents (National Institutes of Health Fogarty International Center (NIH)), Rebecca K. Borchering (PSU, Centers for Disease Control and Prevention (CDC)), Sung-mok Jung (University of North Carolina (UNC)) Sara L. Loo (Johns Hopkins University Infectious Disease Dynamics (JHU-IDD)) Claire P. Smith (JHU-IDD), John Levander (University of Pittsburgh (UPitt)), Jessica Kerr (UPitt), J. Espino (UPitt), Willem G. van Panhuis (National Institute of Allergy and Infectious Diseases), Harry Hochheiser (UPitt), Marta Galanti (Columbia University (CU)), Teresa Yamana (CU), Sen Pei (CU), Jeffrey Shaman (CU)**,** Kaitlin Rainwater-Lovett (JHU/APL), Matt Kinsey (JHU/APL), Kate Tallaksen (JHU/APL), Shelby Wilson (JHU/APL), Lauren Shin (JHU/APL), Joseph C. Lemaitre (UNC), Joshua Kaminsky (JHU-IDD), Juan Dent Hulse (JHU-IDD), Elizabeth C. Lee (JHU-IDD), Clif McKee (JHU-IDD), Alison Hill (JHU-IDD), Dean Karlen (University of Victoria (UVic)), Matteo Chinazzi (Northeastern University (NEU)), Jessica T. Davis (NEU), Kunpeng Mu (NEU), Xinyue Xiong (NEU), Ana Pastore y Piontti (NEU), Alessandro Vespignani (NEU), Erik T. Rosenstrom (North Carolina State University (NCSU)), Julie S. Ivy (NCSU), Maria E. Mayorga (NCSU), Julie L. Swann (NCSU), Guido España (University of Notre Dame (UND)), Sean Cavany (UND), Sean Moore (UND), Alex Perkins (UND)**,** Thomas Hladish (University of Florida (UF)), Alexander Pillai (UF), Kok Ben Toh (Northwestern University), Ira Longini Jr. (UF), Shi Chen (University of North Carolina at Charlotte (UNCC)), Rajib Paul (UNCC), Daniel Janies (UNCC), Jean-Claude Thill (UNCC), Anass Bouchnita (University of Texas at El Paso (UTEP)), Kaiming Bi (University of Texas at Austin (UTA)), Michael Lachmann (Santa Fe Institute (SFI)), Spencer Fox (University of Georgia (UGA)), Lauren Ancel Meyers (UTA), UT COVID-19 Modeling Consortium**,** Ajitesh Srivastava (University of Southern California), Przemyslaw Porebski (University of Virginia (UVA)), Srini Venkatramanan (UVA), Aniruddha Adiga (UVA), Bryan Lewis (UVA), Brian Klahn (UVA), Joseph Outten (UVA), Benjamin Hurt (UVA), Jiangzhuo Chen (UVA), Henning Mortveit (UVA), Amanda Wilson (UVA), Madhav Marathe (UVA), Stefan Hoops (UVA), Parantapa Bhattacharya (UVA), Dustin Machi (UVA), Betsy L. Cadwell (CDC), Jessica M. Healy (CDC), Rachel B. Slayton (CDC), Michael A. Johansson (CDC), Matthew Biggerstaff (CDC)**,** Shaun Truelove (JHU-IDD), Michael C. Runge (U.S. Geological Survey), Katriona Shea (PSU), Cécile Viboud* (NIH), Justin Lessler* (UNC)
*co-senior authors

Abstract

Our ability to forecast epidemics more than a few weeks into the future is constrained by the complexity of disease systems, our limited ability to measure the current state of an epidemic, and uncertainties in how human action will affect transmission. Realistic longer-term projections (spanning more than a few weeks) may, however, be possible under defined scenarios that specify the future state of critical epidemic drivers, with the additional benefit that such scenarios can be used to anticipate the comparative effect of control measures. Since December 2020, the U.S. COVID-19 Scenario Modeling Hub (SMH) has convened multiple modeling teams to make 6-month ahead projections of the number of SARS-CoV-2 cases, hospitalizations and deaths. The SMH released nearly 1.8 million national and state-level projections between February 2021 and November 2022. SMH performance varied widely as a function of both scenario validity and model calibration. Scenario assumptions were periodically invalidated by the arrival of unanticipated SARS-CoV-2 variants, but SMH still provided projections on average 22 weeks before changes in assumptions (such as virus transmissibility) invalidated scenarios and their corresponding projections. During these periods, before emergence of a novel variant, a linear opinion pool ensemble of contributed models was consistently more reliable than any single model, and projection interval coverage was near target levels for the most plausible scenarios (e.g., 79% coverage for 95% projection interval). SMH projections were used operationally to guide planning and policy at different stages of the pandemic, illustrating the value of the hub approach for long-term scenario projections.

Main text

Since SARS-CoV-2 was detected in December 2019, there have been numerous disease modeling efforts aiming to inform the pandemic response. These activities have had a variety of goals, including measuring transmissibility, estimating rates of unobserved infections and evaluating control measures (*1*, *2*). Particular attention has been paid to models that attempt to predict the course of the pandemic weeks or months into the future.

These predictive models can, roughly, be divided into two categories: (1) forecasting models that attempt to predict *what will happen* over the future course of the epidemic, encompassing all current knowledge and future uncertainties, and (2) scenario planning models that aim to capture *what would happen if* the future unfolded according to a particular set of circumstances (e.g., intervention policies). While there is no bright line between the two approaches, there are often differences in how they are implemented. Forecasts are typically limited to shorter time horizons, as key drivers of disease dynamics (e.g., human behavior, variant virus emergence) can become highly uncertain at longer horizons. In contrast, scenario projections often attempt to provide longer term guidance by making explicit assumptions about future changes in those drivers (*3*), potentially at the expense of predicting *what will happen*. These approaches support decision making in different ways; for instance, forecasts can inform near-term resource allocation and situational awareness (*4*), while a scenario approach can inform longer-term resource planning and to compare potential control strategies (*5*, *6*).

Ensembles of independent models consistently outperform individual models in a number of fields (*7*, *8*), including infectious disease forecasting (*9*–*12*). Leveraging this multi-model approach, the US COVID-19 Forecast Hub was formed in April 2020, to predict the number of US cases, hospitalizations, and deaths 1-4 weeks into the future (*13*). Recognizing that longer term planning scenarios could benefit from a similar multi-model approach (*14*–*16*), the US COVID-19 Scenario Modeling Hub (SMH) was formed in December 2020 to produce scenario based projections months into the future.

Between February 2021 and November 2022 SMH produced 16 rounds of projections, 14 of which were released to the public (*17*) (Round 8 was a “practice round”, and the emergence of the Omicron variant invalidated Round 10 projections before their release) (Figure 1). The focus of each round was guided by ongoing discussions with public health partners at the state and federal level and reflected shifting sources of uncertainty in the epidemiology of, and response to, the COVID-19 pandemic. Each round included four scenarios, with early rounds focusing on vaccine availability and use of non-pharmaceutical interventions (NPIs), and later rounds addressing vaccine uptake and the effect of new variants.

In each round, 4-9 modeling teams provided 12 to 52 weeks (depending on the round’s goals) of probabilistic projections for each scenario for weekly cases, hospitalizations, and deaths at the state and national level. Projections were aggregated using the linear opinion pool method (*18*). Thirteen teams have participated overall, with some teams providing projections only for certain rounds or states.

To assess the performance and added value of the SMH we compared projections to real world epidemic trajectories. Whether scenario projections accurately matched those trajectories depends on both how well scenario definitions matched reality, and the calibration of the projections made conditional on those scenarios. Here we attempt to evaluate both (Figure 2), while acknowledging that there may exist complementary evaluations more specific to the many ways SMH projections were used, ranging from informing national vaccine recommendations (*5*, *19*) to planning for future COVID-19 surges (*20*, *21*).

SMH scenarios usually bracketed future epidemic drivers

In each SMH round (except Round 1), the four scenarios considered represented the cells of a 2x2 table, with each of the two axes of this table including two different levels of key sources of uncertainty (e.g., low vs. high variant transmissibility) or intervention (e.g., authorization or not of childhood vaccines) (Figure 2). Typically, these levels aimed to bracket the future values of key epidemic drivers using available information (about, e.g., vaccine hesitancy (*22*–*25*), or characteristics of emerging viral variants (*26*, *27*)).

We first assessed whether scenario assumptions achieved their goal of bracketing epidemic drivers, as compared to the eventually observed data for those assumptions at the national level (Figure S2-Figure S4, Table S3). For instance, say one uncertainty axis in a round's scenarios stipulated vaccine coverage would increase up to a low

value of 70% and a high value of 80% (depending on the scenario) at the end of the projection period. We say this uncertainty axis "brackets" observations if observed vaccination coverage fell within this range (see Figure S3 for an example).

Over the 14 publicly released rounds each with two primary axes of uncertainty (i.e., 28 total uncertainty axes), 19 were considered to be evaluable against available observed data (Table 1, see Methods). We succeeded in bracketing at least one axis for the majority of the projection period in 9 of 14 publicly released rounds (14 of 19 evaluable axes). In rounds where one axis specified monthly national vaccine uptake (Rounds 1-4 and 9 for primary series, Rounds 14-15 for boosters), scenarios successfully bracketed observations in 55% of projection weeks (31% Round 1, 100% Round 2, 54% Round 3 and 12% Round 4, 100% for Round 9, 100% Round 14, 38% Round 15 Figure S2-Figure S5). In other rounds, scenarios specified vaccination coverage at the end of the projection period (Rounds 5-7 for primary series, and 16 for boosters), which bracketed observed coverage in 2 of 4 rounds. There were 6 rounds with a scenario axis that attempted to bracket the transmission characteristics (inherent transmissibility or immune escape) of one or more known SARS-CoV-2 variants of concern (Rounds 2, 6, 7, 11, 12, 16). Scenario specifications bracketed most estimates of transmissibility now available in the literature (*28*, *29*) (though one study offers an estimate above the bracketing range for the Delta variant (*30*)) (Table S3). All rounds including assumptions about variant severity (Rounds 11 and 12) or waning immunity (Round 13) bracketed currently available literature estimates (*31*–*33*) (Figure S6).

The emergence of new variants was a significant challenge in designing scenarios with long term relevance. Changes in the predominantly circulating variant resulted in major divergences from scenario assumptions in 7 of 14 publicly released rounds. Unanticipated variants emerged, on average, 22 weeks into the projection period (median 16 weeks) (Figure S1), substantially limiting the horizon at which our scenarios remained plausible. This challenge was exacerbated by the lag between when scenarios were defined and when projections were released (5 weeks on average, range 2-10 weeks; Figure 1), and even led us to cancel release of one SMH round (Round 10) when the Omicron variant emerged. However, in the post Omicron period (Rounds 13-16) SMH scenarios consistently devoted an axis to the emergence of immune escape variants that were deemed consistent enough with observations that projections were considered to remain plausible throughout.

Conditioning on scenario plausibility as a pathway to evaluating projections

Next we evaluated the performance of SMH projections using prediction interval (PI) coverage and weighted interval score (WIS) (*34*) (see Methods). PI coverage measures the percent of observations that fall in a prediction interval (so coverage of a 95% PI would ideally be 95%). WIS summarizes calibration across all projection intervals, measuring whether a projection interval captures an observation while penalizing for wider intervals. These standard metrics for evaluating probabilistic forecasts directly compare predictions to observations. In the context of scenario modeling, however, divergences between prediction and observation are the product of two distinct factors: (1) how well the underlying scenario assumptions matched reality (here, scenario

plausibility), and (2) how well models would perform in a world where those scenario assumptions are perfectly correct (i.e., model calibration). For instance, if a scenario's definition is highly divergent from real world events, poor predictive accuracy is not necessarily a sign of poor model calibration, and vice versa. Hence, to assess the calibration of SMH models and the ensemble, we need to identify those scenarios and projected weeks where the majority of observed error is likely driven by model miscalibration (i.e., when scenarios are close to reality). We refer to this intersection of scenarios and projected weeks as "plausible scenario-weeks".

To identify the set of plausible scenario-weeks, we first excluded weeks where an emergent variant that was unanticipated in the scenario specifications reached at least 50% prevalence nationally. For evaluation purposes, we considered this to be an invalidation of all remaining scenario-weeks in the round, and thereby removed 79 out of 400 (20%) projection weeks from the plausible set. Then we compared scenario specifications to data on US vaccination coverage and variant characteristics, this time to identify those scenarios that were closest to realized values during non-excluded weeks (see Methods, Table S3 for details). This yielded a total of 292 plausible weeks for calibration analysis (31% of all scenario-weeks), 173 of which (from Rounds 2-4, 13-16) had two plausible scenarios for the same week, which were equally weighted during evaluation.

SMH ensemble consistently outperformed component and reference models

An initial question is whether we benefit from aggregating multiple models. To answer this, we assessed the relative calibration of individual models and various ensembling methods across projections from plausible scenario-weeks using overall relative WIS, a metric of performance relative to other models which adjusts for varying projection difficulty across targets (from Cramer *et al.* (*11*), see Methods). We assessed variations of two common ensembling techniques: the linear opinion pool (LOP) (*18*) and the Vincent average (*35*, *36*). The LOP assumes that individual model projections represent different hypotheses about the world and preserves variation between these differing projections (*37*). In contrast, the Vincent average assumes that each prediction is an imperfect representation of some common distribution of interest (like a sample), and accordingly cancels away much of the variation. In practice, we believed the former assumption better represented the pool of SMH models and chose to use a variation of the LOP as our primary approach in Round 4 (where the highest and lowest values are excluded, called the "trimmed-LOP", see Methods). Hereafter, the trimmed-LOP will be referred to as the "SMH ensemble".

We found that the SMH ensemble consistently outperformed component models (Figure 3C, Figure S48). This ensemble performed better than average, with an overall relative WIS < 1 for all targets), and was the top performer more frequently than any individual model (19 of 42 targets, across 14 rounds with 3 targets per round). It was best or second best 69% of the time (29/42), and in the top 3 performers 93% of the time (39/42). Further, the SMH ensemble partially compensated for the overconfidence of individual models. Across all locations and rounds, overall 95% PI coverage was 79%

compared to the ideal 95% for the SMH ensemble versus a median of 40% (interquartile range (IQR) 31-49%) across individual models for incident cases, 80% versus 42% (IQR 31-54%) for incident hospitalizations, and 78% versus 42% (IQR 31-49%) for incident deaths. The trimmed-LOP SMH ensemble also outperformed the two alternative ensembling methods considered (untrimmed LOP and median Vincent average, Figure S58).

To assess the added value of SMH, it is important that we compare projections to possible alternatives (*38*). In many settings (e.g., weather forecasting) past observations for a similar time of year can be used as a "null" comparator (*9*, *39*). Lacking such historical data for SARS-CoV-2, we chose to compare our projections to two alternate models: (1) a naive model that assumes cases will remain at current levels for the entire projection period with historical variance (the same null model used by the COVID-19 Forecast Hub (*11*)), and (2) a model based on the set of 4-week ahead ensemble predictions from the COVID-19 Forecast Hub (i.e., for any given week predictions from the SMH ensemble were compared to those of the COVID-19 Forecast Hub ensemble made 4-weeks prior). It should be noted that the naive model uses information available at the time of projection, while the 4-week ahead forecast uses more recent observations for most of the projection period.

The SMH ensemble outperformed the naive model across all targets, by 46% for incident cases (relative WIS 0.54, range across rounds 0.14-3.33), 39% for hospitalizations (relative WIS 0.61, range 0.19-1.69) and 58% for deaths (relative WIS

0.42, range 0.07-1.46) (Figure S22). As expected, the SMH ensemble performed worse than the 4-week forecast model overall (relative WIS 1.48, range 0.34-5.79 for cases, 1.41, 0.40-2.85 for hospitalizations and 2.04, 0.92-3.55 for deaths) (Figure 3A,B, Figure S22). Occasionally, the SMH ensemble outperformed the 4-week ahead forecast model for cases and hospitalizations, for instance in the highly truncated Round 5 addressing the Alpha variant and the two Omicron rounds (Rounds 11, 12) (Figure 3 A,B, Figure 5, Figure S14). Some teams that contributed projections to SMH also submitted forecasts to the COVID-19 Forecast Hub. Although modeling methodology varies by intended use, e.g., model projections for SMH are conditioned on specific assumptions that would not necessarily be accounted for in forecasting models.

To better understand the interaction between scenario assumptions and projection performance, we compared average WIS for projections from plausible scenario-weeks with (A) truncated projections from scenarios that were not selected as “most plausible” and (B) all projections, not truncated based on variant emergence. If the ensemble was well calibrated and our selected most plausible scenarios were closest to reality, we would expect projections from plausible scenario-weeks (with truncation) to have the best performance. We found this expectation to be correct in 57% (24/42) of round-target combinations (the other 43% suggesting that SMH ensemble was sometimes “right” for the wrong reasons). Occasionally scenario selection had little effect on performance (e.g., Round 9 and Round 12, Figure 5). In general, performance for truncated scenarios was better than if we had not truncated (normalized WIS was the same or lower in 64 of 84 scenario-round-target combinations with truncation), though

some of this difference may be attributable to longer projection horizons. Similar conclusions hold for of 95% PI coverage (Figure S51-Figure S53).

While adding value over null alternatives, SMH projections struggled to anticipate changing disease trends

Projections may have utility beyond their ability to predict weekly incidence. For instance, projections that predict whether incidence will increase, decrease or stay the same may be useful even if they are inaccurate in predicting the magnitude of those changes. Based on a method proposed by McDonald *et al.* (*40*), we classified projected and observed incident cases, hospitalizations, and deaths in each week and jurisdiction as "increasing", "flat", or "decreasing" using the percent change from two weeks prior (Figure 4, see Methods).

The median of the SMH ensemble correctly identified the observed trend in 43% of plausible scenario-weeks, comparable to the 4-week forecast model (43%) and better than randomly assigning categories (33%) or assuming continuation of the current trend (34%) (Figure 4). A classification can also be assessed by the number correctly classified relative to the number predicted (precision) or the number observed (recall, see Methods) (*41*). Performance on these metrics was similar across targets and classifications, with the exception of correctly anticipating periods of increasing incidence (48% precision and 44% recall for decreasing, 39%/57% for flat, and 45%/24% for increasing, where lower numbers are worse). Although increases were challenging to predict, they have particular public health importance, as these are the

periods when interventions or additional resources may be needed. While misses were common, it was relatively rare for the SMH ensemble to predict a decrease when incidence increased (23% of increases) or vice-versa (10% of decreases). Alternate quantiles (than the median) had similar overall performance, though upper quantiles were better at capturing increasing phases (e.g., 95th quantile had 38% precision and 46% recall for increases), at the expense of reduced performance in flat periods (37%/46%, Figure S36-Figure S37).

Performance and goals varied over a changing pandemic

SMH performance varied across different stages of the pandemic. The earliest SMH scenarios (Rounds 1-4) confronted a period of high uncertainty about vaccine supply and the ongoing effect of NPIs. Still, ensemble performance on forecast metrics (WIS, coverage) for plausible scenario-weeks was comparable to average performance across rounds (Figure 3, Figure S42). Of note, the ensemble did not anticipate the increasing and decreasing trends of the Alpha wave well despite including the variant in scenario definitions, with Round 3 missing increases and Round 4 anticipating an overly long and large wave (Figure S38).

In Rounds 6-7, SMH projections missed the timing and magnitude of the Delta wave, despite scenario assumptions bracketing Delta's transmissibility in both rounds, and vaccine assumptions in Round 6. SMH ensemble performance on forecast metrics was the worst of any period, and trend classification was below par. This miss is likely the result of multiple factors: unexpectedly rapid waning of vaccine protection, differences in

the epidemiology of the Delta variant (serial interval, intrinsic severity), and changing human behavior in response to the early-summer lull in cases. As more information became available about the Delta variant, SMH projections improved in Round 9 both for forecast metrics and anticipation of epidemic trends (Figure S45).

During the initial Omicron wave (Rounds 11-12), SMH scenarios anticipated properties of the Omicron variant (all axes bracketed reality), and projections captured weekly trajectories and trends particularly well over the 3 month time horizon. Notably, these were the only rounds without significant truncation where the SMH ensemble outperformed the 4-week ahead forecast for cases and hospitalizations. It is not completely clear why the SMH was able to perform so well during this period. However, scenario designs were well informed by preliminary data from South Africa and heterogeneity in epidemic drivers was low over the projection period (due to high immune escape and relatively stable human behavior), mitigating many of the types of uncertainty that cause particular difficulties for long term epidemic projections.

The first SMH round of the post-Omicron era, Round 13, considered uncertainties about waning immunity and the emergence of a hypothetical immune escape variant. Performance was poor on all statistics and degraded quickly with projection horizon, despite waning assumptions that were consistent with later literature (*33*). There was substantial disagreement between models, and projections from some models were highly sensitive to subtle differences in assumptions about the exact trajectory of waning immunity, even when average duration and final protection levels were held

constant (Figure S60). Model disagreement and poor performance may have been further driven by low incidence (hence low information) at the time of calibration.

In contrast, the last three rounds considered here (Rounds 14-16) performed well on forecast metrics over the 18-41 evaluable weeks (key data sources became unavailable in March 2023, truncating evaluation). These rounds considered variants with different levels of immune escape and the approval and uptake of bivalent boosters. In these rounds, the SMH ensemble anticipated the occurrence of subsequent waves, was roughly accurate as to their scale, but was less accurate in projecting their timing. Of note, in Round 16 the focus shifted from individual new variants to broad categories of variants with similar levels of immune escape, in an attempt to account for the increasingly complex landscape of SARS-CoV-2 genetic diversity. Still, competition between variants and the resulting dynamics of strain replacement presented challenges for scenario design.

Building on the SMH experience

Since December 2020, SMH has convened multiple modeling teams to produce frequent, real-time, probabilistic projections of COVID-19 outcomes over a 3-12 month horizon based on well-defined scenarios. Scenario assumptions bracketed future conditions (where evaluable) the majority of the time, but the relevance of scenarios was frequently truncated by the emergence of unanticipated variants. For projected weeks where scenario assumptions were considered closest to subsequently observed reality, a trimmed linear opinion pool ensemble was far more reliable than any individual

model, though anticipating epidemic trends, especially in periods of increasing incidence, remained a challenge. The broad reliability of the ensemble, combined with the alignment of multiple teams on shared questions, helped SMH to become an important source of information for a variety of groups ranging from the media(*42*) to federal and local public health agencies (e.g., (*1*, *5*, *19*)).

SMH projections have played an important role in informing the pandemic response to new variants (*20*, *21*) and vaccine interventions (*5*, *19*). For example, projections from Rounds 6 and 7 sounded an important warning about likely resurgences due to the Delta variant (*21*), even though performance was poor. Similarly, Round 11 provided important (and ultimately accurate) information about the size and speed of the coming Omicron wave. Notably, SMH projections have provided key information to guide policy recommendations. Round 9 addressed potential population-level benefits of childhood vaccination (*20*), and Rounds 14 and 15 directly informed the decision to recommend bivalent boosters for a wide age range starting September 2022 (*19*). These public health impacts depended on the timely release (Figure 1) of projections from scenarios that were both relevant to emergent policy questions and tractable to modeling teams. Consistently fulfilling these goals required frequent meetings and conversations between the coordination team, public health collaborators and modeling teams. This process fostered a vibrant scientific community that has been critical to the SMH's success.

Here we have evaluated how SMH scenarios and projections compare to real-world events, with a specific focus on incident cases, hospitalizations and deaths. However, scenario projections may be used in a myriad of ways, and the value of SMH outputs for many of these uses may not directly depend on scenario bracketing or calibration to incident outcomes in plausible scenario-weeks as assessed here. For instance, if the primary goal is to inform a decision about whether or how to implement some intervention, it is the contrast of scenarios with and without that intervention that is important (*5*, *15*, *16*). Alternatively, one might use the full set of scenarios to allocate resources or inform response plans to potential surges in disease incidence; in this case, we might evaluate the extent to which planning around extremes from pessimistic projections would have led to over- or under-allocation of resources. Our current analysis makes no attempt to directly assess SMH performance for either of these goals (nor to the many other possibilities). Assessing the value added by SMH in these settings would require targeted analyses, and remains an important avenue for future research.

Our analysis of scenario bracketing and model calibration has methodological limitations. We lacked data to evaluate scenario definitions regarding NPIs and certain characteristics of emergent variants, limiting our ability to identify a single most plausible scenario. Teams also had discretion on how to apply vaccination specifications and other scenario assumptions at finer spatial scales; consequently, we did not evaluate scenario plausibility at the state level, although it may have varied dramatically there. We chose to evaluate SMH projections based on a set of plausible scenario-weeks, but

did not account for variability in how closely these plausible weeks matched reality. A complementary approach that may offer better assessments of model calibration would be to re-run scenario projections retrospectively with updated assumptions based on later subsequently observed data. Lastly, without a good "null" model it is hard to evaluate the added value of the SMH projections, and lack of historic data and the nature of planning scenarios makes design of such a null model difficult.

The scenario approach is an attempt to provide useful projections in the face of the many complexities that make predicting epidemics difficult. One of the most important complexities is the multiple, interacting drivers of disease dynamics that are themselves difficult to predict, such as ever evolving pathogen characteristics and human behavior. Although the scenario approach allows us to provide projections despite these complexities, only a subset of possible futures are explored. Therefore, it is essential to design scenarios that are useful – narrowing in on the possible futures that will best inform present actions. The fast timescale and multi-wave nature of infectious disease outbreaks often means we have little time to deeply consider both scenario design and model implementation in real-time, but it allows us to learn about the system and refine our approaches to scenario design and epidemic modeling more quickly than is possible in other systems (e.g., climate (*43*)).

Since its inception, SMH has disseminated nearly 1.8 million unique projections, making it one of the largest multi-team infectious disease scenario modeling efforts to date (other notable efforts have been documented, including multi-model estimation of

vaccination impact (*44*)). The SMH process has already been replicated in other settings (*45*) and for other pathogens (*46*). Looking to the future, the lessons learned and developing hub infrastructure (*47*), help to provide a more effective, coordinated, and timely response to emerging pandemic threats. It will be advantageous to launch multi-model efforts for scenario planning, forecasting, and inference in the early stages of future pandemics, when the most critical, time-sensitive decisions need to be made and uncertainty is high. To do this effectively, we can build on the SMH and other efforts from the COVID-19 response, by continuing “peace time” research into how to better collect and use data, construct scenarios, build models and ensemble results. As part of an evidence-based pandemic response, scenario modeling efforts like SMH can support decision making through improved predictive performance of multi-model ensembles and well-defined shared scenarios.

Methods

Overview of evaluation approaches for scenario projections

When evaluating the distance between a scenario projection and an observation, there are two potential factors at play: (1) the scenario assumptions may not match reality (e.g., scenario-specified vaccine uptake may underestimate realized uptake), and (2) if there were to be alignment between the scenario specifications and reality, model predictions may be imperfect due to miscalibration. The difference between projections and observations is a complex combination of both sources of disagreement, and importantly, observing projections that are close to observations does not necessarily

imply projections are well-calibrated (i.e., for scenarios very far from reality, we might expect projections to deviate from observations). To address both components, we evaluated the plausibility of COVID-19 Scenario Modeling Hub (SMH) scenarios and the performance of SMH projections (ensemble and component models). A similar approach has been proposed by Hausfather *et al.* (*43*). Below, we describe in turn the component models contributing to SMH, the construction of the ensemble, the evaluation of scenario assumptions, and our approaches to estimating model calibration and SMH performance.

Models submitting projections to SMH

Over the course of the first sixteen rounds of SMH, thirteen independent models submitted projections, with most submitting to multiple rounds. The majority of submitting models were mechanistic compartmental models, though there was one semi-mechanistic model and two agent-based models. Some models were calibrated to, and made projections at, the county level, whereas others were calibrated and made projections at the state level; many, but not all, had age structure. We have provided an overview of each model in Table S1. As models change each round to accommodate different scenarios and adapt to the evolving pandemic context, we chose not to focus here on model-specific differences (in structure, parameters, or performance). For more information on round-specific implementations, we direct readers to other publications with details (*5*, *21*).

Inclusion criteria and projections used for evaluation

Our analysis included state- and national-level projections of weekly incident cases, hospitalizations, and deaths from individual models and various ensembles for fourteen of the first sixteen rounds of SMH (Rounds 8 and 10 were not released publicly, and therefore are not included; see also Table S2 for a list of jurisdictions included). Each round included projections from between 4 and 9 individual models as well as ensembles. For a given round, modeling teams submitted projections for all weeks of the projection period, all targets (i.e., incident or cumulative cases, hospitalizations, and deaths), all four scenarios, and at least one location (i.e., states, territories, and national). Here, we evaluated only individual models that provided national projections in addition to state-level projections (i.e., excluding individual models that did not submit a national projection, though projections from these models are still included in the state-level ensembles that were evaluated). Submitted projections that did not comply with SMH conditions (e.g., for quantifying uncertainty or defining targets) were also excluded (0.8% of all submitted projections). Detailed description of exclusions can be found in Table S2.

Probabilistic projections and aggregation approaches

Modeling teams submitted probabilistic projections for each target via 23 quantiles (e.g., teams provided projected weekly incident cases for Q1, Q2.5, Q5, Q10, Q20, …, Q80, Q90, Q95, Q97.5, and Q99). We evaluated 3 methods for aggregating projections: untrimmed-LOP, trimmed-LOP (variations of probability averaging or linear opinion pool (*18*), LOP), and median-Vincent (variation of quantile or Vincent averaging (*35*, *36*) which is also used by other hubs (*11*)).

The untrimmed-LOP is calculated by taking an equally weighted average of cumulative probabilities across individual models at a single value. Because teams submitted projections for fixed quantiles, we used linear interpolation between these value-quantile pairs to ensure that all model projections were defined for the same values. We assumed that all projected cumulative probabilities jump to 0 and 1 outside of the defined value-quantile pairs (i.e., Q1-Q99). In other words, for a projection defined by cumulative distribution $F(x)$ with quantile function $F^{-1}(x)$, we assume that $F(x) = 0$ for all $x < F^{-1}(0.01)$ and $F(x) = 1$ for all $x > F^{-1}(0.99)$.

The trimmed-LOP uses exterior cumulative distribution function (CDF) trimming(*48*) of the two outermost values to reduce the variance of the aggregate, compared to the untrimmed-LOP (i.e., the prediction intervals are narrower). To implement this method, we follow the same procedure as the untrimmed-LOP, but instead of using an equally-weighted average, we exclude the highest and lowest quantiles at a given value and equally weight all remaining values in the average. Under this trimming method, the exclusions at different values may be from different teams.

The median-Vincent aggregate is calculated by taking the median value for each specified quantile. These methods were implemented using the CombineDistributions package (*37*) for the R statistical software (*49*).

Scenario plausibility

Projections in each SMH round were made for 4 distinct scenarios that detailed potential interventions, changes in behavior, or epidemiologic situations (Figure 1). These scenarios were designed approximately one month before projections were

submitted, and therefore 4-13 months before the end of the projection period, depending on the round's projection horizon. Scenario assumptions, especially those about vaccine efficacy or properties of emerging viral variants, were based on the best data and estimates available at the time of scenario design (these were often highly uncertain). Here, our purpose was to evaluate SMH scenario assumptions using the best data and estimates currently available, after the projection period has passed. We assessed SMH scenarios from two perspectives:

1. based on their *prospective* purpose: we identified whether scenarios "bracketed" reality along each uncertainty axis (i.e., one axis of the 2x2 table defining scenarios, based on one key source of uncertainty for the round). Scenarios in most SMH rounds were designed to bracket true values of key epidemic drivers (though the true value was not known at the time of scenario design). In other words, along each uncertainty axis in an SMH round, scenarios specified two levels along this axis (e.g., "optimistic" and "pessimistic" assumptions). Here we tested whether the realized value fell between those two assumptions (if so, we call this "bracketing").
2. for *retrospective* evaluation of calibration: we identified the set of plausible scenario-weeks for each round. One of our primary goals in this analysis was to assess and compare the calibration of different approaches (e.g., individual models, SMH ensemble, null comparator models). To assess this in the most direct way possible, we chose scenarios and projection weeks that were close to what actually happened (i.e., we isolated error due to calibration by minimizing

deviation between scenarios and reality; see *Overview of evaluation approaches for scenario projections* for details).

An "evaluable" scenario axis was defined as an axis for which assumptions could be confronted with subsequently observed data on epidemic drivers; in some instances, we could not find relevant data and the axis was not considered evaluable (e.g., NPI, see below). To evaluate scenario assumptions, we used external data sources and literature (Table S3). Due to differences across these sources, we validated each type of scenario assumption differently (vaccination, NPI, and variant characteristics; Figure 2), as described in detail below and in Table S3. Vaccine specifications and realized coverage are shown in Figure S2-Figure S5, while details of our round-by-round evaluation are provided below.

Rounds 1-4 concentrated on the early roll-out of the vaccine in the US and compliance with NPIs. To evaluate our vaccine assumptions in these rounds, we used data on reported uptake from the US Centers for Disease Control and Prevention database (*50*). For these rounds, scenarios prescribed monthly national coverage (state-specific uptake was intentionally left to the discretion of the modeling teams), so we only used national uptake to evaluate the plausibility of each vaccination scenario (Figure S2). In these scenarios, "bracketing" was defined as reality falling between cumulative coverage in optimistic and pessimistic scenarios for 50% or more of all projection weeks. The "plausible" scenario was that scenario with the smallest absolute difference between cumulative coverage in the final projection week (or in cases of variant emergence, the last week of projections before emergence; details below) and the observed cumulative coverage. We also considered choosing the plausible scenario

via the cumulative difference between observed and scenario-specified coverage over the entire projection period; this always led to selecting the same scenario as plausible.

When scenarios specified a coverage threshold, we compared assumptions with the reported fraction of people vaccinated at the end of the projection period. For instance, in round 2 scenario C and D, we stipulated that coverage would not exceed 50% in any priority group, but reported vaccination exceeded this threshold. In Rounds 3-4, the prescribed thresholds were not exceeded during the truncated projection period.

By Round 5 (May 2021), vaccine uptake had started to saturate. Accordingly, in rounds 5-7, vaccine assumptions were based on high and low saturation thresholds that should not be exceeded for the duration of the projection period, rather than monthly uptake curves. For these rounds, we evaluated which of the prescribed thresholds was closest to the reported cumulative coverage at the end of the projection period (Figure S3). Later rounds took similar approaches to specifying uptake of childhood vaccination (Round 9) and bivalent boosters (Round 14-16). Rounds 9 (Figure S4), and 14-15 (Figure S5) specified weekly coverage and Round 16 specified a coverage threshold; we followed similar approaches in evaluating these scenarios.

For vaccine efficacy assumptions, we consulted population-level studies conducted during the period of the most prevalent variant during that round (Table S3). Similarly, for scenarios about emerging viral variants (regarding transmissibility increases, immune escape, and severity) and waning immunity, we used values from the literature as a ground truth for these scenario assumptions. We identified the most

realistic scenario as that with the assumptions closest to the literature value (or average of literature values if multiple were available, Table S3).

Rounds 1-4 included assumptions about NPIs. We could not identify a good source of information on the efficacy and compliance to NPIs that would match the specificity prescribed in the scenarios (despite the availability of mobility and policy data, e.g., Hallas *et al.* (*51*)). Rounds 13-15 included assumptions about immune escape and severity of hypothetical variants that may have circulated in the post-Omicron era. Round 16 considered broad variant categories based on similar levels of immune escape, in response to the increasing genetic diversity of SARS-CoV-2 viruses circulating in fall 2022. There were no data available for evaluation of immune escape assumptions after the initial Omicron BA1 wave. As such, NPI scenarios in Rounds 1-4 and immune escape variant scenarios in Rounds 13-16 were not “evaluable” for bracketing analyses, and therefore we considered all scenarios realistic in these cases. Overall, across 14 publicly released rounds, we identify a single most realistic scenario in 7 rounds, and two most realistic scenarios in the other 7.

Finally, in some rounds, a new viral variant emerged during the projection period that was not specified in the scenarios for that round. We considered this emergence to be an invalidation of scenario assumptions, and removed these weeks from the set of plausible scenario-weeks. Specifically, emergence was defined as the week after prevalence exceeded 50% nationally according to outbreak.info variant reports (*52*–*54*), accessed via outbreak.info R client (*55*). Accordingly, the Alpha variant (not anticipated in Round 1 scenarios) emerged on 3 April 2021, the Delta variant (not anticipated in

Rounds 2-5) emerged on 26 June 2021, and the Omicron variant (not anticipated in Round 9) emerged on 25 December 2021.

Comparator models

To assess the added value of SMH projections against plausible alternative sources of information, we also assessed null models or other benchmarks. Null models based on historical data were not available here (e.g., there was no prior observation of COVID-19 in February in the US when we projected February 2021). There are many potential alternatives, and here we used three comparative models: naive, 4-week forecast, and trend-continuation.

The baseline "naive" model was generated by carrying recent observations forward, with variance based on historical patterns (Figure S13-Figure S15). We used the 4-week ahead "baseline" model forecast from the COVID-19 Forecast Hub (*11*) for the first week of the projection period as the naive model, and assumed this projection held for the duration of the projection period (i.e., this forecast was the "naive" projection for all weeks during the projection period). Because the COVID-19 Forecast Hub collects daily forecasts for hospitalizations, we drew 1,000 random samples from each daily distribution in a given week and summed those samples to obtain a prediction for weekly hospitalizations. The naive model is flat and has relatively large prediction intervals in some instances.

As a forecast-based comparator, we used the COVID-19 Forecast Hub "COVIDhub-4_week_ensemble" ensemble model (Figure S7-Figure S9). This model includes forecasts (made every week) from multiple component models (e.g., on

average 41 component models between January and October 2021 (*11*)). We obtained weekly hospitalization forecasts from the daily forecasts of the COVID-19 Forecast Hub using the same method as the naive model. This 4-week forecast model is particularly skilled at death forecasts (*11*); however, in practice, there is a mismatch in timing between when these forecasts were made and when SMH projections were made. For most SMH projection weeks, forecasts from this model would not yet be available (i.e., projection horizons more than 4 weeks into the future); yet, for the first 4 weeks of the SMH projection period, SMH projections may have access to more recent data. It should also be noted that the team running the COVID-19 Forecast Hub has flagged the 4-week ahead predictions of cases and hospitalizations as unreliable (*56*). Further, SMH may be given an "advantage" by the post-hoc selection of plausible scenario-weeks based on the validity of scenario assumptions.

Finally, the trend-continuation model was based on a statistical generalized additive model (Figure S10-Figure S12). The model was fit to the square root of the 14-day moving average with cubic spline terms for time, and was fit separately for each location. We considered inclusion of seasonal terms, but there were not enough historic data to meaningfully estimate any seasonality. For each round, we used only one year of data to fit the model, and projected forward for the duration of the projection period. The SMH ensemble consistently outperformed this alternative comparator model (see Figure S16-Figure S21).

Projection performance

Prediction performance is typically based on a measure of distance between projections and "ground truth" observations. We used the Johns Hopkins CSSE dataset (*57*) as a source of ground truth data on reported COVID-19 cases and deaths, and U.S. Health and Human Services Protect Public Data Hub (*58*) as a source of reported COVID-19 hospitalizations. These sources were also used for calibration of the component models. CSSE data were only produced through 4 March 2023, so our evaluation of Rounds 13-16 ended at this date (1 week before the end of the 52 week projection period in Round 13, 11 weeks before the end of the 52 week projection period in Round 14, 9 weeks before the end of the 40 week projection period in Round 15, and 8 weeks before the end of the 26 week projection period in Round 16).

We used two metrics to measure performance of probabilistic projections, both common in the evaluation of infectious disease predictions. To define these metrics, let F be the projection of interest (approximated by a set of 23 quantile-value pairs) and o be the corresponding observed value. The "α% prediction interval" is the interval within which we expect the observed value to fall with α% probability, given reality perfectly aligns with the specified scenario.

1. **Ninety-five percent (95%) coverage** measures the percent of projections for which the observation falls within the 95% projection interval. In other words, 95% coverage is calculated as

$$C_{95\%}(F, o) = \frac{1}{N} \sum_{i=1}^{N} 1(F^{-1}(0.025) \leq o \leq F^{-1}(0.975))$$

where $1(\cdot)$ is the indicator function, i.e., $1(F^{-1}(0.025) \leq o \leq F^{-1}(0.975)) = 1$ if the observation falls between the values corresponding to Q2.5 and Q97.5, and

is 0 otherwise. We calculated coverage over multiple locations for a given week (i.e., $i = 1 \ldots N$ for N locations), or across all weeks and locations.

2. **Weighted interval score (WIS)** measures the extent to which a projection captures an observation, and penalizes for wider prediction intervals (*34*). First, given a projection interval (with uncertainty level α) defined by upper and lower bounds, $u = F^{-1}\left(1 - \frac{\alpha}{2}\right)$ and $l = F^{-1}\left(\frac{\alpha}{2}\right)$, the interval score is calculated as

$$IS_\alpha(F, o) = (u - l) + \frac{2}{\alpha}(l - o)1(o < l) + \frac{2}{\alpha}(o - u)1(u < o)$$

where again, $1(\cdot)$ is the indicator function. The first term of IS_α represents the width of the prediction interval, and the second two terms are penalties for over- and under-prediction, respectively. Then, using weights that approximate the continuous rank probability score (*59*), the weighted interval score is calculated as

$$WIS(F, o) = \frac{1}{K + 1/2}\left(\frac{1}{2}|o - F^{-1}(0.5)| + \sum_{i=1}^{K} \frac{\alpha_K}{2} IS_\alpha\right)$$

Each projection is defined by 23 quantiles comprising 11 intervals (plus the median), which we used for the calculation of WIS (i.e., we calculated IS_α for $\alpha = 0.02, 0.05, 0.1, 0.2, \ldots, 0.8, 0.9$ and $K = 11$).

It is worth noting that these metrics do not account for measurement error in the observations.

WIS values are on the scale of the observations, and therefore comparison of WIS across different locations or phases of the pandemic is not straightforward (e.g., the scale of case counts is very different between New York and Vermont). For this reason, we generated multiple variations of WIS metrics to account for variation in the

magnitude of observations. First, for **average normalized WIS** (Figure 3B)**,** we calculated the standard deviation of WIS, $\sigma_{s,w,t,r}$, across all scenarios and models for a given week, location, target, and round and divided the WIS by this standard deviation (i.e., WIS/$\sigma_{s,w,t,r}$). Doing so accounts for the scale of that week, target, and round, a procedure implemented in analyses of climate projections (*60*). Then, we averaged normalized WIS values across strata of interest (e.g., across all locations, or all locations and weeks). Other standardization approaches that compute WIS on a log scale have been proposed (*61*), though may not be as well suited for our analysis which focuses on planning and decision making.

An alternative rescaling introduced by Cramer *et al.* (*11*), **relative WIS,** compares the performance of a set of projections to an "average" projection. This metric is designed to compare performance across predictions from varying pandemic phases. The relative WIS for model *i* is based on pairwise comparisons (to all other models, j) of average WIS. We calculated the average WIS across all projections in common between model i and model j, where $WIS(i)$ and $WIS(j)$ are the average WIS of these projections (either in one round, or across all rounds for "overall") for model i and model j, respectively. Then, relative WIS is the geometric average of the ratio, or

$$relative\ WIS = \left(\prod_{j=1}^{N} \frac{WIS(i)}{WIS(j)} \right)^{1/N}$$

When comparing only two models that have made projections for all the same targets, weeks, locations, rounds, etc. the relative WIS is equivalent to a simpler metric, the ratio of average WIS for each model (i.e., $\frac{WIS(i)}{WIS(j)}$). We used this metric to compare each scenario from SMH ensemble to the 4-week forecast model (Figure 5). For this

scenario comparison, we provided bootstrap intervals by recalculating the ratio with an entire week of projections excluded (all locations, scenarios). We repeated this for all weeks, and randomly drew from these 1,000 times. From these draws we calculated the 5th and 95th quantiles to derive the 90% bootstrap interval, and we assumed performance is significantly better for one scenario over the others if the 90% bootstrap intervals do not overlap. We also used this metric to compare the ensemble projections to each of the comparative models (Figure S22).

Trend classification

In addition to traditional forecast evaluation metrics, we assessed the extent to which SMH projections predict the qualitative shape of incident trajectories (whether trends will increase or decrease). We modified a method from McDonald *et al.* (*40*) to classify observations and projections as “increasing”, “flat” or “decreasing”. First, we calculated the percent change in observed incident trajectories on a two week lag (i.e., $log(o_T + 1) - log(o_{T-2} + 1)$ for each state and target). We took the distribution of percent change values across all locations for a given target and set the threshold for a decrease or increase assuming that 33% of observations will be flat (Figure S23). Based on this approach, decreases were defined as those weeks with a percent change value below -23% for incident cases, -17% for incident hospitalizations, and -27% for incident deaths, respectively. Increases have a percent change value above 14%, 11%, 17%, respectively. See Figure S34 for classification results with a one week lag and different assumptions about the percent of observations that are flat.

Then, to classify trends in projections, we again calculated the percent change on a two week lag of the projected median (we also consider the 75th and 95th quantiles because our aggregation method is known to generate a flat median when asynchrony between component models is high). For the first two projection weeks of each round, we calculated the percent change relative to the observations one and two weeks prior (as there are no projections to use for reference in the week prior, and two weeks prior, projection start date). We applied the same thresholds from the observations to classify a projection, and compared this classification to the observed classification. This method accounts for instances when SMH projections anticipate a change in trajectory but not the magnitude of that change (see Figure S44), and it does not account for instances when SMH projections anticipate a change but miss the timing of that change (this occurred to some extent in Rounds 6 and 7, Delta variant wave). See Figure S24-Figure S33 for classifications of all observations and projections.

We assessed how well SMH projections captured incident trends using precision and recall, two common metrics in evaluating classification tasks with three classes: "increasing", "flat", and "decreasing" (*41*). To calculate these metrics, we grouped all projections by the projected and the observed trend (as in Figure 4D). Let N_{po} be the number of projections classified by SMH as trend p (rows of Figure 4D) and the corresponding observation was trend o (columns of Figure 4D). Then, for class i,

1. *precision* is the fraction of projections correctly classified as i, out of the total number of projections classified as i, or

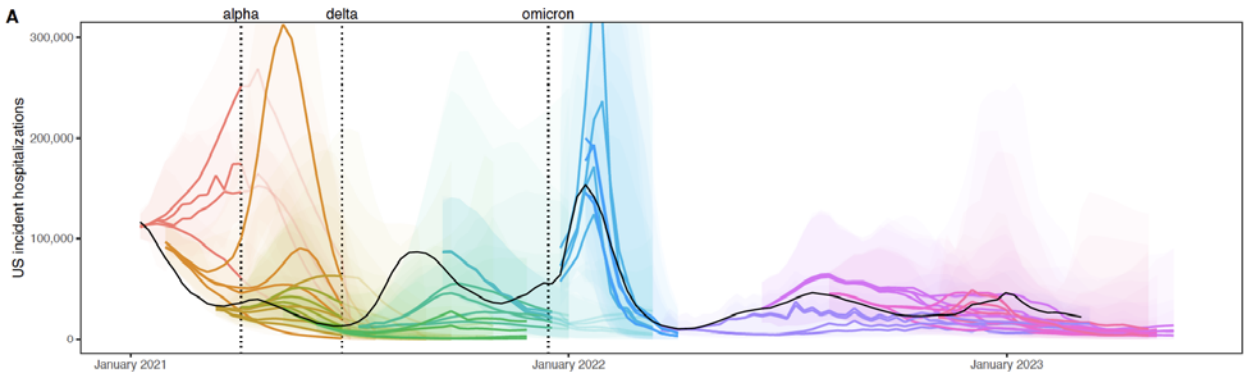
$$precision = \frac{N_{ii}}{\sum_{j=1}^{3} N_{ij}}$$

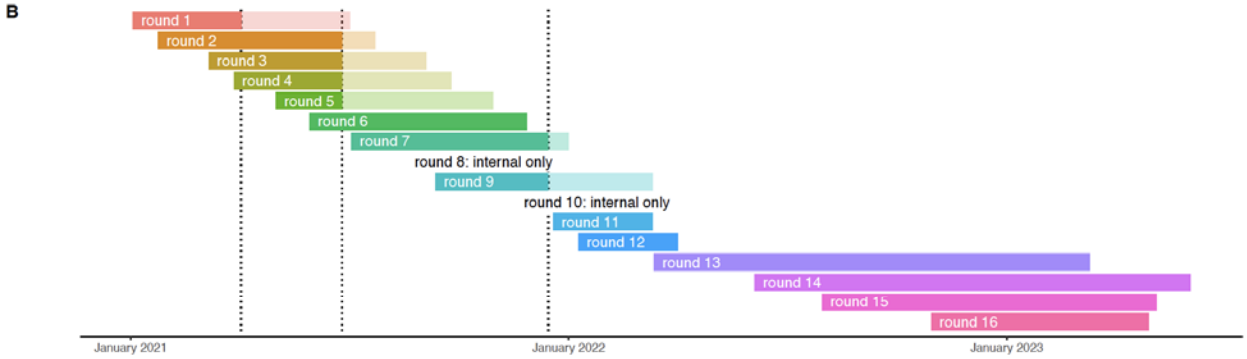
For example, the precision of increasing trends is the number of correctly classified increases divided by the total number of projections classified as increasing.

2. *recall* is the fraction of projections correctly classified as i, out of the total number of projections observed as i, or

$$recall = \frac{N_{ii}}{\sum_{j=1}^{3} N_{ji}}$$

For example, the recall of increasing trends is the number of correctly classified increases divided by the total number of observations that increased.





C

	round 1	round 2	round 3	round 4	round 5	round 6	round 7	round 9	round 11	round 12	round 13	round 14	round 15	round 16
number of models	5	5	4	6	8	9	9	9	8	8	8	6	5	6
turnaround time	58 days	27 days	31 days	34 days	29 days	23 days	16 days	23 days	21 days	11 days	51 days	65 days	25 days	39 days
NPI assumptions	NPI compliance	NPI compliance	NPI compliance	NPI compliance	NPI compliance									
vaccination assumptions	vaccine supply	vaccine supply	vaccine supply	vaccine supply	vaccine uptake	vaccine uptake	vaccine uptake	childhood vaccination				booster uptake	booster uptake	booster uptake
variant assumptions		emergence (alpha)				transmissibility (delta)	transmissibility (delta)	emergence (hypothetical)	severity/ immune escape (omicron)	severity/ immune escape (omicron)	immune escape (hypothetical)	immune escape (hypothetical)	immune escape (hypothetical)	immune escape (mix of variants)

Figure 1: Sixteen rounds of U.S. COVID-19 Scenario Modeling Hub (SMH) projections. Between February 2021 and November 2022, SMH publicly released fourteen rounds of projections with four scenarios per round. Each round is shown in a different color (internal Rounds 8 and 10 not shown). (A) Median (line) and 95% projection interval (ribbon, the interval within which we expect the observed value to fall with 95% probability, given reality perfectly aligns with the scenario) for U.S. weekly incident hospitalizations for four scenarios per round from the SMH ensemble. Observed weekly U.S. incident hospitalizations are represented by the solid black line. (B) Timing of each round of SMH projections is represented by a projection start date and end date (start and end of bar). In panels (A) and (B), scenario specifications were invalidated by the emergence of Alpha, Delta, and Omicron variants in rounds that did not anticipate emergence. Variant emergence dates (estimated as the day after which national prevalence exceeded 50%) are represented by dotted vertical lines. (C) For each round, the table specifies the number of participating modeling teams, the turnaround time from finalization of scenarios to publication of projections, and scenario specifications about non-pharmaceutical interventions (NPIs), vaccination, and variant characteristics. Scenario specifications are shaded gray if scenarios “bracketed” the true values in our retrospective analysis (i.e., the true value fell between the two scenario assumptions on that uncertainty axis). Note, in Rounds 11 and 12 both scenario axes specified assumptions about variants, and both are included in the “variant assumptions'' cell. Not shown here, the second scenario axis for Round 13 specified assumptions about waning immunity, which bracketed waning estimates from a meta-analysis.

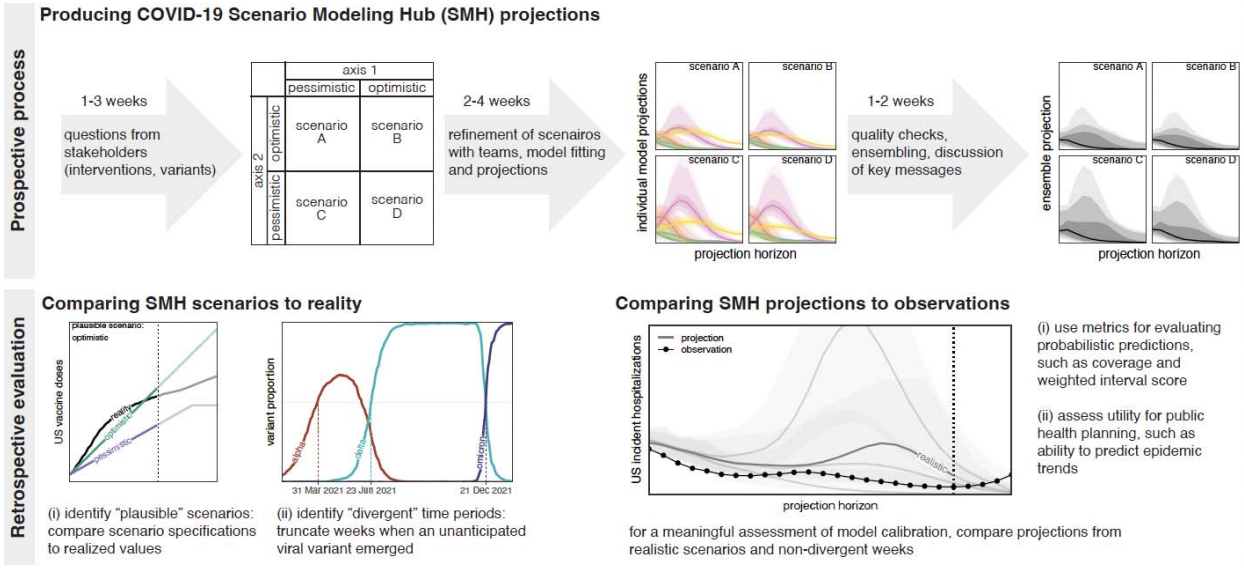


Figure 2: COVID-19 Scenario Modeling Hub (SMH) process. (top) Prospective SMH process: The SMH coordination team takes input from public health partners on key questions to design scenarios. Scenarios have a 2x2 structure (with the exception of Round 1), where two levels are specified along each of two axes of uncertainty or interventions, and all four combinations of these possibilities are considered (scenarios A-D). Scenarios are refined in discussion with modeling teams, after which teams each fit their model and make projections independently. Then, after quality checks, individual model projections are aggregated using linear opinion pool (i.e., probability averaging), and in discussion with the teams, key messages are determined. A report is shared with public health partners and projections are released on the public SMH website (https://covid19scenariomodelinghub.org). (bottom) Retrospective evaluation: Evaluating the SMH effort involves comparing SMH *scenario assumptions* to reality, and comparing SMH *projections* to observations. Comparing scenarios to reality is used to identify the most plausible scenario-weeks, namely the set of "plausible" scenarios in projection weeks where scenario specifications about variants did not diverge from actual variant prevalence. Horizontal dotted lines represent emergence of an unanticipated variant.

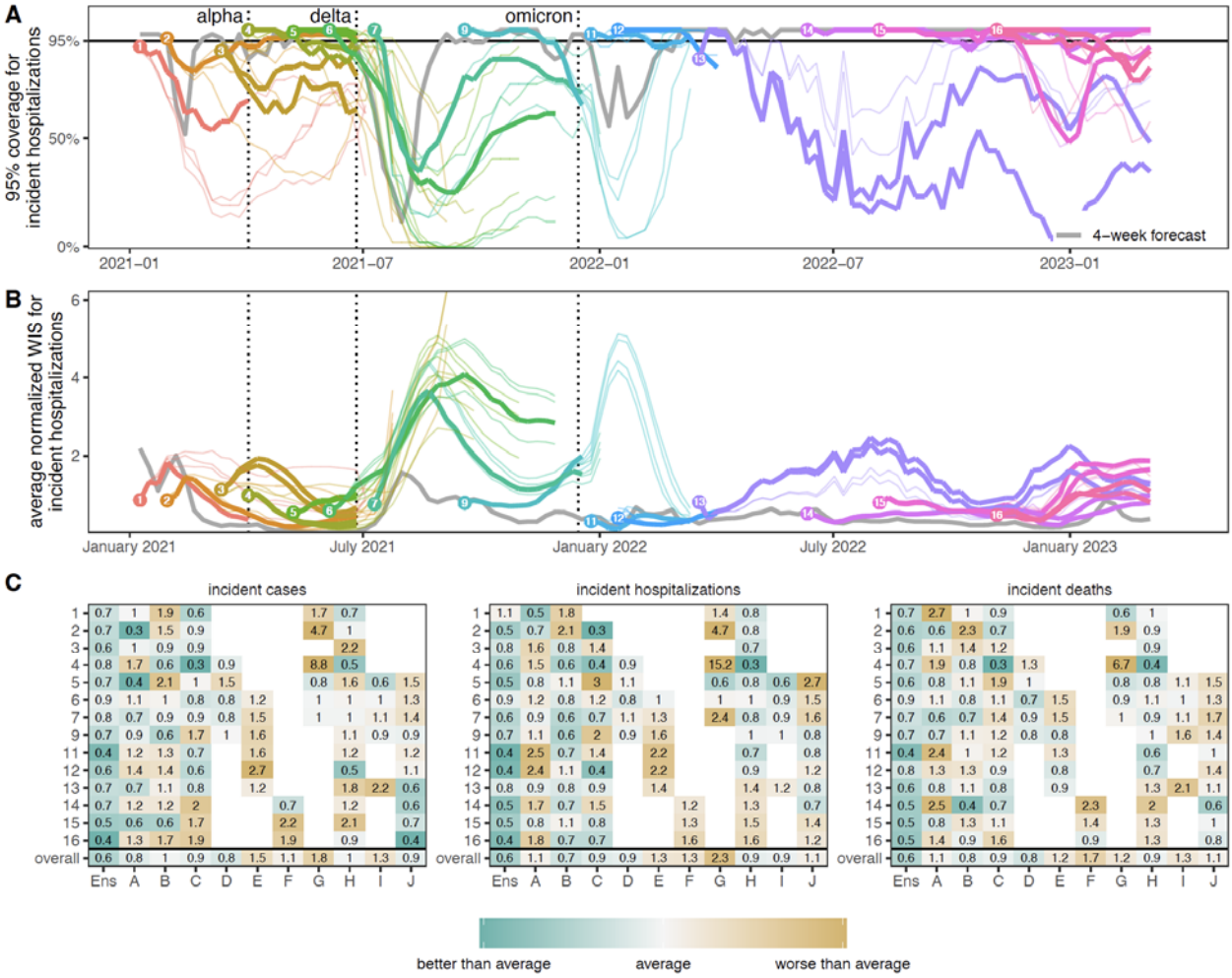


Figure 3: Performance of U.S. COVID-19 Scenario Modeling Hub (SMH) ensemble projections for weekly incident cases, hospitalizations, and deaths. (A) Coverage of SMH ensemble 95% projection interval across locations by round and scenario. Ideal coverage of 95% is shown as a horizontal black line. (B) Normalized weighted interval score (WIS) for SMH ensemble by round and scenario. Normalized WIS is calculated by dividing WIS by the standard deviation of WIS across all scenarios and models for a given week, location, target, and round. This yields a scale-free value, and we averaged normalized WIS across all locations for a given projection week and scenario. For (A) and (B), the round is indicated by color and a number at the start of the projection period. Each scenario is represented by a different line, with plausible scenario-weeks bolded (see Methods). Performance of the 4-wk ahead COVID-19 Forecast Hub ensemble is shown in gray. Vertical dotted lines represent emergence dates of Alpha, Delta, and Omicron variants. Evaluation ended on 10 March 2023, as the source of ground truth observations were no longer produced. (C) Relative WIS comparison of individual models (letters A-I) and SMH ensemble ("Ens") within rounds and overall. A relative WIS of 1 indicates performance equivalent to the "average" model (yellow colors indicate performance worse than average, and greens indicate performance better than average; the color scale is on a log scale and truncated at ±1, representing 2 standard deviations of relative WIS values). See Figure S46-Figure S47 for 50% and 95% coverage of all targets.

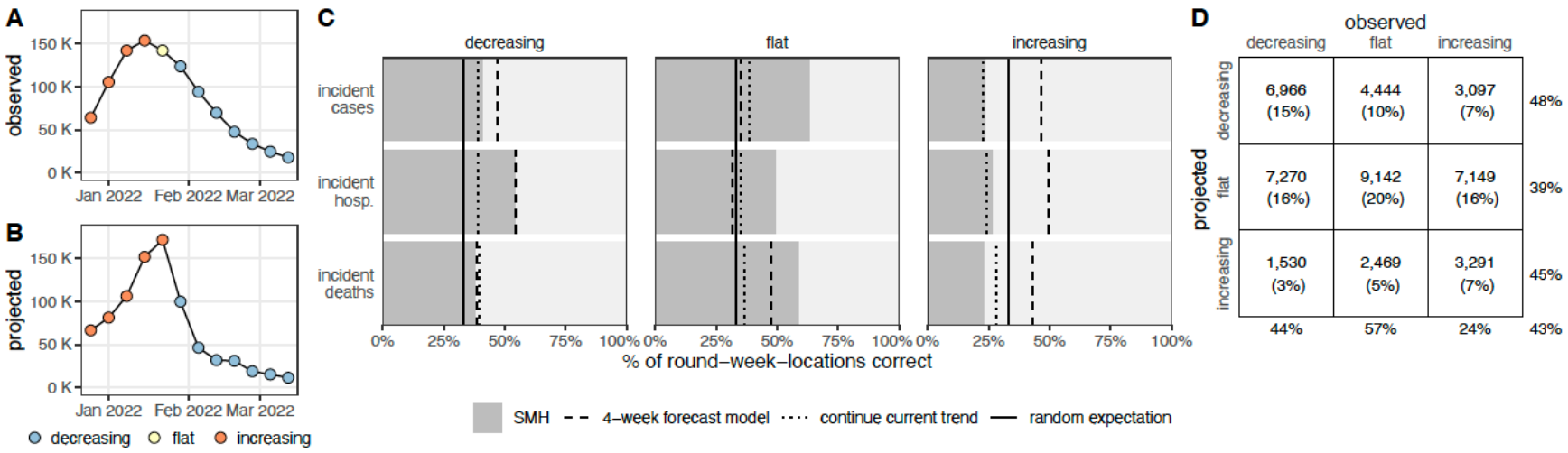


Figure 4: Evaluation of scenario projections to anticipate disease trends. Illustration of classification of increasing (orange), flat (yellow), and decreasing (blue) trends for observed United States incident hospitalizations (A) and U.S. COVID-19 Scenario Modeling Hub (SMH) ensemble projection median for the plausible scenario (B) using Round 11, at the start of the Omicron wave. Evaluation of trends across all rounds and locations for plausible scenario-weeks: (C) For decreasing, flat and increasing observations, percent of incident cases, hospitalizations and deaths correctly identified by SMH ensemble projection median (gray), the 4-week forecast model (dashed line), a model that continues current trend (dotted), and the expectation if observations are classified randomly (solid). (D) For decreasing, flat, and increasing observations in plausible scenario-weeks, the number (and percentage) of observations that are classified as decreasing, flat, or increasing by the SMH ensemble projection median. Totals are calculated across all targets and rounds (meaning that some weeks are included multiple times, and therefore although 33% of observations are in each category, 33% of projections may not be in each category) and weighted by the plausibility of the scenario and week (for rounds with multiple plausible scenarios, this could introduce decimal totals; we rounded values down in these cases). Percentages on the outside show the percent correct for a given observed classification (precision, columns) or projected classification (recall, rows). Projection classifications were also calculated for all scenarios and weeks, regardless of plausibility (Figure S35), using SMH ensemble projection Q75 (Figure S36) and SMH ensemble projection Q97.5 (Figure S37); see supplement for additional stratification of results (by round, Figure S39; by location, Figure S40; by projection horizon, Figure S41-Figure S42; and by variant period, Figure S43).

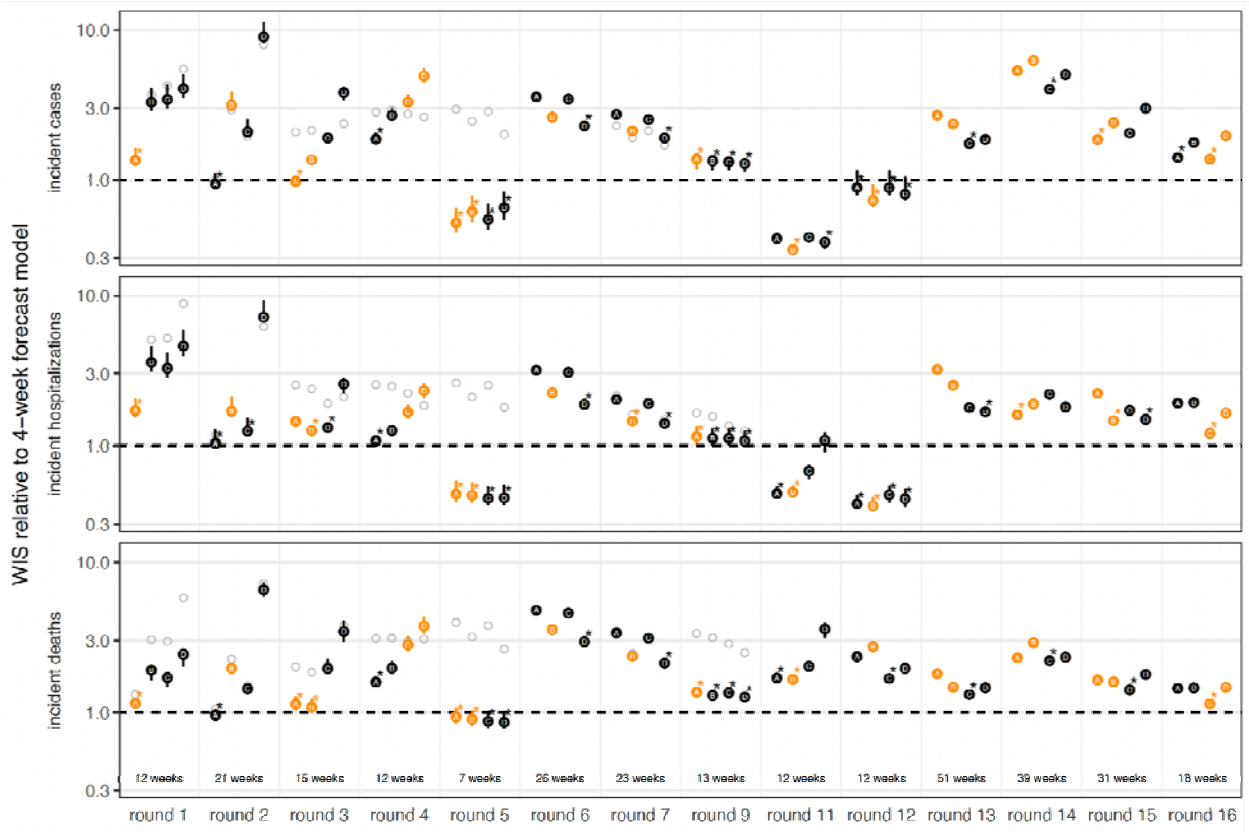


Figure 5: Relative performance of the four U.S. COVID-19 Scenario Modeling Hub (SMH) scenarios (A, B, C, D) across rounds. Weighted interval score (WIS) for SMH ensemble projections in plausible scenario-weeks relative to the 4-week forecast model (4-week ahead COVID-19 Forecast Hub ensemble). WIS is averaged across all locations and plausible scenario-weeks for a given target, round, and scenario. Scenarios deemed plausible are highlighted in orange (see Methods). The number of plausible weeks included in the average is noted at the bottom of the incident death panel. Results for all weeks are shown with gray open circles for comparison. A WIS ratio of one (dashed line) indicates equal average WIS, or equal performance, between the SMH ensemble and 4-week forecast model. Ninety percent (90%) bootstrap intervals (vertical lines around each point) are calculated by leaving out WIS for all locations in a given week (over 1,000 random draws, though most are very narrow and therefore not visible). In each round, the scenario with the lowest WIS ratio is denoted with an asterisk. Any scenario with a 90% bootstrap interval that overlaps the bootstrap interval of the scenario with the lowest WIS ratio is also denoted with an asterisk. WIS ratio is shown on the log scale.

Table 1: Scenario bracketing. For each of two axes per round, bracketing (or not) of reality by U.S. COVID-19 Scenario Modeling Hub (SMH) scenarios. Text color denotes successful (blue) or unsuccessful (red) bracketing. When vaccination scenarios specified coverage weekly, we considered bracketing in 50% or more of all projection weeks to be bracketing overall. For round 4, we use coverage of mRNA doses only to determine bracketing, as this makes up almost all of the assumed doses (i.e., we do not consider coverage of Johnson & Johnson). NPI = non-pharmaceutical intervention.

Round	Axis 1	Axis 2
1	bracket weekly vaccination coverage in 8 weeks out of 26 weeks (31%) and 8 out of 13 plausible weeks (61%)	no second bracketing axis
2	bracket weekly vaccination coverage in 26 out of 26 weeks (100%) and 22 out of 22 plausible weeks (100%)	bracket variant transmissibility estimates
3	bracket weekly vaccination coverage in 14 out of 26 weeks (54%) and 4 out of 16 plausible weeks (25%)	unable to assess NPI scenarios
4	bracket weekly vaccination coverage in 3 out of 26 weeks (12%) and 3 out of 13 plausible weeks (23%)	unable to assess NPI scenarios
5	bracket vaccination coverage at end of projection period	unable to assess NPI scenarios
6	bracket vaccination coverage at end of projection period	bracket variant transmissibility estimates
7	underestimate vaccination coverage in both scenarios	bracket variant transmissibility estimates
9	bracket weekly vaccination coverage in 19 out of 19 weeks (100%) and 13 out of 13 plausible weeks (100%)	no second bracketing axis
11	bracket variant transmissibility estimates	bracket variant severity estimates
12	bracket variant transmissibility estimates	bracket variant severity estimates
13	bracket immune waning estimates	unable to assess immune-escape variant scenarios
14	bracket vaccination coverage in 23 of 23 (100%) evaluated weeks (through March 20, 2023)	unable to assess immune-escape variant scenarios
15	bracket vaccination coverage in 9 of 24 (38%) evaluated weeks (through March 20, 2023)	unable to assess immune-escape variant scenarios
16	overestimate vaccination coverage in both scenarios	unable to assess immune-escape variant scenarios

Funding Acknowledgments

L. Contamin, J. Levander, J. Kerr, J. Espino, and H. Hochheiser were supported by NIGMS 5U24GM132013. E. Howerton, K. Shea and R. Borchering were supported by NSF RAPID awards DEB-2028301, DEB-2037885, DEB-2126278 and DEB-2220903. E. Howerton was supported by the Eberly College of Science Barbara McClintock Science Achievement Graduate Scholarship in Biology at the Pennsylvania State University. M. Chinazzi, J. T. Davis, K. Mu, X. Xiong, A. Pastore y Piontti, and A. Vespignani were supported by HHS/CDC 6U01IP001137, HHS/CDC 5U01IP0001137 and the Cooperative Agreement no. NU38OT000297 from the Council of State and Territorial Epidemiologists (CSTE). P. Porebski, S. Venkatramanan, A. Adiga, B. Lewis, B. Klahn, J. Outten, B. Hurt, H. Mortveit, A. Wilson, M. Marathe, J. Chen, S. Hoops, P. Bhattacharya, D. Machi acknowledge support from NIH Grant R01GM109718, VDH Grant PV-BII VDH COVID-19 Modeling Program VDH-21-501-0135, NSF Grant No. OAC-1916805, NSF Expeditions in Computing Grant CCF-1918656, NSF RAPID CCF-2142997, NSF RAPID OAC-2027541, US Centers for Disease Control and Prevention 75D30119C05935, DTRA subcontract/ARA S-D00189-15-TO-01-UVA, and UVA strategic funds. Model computation was supported by NSF XSEDE TG-BIO210084 and UVA; and used resources, services, and support from the COVID-19 HPC Consortium (https://covid19-hpc-consortium.org). A. Bouchnita, K. Bi, M. Lachmann, S. Fox and L. Meyers were supported by CSTE NU38OT000297, CDC Supplement U01IP001136-Suppl, and NIH Supplement R01AI151176-Suppl. E. Rosenstrom, J. Ivy, M. Mayorga, and J. Swann were supported by TRACS/NIH grant UL1TR002489; CSTE and CDC cooperative agreement no. NU38OT000297. **Disclaimer.** The findings and conclusions

in this report are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

This activity was reviewed by CDC and was conducted consistent with applicable federal law and CDC policy.[§]

[§]See e.g., 45 C.F.R. part 46, 21 C.F.R. part 56; 42 U.S.C. §241(d); 5 U.S.C. §552a; 44 U.S.C. §3501 et seq.

References

1. M. Biggerstaff, R. B. Slayton, M. A. Johansson, J. C. Butler, Improving Pandemic Response: Employing Mathematical Modeling to Confront Coronavirus Disease 2019. *Clinical Infectious Diseases*. **74**, 913–917 (2022).

2. C. J. E. Metcalf, D. H. Morris, S. W. Park, Mathematical models to guide pandemic response. *Science*. **369**, 368–369 (2020).

3. US Centers for Disease Control and Prevention, COVID-19 Pandemic Planning Scenarios. *Centers for Disease Control and Prevention* (2020), (available at https://www.cdc.gov/coronavirus/2019-ncov/hcp/planning-scenarios.html).

4. J. Taghia, V. Kulyk, S. Ickin, M. Folkesson, C. Nyström, K. gren, T. Brezicka, T. Vingare, J. Karlsson, I. Fritzell, R. Harlid, B. Palaszewski, M. Kjellberg, J. Gustafsson, Development of forecast models for COVID-19 hospital admissions using anonymized and aggregated mobile network data. *Sci Rep*. **12**, 17726 (2022).

5. R. K. Borchering, L. C. Mullany, E. Howerton, M. Chinazzi, C. P. Smith, M. Qin, N. G. Reich, L. Contamin, J. Levander, J. Kerr, J. Espino, H. Hochheiser, K. Lovett, M. Kinsey, K. Tallaksen, S. Wilson, L. Shin, J. C. Lemaitre, J. D. Hulse, J. Kaminsky, E. C. Lee, J. T. Davis, K. Mu, X. Xiong, A. P. y Piontti, A. Vespignani, A. Srivastava, P. Porebski, S. Venkatramanan, A. Adiga, B. Lewis, B. Klahn, J. Outten, B. Hurt, J. Chen, H. Mortveit, A. Wilson, M. Marathe, S. Hoops, P. Bhattacharya, D. Machi, S. Chen, R. Paul, D. Janies, J.-C. Thill, M. Galanti, T. Yamana, S. Pei, J. Shaman, G. Espana, S. Cavany, S. Moore, A. Perkins, J. M. Healy, R. B. Slayton, M. A. Johansson, M. Biggerstaff, K. Shea, S. A. Truelove, M.

C. Runge, C. Viboud, J. Lessler, Impact of SARS-CoV-2 vaccination of children ages 5–11 years on COVID-19 disease burden and resilience to new variants in the United States, November 2021–March 2022: A multi-model study. *Lancet Reg Health Am*. **17**, 100398 (2023).

6. H. Yang, Ö. Sürer, D. Duque, D. P. Morton, B. Singh, S. J. Fox, R. Pasco, K. Pierce, P. Rathouz, V. Valencia, Z. Du, M. Pignone, M. E. Escott, S. I. Adler, S. C. Johnston, L. A. Meyers, Design of COVID-19 staged alert systems to ensure healthcare capacity with minimal closures. *Nat Commun*. **12**, 3767 (2021).

7. R. T. Clemen, Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*. **5**, 559–583 (1989).

8. A. Timmermann, "Chapter 4 Forecast Combinations" in *Handbook of Economic Forecasting*, G. Elliott, C. W. J. Granger, A. Timmermann, Eds. (Elsevier, 2006; https://www.sciencedirect.com/science/article/pii/S1574070605010049), vol. 1, pp. 135–196.

9. M. A. Johansson, K. M. Apfeldorf, S. Dobson, J. Devita, A. L. Buczak, B. Baugher, L. J. Moniz, T. Bagley, S. M. Babin, E. Guven, T. K. Yamana, J. Shaman, T. Moschou, N. Lothian, A. Lane, G. Osborne, G. Jiang, L. C. Brooks, D. C. Farrow, S. Hyun, R. J. Tibshirani, R. Rosenfeld, J. Lessler, N. G. Reich, D. A. T. Cummings, S. A. Lauer, S. M. Moore, H. E. Clapham, R. Lowe, T. C. Bailey, M. García-Díez, M. S. Carvalho, X. Rodó, T. Sardar, R. Paul, E. L. Ray, K. Sakrejda, A. C. Brown, X. Meng, O. Osoba, R. Vardavas, D. Manheim, M. Moore, D. M. Rao, T. C. Porco, S. Ackley, F. Liu, L. Worden, M. Convertino, Y. Liu, A. Reddy, E. Ortiz, J. Rivero, H. Brito, A. Juarrero, L. R. Johnson, R. B. Gramacy, J. M. Cohen, E. A. Mordecai, C. C. Murdock, J. R. Rohr, S. J. Ryan, A. M. Stewart-Ibarra, D. P. Weikel, A. Jutla, R. Khan, M. Poultney, R. R. Colwell, B. Rivera-García, C. M. Barker, J. E. Bell, M. Biggerstaff, D. Swerdlow, L. Mier-Y-Teran-Romero, B. M. Forshey, J. Trtanj, J. Asher, M. Clay, H. S. Margolis, A. M. Hebbeler, D. George, J.-P. Chretien, An open challenge to advance probabilistic forecasting for dengue epidemics. *Proc. Natl. Acad. Sci.* **116**, 24268–24274 (2019).

10. N. G. Reich, C. J. McGowan, T. K. Yamana, A. Tushar, E. L. Ray, D. Osthus, S. Kandula, L. C. Brooks, W. Crawford-Crudell, G. C. Gibson, E. Moore, R. Silva, M. Biggerstaff, M. A. Johansson, R. Rosenfeld, J. Shaman, Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the U.S. *PLOS Computational Biology*. **15**, e1007486 (2019).

11. E. Y. Cramer, E. L. Ray, V. K. Lopez, J. Bracher, A. Brennen, A. J. Castro Rivadeneira, A. Gerding, T. Gneiting, K. H. House, Y. Huang, D. Jayawardena, A. H. Kanji, A. Khandelwal, K. Le, A. Mühlemann, J. Niemi, A. Shah, A. Stark, Y. Wang, N. Wattanachit, M. W. Zorn, Y. Gu, S. Jain, N. Bannur, A. Deva, M. Kulkarni, S. Merugu, A. Raval, S. Shingi, A. Tiwari, J. White, N. F. Abernethy, S. Woody, M. Dahan, S. Fox, K. Gaither, M. Lachmann, L. A. Meyers, J. G. Scott, M. Tec, A. Srivastava, G. E. George, J. C. Cegan, I. D. Dettwiller, W. P. England, M.

W. Farthing, R. H. Hunter, B. Lafferty, I. Linkov, M. L. Mayo, M. D. Parno, M. A. Rowland, B. D. Trump, Y. Zhang-James, S. Chen, S. V. Faraone, J. Hess, C. P. Morley, A. Salekin, D. Wang, S. M. Corsetti, T. M. Baer, M. C. Eisenberg, K. Falb, Y. Huang, E. T. Martin, E. McCauley, R. L. Myers, T. Schwarz, D. Sheldon, G. C. Gibson, R. Yu, L. Gao, Y. Ma, D. Wu, X. Yan, X. Jin, Y.-X. Wang, Y. Chen, L. Guo, Y. Zhao, Q. Gu, J. Chen, L. Wang, P. Xu, W. Zhang, D. Zou, H. Biegel, J. Lega, S. McConnell, V. P. Nagraj, S. L. Guertin, C. Hulme-Lowe, S. D. Turner, Y. Shi, X. Ban, R. Walraven, Q.-J. Hong, S. Kong, A. van de Walle, J. A. Turtle, M. Ben-Nun, S. Riley, P. Riley, U. Koyluoglu, D. DesRoches, P. Forli, B. Hamory, C. Kyriakides, H. Leis, J. Milliken, M. Moloney, J. Morgan, N. Nirgudkar, G. Ozcan, N. Piwonka, M. Ravi, C. Schrader, E. Shakhnovich, D. Siegel, R. Spatz, C. Stiefeling, B. Wilkinson, A. Wong, S. Cavany, G. España, S. Moore, R. Oidtman, A. Perkins, D. Kraus, A. Kraus, Z. Gao, J. Bian, W. Cao, J. Lavista Ferres, C. Li, T.-Y. Liu, X. Xie, S. Zhang, S. Zheng, A. Vespignani, M. Chinazzi, J. T. Davis, K. Mu, A. Pastore y Piontti, X. Xiong, A. Zheng, J. Baek, V. Farias, A. Georgescu, R. Levi, D. Sinha, J. Wilde, G. Perakis, M. A. Bennouna, D. Nze-Ndong, D. Singhvi, I. Spantidakis, L. Thayaparan, A. Tsiourvas, A. Sarker, A. Jadbabaie, D. Shah, N. Della Penna, L. A. Celi, S. Sundar, R. Wolfinger, D. Osthus, L. Castro, G. Fairchild, I. Michaud, D. Karlen, M. Kinsey, L. C. Mullany, K. Rainwater-Lovett, L. Shin, K. Tallaksen, S. Wilson, E. C. Lee, J. Dent, K. H. Grantz, A. L. Hill, J. Kaminsky, K. Kaminsky, L. T. Keegan, S. A. Lauer, J. C. Lemaitre, J. Lessler, H. R. Meredith, J. Perez-Saez, S. Shah, C. P. Smith, S. A. Truelove, J. Wills, M. Marshall, L. Gardner, K. Nixon, J. C. Burant, L. Wang, L. Gao, Z. Gu, M. Kim, X. Li, G. Wang, Y. Wang, S. Yu, R. C. Reiner, R. Barber, E. Gakidou, S. I. Hay, S. Lim, C. Murray, D. Pigott, H. L. Gurung, P. Baccam, S. A. Stage, B. T. Suchoski, B. A. Prakash, B. Adhikari, J. Cui, A. Rodríguez, A. Tabassum, J. Xie, P. Keskinocak, J. Asplund, A. Baxter, B. E. Oruc, N. Serban, S. O. Arik, M. Dusenberry, A. Epshteyn, E. Kanal, L. T. Le, C.-L. Li, T. Pfister, D. Sava, R. Sinha, T. Tsai, N. Yoder, J. Yoon, L. Zhang, S. Abbott, N. I. Bosse, S. Funk, J. Hellewell, S. R. Meakin, K. Sherratt, M. Zhou, R. Kalantari, T. K. Yamana, S. Pei, J. Shaman, M. L. Li, D. Bertsimas, O. Skali Lami, S. Soni, H. Tazi Bouardi, T. Ayer, M. Adee, J. Chhatwal, O. O. Dalgic, M. A. Ladd, B. P. Linas, P. Mueller, J. Xiao, Y. Wang, Q. Wang, S. Xie, D. Zeng, A. Green, J. Bien, L. Brooks, A. J. Hu, M. Jahja, D. McDonald, B. Narasimhan, C. Politsch, S. Rajanala, A. Rumack, N. Simon, R. J. Tibshirani, R. Tibshirani, V. Ventura, L. Wasserman, E. B. O'Dea, J. M. Drake, R. Pagano, Q. T. Tran, L. S. T. Ho, H. Huynh, J. W. Walker, R. B. Slayton, M. A. Johansson, M. Biggerstaff, N. G. Reich, Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. *Proc. Natl. Acad. Sci.* **119**, e2113561119 (2022).

12. C. Viboud, K. Sun, R. Gaffey, M. Ajelli, L. Fumanelli, S. Merler, Q. Zhang, G. Chowell, L. Simonsen, A. Vespignani, The RAPIDD ebola forecasting challenge: Synthesis and lessons learnt. *Epidemics*. **22**, 13–21 (2018).

13. E. Y. Cramer, Y. Huang, Y. Wang, E. L. Ray, M. Cornell, J. Bracher, A. Brennen, A. J. C. Rivadeneira, A. Gerding, K. House, D. Jayawardena, A. H. Kanji, A. Khandelwal, K. Le, V. Mody, V. Mody, J. Niemi, A. Stark, A. Shah, N. Wattanchit,

M. W. Zorn, N. G. Reich, The United States COVID-19 Forecast Hub dataset. *Sci Data*. **9**, 462 (2022).

14. N. G. Reich, J. Lessler, S. Funk, C. Viboud, A. Vespignani, R. J. Tibshirani, K. Shea, M. Schienle, M. C. Runge, R. Rosenfeld, E. L. Ray, R. Niehus, H. C. Johnson, M. A. Johansson, H. Hochheiser, L. Gardner, J. Bracher, R. K. Borchering, M. Biggerstaff, Collaborative Hubs: Making the Most of Predictive Epidemic Modeling. *American Journal of Public Health*. **112**, 839–842 (2022).

15. K. Shea, M. C. Runge, D. Pannell, W. J. M. Probert, S.-L. Li, M. Tildesley, M. Ferrari, Harnessing multiple models for outbreak management. *Science*. **368**, 577–579 (2020).

16. K. Shea, R. K. Borchering, W. J. M. Probert, E. Howerton, T. L. Bogich, S.-L. Li, W. G. van Panhuis, C. Viboud, R. Aguás, A. A. Belov, S. H. Bhargava, S. M. Cavany, J. C. Chang, C. Chen, J. Chen, S. Chen, Y. Chen, L. M. Childs, C. C. Chow, I. Crooker, S. Y. Del Valle, G. España, G. Fairchild, R. C. Gerkin, T. C. Germann, Q. Gu, X. Guan, L. Guo, G. R. Hart, T. J. Hladish, N. Hupert, D. Janies, C. C. Kerr, D. J. Klein, E. Y. Klein, G. Lin, C. Manore, L. A. Meyers, J. E. Mittler, K. Mu, R. C. Núñez, R. J. Oidtman, R. Pasco, A. Pastore y Piontti, R. Paul, C. A. B. Pearson, D. R. Perdomo, T. A. Perkins, K. Pierce, A. N. Pillai, R. C. Rael, K. Rosenfeld, C. W. Ross, J. A. Spencer, A. B. Stoltzfus, K. B. Toh, S. Vattikuti, A. Vespignani, L. Wang, L. J. White, P. Xu, Y. Yang, O. N. Yogurtcu, W. Zhang, Y. Zhao, D. Zou, M. J. Ferrari, D. Pannell, M. J. Tildesley, J. Seifarth, E. Johnson, M. Biggerstaff, M. A. Johansson, R. B. Slayton, J. D. Levander, J. Stazer, J. Kerr, M. C. Runge, Multiple models for outbreak decision support in the face of uncertainty. *Proc Natl Acad Sci U S A*. **120**, e2207537120 (2023).

17. COVID-19 Scenario Modeling Hub, COVID-19 Scenario Modeling Hub. *COVID-19 Scenario Modeling Hub*, (available at https://covid19scenariomodelinghub.org/).

18. M. Stone, The Opinion Pool. *The Annals of Mathematical Statistics*. **32**, 1339–1342 (1961).

19. H. G. Rosenblum, Interim Recommendations from the Advisory Committee on Immunization Practices for the Use of Bivalent Booster Doses of COVID-19 Vaccines — United States, October 2022. *MMWR Morb Mortal Wkly Rep*. **71**, 1436–1441 (2022).

20. R. K. Borchering, C. Viboud, E. Howerton, C. P. Smith, S. Truelove, M. C. Runge, N. G. Reich, L. Contamin, J. Levander, J. Salerno, W. van Panhuis, M. Kinsey, K. Tallaksen, R. F. Obrecht, L. Asher, C. Costello, M. Kelbaugh, S. Wilson, L. Shin, M. E. Gallagher, L. C. Mullany, K. Rainwater-Lovett, J. C. Lemaitre, J. Dent, K. H. Grantz, J. Kaminsky, S. A. Lauer, E. C. Lee, H. R. Meredith, J. Perez-Saez, L. T. Keegan, D. Karlen, M. Chinazzi, J. T. Davis, K. Mu, X. Xiong, P. Porebski, S. Venkatramanan, A. Adiga, B. Lewis, B. Klahn, J. Outten, J. Schlitt, P. Corbett, P. A. Telionis, L. Wang, A. S. Peddireddy, B. Hurt, J. Chen, A. Vullikanti, M. Marathe, J.

M. Healy, R. B. Slayton, M. Biggerstaff, M. A. Johansson, K. Shea, J. Lessler, Modeling of Future COVID-19 Cases, Hospitalizations, and Deaths, by Vaccination Rates and Nonpharmaceutical Intervention Scenarios — United States, April–September 2021. *MMWR Morb Mortal Wkly Rep*. **70**, 719–724 (2021).

21. S. Truelove, C. P. Smith, M. Qin, L. C. Mullany, R. K. Borchering, J. Lessler, K. Shea, E. Howerton, L. Contamin, J. Levander, J. Salerno, H. Hochheiser, M. Kinsey, K. Tallaksen, S. Wilson, L. Shin, K. Rainwater-Lovett, J. C. Lemairtre, J. Dent Hulse, J. Kaminsky, E. C. Lee, J. Perez-Saez, A. Hill, D. Karlen, M. Chinazzi, J. T. Davis, K. Mu, X. Xiong, A. Pastore y Piontti, A. Vespignani, A. Srivastava, P. Porebski, S. Venkatramanan, A. Adiga, B. Lewis, B. Klahn, J. Outten, M. Orr, G. Harrison, B. Hurt, J. Chen, A. Vullikanti, M. Marathe, S. Hoops, P. Bhattacharya, D. Machi, S. Chen, R. Paul, D. Janies, J.-C. Thill, M. Galanti, T. K. Yamana, S. Pei, J. L. Shaman, J. M. Healy, R. B. Slayton, M. Biggerstaff, M. A. Johansson, M. C. Runge, C. Viboud, Projected resurgence of COVID-19 in the United States in July-December 2021 resulting from the increased transmissibility of the Delta variant and faltering vaccination. *eLife*. **11**, e73584 (2022).

22. F. Kreuter, N. Barkay, A. Bilinski, A. Bradford, S. Chiu, R. Eliat, J. Fan, T. Galili, D. Haimovich, B. Kim, others, Partnering with Facebook on a university-based rapid turn-around global survey. *Survey Research Methods: SRM*. **14**, 159–163 (2020).

23. N. Barkay, C. Cobb, R. Eilat, T. Galili, D. Haimovich, S. LaRocca, K. Morris, T. Sarig, Weights and Methodology Brief for the COVID-19 Symptom Survey by University of Maryland and Carnegie Mellon University, in Partnership with Facebook (2020), , doi:10.48550/arXiv.2009.14675.

24. J. Fields, J. Hunter-Childs, A. Tersine, E. Parker, V. Velkoff, C. Logan, H. Shin, Design and operation of the 2020 Household Pulse Survey (2020).

25. U.S. Census Bureau, Vaccine Hesitancy for COVID-19, (available at https://data.cdc.gov/stories/s/Vaccine-Hesitancy-for-COVID-19/cnd2-a6zw/).

26. Public Health England, Investigation of novel SARS-CoV-2 variant: Variant of Concern 202012/01, 19 (2020).

27. J. R. C. Pulliam, C. van Schalkwyk, N. Govender, A. von Gottberg, C. Cohen, M. J. Groome, J. Dushoff, K. Mlisana, H. Moultrie, Increased risk of SARS-CoV-2 reinfection associated with emergence of Omicron in South Africa. *Science*. **376**, eabn4947 (2022).

28. Z. Du, C. Liu, C. Wang, L. Xu, M. Xu, L. Wang, Y. Bai, X. Xu, E. H. Y. Lau, P. Wu, B. J. Cowling, Reproduction Numbers of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Variants: A Systematic Review and Meta-analysis. *Clinical Infectious Diseases*. **75**, e293–e295 (2022).

29. K. Sun, S. Tempia, J. Kleynhans, A. von Gottberg, M. L. McMorrow, N. Wolter, J. N. Bhiman, J. Moyes, M. Carrim, N. A. Martinson, K. Kahn, L. Lebina, J. D. du Toit,

T. Mkhencele, C. Viboud, C. Cohen, the P. Group, Rapidly shifting immunologic landscape and severity of SARS-CoV-2 in the Omicron era in South Africa (2022), p. 2022.08.19.22278993, , doi:10.1101/2022.08.19.22278993.

30. R. Earnest, R. Uddin, N. Matluk, N. Renzette, S. E. Turbett, K. J. Siddle, C. Loreth, G. Adams, C. H. Tomkins-Tinch, M. E. Petrone, J. E. Rothman, M. I. Breban, R. T. Koch, K. Billig, J. R. Fauver, C. B. F. Vogels, K. Bilguvar, B. De Kumar, M. L. Landry, D. R. Peaper, K. Kelly, G. Omerza, H. Grieser, S. Meak, J. Martha, H. B. Dewey, S. Kales, D. Berenzy, K. Carpenter-Azevedo, E. King, R. C. Huard, V. Novitsky, M. Howison, J. Darpolor, A. Manne, R. Kantor, S. C. Smole, C. M. Brown, T. Fink, A. S. Lang, G. R. Gallagher, V. E. Pitzer, P. C. Sabeti, S. Gabriel, B. L. MacInnis, A. Altajar, A. DeJesus, A. Brito, A. E. Watkins, A. Muyombwe, B. S. Blumenstiel, C. Neal, C. C. Kalinich, C. Liu, C. Loreth, C. Castaldi, C. Pearson, C. Bernard, C. M. Nolet, D. Ferguson, E. Buzby, E. Laszlo, F. L. Reagan, G. Vicente, H. M. Rooke, H. Munger, H. Johnson, I. R. Tikhonova, I. M. Ott, J. Razeq, J. C. Meldrim, J. Brown, J. Wang, J. Vostok, J. P. Beauchamp, J. L. Grimsby, J. Hall, K. S. Messer, K. L. Larkin, K. Vernest, L. C. Madoff, L. M. Green, L. Webber, L. Gagne, M. A. Ulcena, M. C. Ray, M. E. Fisher, M. Barter, M. D. Lee, M. T. DeFelice, M. C. Cipicchio, N. L. Smith, N. J. Lennon, N. A. Fitzgerald, N. Kerantzas, P. Hui, R. Harrington, R. Downing, R. Haye, R. Lynch, S. E. Anderson, S. Hennigan, S. English, S. Cofsky, S. Clancy, S. Mane, S. Ash, S. Baez, S. Fleming, S. Murphy, S. Chaluvadi, T. Alpert, T. Rivard, W. Schulz, Z. M. Mandese, R. Tewhey, M. D. Adams, D. J. Park, J. E. Lemieux, N. D. Grubaugh, Comparative transmissibility of SARS-CoV-2 variants Delta and Alpha in New England, USA. *Cell Reports Medicine*. **3**, 100583 (2022).

31. I. L. Ward, C. Bermingham, D. Ayoubkhani, O. J. Gethings, K. B. Pouwels, T. Yates, K. Khunti, J. Hippisley-Cox, A. Banerjee, A. S. Walker, V. Nafilyan, Risk of covid-19 related deaths for SARS-CoV-2 omicron (B.1.1.529) compared with delta (B.1.617.2): retrospective cohort study. *BMJ*. **378**, e070695 (2022).

32. J. A. Lewnard, V. X. Hong, M. M. Patel, R. Kahn, M. Lipsitch, S. Y. Tartof, Clinical outcomes associated with SARS-CoV-2 Omicron (B.1.1.529) variant and BA.1/BA.1.1 or BA.2 subvariant infection in Southern California. *Nat Med*. **28**, 1933–1943 (2022).

33. N. Bobrovitz, H. Ware, X. Ma, Z. Li, R. Hosseini, C. Cao, A. Selemon, M. Whelan, Z. Premji, H. Issa, B. Cheng, L. J. A. Raddad, D. L. Buckeridge, M. D. V. Kerkhove, V. Piechotta, M. M. Higdon, A. Wilder-Smith, I. Bergeri, D. R. Feikin, R. K. Arora, M. K. Patel, L. Subissi, Protective effectiveness of previous SARS-CoV-2 infection and hybrid immunity against the omicron variant and severe disease: a systematic review and meta-regression. *The Lancet Infectious Diseases*. **0** (2023), doi:10.1016/S1473-3099(22)00801-5.

34. J. Bracher, E. L. Ray, T. Gneiting, N. G. Reich, Evaluating epidemic forecasts in an interval format. *PLOS Computational Biology*. **17**, e1008618 (2021).

35. S. B. Vincent, thesis, Holt, Cambridge MA (1912).

36. R. Ratcliff, Group reaction time distributions and an analysis of distribution statistics. *Psychological Bulletin*. **86**, 446–461 (1979).

37. E. Howerton, M. C. Runge, T. L. Bogich, R. K. Borchering, H. Inamine, J. Lessler, L. C. Mullany, W. J. M. Probert, C. P. Smith, S. Truelove, C. Viboud, K. Shea, Context-dependent representation of within- and between-model uncertainty: aggregating probabilistic predictions in infectious disease epidemiology. *J R Soc Interface*. **20**, 20220659 (2023).

38. S. Pollett, M. A. Johansson, N. G. Reich, D. Brett-Major, S. Y. D. Valle, S. Venkatramanan, R. Lowe, T. Porco, I. M. Berry, A. Deshpande, M. U. G. Kraemer, D. L. Blazes, W. Pan-ngum, A. Vespigiani, S. E. Mate, S. P. Silal, S. Kandula, R. Sippy, T. M. Quandelacy, J. J. Morgan, J. Ball, L. C. Morton, B. M. Althouse, J. Pavlin, W. van Panhuis, S. Riley, M. Biggerstaff, C. Viboud, O. Brady, C. Rivers, Recommended reporting items for epidemic forecasting and prediction research: The EPIFORGE 2020 guidelines. *PLOS Medicine*. **18**, e1003793 (2021).

39. N. G. Reich, L. C. Brooks, S. J. Fox, S. Kandula, C. J. McGowan, E. Moore, D. Osthus, E. L. Ray, A. Tushar, T. K. Yamana, M. Biggerstaff, M. A. Johansson, R. Rosenfeld, J. Shaman, A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. *Proc. Natl. Acad. Sci.* **116**, 3146–3154 (2019).

40. D. J. McDonald, J. Bien, A. Green, A. J. Hu, N. DeFries, S. Hyun, N. L. Oliveira, J. Sharpnack, J. Tang, R. Tibshirani, V. Ventura, L. Wasserman, R. J. Tibshirani, Can auxiliary indicators improve COVID-19 forecasting and hotspot prediction? *Proceedings of the National Academy of Sciences*. **118**, e2111453118 (2021).

41. M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks. *Information Processing & Management*. **45**, 427–437 (2009).

42. R. Stein, S. Simmons-Duffin, The Delta Variant Will Drive A Steep Rise In U.S. COVID Deaths, A New Model Shows. *NPR* (2021), (available at https://www.npr.org/sections/health-shots/2021/07/22/1019475669/delta-variant-will-drive-a-steep-rise-in-covid-deaths-model-shows).

43. Z. Hausfather, H. F. Drake, T. Abbott, G. A. Schmidt, *Geophysical Research Letters*, in press, doi:10.1029/2019GL085378.

44. X. Li, C. Mukandavire, Z. M. Cucunubá, S. E. Londono, K. Abbas, H. E. Clapham, M. Jit, H. L. Johnson, T. Papadopoulos, E. Vynnycky, M. Brisson, E. D. Carter, A. Clark, M. J. de Villiers, K. Eilertson, M. J. Ferrari, I. Gamkrelidze, K. A. M. Gaythorpe, N. C. Grassly, T. B. Hallett, W. Hinsley, M. L. Jackson, K. Jean, A. Karachaliou, P. Klepac, J. Lessler, X. Li, S. M. Moore, S. Nayagam, D. M. Nguyen, H. Razavi, D. Razavi-Shearer, S. Resch, C. Sanderson, S. Sweet, S. Sy, Y. Tam, H. Tanvir, Q. M. Tran, C. L. Trotter, S. Truelove, K. van Zandvoort, S. Verguet, N.

Walker, A. Winter, K. Woodruff, N. M. Ferguson, T. Garske, Estimating the health impact of vaccination against ten pathogens in 98 low-income and middle-income countries from 2000 to 2030: a modelling study. *The Lancet*. **397**, 398–408 (2021).

45. European Covid-19 Scenario Hub, (available at https://covid19scenariohub.eu/).

46. Flu scenario model hub, (available at https://fluscenariomodelinghub.org/).

47. Consortium of Infectious Disease Modeling Hubs. *GitHub*, (available at https://github.com/Infectious-Disease-Modeling-Hubs).

48. V. R. R. Jose, Y. Grushka-Cockayne, K. C. Lichtendahl, Trimmed Opinion Pools and the Crowd's Calibration Problem. *Management Science*. **60**, 463–475 (2014).

49. R Core Team, "R: A language and environment for statistical computing" (manual, R Foundation for Statistical Computing, Vienna, Austria, 2018), (available at https://www.R-project.org/).

50. Centers for Disease Control and Prevention, COVID-19 Vaccinations in the United States, Jurisdiction, (available at https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-Jurisdi/unsk-b7fc).

51. L. Hallas, A. Hatibie, S. Majumdar, M. Pyarali, R. Koch, A. Wood, T. Hale, Variation in US states' responses to COVID-19. *Blavatnik School of Government* (2020) (available at https://www.bsg.ox.ac.uk/research/publications/variation-us-states-responses-covid-19).

52. K. Gangavarapu, A. A. Latif, J. Mullen, M. Alkuzweny, E. Hufbauer, G. Tsueng, E. Haag, M. Zeller, C. M. Aceves, K. Zaiets, M. Cano, J. Zhou, Z. Qian, R. Sattler, N. L. Matteson, J. I. Levy, R. T. Lee, L. Freitas, S. Maurer-Stroh, GISAID core and curation team, M. A. Suchard, C. Wu, A. I. Su, K. G. Andersen, L. D. Hughes, Center for Viral Systems Biology, "Alpha Variant Report," (available at https://outbreak.info/situation-reports/alpha?loc=GBR&loc=USA&loc=USA_US-CA&selected=GBR).

53. K. Gangavarapu, A. A. Latif, J. Mullen, M. Alkuzweny, E. Hufbauer, G. Tsueng, E. Haag, M. Zeller, C. M. Aceves, K. Zaiets, M. Cano, J. Zhou, Z. Qian, R. Sattler, N. L. Matteson, J. I. Levy, R. T. Lee, L. Freitas, S. Maurer-Stroh, GISAID core and curation team, M. A. Suchard, C. Wu, A. I. Su, K. G. Andersen, L. D. Hughes, Center for Viral Systems Biology, "Delta Variant Report," (available at https://outbreak.info/situation-reports/delta?loc=IND&loc=GBR&loc=USA&selected).

54. K. Gangavarapu, A. A. Latif, J. Mullen, M. Alkuzweny, E. Hufbauer, G. Tsueng, E. Haag, M. Zeller, C. M. Aceves, K. Zaiets, M. Cano, J. Zhou, Z. Qian, R. Sattler, N. L. Matteson, J. I. Levy, R. T. Lee, L. Freitas, S. Maurer-Stroh, GISAID core and curation team, M. A. Suchard, C. Wu, A. I. Su, K. G. Andersen, L. D. Hughes, Center for Viral Systems Biology, "Omicron Variant Report," (available at

https://outbreak.info/situation-reports/omicron?loc=ZAF&loc=GBR&loc=USA&selected).

55. M. Alkuzweny, K. Gangavarapu, L. Hughes, *outbreakinfo: outbreak.info R Client* (2023; https://outbreak-info.github.io/R-outbreak-info/).

56. E. L. Ray, L. C. Brooks, J. Bien, M. Biggerstaff, N. I. Bosse, J. Bracher, E. Y. Cramer, S. Funk, A. Gerding, M. A. Johansson, A. Rumack, Y. Wang, M. Zorn, R. J. Tibshirani, N. G. Reich, Comparing trained and untrained probabilistic ensemble forecasts of COVID-19 cases and deaths in the United States. *International Journal of Forecasting* (2022), doi:10.1016/j.ijforecast.2022.06.005.

57. E. Dong, H. Du, L. Gardner, An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*. **20**, 533–534 (2020).

58. U.S. Department of Health and Human Services, HHS Protect Public Data Hub, (available at https://public-data-hub-dhhs.hub.arcgis.com/).

59. T. Gneiting, R. Ranjan, Comparing Density Forecasts Using Threshold-and Quantile-Weighted Scoring Rules. *Journal of Business & Economic Statistics*. **29**, 411–422 (2011).

60. C. Pennell, T. Reichler, On the Effective Number of Climate Models. *Journal of Climate*. **24**, 2358–2367 (2011).

61. N. I. Bosse, S. Abbott, A. Cori, E. van Leeuwen, J. Bracher, S. Funk, Transformation of forecasts for evaluating predictive performance in an epidemiological context. *medRxiv* (2023), doi:10.1101/2023.01.23.23284722.