Title: Evaluation of FluSight influenza forecasting in the 2021-22 and 2022-23 seasons with a new target laboratory-confirmed influenza hospitalizations

Abstract:

Accurate forecasts can enable more effective public health responses during seasonal influenza epidemics. Forecasting teams were asked to provide national and jurisdiction-specific probabilistic predictions of weekly confirmed influenza hospital admissions for one through four weeks ahead for the 2021-22 and 2022-23 influenza seasons.

Across both seasons, 26 teams submitted forecasts, with the submitting teams varying between seasons. Forecast skill was evaluated using the Weighted Interval Score (WIS), relative WIS, and coverage.

Six out of 23 models outperformed the baseline model across forecast weeks and locations in 2021-22 and 12 out of 18 models in 2022-23. Averaging across all forecast targets, the FluSight ensemble was the 2nd most accurate model measured by WIS in 2021-22 and the 5th most accurate in the 2022-23 season. Forecast skill and 95% coverage for the FluSight ensemble and most component models degraded over longer forecast horizons and during periods of rapid change.

Current influenza forecasting efforts help inform situational awareness, but research is needed to address limitations, including decreased performance during periods of changing epidemic dynamics.

Authors:

Sarabeth M. Mathis[1†*], Alexander E. Webber[1†], Tomás M. León[2], Erin L. Murray[2], Monica Sun[2], Lauren A. White[2], Logan C. Brooks[3,5], Alden Green[3], Addison J. Hu[3], Daniel J. McDonald[4], Roni Rosenfeld[3], Dmitry Shemetov[3], Ryan J. Tibshirani[3,5], Sasikiran Kandula[6], Sen Pei[7], Jeffrey Shaman[7,8], Rami Yaari[7], Teresa K. Yamana[7], Pulak Agarwal[9], Srikar Balusu[9], Gautham Gururajan[9], Harshavardhan Kamarthi[9], B. Aditya Prakash[9], Rishi Raman[9], Alexander Rodríguez[10], Zhiyuan Zhao[9], Akilan Meiyappan[11], Shalina Omar[11], Prasith Baccam[12], Heidi L. Gurung[12], Steve A. Stage[13], Brad T. Suchoski[12], Marco Ajelli[14], Allisandra G. Kummer[14], Maria Litvinova[14], Paulo C. Ventura[14], Spencer Wadsworth[15], Jarad Niemi[15], Erica Carcelen[16], Alison L Hill[16], Sung-mok Jung[17], Joseph C. Lemaitre[17], Justin Lessler[17], Sara L Loo[16], Clifton D. McKee[16], Koji Sato[16], Claire Smith[16], Shaun Truelove[16], Thomas McAndrew[18], Wenxuan Ye[18], Nikos Bosse[19], William S. Hlavacek[20], Yen Ting Lin[20], Abhishek Mallela[20], Ye Chen[21], Shelby M. Lamm[21], Jaechoul Lee[21], Richard G. Posner[21], Amanda C. Perofsky[22], Cécile Viboud[22], Leonardo Clemente[23], Fred Lu[23], Austin G Meyer[23], Mauricio Santillana[23], Matteo Chinazzi[23], Jessica T. Davis[23], Kunpeng Mu[23], Ana Pastore y Piontti[23], Alessandro Vespignani[23], Xinyue Xiong[23], Michal Ben-Nun[24], Pete Riley[24], James Turtle[24], Chis Hulme-Lowe[25], Shakeel Jessa[25], V.P. Nagraj[26], Stephen D. Turner[26], Desiree Williams[26], Avranil Basu[27], John M. Drake[27], Spencer J. Fox[28], Graham C. Gibson[20], Ehsan Suez[28], Edward W. Thommes[29,30], Monica G. Cojocaru[29], Estee Y. Cramer[31], Aaron Gerding[31], Ariane Stark[31], Evan L. Ray[31], Nicholas G. Reich[31], Li Shandross[31], Nutcha Wattanachit[31], Yijin Wang[31], Martha W. Zorn[31], Majd Al Aawar[32], Ajitesh Srivastava[32], Lauren A. Meyers[33], Aniruddha Adiga[34], Benjamin Hurt[34], Gursharn Kaur[34], Bryan L. Lewis[34], Madhav Marathe[34], Srinivasan

Venkatramanan[34], Patrick Butler[35], Andrew Farabow[35], Nikhil Muralidhar[36], Naren Ramakrishnan[35], Carrie Reed[1], Matthew Biggerstaff[1], Rebecca K. Borchering[1*]

[1]Centers for Disease Control and Prevention, Atlanta, Georgia, 30329, USA.

[2]California Department of Public Health, Richmond, CA, 95899

[3]Carnegie Mellon University, Pittsburgh, PA, 15213

[4]University of British Columbia, Vancouver, BC, V6T 1Z4, Canada

[5]University of California, Berkeley, Berkeley, CA 94720

[6]Norwegian Institute of Public Health, 0213 Oslo, Norway

[7]Columbia University, New York, NY, 10032

[8]Columbia University School of Climate, New York, NY 10025

[9]Georgia Institute of Technology, Atlanta, GA, 30318

[10]University of Michigan, Ann Arbor, MI, 48109

[11]Guidehouse Advisory and Consulting Services, McClean VA, 22102

[12]IEM, Bel Air, MD, 21015

[13]IEM, Baton Rouge, LA, 70809

[14]Indiana University School of Public Health, Bloomington, IN, 47405

[15]Iowa State University, Ames, IA, 50011

[16]Johns Hopkins University, Baltimore, MD, 21205

[17]University of North Carolina at Chapel Hill, Chapel Hill, NC,

[18]Lehigh University, Bethlehem, PA, 18015

[19]London School of Health and Tropical Medicine, London, UK, WC1E 7HT

[20]Los Alamos National Laboratory, Los Alamos, NM, 87545

[21]Northern Arizona University, Flagstaff, AZ, 86011

[22]Fogarty International Center, National Institutes of Health, Bethesda, MD, 20892

[23]Northeastern University, Boston, MA, 02115

[24]Predictive Science Inc, San Diego, CA 92121

[25]Signature Science, LLC, Austin, TX, 78759

[26]Signature Science, LLC, Charlottesville, VA, 22911

[27]University of Georgia, Athens, GA, 30602

[28]University of Georgia, Athens, GA, 30609

[29]University of Guelph, Guelph, ON N1G 2W1, Canada

[30]Sanofi, Toronto, ON, M2R 3T4

[31]University of Massachusetts Amherst, Amherst, MA, 01003

[32]University of Southern California, Los Angeles, CA, 90089

[33]University of Texas Austin, Austin, TX, 78712

[34]University of Virginia, Charlottesville, VA, 22903

[35]Virginia Tech, Arlington, VA 22203

[36] Stevens Institute of Technology, Hoboken, NJ, 07030

† Indicates co-first authors

* Indicates co-corresponding authors

Introduction

Traditional influenza surveillance systems provide a comprehensive picture of influenza activity in the United States [1, 2, 3] and are fundamental for situational awareness and risk communication. However, they measure influenza activity after it has occurred, and do not directly anticipate future trends to inform risk assessment and healthcare preparedness. To address these limitations, the Centers for Disease Control and Prevention (CDC) has supported open influenza forecasting challenges since the 2013–14 season [4]. This collaborative process (named FluSight) has ensured that forecasting targets are relevant to public health. Additionally, forecast data are openly available, which enables transparent evaluation of forecast performance [5, 6].

Originally the FluSight collaboration focused on short-term forecasts of outpatient influenza-like-illness (ILI) rates from ILINet [2] and corresponding results have been summarized previously [4, 5, 6]. However, the COVID-19 pandemic resulted in changes in outpatient care-seeking behavior, and the continued co-circulation of SARS-CoV-2 has further complicated the interpretation of ILI data. In the 2021–22 influenza season, the FluSight forecast target shifted to the weekly number of hospital patients admitted with laboratory-confirmed influenza from the Health and Human Services (HHS) Patient Impact and Hospital Capacity Data System [7]. This system was created during the COVID-19 pandemic to gather a complete and unified representation of COVID-19 disease outcomes along with other metrics related to health care capacity. Hospitals registered with Centers for Medicare and Medicaid Services (CMS) are required to report daily COVID-19 and influenza information [8]. Reporting of the influenza data elements, including the previous day's number of admissions with laboratory-confirmed influenza virus infection, became mandatory on February 2, 2022, [8].

The COVID-19 pandemic disrupted the typical timing, intensity, and duration of seasonal influenza activity in the United States and many parts of the world [9, 10]. Influenza activity was very low during the 2020–21 season in the U.S., but activity increased during the 2021–22 season, with activity peaking later in April, May, and early June 2022 and remaining at higher levels than had been reported during these months in previous seasons [10]. In the 2022-23 influenza season, activity began increasing nationally in early October, earlier than previous seasons [2,3,11], and peaked in early December 2022. In this analysis, we summarize the accuracy and reliability of ensemble and component 1- to 4-week ahead forecasts submitted in real-time during the 2021–22 and 2022-23 influenza seasons and identify areas for forecast improvement.

Methods

Forecasts of weekly influenza hospital admissions were openly solicited from existing COVID-19 and influenza forecasting networks every Monday from January 10, 2022, through June 20, 2022, for the 2021-22 season. For the 2022-23 season, forecasts were solicited every Monday from October 17, 2022, through January 9, 2023, then every Tuesday from January 17, 2023, through May 17, 2023. Weeks were defined in terms of MMWR Epiweeks (EW) spanning Sunday to Saturday [12]. Forecasted jurisdictions included the U.S. national level, all fifty states, Washington D.C., and Puerto Rico. Forecasts for the Virgin Islands, while requested, were not included in this evaluation due to low reported hospitalization counts and irregular data submission. Each week, forecasting teams were asked to provide jurisdiction-specific point estimates and probabilistic predictions for 1-, 2-, 3-, and 4-week ahead weekly counts of confirmed influenza hospital admissions. A total of 23 quantiles were requested for the probabilistic forecasts: 0.010, 0.025, 0.050, 0.100, 0.150, …, 0.950, 0.975, and 0.990. Teams were not required to submit forecasts for all four weeks ahead or for all locations. Additional details of the forecast submission process (e.g., file formatting, submission procedures, and required metadata) are provided in the FluSight-forecast-data GitHub Repository [13].

The FluSight Ensemble model was generated for all forecasted jurisdictions each week using the unweighted median of each quantile among eligible forecasts. Forecasts were considered eligible for inclusion in the ensemble if they were submitted by 11:59 PM ET on the due date and if all requested quantiles were provided. Modeling teams could further designate whether a particular model's forecasts should be included in the ensemble. If a forecast was designated as "other", it was not included in the FluSight ensemble and not evaluated in this manuscript.

Baseline forecasts and their prediction intervals were generated each week using the simplets R package [14] based on the incident hospitalizations reported in the previous week. The median prediction of the baseline forecasts is the corresponding target value observed in the previous week, and noise around the median prediction is generated using positive and negative 1-week differences (i.e., differences between consecutive reports) for all prior observations, separately for each jurisdiction. Sampling distributions were truncated to prevent negative values. The same median prediction is used for the 1-through 4-week ahead forecasts. Further details on the generation of the baseline model's prediction intervals from a smoothed version of this distribution of differences have been described previously [15,16].

For inclusion in this analysis, forecasting teams must have submitted greater than or equal to 75% of the requested targets between the forecast evaluation period of February 21, 2022, to June 20, 2022 (total of 18 weeks) for 2021-22 or October 17, 2022, to May 15, 2023 (total of 30 weeks) for 2022-23. These periods translate to 4-week ahead forecast target end dates of March 19, 2022, to July 16, 2022 for the 2021-22 season and November 11, 2022, to June 10, 2023 for the 2022-23 season. The start date of the evaluation period for the 2021-22 season was chosen to be the first forecast date following two weeks of mandatory reporting of confirmed influenza hospitalizations [8] to minimize potential effects of reporting changes on forecasts. For 2021-22 and 2022-23, three and 12 models were excluded from the primary analysis, respectively, for not meeting the inclusion criteria.

Forecasts were evaluated against the reported number of the previous day's laboratory confirmed influenza admissions (Field #34) from the COVID-19 Reported Patient Impact and Hospital Capacity by State Timeseries [17], with data shifted one day earlier to align with admission date and then aggregated to the weekly scale (from Sunday to Saturday) [13], using data as of September 12, 2022, for 2021-22 and June 13, 2023, for 2022-23. This dataset is subject to revision by submitting facilities; therefore, we analyzed backfill and revision for each season (Supplemental Analysis 1). For each of the contributed forecasts included in the analysis, values were rounded to more closely relate the values of prediction intervals of forecasts to the reported numbers of hospital admissions. In particular, forecast values for quantiles less than 0.5 were rounded down, values for quantiles greater than 0.5 were rounded up, and values for the 0.5 quantile were rounded normally. This rounding procedure ensured that teams were not penalized for missing the prediction interval by less than one hospital admission.

To evaluate forecast performance across all states, D.C., and Puerto Rico, we primarily used the Weighted Interval Score (WIS). The WIS is a proper score that generates interval scores for probabilistic forecasts provided in the quantile format [15,18]. Briefly, interval scores are used to account for dispersion, underprediction, and overprediction. Forecasts with lower absolute WIS values are considered more accurate than forecasts with higher absolute WIS values. The relative WIS compares forecast WIS values from those of the baseline model. Simple means were calculated for absolute and relative WIS to get a score for each model, location, and season. Mean absolute error (MAE) values are also considered for characterizing differences between forecasted and reported weekly hospitalizations [15]. Unless otherwise specified, forecasts of national hospitalizations were not included in summary metrics for accuracy (e.g., absolute WIS) since these forecasts can have a disproportionate impact on the overall score. To address concerns related to assessing measures of absolute error on a natural scale when forecasts span multiple orders of magnitude [19], we performed an analogous analysis on log-transformed hospitalization counts after adding one to all counts to account for zero counts (Supplemental Analysis 2). We also performed a separate analysis including only national forecasts (Supplemental Analysis 3).

In addition, we considered coverage values of the quantile-based prediction intervals to assess each model's ability to appropriately capture uncertainty in forecasts. Coverage values are defined as the percent of observed values that fall within the 50% or 95% prediction intervals for the corresponding date. Ideally, the percent coverage values will be equal to the corresponding prediction interval, e.g., 95% percent prediction intervals should contain the reported value 95% of the time.

Comparing model forecasts is complicated by the fact that not all models submit forecasts for each of the forecast targets and for each forecast week in the evaluation period. To partially account for this, we consider the percent of forecasts submitted as an indicator of how often and how many different types of forecasts were submitted by each team. Following Cramer et al. 2022 [15], we also consider a standardized rank score that uses the number of models forecasting a particular location and target and then ranks these forecasts. Ranks were determined by relative WIS performance, with the best performing model for each observation being assigned a rank of 1 and the worst performing model receiving a rank equal to the number of models submitting a forecast for the observation. These ranks were standardized by rescaling so that 0 corresponds to the worst rank and 1 corresponds to the best rank.

Results

The 2021-22 influenza season was characterized by two distinct waves of activity. The first occurred between November 2021 and January 2022 and the second between February and June 2022, though reporting of influenza hospitalizations was not mandatory in the HHS system until February 2, 2022 (see observed data in Figure 1a). Reported national weekly influenza hospital admissions exceeded 1000 for 22 out of 25 of the forecast weeks (Figure 1a). Updates to weekly counts from the forecast evaluation period were generally minimal (Figures S2 – S4), with 94% of updates during the 2021-22 season resulting in changes of under 10 hospitalizations for subnational jurisdictions.

The 2022-23 influenza season was characterized by an early start, reaching 1000 hospital admissions nationally before October 2022. A sharp increase nationally through October and November led to a peak of 26,600 hospital admissions in early December. Hospital admissions decreased rapidly after December, with 3,000 weekly hospital admissions by the end of January, and eventually dropped below 1000 confirmed weekly admissions nationally by May 2023. Weekly numbers of admissions exceeded 1000 for 27 out of 34 of the forecast weeks (Figure 1b, Figure S4). In the 2022-23 season, 83% of updates for weekly admissions resulted in changes of under 10 hospitalizations for subnational jurisdictions.

Models Included

For both the 2021-22 and 2022-23 influenza seasons, 26 modeling teams submitted forecasts and 21 and 16 respectively, were eligible for end-of-season evaluation, not including the FluSight baseline and ensemble models. The number and types of models submitted varied across weeks with a range of methodological approaches (see Table S1). For the 2021-22 season, a median of 21 models were submitted (range: 15-22), with most having a statistical component, three mechanistic, and six ensembles of component models. In 2022-23 there was a median of 20 models (range: 15 to 26) submitted each week, with many having a statistical component, three mechanistic, and four ensemble models. Modeling teams varied across seasons, with 13 modeling groups having submitted eligible forecasts for both seasons. When only national forecasting targets were considered, no additional teams were included for the 2021-22 season, but two teams, NIH-Flu_ARIMA and ISU_NiemiLab-Flu met inclusion criteria for 2022-23 (Supplemental Analysis 3).

Relative WIS

Over the evaluation period more models outperformed the FluSight baseline model in 2022-23 (12) than in 2021-22 (6) based on relative WIS (Table 1). Within each season, the models that achieved an overall relative WIS less than or equal to one represent a variety of modeling strategies, including a basic quantile autoregression fit, a mechanistic compartmental model with stochastic simulations, an ensemble of time-series baseline models, a random walk model, a random forest ensemble, and the FluSight hub ensemble (Table S1). Similar results were observed when models were evaluated based on their point forecasts alone (see MAE estimates in Table 1).

Few teams outperformed the FluSight Ensemble in relative WIS for both seasons. The CMU-TimeSeries model was the only model that outperformed the ensemble for both the 2021-22 and 2022-23 seasons while the MOBS-GLEAM_FLUH, PSI-DICE and MIGHTE-Nsemble models outperformed the ensemble only in the 2022-23 season.

For both seasons, forecasts from the FluSight Ensemble were ranked among the top 50% of all model forecasts for the same location, date, and target, more than three-fourths of the time (79.89% in 2021-22 and 79.02% in 2022-23) (Figure 2). Three models consistently ranked in the top 25% for 2021-22 and 2022-23 seasons, respectively: CMU-TimeSeries (42.49%, 36.32%), PSI-DICE (39.24%, 39.84%), and MOBS-GLEAM_FLUH (38.89%, 50.31%). Several models, ten in 2021-22 and eleven in 2022-23, had bimodal rank distributions, with a combined majority of their forecasts falling in either the bottom 25% or top 25% (Figure 2).

Log-Transformed Analysis

For both seasons, the analysis using log-transformed hospitalization counts resulted in the same top five performing teams in terms of absolute and relative WIS. For the 2021-22 season, all teams were ranked the same for the log-transformed and non-transformed analyses. In 2022-23, MIGHTE-Nsemble and PSI-DICE performed better than CMU-TimeSeries for the log-transformed analysis (Table 1 and Supplemental Analysis 2).

Relative WIS and Spatial Variation

Model performance varied by spatial jurisdiction. For individual states, relative WIS values varied across models ranging from 0.46 to 12.46 in 2021-22 and 0.31 to 12.28 in 2022-23 (Figure 3). More models, including the ensemble, performed better at the state-level than the baseline in 2022-23 compared to 2021-22. The relative WIS of the FluSight Ensemble had the smallest range of values across all locations from 0.58 to 1.06 in 2021-22 to 0.63 to 1 in 2022-23 (Figure 3 and Figure S1). To further examine forecast performance across jurisdictions, we considered the percent of jurisdictions that the relative WIS value for a given model and location pair was less than the baseline (i.e., lower than 1). The FluSight Ensemble performed as well as or better than the baseline for all forecast jurisdictions for 2022-23 and 51 out of 52 forecast jurisdictions for 2021-22, a larger number of jurisdictions than any submitted model (Figure 3). In 2022-23, 12 models performed better than the baseline at the jurisdiction-level at least 50% of the time, compared to six models in 2021-22. In general, the models with lower (better) relative WIS values were consistent between the analysis with all spatial jurisdictions and the analysis considering only national forecast targets for both seasons (Supplemental Analysis 3).

Absolute WIS

Across forecasted weeks, the FluSight Ensemble's worst performance in terms of absolute WIS (maximum values) for 1-week ahead targets on March 19, 2022 for 2021-22 and on November 26, 2022 for 2022-23 (Figure 4a). For the 4-week ahead horizon, maximum absolute values, indicating the worst performance, for each season occurred on June 04, 2022, and December 03, 2022, respectively (Figure 4a). Minimum, or best, absolute WIS values for each season occurred on July 16, 2022, and May 13, 2023, respectively, both during periods of low flu activity.

Coverage

Model performance for the FluSight Ensemble dropped during periods of relatively rapid change (see Figures 1 and 3). The lowest 1-week horizon 95% value occurred for forecasts with target end dates of March 14, 2022, for 2021-22 and on November 21, 2022, for 2022-23 (Figure 4b, c). Across forecasted weeks in the 2021-22 season, the FluSight Ensemble had a minimum 95% coverage value at the 1-week horizon of 75%. Lower 95% coverage for the 1-week horizon was observed in the 2022-23 season with a minimum of 29%. The maximum coverage rate achieved by the FluSight Ensemble in any individual week was 100% in both seasons. Minimum FluSight Ensemble 95% coverage values for forecasts at the 4-week horizon in any individual week were 62% for 2021-22 and 15% for 2022-23.

Model performance, in terms of coverage, tended to decline at longer time horizons for the FluSight Ensemble, baseline, and individual contributed models (see Table 2). Over the forecast weeks, the 2021-22 FluSight ensemble had slightly higher overall 95% coverage values of 89.32%, 86.11%, 85.15%, and 83.33% for the 1- to 4-week ahead horizons respectively, compared to the 2022-23 season during which the FluSight Ensemble had 95% coverage values of 85.79%, 81.64%, 78.78%, and 77.85% for the 1- to 4-week ahead horizons respectively. A similar proportion of models had higher overall 95% coverage values at the 1-week ahead horizon than at the 4-week ahead horizon for 2022-23 (14 of 18 models) and 2021-22 (18 out of 23 models) (Table 2). Out of the forecast targets and across forecast weeks, the FluSight Ensemble's 95% prediction interval contained at least 90% of the corresponding observed values only 55.56% and 64.52% of the time, for 2021-22 and 2022-23 respectively (Table 2). Ideally 95% prediction intervals are just wide enough to capture 95% of eventually observed values.

Discussion

The 2021-22 influenza season marked the return of from very low levels of seasonal influenza activity observed in the U.S. following the first years of the COVID-19 pandemic, and many components of the 2021-22 and 2022-23 FluSight Forecasting Challenges were new. One of the most substantial changes was the shift from the original FluSight forecasting targets of weekly influenza-like-illness (ILI) percentages to weekly counts of confirmed influenza hospitalizations. The COVID-19 pandemic resulted in the availability of a new data source, the unified HHS-Protect dataset [17], which provided information on laboratory confirmed daily influenza hospitalizations from all 50 states, D.C., and Puerto Rico. Confirmed influenza hospital admissions may more directly inform influenza preparedness and response efforts. During the time period that these forecasting results cover, data were reported daily, with mandatory reporting for influenza admissions from most hospitals in each state, U.S. territories, and D.C starting February 2, 2022. Despite challenges accompanying the shift to the new target of influenza hospitalizations, such as limited historic data from this system for model training, these forecasts provided substantial utility and reinforced a number of lessons learned over the course of previous forecasting activities, both during the pre-pandemic influenza seasons and the COVID-19 pandemic.

Forecast performance - accuracy

As demonstrated in this analysis, collaborative forecasting hub approaches provide opportunities to systematically evaluate performance across multiple modeling strategies and enable the creation of ensemble models. Since a particular model's performance often varies within and across seasons [20], it is helpful to have a unified representation of model inputs that can be used to quickly assess expected upcoming trends. Additionally, this work indicates that ensemble models may also provide more consistently reliable and well-calibrated forecasts across spatial jurisdictions.

Across the evaluation period for both seasons and all forecast jurisdictions, the FluSight ensemble was among the top 5 performing models in terms of Absolute WIS and Relative WIS.

Additionally, when considering forecast performance by rank (Figure 2), the FluSight ensemble more accurately predicted weekly influenza hospital admissions than most contributed models with the majority of the FluSight ensemble forecasts falling within the top 50% of submitted forecasts (Table 1, Figure 2). While the PSI-DICE, CMU-TimeSeries, and MOBS-GLEAM_FLUH models have more forecasts in the top 25%, they exhibit higher spatial heterogeneity than the FluSight ensemble in forecast performance (Figure 3). The generally high accuracy of the FluSight Ensemble relative to that of individual models is consistent with previous findings that ensemble models, that utilize the outputs from multiple teams, generally outperform individual models on average [15,21,22,23]. Like most models, ensembles may have decreased performance during periods of rapid change when some individual models may have higher accuracy; however, identifying these time frames and corresponding high-performing models has been difficult a priori [5,6].

One option to better evaluate forecast performance during periods of change and across multiple magnitudes is to evaluate transformed counts [19]. We did not find notable differences in model performance using this approach in either season. We expected that there might be a stronger influence on performance in the 2022-2023 season which saw a sharp increase in hospitalizations in fall 2022, but it is possible that models were not able to capture this initial rise and thus did not accrue additional benefit in the log transform score. The long tail of the season may also have elevated scores across all models.

Forecast model performance tended to decline over longer time horizons. For both the 2021-22 and 2022-23 FluSight seasons, accuracy declined across the 1- to 4-week ahead horizons. This trend has been observed previously in multiple forecast activities. The U.S. COVID-19 Forecast Hub observed declines in accuracy for forecasted deaths over periods of 1- to 4- weeks ahead, and German and Polish COVID-19 forecast efforts also showed declines in performance at the 3- and 4-week ahead horizons [18]. Accuracy scores were also shown to decline over longer time horizons for influenza-like-illness forecasts [20].

Across the forecast weeks, individual models often showed larger increases in absolute WIS, while the FluSight ensemble had the smallest range of absolute WIS for each season, demonstrating one aspect of stability for the FluSight ensemble. In terms of state-level performance, the FluSight ensemble tended to be more robust than individual models, as measured by relative WIS scores (Figure 3). Similarly, the COVID-19 Forecast Hub ensemble [15] performed better across all locations, with the COVID-19 Hub ensemble being the only model to outperform the baseline in each of the forecast locations [15].

Forecast performance – coverage

Our analysis found that, as the forecast horizon moved from 1- to 4-weeks, the FluSight ensemble 95% prediction interval coverage declined from 89.61% to 83.74% in 2021-22 and from 85.69% to 77.85% in 2022-23. These results highlight room for improvement in model calibration, as almost all models (with the exception of the UMass trends ensemble) were overconfident in their predictions (Table 2). The lack of comparable historical data for model fitting may have contributed to poor calibration of 95% prediction intervals.

Consistent with past forecasting efforts, forecasting remains difficult in periods of rapid change and epidemic turning points (e.g., during initial increases or periods of peaking activity). This analysis highlights declines in forecast accuracy and coverage during periods of rapid change in influenza hospitalizations during both the 2021-22 and 2022-23 seasons. For example, the only model that had 95% coverage greater than 80% from October to January 2023 when hospitalizations were rapidly increasing and then peaking was LUCompUncertLab-humanjudgment, which did not end up meeting inclusion criteria for the full season analysis. Analogous declines were also observed for COVID-19 case forecasts [24] and mortality forecasts across different waves of the COVID-19 pandemic [15], where forecasts systematically underpredicted during periods of increase and overpredicted during periods of decrease.

Times of changing dynamics are the most important periods for public health response and communication. While forecasting the magnitude at these times may be less tractable, it is possible that we may be able to provide more reliable information during these difficult forecasting periods so that forecasts are better able to inform critical planning. In general, most ensembles tend to predict less activity than observed when trends are steeply increasing and predict more activity than observed when trends are steeply decreasing, especially when there is between- or within-model uncertainty in the timing of peaks in cases, hospitalizations, or deaths. Thus, it may be possible that an ensemble of forecasts for categorical increases or decreases in activity [25] may have additional utility in terms of preserving valuable information while also maintaining the benefits of the use of ensembles over individual models. As such, the FluSight Forecasting Hub added an experimental target in the 2022-23 season for forecasting categorical rate changes in influenza hospitalizations (e.g., probabilities of increase or decrease) [13]. Assessing the utility of this additional forecast target will be an important area of investigation moving forward. Aside from soliciting a separate forecasting target, it may be possible to determine which forecasting models perform better during different phases of epidemics and then use this information to weight models accordingly when their forecasts are aggregated into an ensemble [26].

Influenza forecasting in the COVID-19 era: challenges and opportunities

Several challenges for forecasting existed during the 2021-22 and 2022-23 influenza seasons. First, as noted earlier, the change in the forecasting target from outpatient ILI percentages to counts of influenza-associated hospitalizations from a data collection system established during the COVID-19 pandemic meant that there was little data for forecast calibration and training. This shift also required changes in data processing for teams that had produced ILI forecasts previously. While previous data on influenza-associated hospitalizations was available through the FluSurv-NET system, differences in reporting and the spatial resolution, of the FluSurv-NET system may have complicated the process of utilizing this dataset for the purpose of forecasting model calibration. In addition, reporting within the unified HHS-Protect hospitalization dataset changed throughout this forecasting endeavor. For example, the confirmed influenza hospital admissions field only became mandatory for the 2021–22 season on February 2, 2022, leading to an increase in the number of reported hospitalizations and a change in hospital reporting practices during a period of increasing influenza activity.

In addition to changing reporting patterns, the COVID-19 pandemic brought other challenges for forecasting influenza, including changing human behavior. The quantity and types of interactions between people likely changed in tandem with perceptions of risk of illness with COVID-19. In addition, the use of nonpharmaceutical interventions (NPIs) aimed at preventing SARS-CoV-2 transmission (e.g., stay-at-home orders, mask wearing) reduced transmission of other respiratory pathogens [9], including influenza. These changes in behavior may be related to the minimal influenza activity observed in the U.S. in the 2020–21 season and the low severity but atypically late influenza season observed in the 2021–22 season. Population-level behavior is difficult to predict, especially in the context of changing public health recommendations and emerging SARS-CoV-2 variants, which complicated the process of forecasting. Despite these challenges, FluSight forecasting teams provided forecasts of confirmed influenza hospitalizations throughout each season, which helped public health officials anticipate trends during the unusually prolonged influenza season in 2021-22, with forecasting efforts extending into June, and then again for the atypically early 2022-23 season.

While the shift to forecasting for a new target presented a modeling challenge, the utility of the corresponding new data source should be recognized. The HHS-Protect dataset [7] provided, in addition to the state-level timeseries, facility-level data, which is at a higher spatial resolution than other indicators of influenza activity. During the forecasting time frame analyzed here, the data were also reported daily with previous day admission data published as soon as the day after their occurrence, providing a timely source of information. As our data update analysis (Figures S2 – S4) shows, these data demonstrated remarkably stable reporting behavior, particularly during the 2021-22 season, with 94% of updates resulting in changes of under 10 hospitalizations for subnational jurisdictions. Stability of reporting decreased slightly during the 2022-23 season, with 83% of updates resulting in changes of under 10 hospitalizations for subnational jurisdictions. Degraded forecast performance has been associated with large revisions to initially observed values [6], and consistency in reporting is an important component of a reliable forecasting target. Additionally, this dataset provided national and jurisdictional-level data for confirmed influenza hospital admissions. In contrast with ILI, this indicator eliminated the need to model outpatient visits associated with co-circulating non-influenza pathogens that can cause ILI. Continued availability of rapid, disease-specific indicators of hospitalization, such as those provided by these data, will facilitate improved forecasting utility and possibly improvements in accuracy [27], particularly when forecasts are informed by mechanistic transmission models.

The FluSight forecasting collaboration adapted quickly in 2021 to utilize a novel laboratory confirmed influenza hospital admission dataset. Even with limited calibration data and atypical influenza seasonality in the 2021-22 and 2022-23 seasons, the FluSight ensemble forecast provided more robust forecasts than individual component models across spatial jurisdictions and time horizons. This result mirrors those of other forecasting hubs. Collaborative hubs also offer the ability for frequent feedback and interaction between modeling teams, providing opportunities for rapidly sharing observations about underlying data and insights for forecast development [28]. We observed poor coverage and general performance especially at the beginning of the 2022-23 season and during other periods of rapid change. Collective insights from these challenges can also inform when forecasts should be interpreted with extra caution. Ongoing availability of the confirmed influenza hospitalization dataset, which covers all states, could improve model calibration and ultimately contribute to the improvement of influenza forecast performance and utility, as well as continued exploration and improvement of forecasting and ensembling methodologies. These improvements are needed, particularly to more accurately capture trends and appropriate levels of uncertainty during times of rapid change.

Tables and Figures

Table 1: Performance metrics for teams submitting at least 75% of weekly FluSight targets.

Model	Absolute WIS	Relative WIS	MAE	50% Coverage (%)	95% Coverage (%)	% of Forecasts Submitted	Log Absolute WIS	Log Relative WIS
2021-22								
CMU-TimeSeries	12.54	0.74	18.92	47	90	100	0.31	0.78
Flusight-ensemble	13.86	0.82	20.79	48	86	100	0.33	0.83
PSI-DICE	14.03	0.83	20.17	43	82	100	0.33	0.84
UMass-trends_ensemble	14.35	0.84	22.24	71	97	100	0.36	0.91
Sgroup-RandomForest	15.45	0.91	23.87	47	95	100	0.38	0.97
CEID-Walk	15.63	0.94	22.19	52	82	89	0.39	0.99
Flusight-baseline	16.99	1.00	24.10	49	83	100	0.40	1.00
GT-FluFNP	17.57	1.02	23.40	39	69	96	0.38	0.98
MOBS-GLEAM_FLUH	17.17	1.03	22.25	32	63	91	0.42	1.08
SigSci-TSENS	17.79	1.03	24.86	38	72	96	0.40	1.01
IEM_Health-FluProject	17.69	1.04	23.98	50	85	100	0.40	1.02
CU-ensemble	18.32	1.08	25.41	44	77	100	0.39	0.98
LucompUncertLab-TEVA	21.02	1.22	29.99	54	86	89	0.41	1.05
UVAFluX-Ensemble	21.65	1.28	25.76	38	64	99	0.45	1.14
LucompUncertLab-VAR2_plusCOVID	22.03	1.30	28.99	42	74	94	0.42	1.08
UT_FluCast-Voltaire	23.64	1.39	35.19	50	91	99	0.45	1.14
LucompUncertLab-VAR2K_plusCOVID	24.44	1.42	32.43	42	74	89	0.47	1.20
LucompUncertLab-VAR2	25.93	1.53	35.05	39	72	94	0.53	1.35
LucompUncertLab-VAR2K	26.81	1.55	39.35	42	83	89	0.61	1.56
LosAlamos_NAU-Cmodel_Flu	28.69	1.69	36.14	26	59	100	0.63	1.60
Sgroup-SikJalpha	28.94	1.70	38.59	18	46	100	0.49	1.24
GH-Flusight	30.93	1.82	31.89	6	13	94	0.74	1.88
SigSci-CREG	27.36	1.93	31.00	19	44	89	0.80	2.03
2022-23								
MOBS-GLEAM_FLUH	42.20	0.61	57.97	42	78	94	0.37	0.65
CMU-TimeSeries	44.48	0.67	65.94	49	87	94	0.41	0.70
PSI-DICE	47.45	0.70	63.17	48	80	100	0.42	0.71
MIGHTE-Nsemble	48.99	0.72	67.50	53	82	96	0.41	0.70
Flusight-ensemble	51.72	0.76	71.04	56	81	100	0.44	0.74
Umass-trends_ensemble	53.86	0.80	79.40	63	89	100	0.49	0.83

Model	Absolute WIS	Relative WIS	MAE	50% Coverage (%)	95% Coverage (%)	% of Forecasts Submitted	Log Absolute WIS	Log Relative WIS
GT-FluFNP	59.75	0.81	72.88	56	75	89	0.53	0.89
CEPH-Rtrend_fluH	54.20	0.83	70.47	44	78	90	0.58	1.05
Sgroup-RandomForest	54.29	0.83	75.98	53	84	97	0.52	0.88
CU-ensemble	62.23	0.85	75.57	51	70	84	0.51	0.85
UGA_flucast-Okeeffe	62.13	0.94	77.33	50	72	95	0.61	1.02
SigSci-TSENS	64.27	0.96	80.02	58	74	93	0.66	1.09
Flusight-baseline	67.69	1.00	80.05	49	74	100	0.59	1.00
VTSanghani-ExogModel	72.30	1.00	92.56	30	61	81	0.63	1.05
UNC_IDD-InfluPaint	61.14	1.01	77.90	40	75	79	0.52	0.94
UVAFluX-Ensemble	78.71	1.11	94.45	22	41	95	0.61	1.03
SigSci-CREG	79.68	1.36	89.29	38	62	91	0.68	1.16
JHU_IDD-CovidSP	129.16	1.91	174.98	48	80	81	0.49	0.82

The Absolute WIS column refers to the Weighted Interval Score for each model across all fifty states, D.C., and Puerto Rico forecast targets. The Relative WIS compares the WIS value of each model to the Flusight-baseline model. All models with a relative WIS score less than one outperformed the baseline model when evaluated solely upon WIS. 95% and 50% coverage values are provided for the percent of times that reported weekly incidence values were within the 95% or 50% prediction intervals respectively, across all the forecast targets submitted by each team. The percent of forecasts submitted is determined by the number of forecast targets submitted by each team out of all possible forecast targets occurring within the duration of the evaluation period.

Table 2: One-to-four-week coverage and one-to-four-week percent of coverage above 90% for teams meeting inclusion criteria.

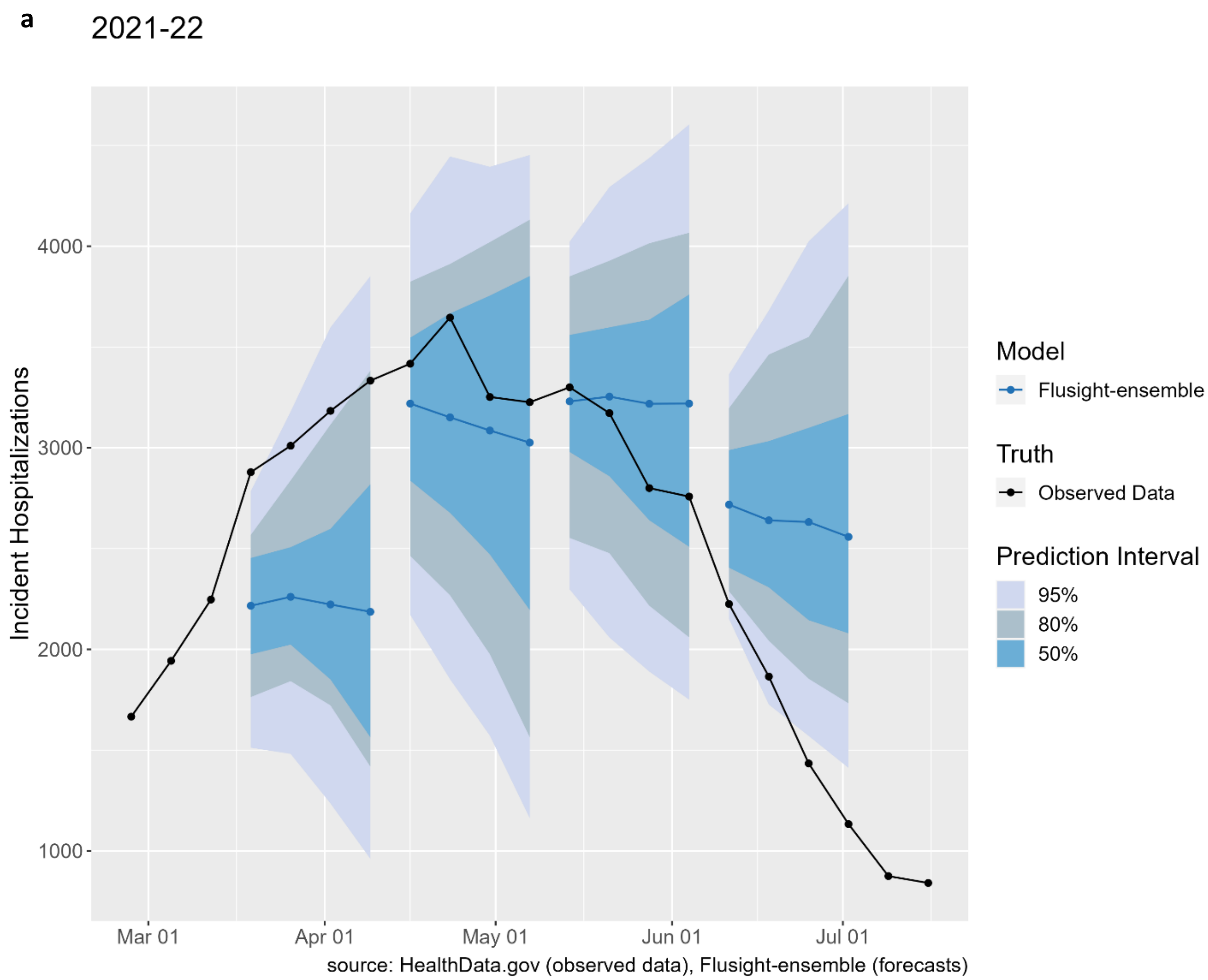
Model	Relative WIS	% WIS Below Baseline	Coverage				% Coverage above 90			
			1 Wk	2 Wk	3 Wk	4 Wk	1 Wk	2 Wk	3 Wk	4 Wk
2021-22										
CMU-TimeSeries	0.74	75.00	90.17	91.45	90.60	86.54	50.00	72.22	61.11	27.78
Flusight-ensemble	0.82	92.31	89.32	86.11	85.15	83.33	55.56	33.33	27.78	38.89
PSI-DICE	0.83	76.92	88.89	83.87	78.31	76.50	38.89	27.78	5.56	0.00
Umass-trends_ensemble	0.84	48.08	96.15	97.65	96.90	96.15	100.00	100.00	100.00	100.00
Sgroup-RandomForest	0.91	44.23	95.41	94.87	94.66	94.12	88.89	88.89	83.33	88.89
CEID-Walk	0.94	80.77	82.09	83.77	81.01	81.85	37.50	37.50	31.25	37.50
Flusight-baseline	1.00	0.00	82.26	84.19	82.48	81.62	27.78	22.22	22.22	22.22
GT-FluFNP	1.02	54.00	70.11	68.67	68.22	70.11	5.56	16.67	16.67	22.22
MOBS-GLEAM_FLUH	1.03	60.00	71.11	65.80	59.79	56.49	0.00	0.00	0.00	0.00
SigSci-TSENS	1.03	46.00	74.11	73.44	70.54	69.20	11.11	5.56	5.56	5.56
IEM_Health-FluProject	1.04	48.08	91.45	86.54	82.59	78.21	72.22	38.89	22.22	22.22
CU-ensemble	1.08	32.69	79.59	80.66	76.50	71.90	16.67	11.11	0.00	0.00
LucompUncertLab-TEVA	1.22	32.69	84.86	85.58	86.06	86.18	25.00	18.75	25.00	31.25
UVAFluX-Ensemble	1.28	25.00	66.05	65.51	62.58	60.95	11.11	0.00	0.00	0.00
LucompUncertLab-VAR2_plusCOVID	1.30	36.54	76.70	74.77	73.30	70.14	17.65	5.88	5.88	5.88
UT_FluCast-Voltaire	1.39	5.77	94.73	90.96	89.13	90.42	83.33	72.22	55.56	61.11
LucompUncertLab-VAR2K_plusCOVID	1.42	25.00	75.72	75.24	74.04	72.72	6.25	0.00	0.00	0.00
LucompUncertLab-VAR2	1.53	9.62	73.87	72.29	72.17	70.81	11.76	5.88	11.76	11.76
LucompUncertLab-VAR2K	1.55	9.62	81.97	81.49	83.05	85.46	6.25	18.75	25.00	37.50
LosAlamos_NAU-Cmodel_Flu	1.69	13.46	65.28	59.29	56.52	54.06	5.56	0.00	0.00	0.00
Sgroup-SikJalpha	1.70	1.92	40.28	45.73	48.08	48.29	0.00	0.00	0.00	0.00
GH-Flusight	1.82	5.77	18.33	12.90	11.99	10.63	0.00	0.00	0.00	0.00
SigSci-CREG	1.93	12.00	46.87	43.98	43.86	43.13	0.00	0.00	0.00	0.00
2022-23										
MOBS-GLEAM_FLUH	0.61	94.12	81.34	77.50	76.84	77.67	41.94	29.03	29.03	23.33
CMU-TimeSeries	0.67	86.54	86.27	87.12	87.25	86.31	58.06	64.52	70.97	70.00
PSI-DICE	0.70	92.31	88.03	81.27	77.17	74.87	64.52	67.74	64.52	60.00
MIGHTE-Nsemble	0.72	90.38	86.16	84.22	81.71	76.00	63.33	60.00	66.67	58.62
Flusight-ensemble	0.76	100.00	85.79	81.64	78.78	77.12	64.52	67.74	64.52	60.00

Model	Relative WIS	% WIS Below Baseline	Coverage				% Coverage above 90			
			1 Wk	2 Wk	3 Wk	4 Wk	1 Wk	2 Wk	3 Wk	4 Wk
Umass-trends_ensemble	0.80	92.31	90.88	89.89	87.41	85.83	77.42	74.19	70.97	70.00
GT-FluFNP	0.81	96.00	75.98	72.70	75.00	77.30	55.17	55.17	55.17	65.52
CEPH-Rtrend_fluH	0.83	67.31	75.82	80.22	79.33	77.21	46.43	50.00	57.14	44.44
Sgroup-RandomForest	0.83	92.31	90.06	84.49	81.86	79.71	73.33	70.00	70.00	65.52
CU-ensemble	0.85	75.00	71.60	71.38	69.90	66.85	46.15	53.85	53.85	52.00
UGA_flucast-Okeeffe	0.94	58.82	80.20	73.07	69.02	65.86	50.00	46.67	40.00	37.93
SigSci-TSENS	0.96	42.00	76.31	74.12	72.93	70.35	54.84	54.84	54.84	56.67
Flusight-baseline	1.00	0.00	78.72	74.26	71.34	68.85	58.06	58.06	58.06	56.67
VTSanghani-ExogModel	1.00	51.92	65.62	61.54	58.00	57.29	0.00	0.00	0.00	4.17
UNC_IDD-InfluPaint	1.01	64.71	75.20	74.25	75.12	75.14	52.00	44.00	64.00	54.17
UVAFluX-Ensemble	1.11	11.76	42.81	43.53	39.35	39.42	0.00	0.00	0.00	0.00
SigSci-CREG	1.36	6.00	68.28	62.27	58.85	55.34	48.39	48.39	45.16	43.33
JHU_IDD-CovidSP	1.91	33.33	86.74	81.67	78.18	73.60	65.38	61.54	53.85	48.00

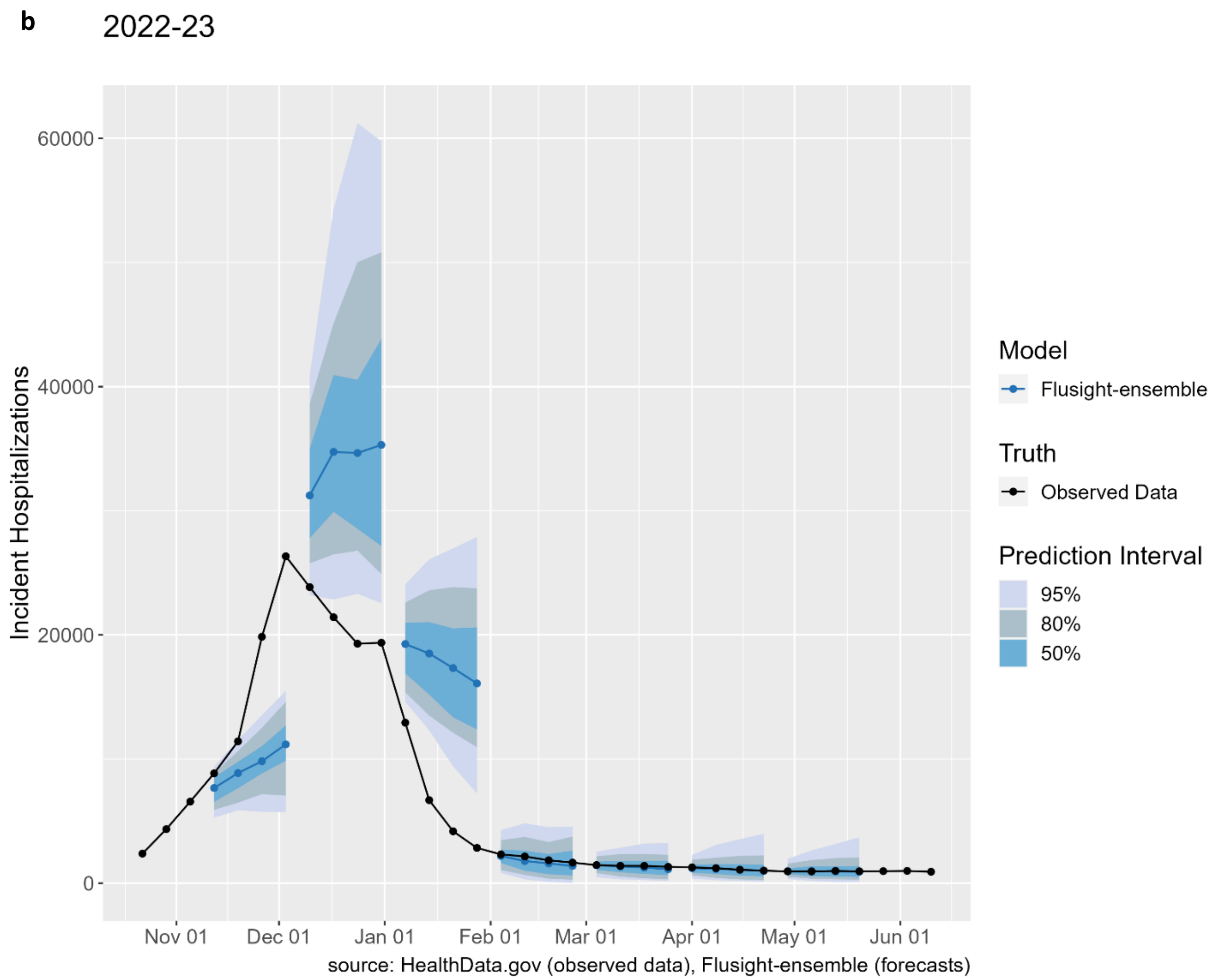
Table 2: % WIS Below Baseline shows the percent of WIS values for each model below the corresponding Flusight-baseline WIS. The '% Coverage above 90' columns show the percent of weekly 95% coverage values that are greater than or equal to 90% for each model by horizon. Modeling teams are ordered within each season by their relative WIS performance.

Figure 1: National weekly observed hospitalizations (black points) along with FluSight ensemble forecasts for four weeks of submissions in the 2021-22 season (panel a) and seven weeks of submissions in the 2022-23 season (panel b). The median FluSight ensemble forecast values (blue points) are shown with the corresponding 50%, 80%, and 95% prediction intervals (blue shaded regions).

a



b



2022-23
Incident Hospitalizations
60000
40000
20000
0
Nov 01
Dec 01
Jan 01
Feb 01
Mar 01
Apr 01
May 01
Jun 01
source: HealthData.gov (observed data), Flusight-ensemble (forecasts)
Model
Flusight-ensemble
Truth
Observed Data
Prediction Interval
95%
80%
50%

Figure 2: Standardized rank of weighted interval score (WIS) over all forecast jurisdictions and horizons (1- to 4-week ahead), for the FluSight ensemble and each team submitting at least 75% of the forecast targets (see Table 1 for qualifying teams and season metrics).

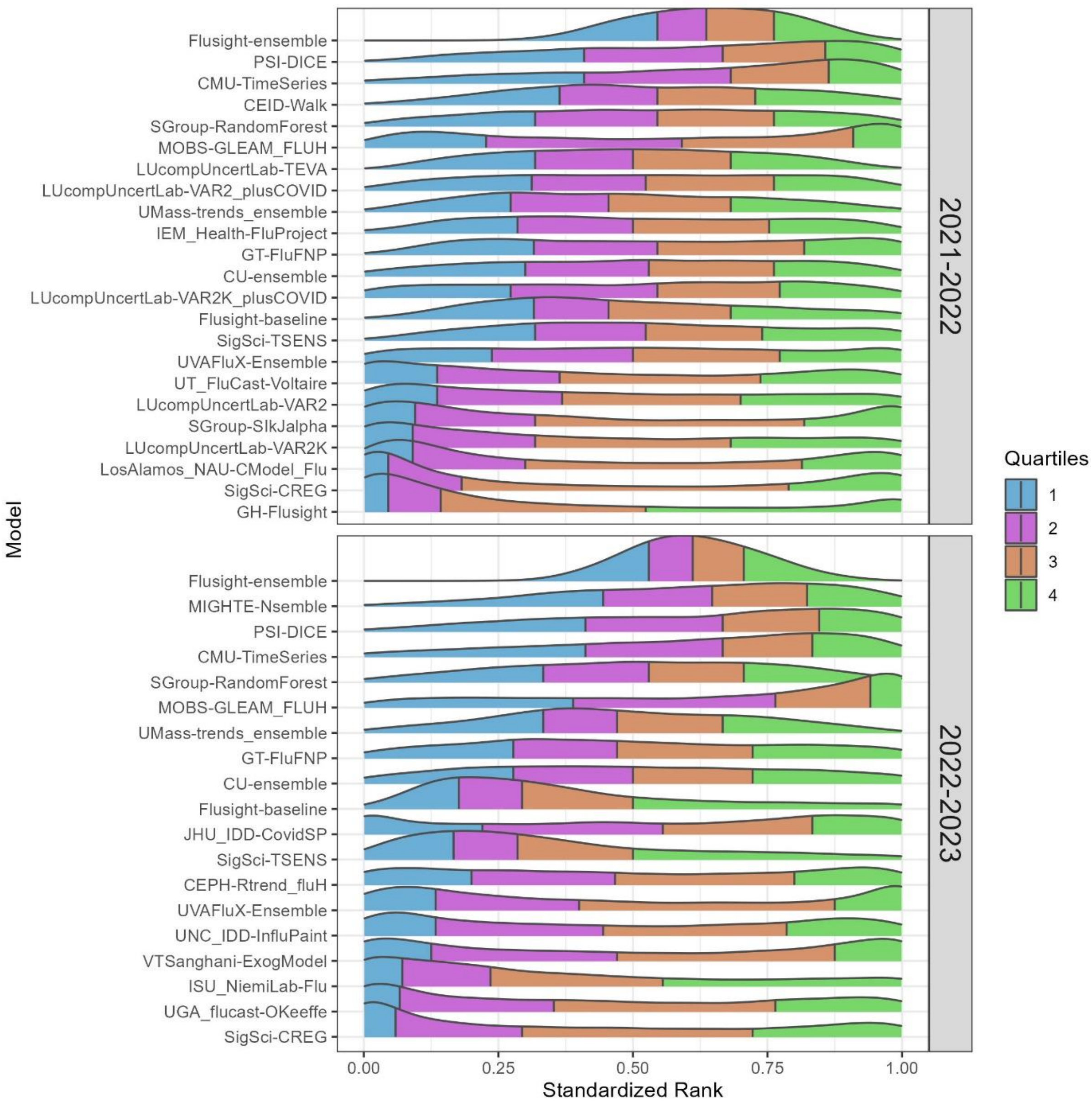


Figure 3: State-level WIS values for each team relative to the FluSight baseline model. The range of Relative WIS values below 1, in blue, indicate better performance than the FluSight baseline (white). Relative WIS values above 1, in red, indicate poor performance relative to the FluSight baseline. Teams are ordered on horizontal axis from lowest to highest Relative WIS values for each season. Analogous jurisdiction-specific relative WIS scores on log transformed counts are displayed in Figure S7.

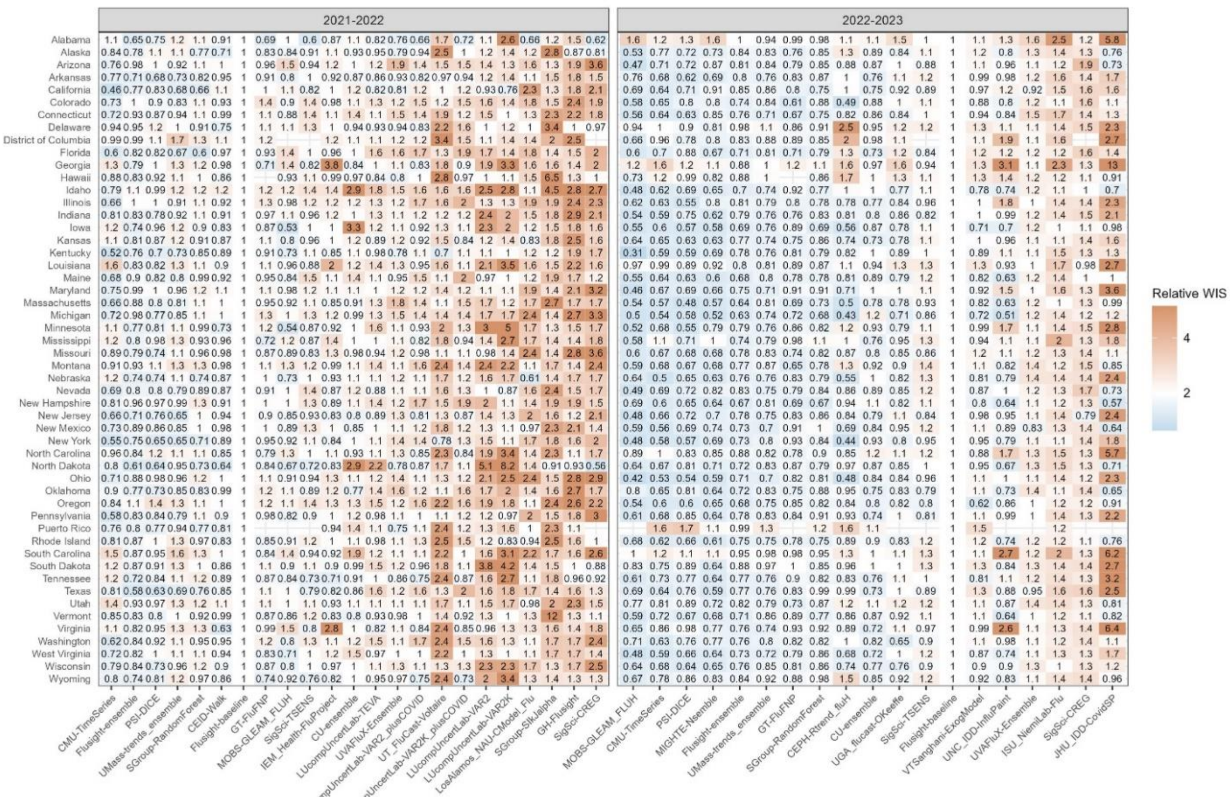
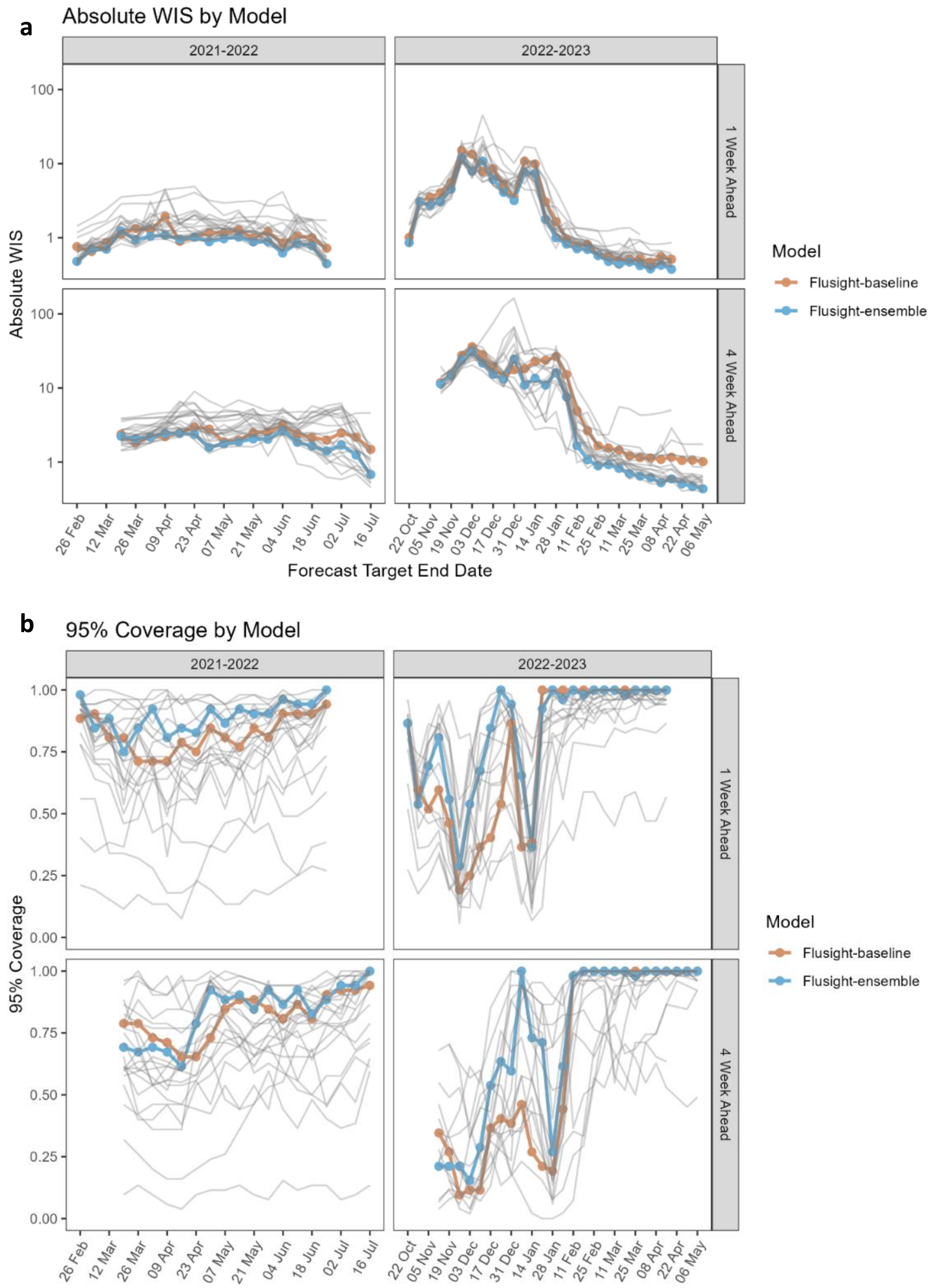


Figure 4: Time series of log transformed absolute WIS (panel a) and 1- and 4-week ahead 95% coverage (panel b) for state and territory targets. Note that the forecast evaluation period translates to 1-week ahead forecast target end dates from February 26 to June 25, 2022, and October 22, 2022, to May 20, 2023, and 4-week ahead forecast target end dates from March 19 to July 16, 2022, and November 5, 2022, to June 10, 2023. Weekly results for the FluSight baseline and ensemble models are shown in red and blue respectively. Results for individual contributing models are shown in light gray.



References

1. Weekly U.S. Influenza Surveillance Report. https://www.cdc.gov/flu/weekly/index.htm. (2023)

2. U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet). https://wwwn.cdc.gov/ILINet/. (2023)

3. Influenza Hospitalization Surveillance Network (FluSurv-NET). https://www.cdc.gov/flu/weekly/influenza-hospitalization-surveillance.htm. (2023)

4. Lutz C.S., et al. Applying infectious disease forecasting to public health: a path forward using influenza forecasting examples. *BMC Public Health*, **19-1659** (2019).

5. McGowan C. J., et al.; Influenza forecasting working group, Collaborative efforts to forecast seasonal influenza in the United States, 2015-2016. *Sci. Rep. 9*, **683** (2019).

6. Reich N. G., et al., A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. *Proc. Natl. Acad. Sci. U.S.A. 116*, **3146–3154** (2019).

7. COVID-19 Reported Patient Impact and Hospital Capacity by State Timeseries (RAW). https://healthdata.gov/Hospital/COVID-19-Reported-Patient-Impact-and-Hospital-Capa/g62h-syeh. (2023).

8. COVID-19 Guidance for Hospital Reporting and FAQs For Hospitals, Hospital Laboratory, and Acute Care Facility Data Reporting, 11 June 2023, www.hhs.gov/sites/default/files/covid-19-faqs-hospitals-hospital-laboratory-acute-care-facility-data-reporting.pdf.

9. Olsen S.J., et al., Changes in influenza and other respiratory virus activity during the COVID-19 pandemic—United States, 2020–2021. *Morbidity and Mortality Weekly Report*. **Jul 7, 70(29)-1013** (2021).

10. Merced-Morales A., et al. Influenza activity and composition of the 2022–23 influenza vaccine—United States, 2021–22 season. *Morbidity and Mortality Weekly Report*. **Jul 7, 71(29)-913** (2022).

11. Centers for Disease Control and Prevention. https://www.cdc.gov/flu/spotlights/2023-2024/22-23-summary-technical-report.htm. (2023).

12. Centers for Disease Control and Prevention - MMWR Weeks definition. https://ndc.services.cdc.gov/wp-content/uploads/MMWR_Week_overview.pdf. (2023).

13. FluSight Forecast Hub. https://github.com/cdcepi/Flusight-forecast-data. (2023).

14. Simple models for time series forecasting version 0.0.0.1000. https://rdrr.io/github/reichlab/simplets/. (2023).

15. Cramer E.Y., et al., Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the US. *Proc. Natl. Acad. Sci. U.S.A*., https://doi.org/10.1073/pnas.2113561119 (2022).

16. Cramer E.Y., et al. The United States COVID-19 forecast hub dataset. *Scientific Data*. **Aug 1, 9(1), 1-5** (2022).

17. HHS Protect Public Data Hub. https://public-data-hub-dhhs.hub.arcgis.com/. (2023)

18. Bracher J., Ray E. L., Gneiting T., Reich N. G.,. Evaluating epidemic forecasts in an interval format. *arXiv preprint* **arXiv:2005.12881** (2020).

19. Bosse N. I., et al.. Transformation of forecasts for evaluating predictive performance in an epidemiological context. *PLOS Comput Biol*, **19(8):e1011393** (2023).

20. Reich N.G., et al. Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the U.S. *PloS Comput Biol*, **15:e1007486** (2019).

21. Biggerstaff M., Slayton R. B., Johansson M. A., Butler J. C., Improving pandemic response: employing mathematical modeling to confront COVID-19. *Clin. Infect. Dis.*, **10.1093/cid/ciab673** (2021).

22. Reich N.G., et al. Collaborative hubs: making the most of predictive epidemic modeling. *American Journal of Public Health*. **Apr 13(0):e1-4** (2022).

23. Howerton E., et al. Informing pandemic response in the face of uncertainty. An evaluation of the U.S. COVID-19 scenario modeling hub. m*edRxiv.* (2023).

24. Lopez V. K., et al. Challenges of COVID-19 case forecasting in the US, 2020-2021. *medRxiv*, *Cold Spring Harbor Laboratory Press*, **1 Jan**. (2023).

25. Srivastava A., Singh S., Lee F. Shape-based evaluation of epidemic forecasts. *arXiv preprint* **arXiv:2209.04035**. (2022)

26. Adiga A., et al. Phase-Informed Bayesian ensemble models improve performance of COVID-19 forecasts. *Proceedings of the AAAI Conference on Artificial Intelligence,* **37(13), 15647-15653**. https://doi.org/10.1609/aaai.v37i13.26855. (2023).

27. Fox J., et al., Real-time pandemic surveillance using hospital admissions and mobility data. *Proc. Natl. Acad. Sci. U.S.A. 119*, **10.1073/pnas.2111870119** (2022).

28. Johansson M.A., et al. An open challenge to advance probabilistic forecasting for dengue epidemics. *Proc. Natl. Acad. Sci. U S A*. **116-24, 268–74** (2019).

29. Zoltar Forecast Archive. https://zoltardata.com/project/299. (2023).

30. R Core Team. R: A language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria. <https://www.R-project.org/>. (2023).

Disclaimers

Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government. 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 or the National Institutes of Health.

Data Availability

The forecasts from models used in this paper are available from the FluSight Forecast Hub GitHub repository (https://github.com/cdcepi/Flusight-forecast-data) [13] and the Zoltar forecast archive (https://zoltardata.com/project/299) [29]. These are both publicly accessible. The code used to generate all figures and tables in the manuscript will be available in a public repository (https://github.com/cdcepi/FluSight-manuscripts) at the time of publication. All analyses were conducted using the R language for statistical computing (version 4.0.3) [30].

Acknowledgments

The authors would like to acknowledge Michael A. Johansson and Nicole Samay for their contributions to this work.

M.B.N., P.R., J.T., S.V., A.A., G.K., B.H., B.L.L., M.V.M., M.A.A, A.S. disclose support for the research of this work from the Centers for Disease Control and Prevention (CDC) and Council of State and Territorial Epidemiologists (CSTE), [Cooperative Agreement number NU380T000297]

M.C., J.D, K.M., X.X, APP, AV, P.C.V., A.G.K. M.L., M.A. disclose support for the research of this work from the HHS/CDC 6U01IP001137 and HHS/CDC 5U01IP0001137.

B.A.P., A.R., H.P.K., Z.Z., G.G., P.A., S.S.B., R.V.V., disclose support for the research of this work from NSF (Expeditions CCF-1918770, CAREER IIS-2028586, RAPID IIS-2027862, Medium IIS-1955883, Medium IIS-2106961, PIPP CCF-2200269), CDC MInD program, faculty gifts from Facebook/Meta, and funds/computing resources from Georgia Tech and GTRI.

M.S., L.C. F.L and A.G. M. disclose support for the research of this work from the Centers for Disease Control and Prevention (CDC) and Council of State and Territorial Epidemiologists (CSTE), [Cooperative Agreement number NU380T000297]

M.S. discloses support for the research of this work from the National Institutes of Health (grant number R01GM130668) and (in part) by contract 200-2016-91779 with the Centers for Disease Control and Prevention.

S.V., A.A., G.K., B.H., B.L.L., M.V.M., also disclose support for the research of this work from NSF Expeditions CCF-1918656, VDH Grant VDH-21-501-0135, University of Virginia Strategic Investment Fund Award SIF160.

L.C.B., A.G., A.J.H., D.J.M., R.R., D.S., R.J.T. disclose support for the research of this work from the Centers for Disease Control and Prevention U011P001121 and Centers for Disease Control and Prevention 75D30123C15907.

B.T.S., S.A.S., H.L.G., and P.B. disclose support for the research of this work from the Centers for Disease Control and Prevention (CDC) and Council of State and Territorial Epidemiologists (CSTE), [Cooperative Agreement number NU38OT000297]

N.R. discloses support from National Science Foundation grants CCF-1918770, NRT DGE-1545362, and OAC-1835660.

S.T., C.P.S., A.H. disclose support for the research of this work from the National Science Foundation [2127976]. S.T., C.P.S., A.H., J.L., J.C.L., S.L.L., C.D.M., K.S., S-m.J. disclose support from the Centers for Disease Control and Prevention [200-2016-91781].

J.L. and J.C.L. disclose support from the National Institutes of Health (NIH 5R01AI102939)

A.M., Y.T.L., and W.S.H. disclose support for the research of this work from Laboratory Directed Research and Development Program at Los Alamos National Laboratory [20220268ER].

W.S.H., R. G. P., S. L., Y. C. disclose support for the research of this work from National Institute of Health [R01GM111510].

Author contributions: S.M.M., A.E. W, M.B, and R.K.B. contributed to conceptualization. S.M.M. and R.K.B. wrote the original draft of the manuscript. S.M.M. and A.E.W. performed the formal analysis. M.B. and C.R. performed supervision and project administration. All authors contributed modeling data and T.M.L., E.L.M., M.S., L.A.W., L.C.B., A.G., A.J.H., D.J.M., R.R., D.S., R.J.T., S.K., S.P., J.S., R.Y., T.K.Y., Y.A., S.B., G.G., H.K., B.A.P., R.R., A.R., Z.Z., A.M., S.O., P.B., H.L.G., S.A.S., B.T.S., M.A., A.G.K., M.L., P.C.V., S.W., J.N., E.C., A.L.H., S.J., J.C.L., J.L. S.L.L., C.D.M., K.S., C.S., S.T., T.M., W.Y., N.B., W.S.H., Y.T.L., A.M., Y.C., S.M.L., J.L., R.G.P., A.C.P., C.V., L.C., F.L., A.G.M., M.S., M.C., J.T.D., K.M., A.P.P., A.V., X.X., M.B.N., P.R., J.T., C.H.L., S.J., V.P.N., S.D.T., D.W., A.V., J.M.D., S.J.F., G.C.G., E.S., E.W.T., M.G.C., W.Y.C., A.G., A.S., E.L.R., N.G.R., L.S., N.W., Y.W., M.W.Z., M.A.A., A.S., L.A.M., A.A.., B.H., G.K., B.L.L., M.M., S.V., P.B., A.F., N.M., AND N.R. submitted forecast data for the analysis. All authors contributed to the review and editing of the manuscript.

Competing interests: E.W.T. is an employee of Sanofi, which manufactures influenza vaccines. J.S. and Columbia University disclose partial ownership of SK Analytics. J.S. discloses consulting for BNI.