Epidemics 24 (2018) 26–33

Contents lists available at ScienceDirect

Epidemics

journal homepage: www.elsevier.com/locate/epidemics

Results from the second year of a collaborative effort to forecast influenza seasons in the United States

Matthew Biggerstaff[a,*], Michael Johansson[b], David Alper[c], Logan C. Brooks[d], Prithwish Chakraborty[e], David C. Farrow[f], Sangwon Hyun[g], Sasikiran Kandula[h], Craig McGowan[a], Naren Ramakrishnan[e], Roni Rosenfeld[i], Jeffrey Shaman[h], Rob Tibshirani[j], Ryan J. Tibshirani[k], Alessandro Vespignani[l], Wan Yang[h], Qian Zhang[l], Carrie Reed[a]

[a] Epidemiology and Prevention Branch, Influenza Division, Centers for Disease Control and Prevention, Atlanta, GA, USA
[b] Dengue Branch, Division of Vector-Borne Diseases, Centers for Disease Control and Prevention, Atlanta, GA, USA
[c] Everyday Health, New York City, NY, USA
[d] Department of Computer Science, Carnegie Mellon University, Pittsburg, PA, USA
[e] Discovery Analytics Center, Department of Computer Science, Virginia Tech, Blacksburg, VA, USA
[f] Department of Computational Biology, Carnegie Mellon University, Pittsburg, PA, USA
[g] Deptartment of Statistics, Carnegie Mellon University, Pittsburg, PA, USA
[h] Department of Environmental Health Sciences, Mailman School of Public Health, Columbia University, New York, NY, USA
[i] Deptartment of Machine Learning, Department of Language Technologies, Department of Computational Biology, Department of Computer Science, Carnegie Mellon University, Pittsburg, PA, USA
[j] Department of Health Research and Policy, Department of Statistics, Stanford University, Stanford, CA, USA
[k] Deptartment of Statistics, Department of Machine Learning, Carnegie Mellon University, Pittsburg, PA, USA
[l] Northeastern University, Boston, MA, USA

ARTICLE INFO

Keywords:
Influenza
Epidemics
Forecasting
Prediction
Modeling

ABSTRACT

Accurate forecasts could enable more informed public health decisions. Since 2013, CDC has worked with external researchers to improve influenza forecasts by coordinating seasonal challenges for the United States and the 10 Health and Human Service Regions. Forecasted targets for the 2014–15 challenge were the onset week, peak week, and peak intensity of the season and the weekly percent of outpatient visits due to influenza-like illness (ILI) 1–4 weeks in advance. We used a logarithmic scoring rule to score the weekly forecasts, averaged the scores over an evaluation period, and then exponentiated the resulting logarithmic score. Poor forecasts had a score near 0, and perfect forecasts a score of 1.

Five teams submitted forecasts from seven different models. At the national level, the team scores for onset week ranged from < 0.01 to 0.41, peak week ranged from 0.08 to 0.49, and peak intensity ranged from < 0.01 to 0.17. The scores for predictions of ILI 1–4 weeks in advance ranged from 0.02–0.38 and was highest 1 week ahead. Forecast skill varied by HHS region.

Forecasts can predict epidemic characteristics that inform public health actions. CDC, state and local health officials, and researchers are working together to improve forecasts.

1. Introduction

Preparing for and responding to influenza epidemics and pandemics are critical functions of public health agencies. The Centers for Disease Control and Prevention (CDC) currently tracks influenza activity through a nationwide influenza surveillance system (Centers for Disease Control and Prevention, 2014a). Together with information on historic experiences, these data are used for situational awareness and assessing needs for the near future. However, these data lag behind real-time flu activity and give no direct insight on what might happen next. Accurate, timely, and reliable influenza forecasts could enable more informed public health and emergency response decisions during both influenza seasons and pandemics, including the development and use of pharmaceutical (e.g., vaccine and influenza antivirals) and non-

Abbreviations: CDC, centers for disease control and prevention; HHS, health and human service; ILI, influenza-like illness; MMWR, morbidity and mortality weekly report; ILINet, U.S. outpatient influenza-like illness surveillance network

* Corresponding author at: Centers for Disease Control and Prevention, 1600 Clifton Road NE MS A-32, Atlanta, GA 30333, USA.
E-mail addresses: zmo2@cdc.gov, mbiggerstaff@cdc.gov (M. Biggerstaff).

https://doi.org/10.1016/j.epidem.2018.02.003
Received 1 May 2017; Received in revised form 6 February 2018; Accepted 20 February 2018
Available online 24 February 2018

pharmaceutical (e.g., school closures and social distancing, travel restrictions) countermeasures, communication, deployment of Strategic National Stockpile assets (e.g., ventilators), and hospital resource management (e.g., inventory and staff management) (Chretien et al., 2014).

CDC's Influenza Division began working in 2013 to advance influenza forecasting efforts by engaging with members of the scientific community who were developing innovative methods to predict influenza activity (Brooks et al., 2015; Shaman et al., 2009; Shaman and Karspeck, 2012; Kandula et al., 2017; Tizzoni et al., 2012; Balcan et al., 2009; Nsoesie et al., 2014). This effort launched with the "Predict the Influenza Season Challenge," a contest which encouraged participants to predict the timing, peak, and intensity of the 2013–14 influenza season using social media data (e.g., Twitter, internet search data, web surveys, etc.) along with data from CDC's routine flu surveillance systems (Centers for Disease Control and Prevention, 2013). Eleven teams participated in the original CDC competition, and team members developed their own models to predict flu activity based on a variety of data sources (Biggerstaff et al., 2016). This challenge identified a number of research gaps limiting forecasting model development, evaluation, and adoption by decision-makers, including the need to develop standardized metrics to assess forecast accuracy and standardized ways to communicate forecasts and their uncertainty.

To address these gaps, CDC and original challenge participants worked together through a collaborative challenge to forecast the 2014–15 influenza season. The objectives of this challenge were to continue to improve the accuracy of influenza forecasts, develop standardized metrics to assess and communicate forecast accuracy and uncertainty, and to identify the types of decisions best aided by forecasts. Challenge participants were asked to forecast seasonal milestones (the onset, peak, and intensity) and short-term activity during the 2014–15 influenza season for the United States as a country and for each of the 10 Health and Human Services (HHS) regions. In this report, we present the results and lessons learned from the challenge.

2. Methods

Teams that participated in CDC's 2013–14 Predict the Influenza Season Challenge were invited to continue to work with CDC to provide forecasts for the 2014–15 influenza season in the United States. This group of teams and CDC collaboratively defined a set of forecast targets and established evaluation metrics to assess accuracy prior to the challenge. Participating groups then submitted weekly forecasts for the 2014–2015 influenza season beginning October 20, 2014, and ending May 25, 2015. Forecasting targets were selected to ensure they were feasible for forecasting models and provided information for public health decision making.

All forecasting targets were based on data from the U.S. Outpatient Influenza-like Illness (ILI) Surveillance Network (ILINet). ILINet provides accurate information on the timing and impact of influenza activity each season and consists of more than 2000 outpatient healthcare providers around the country who report data to CDC weekly on the number of patients with ILI and the total number of patients seen in their practices (Centers for Disease Control and Prevention, 2014a; Brammer et al., 2011). ILINet data are based on a Morbidity and Mortality Weekly Report (MMWR) surveillance week that starts on Sunday and ends on Saturday; data are reported online through CDC's FluView surveillance report the following Friday (or Monday if federal holidays delay publication) (Centers for Disease Control and Prevention, 2014b). Further information on ILINet is available elsewhere (Centers for Disease Control and Prevention, 2014a; Brammer et al., 2011). Teams could use any other data sources available to them, including digital (e.g., Twitter data, mining internet search term data, Internet-based surveys), meteorological, and traditional surveillance.

The minimum set of forecasts required of all participants were national-level forecasts of the onset week, peak week, and peak intensity of the influenza season (collectively referred to in the paper as seasonal targets), and short-term forecasts of the weekly percentage of outpatient ILINet visits due to ILI one, two, three, and four weeks after the week most recently reported by ILINet in FluView (collectively referred to in the paper as short-term targets). Participants also had the option of submitting forecasts of the same targets for each of the 10 HHS regions. We defined the onset of the season as the first surveillance week in ILINet where the ILINet percentage was at or above the baseline value (which is developed by calculating the mean percentage of patient visits for ILI during non-influenza weeks for the previous three seasons and adding two standard deviations (Centers for Disease Control and Prevention, 2014a) and remained there for at least two additional weeks. We defined the peak week of the season as the surveillance week that the ILINet percentage was the highest; if more than one week achieved the highest value, all such weeks were considered peak weeks. We defined the peak value as the highest numeric value that the ILINet percentage reached (Centers for Disease Control and Prevention, 2014b).

Each forecast included a point estimate and a probability distribution within pre-defined bins for each target. For onset and peak weeks, each bin represented a single week (e.g., week 1, week 2). For start week, an additional bin was used for the probability that the onset week definition would not be met during the influenza season. For the peak percentage of outpatient visits due to ILI and the weekly percentage of ILI one to four weeks in advance, 11 bins were used; 10 bins represented semi-open 1% intervals (e.g., $3\% <= \text{ILI peak value} < 4.0\%$) from 0% to 10% while the final bin represented all values greater than or equal to 10%. Teams were also required to submit a narrative describing the methodology of the forecasting model. The forecasting methodology could be changed during the course of the season if an updated narrative describing the changes was provided; no team indicated that they changed their methodology during the 2014–15 season.

We used the logarithmic scoring rule to measure the accuracy of the probability distribution of a forecast (Gneiting and Raftery, 2007; Rosenfeld et al., 2018). If $\mathbf{p}$ is the set of probabilities across all bins for a given forecast, and p_i is the probability assigned to the observed outcome, i, the logarithmic score is $S(\mathbf{p},i) = \ln(p_i)$. For example, a forecast that assigned a probability of 0.6 to the correct influenza season onset week would receive a score of $\ln(0.6) = -0.51$. Undefined natural logs (which occur when the probability assigned to the observed outcomes was 0), missing forecasts, and forecasts that summed to probabilities less than 0.9 or greater than 1.1 were assigned a value of -10. Logarithmic scores were averaged across different combinations of seasonal and short-term targets, geographic locations, and time periods. For the seasonal targets, the evaluation period was chosen post hoc to represent periods when the forecasts would be most useful and began with the first forecast submission on October 20, 2014, while the end of the evaluation period varied by seasonal target. The evaluation period end for the onset target was the forecast received after the week in which peak occurred in the final ILINet data, and the evaluation period end for the peak week and peak percent targets was the forecast received after the final week ILINet was above baseline (Table 1 and Supplemental Tables 1–10). For the short-term forecasts, time periods were chosen to represent forecasts that were received during the weeks that ILINet was above baseline (Table 1 and Supplemental Tables 1–10). Evaluation results for national- and regional-level targets using forecasts from the entire forecast period (October 20, 2014 to May 25, 2015) are found in Supplemental Table 11. Because ILINet data for past weeks may change as more reports are received, we used the ILINet data weighted on the basis of state population reported on week 34 of 2015 (the week ending August 29) for forecast evaluation.

To aid in interpretation, we exponentiated the mean log score to indicate forecast skill on a 0–1 scale. Perfect forecasts (i.e. forecasted probability of 1.0 for the observed outcome across all forecasts) have a log score of 0 and a forecast skill of 1. For forecasts with low

Table 1
Onset week, peak week, peak percent, and the forecast evaluation period, as calculated from ILINet during the 2014–15 influenza season, United States.

Baseline value	2.0%
Onset week	WK 47 (week ending Nov. 22)
Publish date	December 1, 2014
Peak week	WK 52 (week ending December 27)
Peak percentage	5.99
Publish date	January 5, 2015
Last week above baseline	WK 13 (week ending April 4)
Publish date	April 10, 2015
Evaluation period for onset forecasts	October 20, 2014–January 5, 2015
Evaluation period for peak week and percent	October 20, 2014–April 13, 2015
Evaluation period for 4-wk ahead forecasts (in season)	December 1, 2014–April 13, 2015

probabilities for the observed outcome, the log score is a low negative number and forecast skill is approximately 0. For example, an average log score of −10 gives a skill of approximately 0.00005.

For comparison purposes, we created a historical average forecast. For peak week, the peak percentage, and the short-term targets, we used ILINet data from the 1997-98 influenza season through the 2013-14 influenza season (excluding the 2009 pandemic) while for the onset week target, we used ILINet data from the 2007–08 influenza season through the 2014-15 flu season (excluding the 2009 pandemic). For each MMWR week that would be predicted by the model, a Gaussian kernel density estimate using bandwidths estimated by the Sheather-Jones method (Sheather and Jones, 1991) was fit to that week's previous observed ILINet values. Approximate probabilities for observing each of the prediction bins were calculated by integrating the kernel density using the bin boundaries, and the point estimate was generated using the median of the estimated distribution. For the onset week target, the probability of no start week (i.e. ILINet never went above baseline for three or more weeks in a season) was calculated as the percentage of seasons in which the criteria for season onset was not met. A Gaussian kernel density estimate was fit to observed onset weeks and probabilistic estimates for each week were calculated as described above and then normalized to reflect the previously calculated probability of no start week. These methods were repeated for each HHS region as well as the United States as a whole.

This study did not involve human participants, and institutional review board approval was not required.

3. Results

Five teams predicted three seasonal targets and four short-term targets at 32 weekly intervals over the influenza season. Teams used Google Flu Trends (n = 4 teams), Twitter (n = 2), and weather data (n = 2) to inform their forecasting models (Table 2). Four (57%) forecasts employed statistical methods, and three (43%) employed mechanistic models that incorporated compartmental modeling (e.g., Susceptible-Exposed-Infected-Recovered [SEIR] models) (Table 2). Four out of 5 teams made forecasts for the HHS regions (Table 2). One team provided the results of three separate forecast models for the United States and the 10 HHS Regions. A total of 7 forecasts for the United States and 6 forecasts for the 10 HHS regions were evaluated.

3.1. National level forecasts

Different forecast models achieved the best average skill for each national-level seasonal target: Forecast E had the highest average forecast skill for season onset, Forecast B had the highest forecast skill for peak week and the highest forecast skill for the seasonal targets combined, and Forecast A had the highest skill for peak ILINet percent. In contrast, for the short-term targets, Forecast E had the highest forecast skill for ILINet forecasts 1–4 weeks in advance and the highest

Table 2
Characteristics of nine forecasts that competed in the Predict the 2014–15 Influenza Season Challenge.

Forecast	Data source	Model type	Regional forecast*	Brief description
A	Google Flu Trends, Healthmap, Wikipedia, weather data, ILINet	Mechanistic**	No	Susceptible-Infected-Recovered-Susceptible (SIRS) using Bayesian data assimilation to drive a prediction ensemble
B	ILINet, specific humidity data	Mechanistic	Yes	SIR, SIRS, Susceptible-Exposed-Infectious-Recovered (SEIR), SEIRS models combined with three different ensemble filter algorithms (12 model-filter combinations)
C	ILINet	Statistical***	Yes	Spline-basis regression to match current observations with typical futures. Bootstrap to generate forecast distribution.
D	Google Flu Trends, Twitter, ILINet	Statistical	Yes	Extrapolation of correlation between Google Flu Trends and Twitter with ILINet data
E	ILINet, crowd-sourced forecasts	Statistical	Yes	Crowdsourcing to collect many different influenza forecasts and generate an aggregate forecast
F	Google Flu Trends, ILINet	Statistical	Yes	An empirical Bayes model of ILI trajectories, with iid Gaussian noise and importance sampling
G	Twitter, ILINet data	Mechanistic	Yes	Combines Twitter data, historical ILINet data and an epidemic stochastic generative model

* Yes denotes forecast for ≥1 region (for all weeks).
** Includes models that incorporate compartmental modeling like Susceptible-Exposed-Infected-Recovered [SEIR] models.
*** Includes models like time series analysis and generalized linear models.

Table 3a
The average forecast skill score[a] over the evaluation period for onset week, peak week, peak percent, the ILINet value 1–4 week(s) ahead, by forecast team, United States.

	Onset	Peak Week	Peak%	Seasonal target (ST) average[b]	1 week	2 week	3 week	4 week	Short-term target (STT) average[c]
Forecast A	< 0.01	0.20	0.17	0.02	0.14	0.13	0.13	0.13	0.13
Forecast B[d]	0.30	0.49	0.07	0.21	0.07	0.04	0.05	0.04	0.05
Forecast C	0.27	0.08	< 0.01	0.04	0.17	0.10	0.05	0.13	0.10
Forecast D	0.01	0.48	0.09	0.07	0.14	0.17	0.12	0.13	0.14
Forecast E[b]	0.41	0.31	< 0.01	0.08	0.43	0.36	0.37	0.35	0.38
Forecast F	0.15	0.32	0.06	0.14	0.35	0.27	0.24	0.33	0.29
Forecast G	0.03	0.18	< 0.01	0.01	0.03	0.02	0.01	0.01	0.02
Best score	0.41	0.49	0.17	0.21	0.43	0.36	0.37	0.35	0.38
Historic avg.	0.07	0.12	0.14	0.12	0.12	0.14	0.15	0.18	0.15
Avg. score	0.04	0.25	0.02	0.06	0.14	0.11	0.08	0.10	0.11

[a] Skill scores range from 0 to 1 with 1 indicating a perfect forecast.
[b] Seasonal-targets (ST) average (average of the skill score for onset week, peak week, and peak percent forecasts).
[c] Short-term-targets (STT) average (average of the skill score for 1–4 week ahead ILINet forecasts).
[d] Winner of the 2014–15 forecasting challenge.

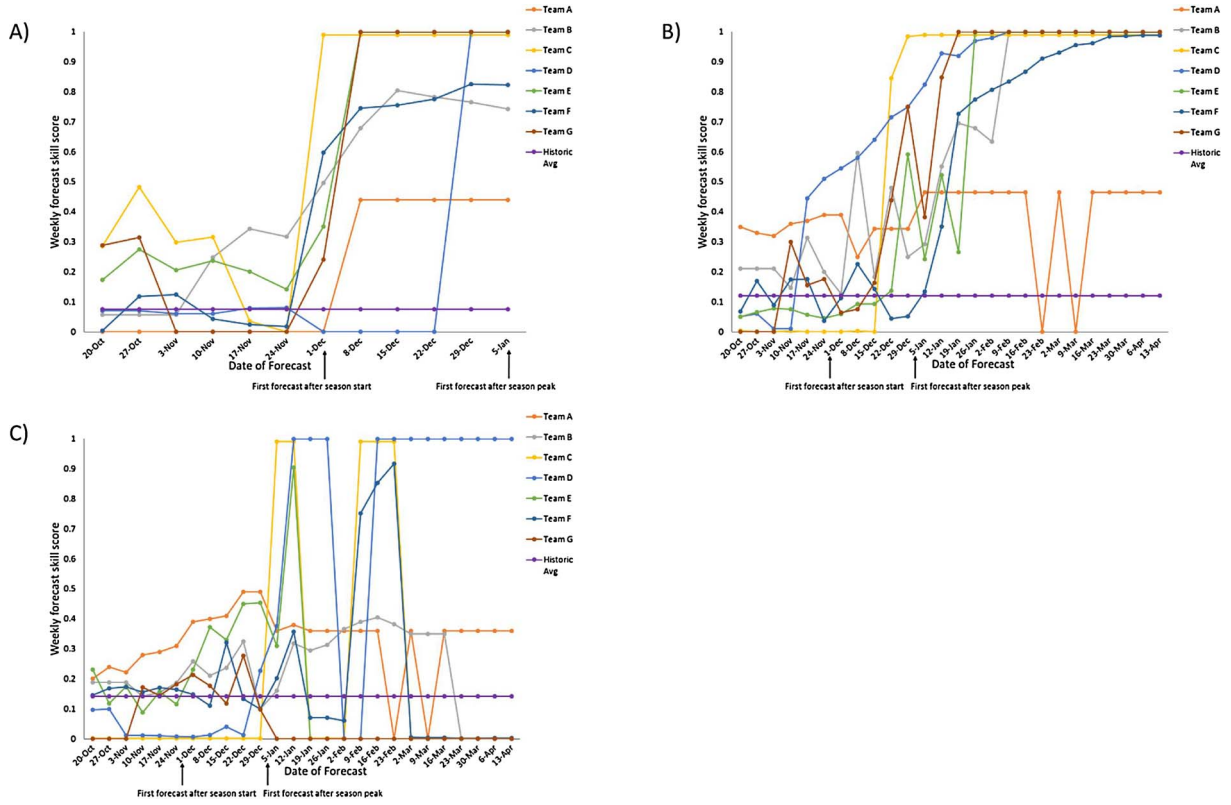

Fig. 1. Weekly forecast skill score[a] for A) onset week, B) peak week, and C) peak percent, as calculated from ILINet data during the 2014–15 influenza season, by the date of forecast, for the evaluation period, United States (n = 7 forecasts).
[a]A forecast skill score of 0 indicates that the forecast assigned a 0% chance of occurrence to the correct outcome while a forecast confidence of 1 indicates that the forecast assigned a 100% chance of occurrence.

short-term forecast skill combined (Table 3a). When compared to the historic average model, four models had higher skill scores for season onset forecasts, six for season peak, one for season intensity, and two for the seasonal milestones combined while five models had higher skill scores for 1-week ahead forecasts, three for 2-week ahead, and two each for 3- and 4-week ahead forecasts and the short-term targets combined. Forecasts with the best skill scores outperformed the historical average model for all national-level forecast targets (Table 3a). The weekly forecast skill score for seasonal targets was generally low for all forecasts in October, November, and December. Large increases in confidence for several season onset forecasts occurred after the publication of the first FluView showing ILINet above the national baseline and for peak week forecasts after the publication of the first FluView showing ILINet decreasing after reaching 6.0% (Fig. 1).

Other forecasts (e.g. Teams B and E for season onset and Teams B and D for peak week) more consistently placed a high confidence on the correct onset or peak week prior to the publication of these data. The skill scores for predictions of ILI 1–4 weeks in advance and the accuracy of point forecasts were highest 1 week ahead and declined for the 2–4 weeks ahead forecasts (Table 3a; Figs. 2 and 3). Short-term forecasts had higher skill scores outside the influenza season than during the influenza season (Fig. 2).

3.2. Regional level forecasts

Average forecast skill scores for the seasonal and short-term targets for the 10 HHS regions are presented in Table 3b. Forecast score varied by region and by forecast model. Forecast B had the highest average

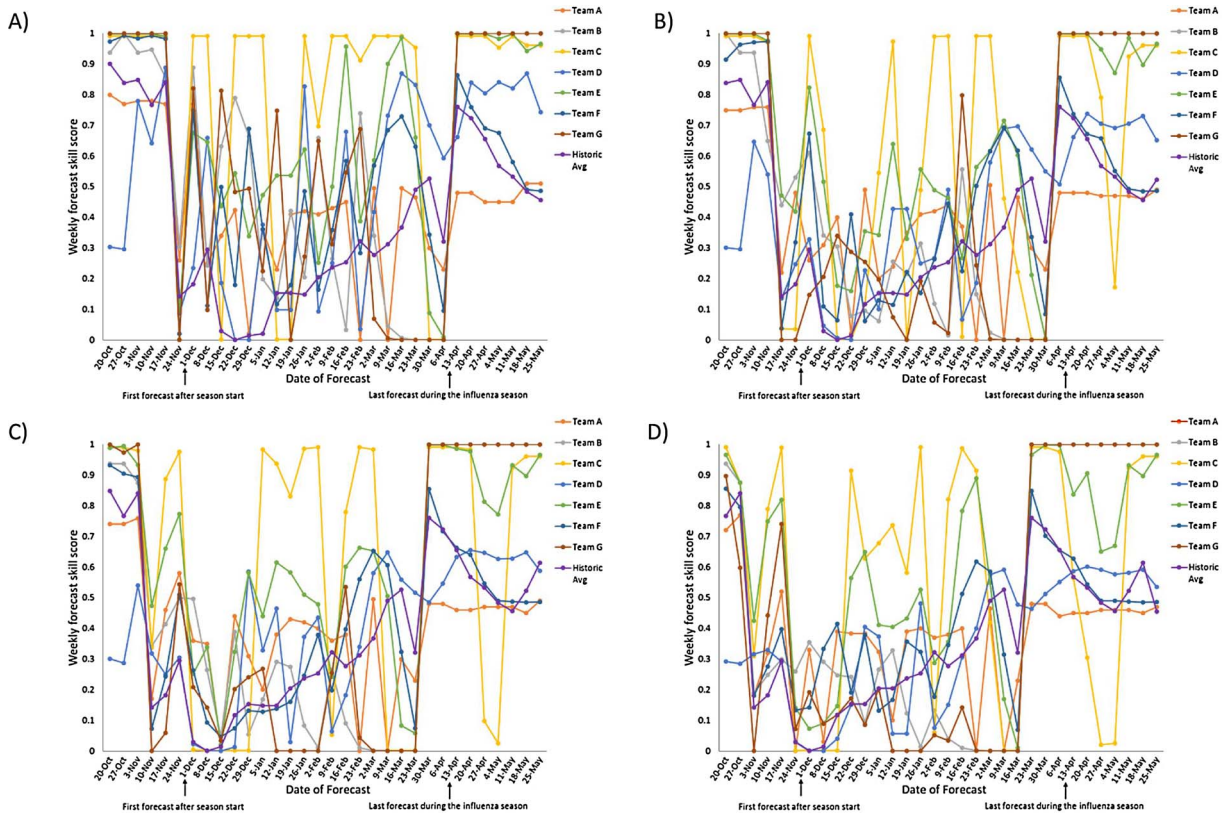

Fig. 2. Weekly forecast skill score[a] for A) ILINet values 1 week ahead, B) ILINet values 2 weeks ahead, C) ILINet values 3 weeks ahead, and D) ILINet values 4 weeks ahead, as calculated from ILINet data during the 2014–15 influenza season, by the date of forecast, for the entire forecast period, United States (n = 7 forecasts).
[a]A forecast skill score of 0 indicates that the forecast assigned a 0% chance of occurrence to the correct outcome while a forecast confidence of 1 indicates that the forecast assigned a 100% chance of occurrence.

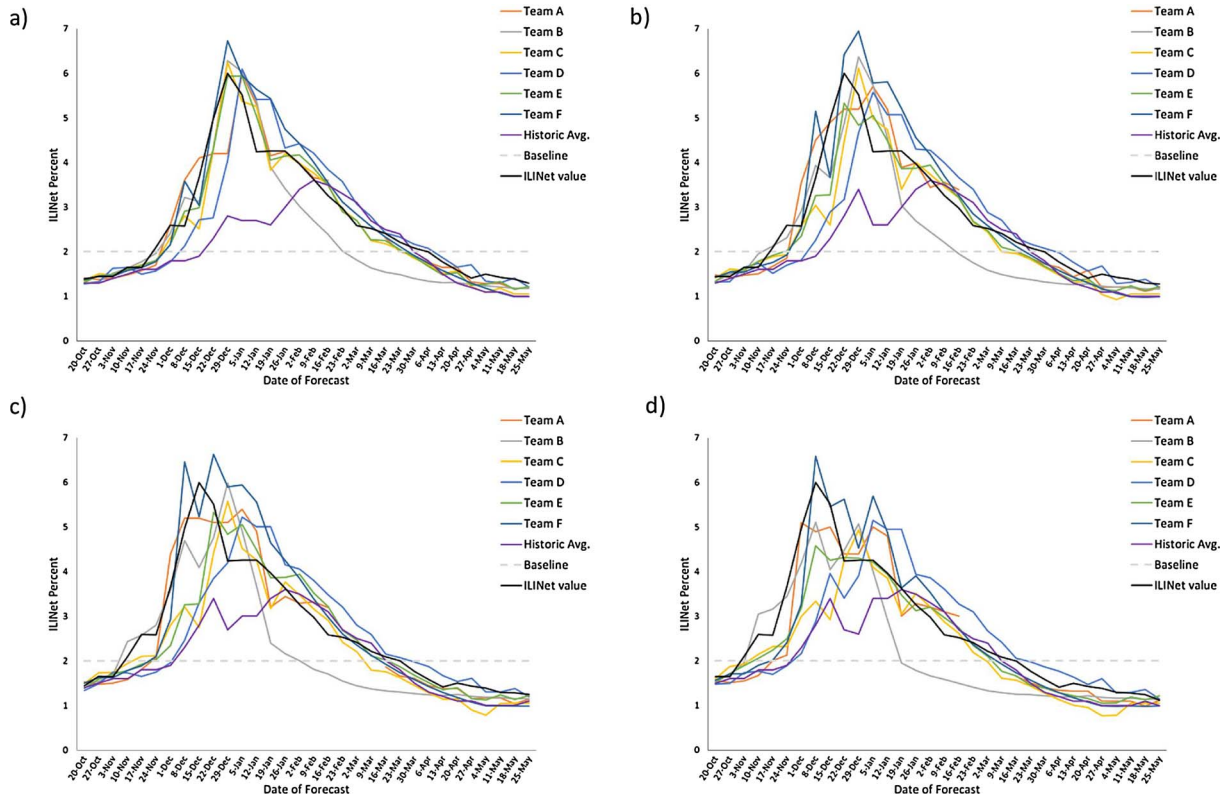

Fig. 3. 1-week- ahead, 2-week- ahead, 3-week- ahead, and 4-week-ahead point forecasts for the percent of visits due to influenza-like illness reported through the U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet) and the actual ILINet value (in black).

Table 3b
The average forecast skill score[a] over the evaluation period for seasonal-target forecasts (onset week, peak week, peak percent) and short-term target forecasts (ILINet value 1–4 week(s)) ahead, by Health and Human Service Region and forecast team.

	Region 1		Region 2		Region 3		Region 4		Region 5		Region 6		Region 7		Region 8		Region 9		Region 10		Average Team Score	
	ST[b]	STT[c]	ST	STT	ST	STT	ST	STT	ST	STT	ST	STT	ST	STT	ST	STT	ST	STT	ST	STT	ST	STT
Forecast B[d]	0.05	0.15	0.09	0.13	0.18	0.08	0.36	0.07	0.33	0.24	0.20	< 0.01	0.34	< 0.01	0.26	0.16	0.14	0.21	0.20	0.28	0.19	0.06
Forecast C	0.02	0.16	0.04	0.05	0.02	0.02	0.08	0.10	0.05	0.08	0.05	0.02	0.06	0.03	0.04	0.10	0.07	0.14	0.16	0.07	0.05	0.06
Forecast D	0.01	0.32	0.01	0.12	< 0.01	0.02	0.05	0.05	0.01	0.02	0.02	0.06	0.01	0.08	0.01	0.24	0.16	0.29	0.06	0.26	0.02	0.10
Forecast E[b]	0.04	0.42	0.10	0.39	0.06	0.19	0.18	0.29	0.13	0.22	0.26	0.17	0.20	0.28	0.26	0.50	0.16	0.38	0.30	0.31	0.14	0.31
Forecast F	0.09	0.30	0.14	0.17	0.09	0.12	0.25	0.27	0.19	0.19	0.20	0.11	0.27	0.17	0.17	0.23	0.13	0.36	0.24	0.29	0.16	0.21
Forecast G	< 0.01	< 0.01	0.02	< 0.01	0.01	< 0.01	0.16	0.02	0.06	0.13	0.04	< 0.01	0.08	0.01	0.10	0.03	0.01	< 0.01	0.02	0.01	0.02	0.01
Best score	0.09	0.42	0.14	0.39	0.18	0.19	0.36	0.29	0.33	0.24	0.26	0.17	0.34	0.28	0.26	0.50	0.16	0.38	0.30	0.31	0.19	0.31
Historic avg.	0.11	0.26	0.07	0.19	0.07	0.09	0.10	0.18	0.09	0.05	< 0.01	0.08	0.06	0.12	0.11	0.24	0.10	0.23	0.08	0.21	0.06	0.13
Average score	0.02	0.10	0.04	0.08	0.02	0.04	0.15	0.08	0.07	0.11	0.09	0.02	0.10	0.04	0.09	0.15	0.08	0.12	0.12	0.12	0.07	0.08

[a] Skill scores range from 0 to 1 with 1 indicating a perfect forecast.
[b] Seasonal-targets (ST) average (average of the skill score for onset week, peak week, and peak percent forecasts).
[c] Short-term-targets (STT) average (average of the skill score for 1–4 week ahead ILINet forecasts).
[d] Winner of the 2014–15 forecasting challenge.

skill score for the seasonal targets in 5 HHS regions while Forecast E had the highest average score for the seasonal targets in 4 HHS regions and the highest average skill score for the short-term targets in 9 HHS regions. The highest average skill for the combined seasonal targets among the 10 HHS regions ranged from 0.09 in Region 1–0.36 in Region 4 while the highest average skill for the combined short-term targets ranged from 0.17 in Region 6–0.50 in Region 8 (Table 3b). The highest average skill score for milestone forecasts (0.19) and near-term forecasts (0.31) for the 10 HHS Regions were similar to the highest average skill score for milestone forecasts (0.21) and near-term forecasts (0.38) for the United States (Tables 3a and 3b). When compared to the historic average model and using the scores for the 10 HHS regions averaged together, three models had higher scores for the combined seasonal targets while two models had higher skill scores for the combined short-term targets (Table 3b).

4. Discussion

This challenge represents the second year of a CDC-coordinated effort to forecast influenza seasons in the United States. CDC and forecasting teams have continued to work together to ensure forecasting targets are feasible for constructing forecasting models and provide information for public health decision making. For the 2014–15 influenza season, the forecast targets were modified to better align them with public health needs and standardized metrics were added to quantitatively assess forecast accuracy. Differences in forecast accuracy were observed between models, with Forecast models B and E generally performing better than other forecasts and the historic average model. Nonetheless, the evaluation of forecast skill indicated that forecasting the different characteristics of an influenza epidemic accurately is challenging, even in the United States where researchers have access to robust traditional and non-traditional influenza surveillance data. Continued work to improve forecast accuracy and communication of accuracy and past forecast performance is needed to improve the contribution of forecasts to public health decision making.

One criterion for influenza forecasts to be useful to public health decision makers is that they must address key public health questions, and identifying the key targets for influenza forecasting was a major focus of the 2014–15 challenge. During the 2013–14 influenza season, the included targets were the onset and peak weeks, the peak intensity, and the duration of the season. For the 2014–15 challenge, we maintained the onset, peak week, and peak intensity targets because of their variation season-to-season and the public health actions that could be informed by these forecasts. Historically, U.S. influenza seasons have begun between November and January, have peaked between December and March, and have had peak ILINet values ranging from 2.4% to 7.7% (Centers for Disease Control and Prevention, 2014b; Appiah et al., 2015). Seasons that onset and peak earlier have less time for vaccination to occur before influenza activity starts, and higher ILINet values indicate seasons with an increased demand on health care services. Therefore, forecasts for the onset and peak week and the intensity can inform the timing of influenza vaccination and treatment efforts, communication around influenza season preparedness, and efforts to manage the influenza-associated healthcare surge (which can include staffing and inventory management). In the event of an influenza pandemic, forecasts with targets such as these could also be used to guide the timing and duration of non-pharmaceutical interventions, like school closures. We eliminated the season duration target because it did not inform a key public health question on its own.

The 2014–15 season was the first year that short-term targets were also included in the challenge. Because the most recent published ILINet data on Friday covers activity occurring during the previous MMWR week (e.g. the data published on November 6, 2016 covered activity that occurred October 25–31, 2016) and forecasts are received the following Monday, forecasts for the short-term targets cover activity that has occurred during the past week, the present week, and the next

two weeks in the future. They were included because they bridge the gap between surveillance data, which describe activity that has occurred in the past, and the seasonal targets, which describe one-time annual events that can be weeks to months away or already have passed. Therefore, short-term forecasts are an important tool for situational awareness because they provide the likelihood that influenza activity will be increasing, decreasing, or staying constant in the near future, which can help inform influenza-associated healthcare surge management and communication efforts.

Another factor identified to make forecasts more useful for decision making was to provide a measure of forecast confidence. During the 2013–14 influenza season challenge, we did not require forecasters to provide any metric of forecast confidence. Some teams provided no metric, others provided a qualitative metric (e.g. high, medium, low) or a confidence interval, while others provided a probability of the forecasted outcome occurring. The lack of a standardized way of communicating forecast uncertainty reduced the utility of the 2013–14 forecasts (Annon, 2013). Therefore, we collectively decided to standardize how forecast confidence was reported by having teams report forecasts as probability distributions in pre-defined bins across the range of potential target values. Much like a weather forecast provides the probability that rain will occur on a given day and allows a person to decide to carry an umbrella, the probability of an influenza outcome occurring communicates both the most likely outcome and the forecast confidence to decision makers and can inform calculations about the potential cost and benefit of a decision against the likelihood of the outcome occurring.

The probabilistic forecast distributions also allow for a quantitative evaluation of accuracy, which can be used to compare and communicate forecast performance. The forecast skill for the 2014–15 influenza season showed wide variation in the accuracy among the forecasts received, with Forecast B and E generally being the most accurate forecasts for both the United States and the 10 HHS regions. A major concern with forecasting is the use of an inaccurate forecast to inform a high consequence or high cost decision, which can have wide ranging consequences like wasted and misdirected resources, increases in morbidity or mortality, and the loss of credibility. In addition, because forecasts that assign little chance to the correct outcome occurring can be especially problematic for decision making, we utilized a skill score that averaged the weekly logarithmic scores before exponentiating instead of averaging the forecast probabilities before exponentiating. This approach penalizes teams more for forecasts that assign very low probabilities to the correct outcome. The goal of adding a standardized measure of forecast confidence and accuracy is to make decision makers as informed as possible when they use forecasts, and decision makers and other public health officials had access to the 2014–15 accuracy information during the 2015–16 and 2016–17 influenza seasons to understand which teams had previously provided the most accurate forecasts.

The use of a standardized metric for forecast accuracy also aids in the comparison of forecast performance among geographic regions and influenza seasons. For example, the highest average skill score for the short-term targets for HHS Regions 3 and 6 were below the highest average skill score for the remaining HHS regions (Tables 3a and 3b). These findings may indicate that certain forecasting targets and geographic regions may be more challenging to forecast and that future forecasts for these targets or geographic regions should be interpreted accordingly until accuracy data from more influenza seasons are available to confirm if this finding is consistent or due to chance. A standardized accuracy metric also provides a benchmark to measure year-to-year changes in forecast accuracy, which can inform broader discussions around the consistency of the most accurate forecasts from season to season, overall accuracy trends, model performance in influenza seasons with certain characteristics (e.g., late seasons vs. early seasons; high severity seasons vs. low severity seasons), the accuracy of forecasts for influenza compared with other infectious diseases, and identifying model and data characteristics associated with more accurate forecasts. These analyses are being conducted as part of future forecasting challenges.

Because of the success of the 2013–14 and 2014–15 forecasting challenges, CDC and forecasting teams continued to work together to forecast subsequent influenza seasons. In January of 2016, CDC released a provisional public website where seasonal influenza forecasts from multiple teams could be accessed in real time (Centers for Disease Control and Prevention, 2016). CDC and forecasting teams published forecasts for the 2016–17 influenza season, and CDC plans to make a number of improvements to the website, including the addition of accuracy information from previous seasons, interactive graphs, and additional data sources that may be helpful for forecasting efforts. CDC and forecasting teams will also continue to engage federal, state, and local health officials to understand how forecasts are being used to inform public health decisions and how forecast accuracy and uncertainty are understood and incorporated by decision makers, which may lead to further refinement of the targets or the presentation of the forecasts.

5. Conclusion

Preparing for and responding to influenza epidemics and pandemics are critical functions of public health. Infectious disease forecasting holds the potential to change the way that public health responds to epidemics and pandemics by providing accurate and timely forecasts, which could be used to make earlier and better decisions on pharmaceutical and non-pharmaceutical countermeasures, communication strategies, and hospital resource management. CDC has collaborated with a group of external researchers to identify actionable forecast targets and better measure forecast accuracy. The results of the 2014-15 influenza season challenge indicated that forecast accuracy varied by model and geographic location but that even in the best models, improvements in forecast accuracy were needed. Infectious disease forecasting is in its early years of development, and work continues between CDC and forecasting teams to fully incorporate it into public health decision making.

Disclaimer

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

Presentations

Selected findings reported in this manuscript were presented during Options IX for the Control of Influenza held August 24–28, 2016, in Chicago, Illinois.

Acknowledgements

The authors would like to thank participating state, territorial, and local health departments and healthcare providers that contribute data to the U.S. Outpatient ILI Surveillance Network.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.epidem.2018.02.003.

References

Annon, 2013. Announcement of Requirements and Registration for the Predict the Influenza Season Challenge, vol. 78. pp. 70303–70305 Fed. Reg.
Appiah, G.D., Blanton, L., D'Mello, T., Kniss, K., Smith, S., Mustaquim, D., et al., 2015.

Influenza activity – United States, 2014–15 season and composition of the 2015–16 influenza vaccine. MMWR. Morb. Mortal. Wkly. Rep. 64 (21), 583–590.
Balcan, D., Colizza, V., Singer, A.C., Chouaid, C., Hu, H., Goncalves, B., et al., 2009. Modeling the critical care demand and antibiotics resources needed during the Fall 2009 wave of influenza A(H1N1) pandemic. PLoS Curr. 1, Rrn1133.
Biggerstaff, M., Alper, D., Dredze, M., Fox, S., Fung, I.C., Hickmann, K.S., et al., 2016. Results from the centers for disease control and prevention's predict the 2013–2014 influenza season challenge. BMC Infect. Dis. 16, 357.
Brammer, L., Blanton, L., Epperson, S., Mustaquim, D., Bishop, A., Kniss, K., et al., 2011. Surveillance for influenza during the 2009 influenza a (H1N1) pandemic-United States, April 2009–March 2010. Clin. Infect. Dis. 52 (Suppl. 1), S27–35.
Brooks, L.C., Farrow, D.C., Hyun, S., Tibshirani, R.J., Rosenfeld, R., 2015. Flexible modeling of epidemics with an empirical bayes framework. PLoS Comput. Biol. 11 (8), e1004382.
Centers for Disease Control and Prevention, 2013. Announcement of Requirements and Registration for the Predict the Influenza Season Challenge, vol. 78. pp. 70303–70305 FR 70303.
Centers for Disease Control and Prevention, 2014a. Overview of Influenza Surveillance in the United States. http://www.cdc.gov/flu/weekly/overview.htm (Accessed 25, September 2014).
Centers for Disease Control and Prevention, 2014b. FluView Interactive. http://www.cdc.gov/flu/weekly/fluviewinteractive.htm (Accessed 08, November 2014).
Centers for Disease Control and Prevention, 2016. Flu Activity Forecasting Website Launched | News (Flu) | CDC. http://www.cdc.gov/flu/news/flu-forecast-website-launched.htm (Accessed 16, March 2016).
Chretien, J.P., George, D., Shaman, J., Chitale, R.A., McKenzie, F.E., 2014. Influenza forecasting in human populations: a scoping review. PLoS One 9 (4), e94130.
Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. J. Am. Stat. Assoc. 102 (477), 359–378.
Kandula, S., Yang, W., Shaman, J., 2017. Type- and subtype-specific influenza forecast. Am. J. Epidemiol. 1–8.
Nsoesie, E.O., Brownstein, J.S., Ramakrishnan, N., Marathe, M.V., 2014. A systematic review of studies on forecasting the dynamics of influenza outbreaks. Influenza Respir. Viruses 8 (3), 309–316.
Rosenfeld, R., Grefenstette, J.J., Burke, D., 2018. A Proposal for Standardized Evaluation of Epidemiological Models. http://delphi.midas.cs.cmu.edu/files/StandardizedEvaluation_Revised_12-11-09.pdf; 2012.
Shaman, J., Karspeck, A., 2012. Forecasting seasonal outbreaks of influenza. Proc. Natl. Acad. Sci. U. S. A. 109 (50), 20425–20430.
Shaman, J., Pitzer, V., Viboud, C., Lipsitch, M., Grenfell, B., 2009. Absolute humidity and the seasonal onset of influenza in the continental U.S. PLoS Curr. 2 RRN1138.
Sheather, S.J., Jones, M.C., 1991. A reliable data-based bandwidth selection method for kernel density estimation. J. Roy. Stat.Soc. 683–690 Series B (Methodological).
Tizzoni, M., Bajardi, P., Poletto, C., Ramasco, J.J., Balcan, D., Goncalves, B., et al., 2012. Real-time numerical forecast of global epidemic spreading: case study of 2009 A/H1N1pdm. BMC Med. 10, 165.