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

Matthew Biggerstaff,⁎ Michael Johansson, David Alper, Logan C. Brooks, Prithwish Chakraborty, David C. Farrow, Sangwon Hyung, Sasikiran Kandula, Craig McGowan, Naren Ramakrishnan, Roni Rosenfeld, Jeffrey Shaman, Rob Tibshirani, Ryan J. Tibshirani, Alessandro Vespignani, Wan Yang, Qian Zhang, Carrie Reed

⁎ Corresponding author at: Centers for Disease Control and Prevention, 1600 Clifton Road NE MS A-32, Atlanta, GA 30333, USA.
E-mail addresses: zm02@cdc.gov, mbiggerstaff@cdc.gov (M. Biggerstaff).

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-
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; Randula et al., 2017; Tizzoni et al., 2012; Balcan et al., 2009; Noesie 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 Surveillance Network (ILIINet). 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% < = 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 pi is the set of probabilities across all bins for a given forecast, and pi is the probability assigned to the observed outcome, i, the logarithmic score is S(pi, i) = ln(pi). 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
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 1**

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

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

---

**Table 2**

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

<table>
<thead>
<tr>
<th>Forecast</th>
<th>Data source</th>
<th>Model type</th>
<th>Region</th>
<th>Forecast</th>
<th>Regional forecast</th>
<th>Brief description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Google Flu Trends, Healthmap, Wikipedia, weather data, ILINet, specific humidity data</td>
<td>Mechanistic***</td>
<td>No</td>
<td>Yes</td>
<td>* Brief description</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>ILINet</td>
<td>Statistical</td>
<td>Yes</td>
<td>SIR, SIRS, SEIR, SEIRS models combined with three different ensemble filter algorithms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>ILINet, crowd-sourced forecasts</td>
<td>Statistical</td>
<td>Yes</td>
<td>Crowdsourcing to collect many different in-flu forecasts and generate an aggregate forecast</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>Google Flu Trends, ILINet</td>
<td>Mechanistic**</td>
<td>No</td>
<td>Yes</td>
<td>** Includes models that incorporate compartmental modeling like Susceptible-Exposed-Infected-Recovered (SEIR) models.</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>Google Flu Trends, ILINet</td>
<td>Mechanistic</td>
<td>Yes</td>
<td>Yes</td>
<td>*** Includes models like time series analysis and generalized linear models.</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>Twitter, ILINet data</td>
<td>Mechanistic</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>F</td>
<td>Statistical</td>
<td>Yes</td>
<td>Combines Twitter data, historical ILINet data and an epidemic stochastic generative model</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

* Yes denotes forecast for ≥ 1 region (for all weeks)
** Includes models like Susceptible-Exposed-Infected-Recovered (SEIR) models
*** Includes models like time series analysis and generalized linear models
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

---

Table 3a

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

<table>
<thead>
<tr>
<th>Forecast</th>
<th>Onset</th>
<th>Peak Week</th>
<th>Peak%</th>
<th>Seasonal target (ST) average</th>
<th>1 week</th>
<th>2 week</th>
<th>3 week</th>
<th>4 week</th>
<th>Short-term target (STT) average</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>&lt; 0.01</td>
<td>0.20</td>
<td>0.17</td>
<td>0.02</td>
<td>0.14</td>
<td>0.13</td>
<td>0.13</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>0.30</td>
<td>0.49</td>
<td>0.07</td>
<td>0.21</td>
<td>0.07</td>
<td>0.04</td>
<td>0.05</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0.27</td>
<td>0.08</td>
<td>&lt; 0.01</td>
<td>0.04</td>
<td>0.17</td>
<td>0.10</td>
<td>0.05</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>0.01</td>
<td>0.48</td>
<td>0.09</td>
<td>0.07</td>
<td>0.14</td>
<td>0.17</td>
<td>0.12</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>0.41</td>
<td>0.32</td>
<td>0.06</td>
<td>0.14</td>
<td>0.43</td>
<td>0.36</td>
<td>0.37</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>0.03</td>
<td>0.18</td>
<td>&lt; 0.01</td>
<td>0.01</td>
<td>0.03</td>
<td>0.02</td>
<td>0.01</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>Best</td>
<td>0.41</td>
<td>0.49</td>
<td>0.17</td>
<td>0.21</td>
<td>0.43</td>
<td>0.36</td>
<td>0.37</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>Historic</td>
<td>0.07</td>
<td>0.12</td>
<td>0.14</td>
<td>0.12</td>
<td>0.12</td>
<td>0.14</td>
<td>0.15</td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td>Avg.</td>
<td>0.04</td>
<td>0.25</td>
<td>0.02</td>
<td>0.06</td>
<td>0.14</td>
<td>0.11</td>
<td>0.18</td>
<td>0.15</td>
<td></td>
</tr>
</tbody>
</table>

---

*Skill scores range from 0 to 1 with 1 indicating a perfect forecast.

Seasonal-targets (ST) average (average of the skill score for onset week, peak week, and peak percent forecasts).

Short-term-targets (STT) average (average of the skill score for 1–4 week ahead ILINet forecasts).

Fig. 2. Weekly forecast skill scorea 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 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 over the evaluation period for seasonal-target forecasts (onset week, peak week, peak percent) and short-term target forecasts (ILI Net value, 1-4 week(s)) ahead, by Health and Human Service Region and forecast team.

<table>
<thead>
<tr>
<th>Region</th>
<th>Region 1</th>
<th>Region 2</th>
<th>Region 3</th>
<th>Region 4</th>
<th>Region 5</th>
<th>Region 6</th>
<th>Region 7</th>
<th>Region 8</th>
<th>Region 9</th>
<th>Region 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>STb</td>
<td>0.02</td>
<td>0.16</td>
<td>0.04</td>
<td>0.05</td>
<td>0.02</td>
<td>0.02</td>
<td>0.08</td>
<td>0.10</td>
<td>0.05</td>
<td>0.08</td>
</tr>
<tr>
<td>STTc</td>
<td>0.01</td>
<td>0.32</td>
<td>0.01</td>
<td>0.12</td>
<td>&lt; 0.01</td>
<td>0.02</td>
<td>0.05</td>
<td>0.05</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td>Forecast C</td>
<td>0.04</td>
<td>0.38</td>
<td>0.10</td>
<td>0.39</td>
<td>0.06</td>
<td>0.19</td>
<td>0.18</td>
<td>0.29</td>
<td>0.13</td>
<td>0.36</td>
</tr>
<tr>
<td>Forecast D</td>
<td>0.03</td>
<td>0.50</td>
<td>0.08</td>
<td>0.74</td>
<td>0.01</td>
<td>0.09</td>
<td>0.26</td>
<td>0.50</td>
<td>0.16</td>
<td>0.38</td>
</tr>
<tr>
<td>Forecast E</td>
<td>0.02</td>
<td>0.50</td>
<td>0.08</td>
<td>0.74</td>
<td>0.01</td>
<td>0.09</td>
<td>0.26</td>
<td>0.50</td>
<td>0.16</td>
<td>0.38</td>
</tr>
<tr>
<td>Average Team Score</td>
<td>0.05</td>
<td>0.15</td>
<td>0.08</td>
<td>0.18</td>
<td>0.09</td>
<td>0.13</td>
<td>0.26</td>
<td>0.33</td>
<td>0.14</td>
<td>0.20</td>
</tr>
</tbody>
</table>

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).


e Combined short-term targets (Table 3b).
two weeks in the future. They were included because they bridge the
gap between surveillance data, which describe activity that has oc-
curred 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 si-
tuational 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 fore-
casted outcome occurring. The lack of a standardized way of commu-
nicating forecast uncertainty reduced the utility of the 2013–14 fore-
casts (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 po-
tential target values. Much like a weather forecast provides the prob-
ability 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 confi-
dence to decision makers and can inform calculations about the po-
tential cost and benefit of a decision against the likelihood of the out-
come occurring.

The probabilistic forecast distributions also allow for a quantitative
evaluation of accuracy, which can be used to compare and commu-
nicate forecast performance. The forecast skill for the 2014–15 influ-
enza season showed wide variation in the accuracy among the forecast
receivers, 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 in-
stead 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 geo-
graphic 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 in-
fluenza 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 accu-
rate 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 re-
leased 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 ac-
curacy information from previous seasons, interactive graphs, and ad-
ditional 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 un-
certainty 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 pharma-
ceutical 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, im-
provements in forecast accuracy were needed. Infectious disease fore-
casting is in its early years of development, and work continues be-
tween 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

References

Annon, 2013. Announcement of Requirements and Registration for the Predict the
M. Biggerstaff et al.
Epidemics 24 (2018) 26–33

32


