A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States

Nicholas G. Reich¹, Logan C. Brooks², Spencer J. Fox³, Sasikiran Kandula⁴, Craig J. McGowan⁵, Evan Moore⁵, Dave Osthus⁵, Evan L. Ray⁵, Abhinav Tushar⁶, Teresa K. Yamana⁶, Matthew Biggerstaff⁷, Michael A. Johansson⁸, Roni Rosenfeld⁹, and Jeffrey Shaman¹⁰

¹Department of Biostatistics and Epidemiology, University of Massachusetts-Amherst, Amherst, MA 01003; ²Computer Science Department, Carnegie Mellon University, Pittsburgh, PA 15213; ³Department of Integrative Biology, University of Texas at Austin, Austin, TX 78712; ⁴Department of Environmental Health Sciences, Columbia University, New York, NY 10032; ⁵Influenza Division, Centers for Disease Control and Prevention, Atlanta, GA 30333; ⁶Statistical Sciences Group, Los Alamos National Laboratory, Los Alamos, NM 87545; ⁷Department of Mathematics and Statistics, Mount Holyoke College, South Hadley, MA 01075; ⁸Division of Vector-Borne Diseases, Centers for Disease Control and Prevention, San Juan, PR 00920; and ⁹Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213

Edited by Sebastian Funk, London School of Hygiene & Tropical Medicine, London, United Kingdom, and accepted by Editorial Board Member Diane E. Griffin December 10, 2018 (received for review July 24, 2018)

Influenza infects an estimated 9–35 million individuals each year in the United States and is a contributing cause for between 12,000 and 56,000 deaths annually. Seasonal outbreaks of influenza are common in temperate regions of the world, with highest incidence typically occurring in colder and drier months of the year. Real-time forecasts of influenza transmission can inform public health response to outbreaks. We present the results of a multinstitution collaborative effort to standardize the collection and evaluation of forecasting models for influenza in the United States for the 2010/2011 through 2016/2017 influenza seasons. For these seven seasons, we assembled weekly real-time forecasts of seven targets of public health interest from 22 different models. We compared forecast accuracy of each model relative to a historical baseline seasonal average. Across all regions of the United States, over half of the models showed consistently better performance than the historical baseline when forecasting incidence of influenza-like illness 1 wk, 2 wk, and 3 wk ahead of available data and when forecasting the timing and magnitude of the seasonal peak. In some regions, delays in data reporting were strongly and negatively associated with forecast accuracy. More timely reporting and an improved overall accessibility to novel and traditional data sources are needed to improve forecasting accuracy and its integration with real-time public health decision making.

Significance

Accurate prediction of the size and timing of infectious disease outbreaks could help public health officials in planning an appropriate response. This paper compares approaches developed by five different research groups to forecast seasonal influenza outbreaks in real time in the United States. Many of the models show more accurate forecasts than a historical baseline. A major impediment to predictive ability was the real-time accuracy of available data. The field of infectious disease forecasting is in its infancy and we expect that innovation will spur improvements in forecasting in the coming years.

Author contributions: N.G.R., L.C.B., S.J.F., S.K., C.J.M., E.M., D.O., E.L.R., A.T., T.K.Y., M.B., M.A.J., R.R., and J.S. designed research; N.G.R., L.C.B., S.J.F., S.K., C.J.M., E.M., D.O., E.L.R., A.T., and T.K.Y. performed research; N.G.R., C.J.M., and E.M. analyzed data; and N.G.R. wrote the paper.

Conflict of interest statement: J.S. and Columbia University disclose partial ownership of SK Analytics. This article is a PNAS Direct Submission. S.F. is a guest editor invited by the Editorial Board.

This open access article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

Data deposition: The data and code for this analysis have been deposited in GitHub, https://github.com/FluSightNetwork/cdc-flusight-ensemble/tree/first-papers and a public interactive data visualization of data presented in this paper can be found at https://reichlab.shinyapps.io/FSN-Model-Comparison/

See Commentary on page 2802.

¹To whom correspondence should be addressed. Email: nick@schoolph.umass.edu.

Published online January 15, 2019.

See Commentary on page 2802.
(and in other global regions with temperate climates), often circulating widely during colder months (i.e., November through April). The social, biological, environmental, and demographic features that contribute to higher-than-usual incidence in a particular season are not fully understood, although contributing factors may include severity of the dominant influenza subtype (13), preexisting population immunity due to prior infections or vaccination (14, 15), temperature and humidity (16), vaccine effectiveness (12), or timing of school vacations (17).

Starting in the 2013/2014 influenza season, the US Centers for Disease Control and Prevention (CDC) has run the “Forecast the Influenza Season Collaborative Challenge” (a.k.a. FluSight) each influenza season, soliciting prospective, real-time weekly forecasts of regional-level weighted influenza-like illness (wILI) measures from teams across the world (Fig. 1) (8, 10). The FluSight challenge focuses on forecasts of the weighted percentage of doctor’s office visits where the patient showed symptoms of an ILI in a particular US Health and Human Services (HHS) region. Weighting is done by state population as the data are aggregated to the regional and the national level. This wILI metric is a standard measure of seasonal flu activity, for which public data are available for the United States back to the 1997/1998 influenza season (Fig. 1A). The FluSight challenge forecasts are displayed together on a website in real time and are evaluated for accuracy at the end of the season (18). This effort has galvanized a community of scientists interested in forecasting, creating a testbed for improving both the technical understanding of how different forecast models perform and the integration of these models into decision making.

Building on the structure of the FluSight challenges [and those of other collaborative forecasting efforts (9, 11)], a subset of FluSight participants formed a consortium in early 2017 to facilitate direct comparison and fusion of modeling approaches. Our work brings together 22 models from five different institutions: Carnegie Mellon University, Columbia University, Los Alamos National Laboratory, University of Massachusetts-Amherst, and University of Texas at Austin (Table 1). In this paper, we provide a detailed analysis of the performance of these different models in forecasting the targets defined by the CDC FluSight challenge organizers (Fig. 1B). Drawing on the different expertise of the five teams allows us to make fine-grained and standardized comparisons of distinct approaches to disease incidence forecasting that use different data sources and modeling frameworks.

In addition to analyzing comparative model performance over multiple seasons, this work identifies key bottlenecks that limit the accuracy and generalizability of current forecasting efforts. Specifically, we present quantitative analyses of the impact that incomplete or partial case reporting has on forecast accuracy. Additionally, we assess whether purely statistical models show similar performance to that of models that consider explicit mechanistic models of disease transmission. Overall, this work shows strong evidence that carefully crafted forecasting models for region-level influenza in the United States consistently outperformed a historical baseline model for targets of particular public health interest.

Results

Performance in Forecasting Week-Ahead Incidence. Influenza forecasts have been evaluated by the CDC primarily using a variation of the log score, a measure that evaluates both the precision and accuracy of a forecast (30). Consistent with the primary evaluation performed by the CDC, we used a modified form

Fig. 1. (A) wILI data downloaded from the CDC website from selected regions. The y axis shows the weighted percentage of doctor’s office visits in which a patient presents with ILI for each week from September 2010 through July 2017, which is the time period for which the models presented in this paper made seasonal forecasts. (B) A diagram showing the anatomy of a single forecast. The seven forecasting targets are illustrated with a point estimate (circle) and an interval (uncertainty bars). The five targets on the wILI scale are shown with uncertainty bars spanning the vertical wILI axis, while the two targets for a time-of-year outcome are illustrated with horizontal uncertainty bars along the temporal axis. The onset is defined relative to a region- and season-specific baseline wILI percentage defined by the CDC (19). Arrows illustrate the timeline for a typical forecast for the CDC FluSight challenge, assuming that forecasts are generated or submitted to the CDC using the most recent reported data. These data include the first reported observations of wILI% from 2 wk prior. Therefore, 1- and 2-wk-ahead forecasts are referred to as nowcasts, i.e., at or before the current time. Similarly, 3- and 4-wk-ahead forecasts are forecasts or estimates about events in the future.
Overall, the forecast scores for both the “nowcast” targets (1 and 2 wk) show improvements over the historical baseline model. For example, in HHS region 1, the forecast score was lower for seasonal targets than for week-ahead targets, which means that the models showed greater relative improvement compared with the baseline model (Fig. 2). The historical baseline model achieved an overall forecast score of 0.12, while the models with highest forecast scores across all 4-wk-ahead targets (CU-EKF, SIRS, and SEIRS) showed the best overall week-ahead predictability of any region. Here, the models showed an average forecast score of 0.54 for week-ahead targets (Fig. 3A). This means that in a typical season, these models assigned an average of 0.54 probability to the accurate wILI percentages. This resulted from having the highest score from the baseline model (0.37) and having the largest improvement upon baseline predictions (0.17) from the other models (Fig. 3B). In HHS region 6, the average week-ahead score was 0.24. While HHS region 6 showed the lowest historical baseline model score of any region (0.12), it also showed the second-highest improvement (0.12) upon baseline predictions (Fig. 3B).

Forecast score declined as the target moved farther into the future relative to the most recent observation. Over half of the models outperformed the historical baseline model in making 1-wk-ahead forecasts, as 15 of 22 models outperformed the historical baseline in at least six of the seven seasons. However, only 7 of 22 models outperformed the historical baseline in at least six seasons when making 4-wk-ahead forecasts. For the model with highest forecast score across all 4-wk-ahead targets (CU-EKF, SIRS), the average scores across regions and seasons for 1- through 4-wk-ahead forecasts were 0.55, 0.44, 0.36, and 0.31. This mirrored an overall decline in score observed across most models. Only in HHS region 1 were the forecast scores from the CU-EKF, SIRS model for both the “nowcast” targets (1 and 2 wk ahead) above 0.5.

Table 1. List of models, with key characteristics

| Team    | Model abbreviation | Model description                              | Ref. | Ext. data | Mech. model | Ens. model |
|---------|--------------------|------------------------------------------------|------|-----------|-------------|------------|
| CU      | EAKFC, SEIRS       | Ensemble adjustment Kalman filter SEIRS        | (20) | x         | x           |            |
|         | EAKFC, SIRS        | Ensemble adjustment Kalman filter SIRS         | (20) | x         | x           |            |
|         | EKF, SEIRS         | Ensemble Kalman filter SEIRS                  | (21) | x         | x           |            |
|         | EKF, SIRS          | Ensemble Kalman filter SIRS                   | (21) | x         | x           |            |
|         | RHF, SEIRS         | Rank histogram filter SEIRS                   | (21) | x         | x           |            |
|         | RHF, SIRS          | Rank histogram filter SIRS                    | (21) | x         | x           |            |
|         | BMA                | Bayesian model averaging                      | (22) |           |             |            |
| Delphi  | BasisRegression*   | Basis regression, epiforecast defaults         | (23) |           |             |            |
|         | DeltaDensity1*     | Delta density, epiforecast defaults            | (24) |           |             |            |
|         | EmpiricalBayes1*   | Empirical Bayes, conditioning on past 4 wk     | (23) |           |             |            |
|         | EmpiricalBayes2*   | Empirical Bayes, epiforecast defaults          | (23) |           |             |            |
|         | EmpiricalFuture*   | Empirical futures, epiforecast defaults        | (23) |           |             |            |
|         | EmpiricalTraj*     | Empirical trajectories, epiforecast defaults   | (23) |           |             |            |
|         | DeltaDensity2*     | Markovian Delta density, epiforecast defaults  | (24) |           |             |            |
|         | Uniform*           | Uniform distribution                           |       |           |             |            |
|         | Stat               | Ensemble, combination of 8 Delphi models       | (24) |           |             | x          |
| LANL    | DBM                | Dynamic Bayesian SIR model with discrepancy    | (26) |   x       |             |            |
| ReichLab| KCDE               | Kernel conditional density estimation          | (27) |           |             |            |
|         | KDE                | Kernel density estimation and penalized splines| (28) |           |             |            |
|         | SARIMA1            | SARIMA model without seasonal differencing     | (28) |           |             |            |
|         | SARIMA2            | SARIMA model with seasonal differencing        | (28) |           |             |            |
| UTAustin| EDM                | Empirical dynamic model or method of analogues | (29) |           |             |            |

Team abbreviations: CU, Columbia University; Delphi, Carnegie Mellon; LANL, Los Alamos National Laboratories; ReichLab, University of Massachusetts-Amherst; SEIRS, Susceptible-Exposed-Infectious-Recovered-Susceptible, and SIRS, Susceptible-Infectious-Recovered-Susceptible, compartmental models of infectious disease transmission; UTAustin, University of Texas at Austin. The “Ext. data” column notes models that use data external to the ILINet data from CDC. The “Mech. model” column notes models that rely to some extent on a mechanistic or compartmental model formulation. The “Ens. model” column notes models that are ensemble models.

*Note that some of these components were not designed as standalone models, so their performance may not reflect the full potential of the method’s accuracy (Materials and Methods).
Fig. 2. Average forecast score by model region and target type, averaged over weeks and seasons. The text within the grid shows the score itself. The white midpoints of the color scale is set to be the target- and region-specific average of the historical baseline model, ReichLab-KDE, with darker blue colors representing models that have better scores than the baseline and darker red colors representing models that have worse scores than the baseline. The models are sorted in descending order from most accurate (top) to least accurate (bottom) and regions are sorted from high scores (right) to low scores (left).

Fig. 3. Absolute and relative forecast performance for week-ahead (A and B) and seasonal (C and D) targets, summarized across all models that on average performed better than the historical baseline. A and C show maps of the United States that illustrate spatial patterns of average forecast accuracy for week-ahead (A) and seasonal (C) targets. Color shading indicates average forecast score for this model subset. B and D compare historical baseline model score (x axis) with the average score (y axis, horizontal dashed line at average across regions) with one point for each region. For example, a y value of 0.1 indicates that the models on average assigned 10% more probability to the eventually observed value than the baseline historical model. The digits in the plot refer to the corresponding HHS region number, with N indicating the US national region. E shows the number of seasons each model had average performance above the historical baseline.

% change from baseline

| week-ahead forecast score | seasonal forecast score |
|--------------------------|------------------------|
|                           |                        |

an overall forecast score of 0.36, more than a twofold increase in score over the baseline.

Of the three seasonal targets, models showed the lowest average score in forecasting season onset, with an overall average score of 0.15. Due to the variable timing of season onset, different numbers of weeks were included in the final scoring for each region-season (Materials and Methods). Of the 77 region-seasons evaluated, 9 had no onset; i.e., the wILI did not remain above a fixed region-specific threshold of influenza activity for 3 or more weeks (see Materials and Methods for details). The best model for onset was LANL-DBM, with an overall average score of 0.33 and region-season-specific scores for onset that ranged from 0.03 to 0.81. The historical baseline model showed an average score of 0.11 in forecasting onset. Overall, 8 of 22 models (36%) had better overall score for onset in at least six of the seven seasons evaluated (Fig. 3E).

Accuracy in forecasting season onset was also impacted by revisions to wILI data. In some region-seasons current data led models to be highly confident that onset had occurred in one week, only to have revised data later in the season change the week that was considered to be the onset. One good example of this is HHS region 2 in 2015/2016. Here, data in early 2016 showed season onset to be epidemic week 2 (EW2) of 2016. Revisions to the data around EW12 led the models to identify EW51 as the onset. A further revision, occurring in EW21 of 2016, showed the onset actually occurred on EW4 of 2016. Networked metapopulation models that take advantage of observed activity in one location to inform forecasts of other locations have shown promise for improving forecasts of season onset (31).

Models showed an overall average score of 0.23 in forecasting peak week. The best model for peak week was ReichLab-KCDE (Table 1), with an overall average score of 0.35. Region-
probability to within 1 wk of the observed peak week while at the peak week, 14 models assigned on average 0.3 or more probability to the eventually observed peak week.

### Comparing Models' Forecasting Performance by Season.

Averaging across all targets and locations, forecast scores varied widely by model and season (Fig. 5). The historical baseline model (ReichLab-KDE) showed an average seasonal score of 0.20, meaning that in a typical season, across all targets and locations, this model assigned on average 0.20 probability to the eventually observed value. The models with the highest average seasonal forecast score (Delphi-Stat) (Table 1) and the lowest one (Delphi-EmpiricalBayes2) (Table 1) had scores of 0.37 and 0.07, respectively. Of the 22 models, 16 models (73%) showed higher average seasonal forecast score than the historical average. Season-to-season variation was substantial, with 10 models having at least one season with greater average forecast score than the Delphi-Stat model did.

The six top-performing models used a range of methodologies, highlighting that very different approaches can result in very similar overall performance. The overall best model was an ensemble model (Delphi-Stat) that used a weighted combination of other models from the Delphi group. Both the ReichLab-KCDE and the Delphi-DeltaDensity1 (Table 1) models used kernel conditional density estimation, a non-parametric statistical methodology that is a distribution-based variation on nearest-neighbors regression. These models used different implementations and different input variables, but showed similarly strong performance across all seasons. The UTAustin-EDM (Table 1) and Delphi-DeltaDensity2 models also used variants of nearest-neighbors regression, although overall scores for these models were not consistent, indicating that implementation details and/or input variables can impact the performance of this approach. The LANL-DBM and CU-EKF-SIRS models both rely on a compartmental model of influenza transmission; however, the methodologies used to fit and forecast were different for these approaches. The ReichLab-SARIMA2 (Table 1) model used a classical statistical time-series model, the seasonal autoregressive integrated moving average (SARIMA), to fit and generate forecasts. Interestingly, several pairs of models, although having strongly contrasting methodological approaches, showed similar overall performance; e.g., CU-EKF-SIRS and ReichLab-SARIMA2, LANL-DBM and ReichLab-KCDE.

### Comparison Between Statistical and Compartmental Models.

On the whole, statistical models achieved similar or slightly higher scores to those of compartmental models when forecasting both week-ahead and seasonal targets, although the differences were small and of minimal practical significance. Using the best three overall models from each category, we computed the average forecast score for each combination of region, season, and target (Table 2). For all targets, except 1-wk-ahead forecasts and peak intensity, the difference in model score was slight and never greater than 0.02. For 1-wk-ahead forecasts, the statistical models had slightly higher scores on average than mechanistic models (0.06, on the probability scale). We note that the 1-wk-ahead forecasts from the compartmental models from the CU team are driven by a statistical nowcast model that uses data from the Google Search application programing interface (API) (32). Therefore, the CU models were not counted as mechanistic models for 1-wk-ahead forecasts. For peak percentage forecasts, the statistical models had slightly higher scores on average than mechanistic models (0.05).

### Delayed Case Reporting Impacts Forecast Score.

In the seven seasons examined in this study, wILI percentages were often revised after first being reported. The frequency and magnitude of revisions varied by region, and the majority of initial values (nearly 90%) are within ±0.5% of the final observed value. For example, in HHS region 9, over 51% of initially reported wILI values ended up being revised by over 0.5 percentage points while in HHS region 5 less than 1% of values were revised that much.

![Fig. 4. Average forecast score by model and week relative to peak. Scores for each location-season were aligned to summarize average performance relative to the peak week on the x axis, zero indicates the peak week and positive values represent weeks after the peak week. In general, models that were updating forecasts based on current data showed improved accuracy for peak targets once the peak had passed. Only several of the models consistently assigned probabilities greater than 0.2 to the eventually observed values before the peak week.](image)

![Fig. 5. Average forecast score, aggregated across targets, regions, and weeks, plotted separately for each model and season. Models are sorted from lowest scores (left) to highest scores (right). Higher scores indicate better performance. Circles show average scores across all targets, regions, and weeks within a given season. The “x” marks the geometric mean of the seven seasons. The names of compartmental models are shown in boldface type. The ReichLab-KDE model (red italics) is considered the historical baseline model.](image)
Across all regions, 10% of observations were ultimately revised by more than 0.5 percentage points.

When the first report of the wILI measurement for a given region–week was revised in subsequent weeks, we observed a corresponding strong negative impact on forecast accuracy. Larger revisions to the initially reported data were strongly associated with a decrease in the forecast score for the forecasts made using the initial, unrevised data. Specifically, among the four top-performing nonensemble models (ReichLab-KCDE, LANL-DBM, Delphi-Density1, and CU-EKF_SIRS), there was an average change in forecast score of $-0.29$ (95% CI: $-0.39$, $-0.19$) when the first observed wILI measurement was between 2.5 and 3.5 percentage points lower than the final observed value, adjusting for model, week of year, and target (Fig. 6; see Materials and Methods for details on regression model). Additionally, we observed an expected change in forecast score of $-0.24$ (95% CI: $-0.29$, $-0.19$) when the first observed wILI measurement was between 1.5 and 2.5 percentage points higher than the final observed value. This pattern is similar for underestimate and overestimate values, although there were more extreme underreported values than there were overreported values. Some of the variation in region-specific performance could be attributed to the frequency and magnitude of data revisions.

**Discussion**

This work presents a large-scale comparison of real-time forecasting models from different modeling teams across multiple years. With the rapid increase in infectious disease forecasting efforts, it can be difficult to understand the relative importance of different methodological advances in the absence of an agreed-upon set of standard evaluations. We have built on the foundational work of CDC efforts to establish and evaluate models against a set of shared benchmarks which other models can use for comparison. Our collaborative, team science approach highlights the ability of multiple research groups working together to uncover patterns and trends of model performance that are harder to observe in single-team studies.

Seasonal influenza in the United States, given the relative accessibility of historical surveillance data and recent history of coordinated forecasting “challenges,” is an important testbed system for understanding the current state of the art of infectious disease forecasting models. Using models from some of the most experienced forecasting teams in the country, this work reveals several key results about forecasting seasonal influenza in the United States: A majority of models consistently showed higher accuracy than historical baseline forecasts, both in regions with more predictable seasonal trends and in those with less consistent seasonal patterns (Figs. 3 B, D, and E); a majority of the presented models showed consistent improvement over the historical baseline for 1- and 2-week-ahead forecasts, although fewer models consistently outperformed the baseline model for 3- and 4-week-ahead forecasts (Fig. 3E); at the presented spatial and temporal resolutions for influenza forecasts, we did not identify substantial or consistent differences between high-performing models that rely on an underlying mechanistic (i.e., compartmental) model of disease transmission and those that are more statistical in nature (Table 2); and forecast accuracy is significantly degraded in some regions due to initial partially reported real-time data (Fig. 6).

As knowledge and data about a given infectious disease system improve and become more granular, a common question among domain-area experts is whether mechanistic models will outperform more statistical approaches. However, the statistical vs. mechanistic model dichotomy is not always a clean distinction in practice. In the case of influenza, mechanistic models simulate a specific disease transmission process governed by the assumed parameters and structure of the model. But observed “influenza-like illness” data are driven by many factors that have little to do with influenza transmission (e.g., clinical visitation behaviors, the symptomatic diagnosis process, the data-reporting process, etc.). Since ILI data represent an impure measure of actual influenza transmission, purely mechanistic models may be at a disadvantage in comparison with more structurally flexible statistical approaches when attempting to model and forecast ILI. To counteract this potential limitation of mechanistic models in modeling noisy surveillance data, many forecasting models that have a mechanistic core also use statistical approaches that explicitly or implicitly account for unexplained discrepancies from the underlying model (20, 26).

There are several important limitations to this work as presented. While we have assembled and analyzed a range of models from experienced influenza-forecasting teams, there are large gaps in the types of data and models represented in our library of models. For example, relatively few additional data sources have been incorporated into these models, no models are included that explicitly incorporate information about circulating strains

| Target                  | Statistical model | Compartmental model | Difference |
|-------------------------|-------------------|---------------------|------------|
| 1 wk ahead              | 0.49              | 0.43                | 0.06       |
| 2 wk ahead              | 0.40              | 0.41                | -0.01      |
| 3 wk ahead              | 0.35              | 0.34                | 0.00       |
| 4 wk ahead              | 0.32              | 0.30                | 0.02       |
| Season onset            | 0.23              | 0.22                | 0.01       |
| Season peak percentage  | 0.32              | 0.27                | 0.05       |
| Season peak week        | 0.34              | 0.32                | 0.02       |

The difference column represents the difference in the average probability assigned to the eventual outcome for the target in each row. Positive values indicate the top statistical models showed higher average score than the top compartmental models.

**Fig. 6.** Model-estimated changes in forecast skill due to bias in initial reports of wILI %.

Shown are estimated coefficient values (and 95% confidence intervals) from a multivariable linear regression using model, week of year, target, and a categorized version of the bias in the first reported value. The x-axis labels show the range of bias (e.g., “$(-0.5, 0.5)$” represents all observations whose first observed wILI % were within $\pm 0.5$ percentage points of the final reported value). Values to the left of the dashed gray line are observations whose first reported value was lower than the final value. y-axis values of less than zero (the reference category) represent decreases in expected forecast skill. The total number of observations in each category is shown above each x-axis label.
of influenza, and no model explicitly includes spatial relationships between regions. Given that several of the models rely on similar modeling frameworks, adding a more diverse set of modeling approaches would be a valuable contribution. Additionally, while seven seasons of forecasts from 22 models is the largest study we know of that compares models from multiple teams, this remains a less-than-ideal sample size to draw strong conclusions about model performance. Since each season represents a set of highly correlated dynamics across regions, few data are available from which to draw strong conclusions about comparative model performance. Finally, these results should not be used to extrapolate hypothetical accuracy in pandemic settings, as these models were optimized specifically to forecast seasonal influenza.

What is the future of influenza forecasting in the United States and globally? While long-run forecast accuracy for influenza will vary based on a variety of factors [including, e.g., data quality, the geographical scale of forecasts, population density of forecasted areas, and consistency of weather patterns over time (33)], we expect to see continued forecast improvement through competition, collaboration, and methodological and technological innovation. Further analyses that help elucidate factors that drive forecast accuracy in specific settings will be particularly instructive. We see particular promise in models that leverage different data sources, such as pathogen-specific and highly localized incidence data. Additionally, building ensemble models that capitalize on the strengths of a diverse set of individual component models will be critical to improving accuracy and consistency of models in all infectious disease forecasting settings. Ensemble forecasting was the motivation behind the creation of the FluSight Network, although it is out of the scope of this paper.

To advance infectious disease forecasting broadly, a complete enumeration and understanding of the challenges facing the field are critical. In this work, we have identified and quantified some of these challenges, specifically focusing on timely reporting of surveillance data. However, other barriers may be of equal or greater importance to continued improvement of forecasts. Often, researchers either lack access to or do not know how best to make use of novel data streams (e.g., Internet data, electronic medical health record data). Increased methodological innovation in models that merge together an understanding of biological drivers of disease transmission (e.g., strain-specific dynamics and vaccination effectiveness) with statistical approaches to model the outbreak phenomenon by incorporating public health activities will be critical to moving this field forward. From a technological perspective, additional efforts to standardize data collection, format, storage, and access will increase interoperability between groups with different modeling expertise, improve accessibility of novel data streams, and continue to provide critical benchmarks and standards for the field. Continuing to refine forecasting targets to more closely align with public health activities will improve integration of forecasts with decision making. Recent work from the CDC has developed standardized algorithms to classify the severity of influenza seasons (19), which could be used to inform the development of new forecasting targets.

Public health officials are still learning how to best integrate forecasts into real-time decision making. Close collaboration between public health policymakers and quantitative modelers is necessary to ensure that forecasts have maximum impact and are appropriately communicated to the public and the broader public health community. Real-time implementation and testing of forecasting methods play a central role in planning and assessing what targets should be forecasted for maximum public health impact.

Materials and Methods

**FluSight Challenge Overview.** Detailed methodology and results from previous FluSight challenges have been published (8, 10), and we summarize the key features of the challenge here.

During each influenza season, the wILI data are updated each week by the CDC. When the most recent data are released, the prior weeks’ reported wILI data may also be revised. The revised data, available at a particular moment in time, are available via the DELPHI real-time epidemiological data API beginning in the 2014/2015 season (34). This API enables researchers to “turn back the clock” to a particular moment in time and use the data available at that time. This tool facilitates more accurate assessment of how models would have performed in real time.

The FluSight challenges have defined seven forecasting targets of particular public health relevance. Three of these targets are fixed scalar values for a particular season: onset week, peak week, and peak intensity (i.e., the maximum observed wILI percentage). The remaining four targets are the observed wILI percentages in each of the subsequent 4 wk (Fig. 1). A season has an onset week when at least 3 consecutive weeks are above a CDC-defined regional baseline for wILI. The first of these weeks is considered to be the onset week.

The FluSight challenges have also required that all forecast submissions follow a particular format. A single submission file (a comma-separated text file) contains the forecast made for a particular EW of a season. Standard CDC definitions of EW are used (35–37). Each file contains binned predictive distributions for seven specific targets across the 10 HHS regions of the United States plus the national level. Each file contains over 8,000 rows and typically is about 400 kB in size.

Included in the model comparison presented here, previous participants in the CDC FluSight challenge were invited to provide out-of-sample forecasts for the 2010/2011 through 2016/2017 seasons. For each season, files were submitted for EW40 of the first calendar year of the season through EW20 of the following calendar year. (For seasons that contained an EW53, an additional file labeled EW53 was included.) For each model, teams submitting creating 23 forecast datasets, one for each of the weeks in the seven training seasons. In total, the forecasts represent over 40 million rows and 2.5 GB of data. Each forecast file represented a single submission file, as would be submitted to the CDC challenge. Each team created submitted forecasts in a prospective, out-of-sample fashion, i.e., fitting or training the model only on data available before the time of the forecast (Fig. 1). All teams used the Delphi epidata API to retrieve ILINet data (34). Some data sources (e.g., wILI data before the time of the forecast) were not archived in a way that made data reliably retrievable in this “real-time” manner. In these situations, teams were still allowed to use these data sources with best efforts made to ensure forecasts were made using only data available at the time forecasts would have been made.

**Summary of Models.** Five teams each submitted between one and nine separate models for evaluation (Table 1). A wide range of methodological approaches and modeling paradigms are included in the set of forecast models. For example, seven of the models use a compartmental structure (e.g., susceptible–infectious–recovered), a model framework that explicitly encodes both the transmission and the susceptible-limiting dynamics of infectious disease outbreaks. Other less directly mechanistic models use statistical approaches to model the outbreak phenomenon by incorporating recent incidence and seasonal trends. One model, DelphiStat, is an ensemble model, a combination of other models from the Delphi team. No team had early access to CDC surveillance data or any other data from sentinel surveillance sites. Every team accessed the data using the same API and baseline datasets. The Columbia University team used data from Google Extended Health Trends API to nowcast for the 1-wk-ahead target in all of its models. In their six mechanistic models, the nowcast was also used as an observation, i.e., as if it were CDC surveillance data. An analysis of results from the 2016/2017 season showed that the forecast quality of these models improved by about 7% (38). Additionally, the Columbia University team used daily specific humidity averaged over 24 y (1979–2002) in their six mechanistic models. These climatological estimates were calculated from the National Land Data Assimilation System (NLDAS) project-2 dataset.

Three models stand out as being reference models. One shared feature of these models is that their forecasts do not depend on observed data from the season being forecasted. The Delphi-Uniform model always provides a forecast that assigns equal probability to all possible outcomes. The ReichLab-KDE model yields predictive distributions based entirely on data from other seasons using kernel density estimation (KDE) for seasonal targets and a generalized additive model with cyclic penalized splines for weekly incidence. The Delphi-EmpiricalTraj model uses KDE for all targets. The “historical baseline” model named throughout this paper refers to the ReichLab-KDE model. Because this model represents a prediction that
essentially summarizes historical data, we consider this model an appropriate baseline model to reflect historical trends. We note that some of the models presented here were developed as standalone forecasting models whereas others were developed as components of a larger ensemble system. We define a standalone model as one that is rigorously validated to show optimal performance on its own. Component models could also be optimized, although they could also be developed solely to provide a specific or supplemental signal as part of a larger system. All of the Delphi group’s models except for Delphi-Stat were developed as components rather than standalone models. Despite this, some of the Delphi models, in particular, Delphi-DeltaDensity1, performed quite well relative to other standalone models. Component models can also provide useful baselines for comparison, e.g., the Delphi-Uniform model, which assigns uniform probability to all possible outcomes, and the Delphi-Empirical/rallaj model, which creates a seasonal average model that is not updated based on current data.

Once submitted to the central repository, the models were not updated or modified except in four cases to fix explicit bugs in the code that yielded numerical problems with the forecasts. (In all cases, the updates did not substantially change the performance of the updated models.) Refitting of models or tuning of model parameters was explicitly discouraged to avoid unintentional overfitting of models.

**Metric Used for Evaluation and Comparison.** The log score for a model m is defined as $\log f_m(x) = \log f_m(x) - \log f(x)$, where $f_m(x)$ is the predicted density function from model m for target Z conditional on some data x, $z$ is the observed value of the target Z, and log is the natural logarithm. The log score is a “proper” scoring rule, which has the practical implication that linear combinations of log scores will also be proper (30).

Following CDC FluSight evaluation procedures, we computed modified log scores for the targets on the wILI percentage scale such that predictions within ±0.5 percentage points are considered accurate; i.e., modified log score = $\log \left(\int_{z-0.5}^{z+0.5} f_m(x)dx \right)$. When this modification means that the resulting score is not formally a proper scoring rule, some have suggested that improper scores derived from proper scoring rules may, with large enough sample size, have negligible differences in practice (30). Additionally, this modified log score has the advantage of having a clear interpretation and was motivated and designed by public health officials to reflect an accuracy of practical significance. Hereafter, we refer to these modified log scores as simply log scores.

Average log scores can be used to compare models’ performance in forecasting for different locations, seasons, targets, or times of season. In practice, each model m has a set of log scores associated with it that are region, target, season, and week specific. We represent one specific scalar log-score value as $\log f_m(z|x)$. These values can be averaged across any of the indexes to create a summary measure of performance. For example,

$$\log f_m(z|x)$$

represents a log score for model m and target t averaged across all regions, seasons, and weeks.

While log scores are not on a particularly interpretable scale, a simple transformation enhances interpretability substantially. Exponentiating an average log score yields a forecast score equivalent to the geometric mean of the probabilities assigned to the eventually observed outcome (or, more specifically for the modified log score, to regions of the distribution eventually considered accurate). The geometric mean is an alternative measure of central tendency to an arithmetic mean, representing the nth root of a product of N numbers. Using the example from Eq. 1 above, we then have that

$$S_{m, t} = \exp \left( \frac{1}{N} \sum_{i,k,w} \log f_m(x|t) \right)$$

In this setting, this score 5 has the intuitive interpretation of being the average probability assigned to the true outcome (where average is considered to be a geometric average). Throughout this paper, we refer to an exponentiated average log score as an average score. In all cases, we compute the averages arithmetically on the log scale and exponentiate only before reporting and interpreting a final number. Therefore, all reported average scores can be interpreted as the corresponding geometric means or as the corresponding average probabilities assigned to the true outcome.

Following the convention of the CDC challenges, we included only certain weeks in the calculation of the average log scores for each target. This focuses model evaluation on periods of time that are more relevant for public health decision making. Forecasts of season onset are evaluated based on the forecasts that are received 4 wk before the onset week. After the observed onset week within a given region, peak week and peak intensity forecasts were scored for all weeks in a specific region–season up until the wILI measure drops below the regional baseline level for the final time. Week-ahead forecasts are evaluated using forecasts received 4 wk before the onset week through forecasts received 3 wk after the wILI goes below the regional baseline for the final time. In a region–season without an onset, all weeks are scored. To ensure all calculated summary measures would be finite, all log scores with values of less than −10 were assigned the value −10, following CDC scoring conventions. This rule was invoked for 2,648 scores or 0.8% of all scores that fell within the scoring period. All scores were based on “ground truth” values of wILI data obtained as of September 27, 2017.

**Specific Model Comparisons. Analysis of data revisions.** The CDC publicly releases data on doctor’s office visits due to ILI each week. These data, especially for the most recent weeks, are occasionally revised, due to new or updated data being reported to the CDC since their last publication. While often these revisions are fairly minor or nonexistent, at other times these revisions can be substantial, changing the reported wILI value by over 50% of the originally reported value. Since the reported data are used to forecast to generate current forecasts, real-time forecasts can be biased by the initially reported, preliminary data. We used a regression model to analyze the impact of these unrevised reports on forecasting. Specifically, for each region and EW we calculated the difference between the first and the last reported wILI values for each EW for which forecasts were generated in the seven seasons under consideration. We then created a binomial distribution with the probability of occurrence for each of the following six categories covering the entire range of observed values: (−3.5, −2.5], (−2.5, −1.5], ... , (1.5, 2.5]. Using the forecasting results from the four most accurate individual nonensemble models (ReichLab-KCDE, LANL-DBM, Delphi-DeltaDensity1, CU-EKF-SIIRS), we then fitted the linear regression model

$$S_t = \beta + \alpha_m + \gamma_t + \lambda_m + \theta \cdot X_t + \epsilon_t$$

where $S_t$ is the score; the index $i$ indexes a specific set of subscripts (m, r, t, s, t, w); and the $\alpha_m$, $\gamma_t$, $\lambda_m$, and $\theta$ are model-, target-, and week-specific fixed effects, respectively. [The notation m(i) refers to the model contained in the ith observation of the dataset.] The error term is assumed to follow a Gaussian distribution with mean zero and an estimated variance parameter. The parameter of interest in the model is the vector $\theta$, which represents the average change in score based on the magnitude of the change in the wILI data for the most recent week. This $\theta$ is the parameter estimated in the regression model. The parameter values were computed as the average of the change in score based on the magnitude of the change in the wILI data for the most recent week. The parameter is interpreted as the average change in score based on the magnitude of the change in the wILI data for the most recent week.

**Mechanistic vs. statistical models.** There is not a consensus on a single best modeling approach or method for forecasting the dynamic patterns of infectious disease outbreaks in both endemic and emergent settings. Semantically, modelers and forecasters often use a dichotomy of mechanistic vs. statistical (or “phenomenological”) models to represent two different approaches to modeling. Mechanistic models for infectious disease consider the biological underpinnings of disease transmission and in practice are implemented as variations on the susceptible–infectious–recovered (SIR) model. Statistical models largely ignore the biological underpinnings and theory of disease transmission and focus instead on using data-driven, empirical, and statistical approaches to make the best forecasts possible of a given dataset or phenomenon. However, in practice, this dichotomy is less clear than it is in theory. For example, statistical models for infectious disease counts may have an autoregressive term for incidence (e.g., as done by the ReichLab SARIMA model). This could be interpreted as representing a transmission process from one time period to another. In another example, the LANL-DBM model has an explicit SIR compartmental model component but also uses a purely statistical model for the discrepancy of the compartmental model with observed
1. Molodecky NA, et al. (2017) Risk factors and short-term projections for serotype-1 poliomyelitis incidence in Pakistan: A spatiotemporal analysis. PLoS Med 14: e1002323.
2. Du X, King AA, Woods RJ, Pascual M (2017) Evolution-informed forecasting of seasonal influenza A (H3N2). Sci Transl Med 9eaan5325.
3. Bansal S, Chowell G, Simonsen L, Vespignani A, Viboud C (2016) Big data for infectious disease surveillance and modeling. J Infect Dis 214:5375–5379.
4. Myers MF, Rogers DJ, Cox J, Flahault A, Hay SI (2000) Forecasting disease risk for increased epidemic preparedness in public health. Adv Parasitol 47: 309–330.
5. World Health Organization (2016) Anticipating emerging infectious disease epidemics (World Health Organization, Geneva). Available at http://apps.who.int/iris/bitstream/handle/10665/256246/WHO-OHE-PED-2016-2-eng.pdf. Accessed January 25, 2018.
6. Chretien JP, et al. (2015) Advancing epidemic prediction and forecasting: A new US government initiative. Online J Public Health Informat 7:e13.
7. Lipstich M, Finelli L, Hefferna RT, Leung GM, Redd SC (2009) H1N1 Surveillance group. Improving the evidence base for decision making during a pandemic: The example of 2009 influenza A/H1N1. Biosurce Bioterror Bioddefense Strategy Pract Sci 9:89–115.
8. Biggerstaff M, Velardi P, Vespignani A, Finelli L (2016) Results from the second year of a collaborative effort to forecast influenza seasons in the United States. Epidemics 24:26–33.
9. Viboud C, et al. (2018) The RAPIDP Ebola forecasting challenge: Synthesis and lessons learnt. Emerg Infect Dis 22:13–21.
10. WolfHaug J, Wallinga J, van Boven M (2017) Variation in the impact of immunity shapes influenza epidemics and the impact of vaccination. BMC Infect Dis 17:632.
11. WOlfHaug J, Wallinga J, van Boven M (2017) Variation in the impact of immunity shapes influenza epidemics and the impact of vaccination. BMC Infect Dis 17:632.
12. Anic Cl, Mubareka S, Steel J, Peter P (2007) Influenza virus transmission is dependent on relative humidity and temperature. PLoS Pathog 3:e151.
13. Simon C, Valleron A-J, Boelle P-Y, Antoine F, Ferguson NM (2008) Estimating the impact of school closure on influenza transmission from Sentinel data. Nature 452:750–754.
14. PHIpredict. (2018) Epidemiologic prediction initiative. Available at https://predict.phiresearchlab.org. Accessed January 25, 2018.
15. BIGGERSTAFF M (2018) Systematic assessment of multiple routine and near-real-time indicators to classify the severity of influenza seasons and pandemics in the United States, 2003-04 through 2015–2016. Am J Epidemiol 187:1040–1050.
16. Pei S, Shaman J (2017) Counterfactually structural errors in ensemble forecast of influenza outbreaks. PLoS One 13:e0199674.
17. Yang W, Karspeck A, Shaman J (2014) Comparison of filtering methods for the modeling and retrospective forecasting of influenza epidemics. PLoS Comput Biol 10:e1003583.