Epidemics 27 (2019) 41–51

Contents lists available at ScienceDirect

Epidemics

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

Near-term forecasts of influenza-like illness
An evaluation of autoregressive time series approaches

Sasikiran Kandula[*], Jeffrey Shaman

Department of Environmental Health Sciences, Columbia University, New York, NY, United States

ARTICLE INFO

Keywords:
Influenza-like illness
Forecasts
Near-term incidence
Auto-regressive models
Neural networks

ABSTRACT

Seasonal influenza in the United States is estimated to cause 9–35 million illnesses annually, with resultant economic burden amounting to $47-$150 billion. Reliable real-time forecasts of influenza can help public health agencies better manage these outbreaks. Here, we investigate the feasibility of three autoregressive methods for near-term forecasts: an Autoregressive Integrated Moving Average (ARIMA) model with time-varying order; an ARIMA model fit to seasonally adjusted incidence rates (ARIMA-STL); and a feed-forward autoregressive artificial neural network with a single hidden layer (AR-NN). We generated retrospective forecasts for influenza incidence one to four weeks in the future at US National and 10 regions in the US during 5 influenza seasons. We compared the relative accuracy of the point and probabilistic forecasts of the three models with respect to each other and in relation to two large external validation sets that each comprise at least 20 other models.

Both the probabilistic and point forecasts of AR-NN were found to be more accurate than those of the other two models overall. An additional sub-analysis found that the three models benefitted considerably from the use of search trends based 'nowcast' as a proxy for surveillance data, and these three models with use of nowcasts were found to be the highest ranked models in both validation datasets. When the nowcasts were withheld, the three models remained competitive relative to models in the validation sets. The difference in accuracy among the three models, and relative to models of the validation sets, was found to be largely statistically significant.

Our results suggest that autoregressive models even when not equipped to capture transmission dynamics can provide reasonably accurate near-term forecasts for influenza. Existing support in open-source libraries make them suitable non-naïve baselines for model comparison studies and for operational forecasts in resource constrained settings where more sophisticated methods may not be feasible.

1. Introduction

Seasonal influenza in the United States is estimated to cause 9–35 million illnesses annually. Significant variations in infection rates occur across seasons due to several factors including differences in population susceptibility to and the contagiousness of circulating strains, and vaccination efficacy and uptake rates (Rolfes et al., 2016). While most influenza-related illnesses are not severe enough to require medical attention, 140–700 thousand hospitalizations and thousands of deaths result from influenza-like illness (ILI) every year. It is also estimated that the annual economic burden from seasonal influenza, either from direct medical costs or loss of earnings and life, amounts to $47-$150 billion (Molinari et al., 2007).

The Centers for Disease Control and Prevention (CDC) has built robust surveillance systems to collect and disseminate near real-time information on ILI outbreaks in the United States. While the value of these observations is undeniable, systems that can reliably forecast the future path of an outbreak can provide additional help to public health agencies, health systems and practitioners as they plan and co-ordinate disease control strategies. Several statistical (Brooks et al., 2015; Farrow et al., 2017; Ray et al., 2017; Viboud et al., 2003; Wang et al., 2015), population-(Osthus et al., 2017; Shaman and Karspeck, 2012) and individual-level (Balcan et al., 2009; Hyder et al., 2013) mechanistic models for forecasting influenza have been developed and are in use operationally (see (Chretien et al., 2014; Nsoesie et al., 2014) for detailed reviews of these methods).

Beginning with the 2013/14 influenza season, the CDC Influenza Division through the Epidemic Prediction Initiative (EPI) has solicited real-time ILI forecasts from modelers (Biggerstaff et al., 2016, 2018). These yearly challenges help systematize model comparison by defining targets, evaluation metrics and submission templates. To participate, teams are required to submit weekly forecasts in real-time for three

[*] Corresponding author at: Dept. of Environmental Health Sciences, 722 West 168th Street, 11th Floor, Columbia University, New York, NY, 10032, United States.
E-mail address: sk3542@cumc.columbia.edu (S. Kandula).

https://doi.org/10.1016/j.epidem.2019.01.002
Received 15 August 2018; Received in revised form 15 December 2018; Accepted 16 January 2019
Available online 17 January 2019

seasonal targets – peak incidence, week of peak incidence and week of onset – and four near-term targets—weekly incidence one to four weeks from the week of forecast—at US National and 10 Health and Human Services (HHS) designated regions (U.S. Department of Health and Human Services Regional Offices, 2019). The near-term targets, which are the focus of this study, can inform decision-making related to school closings, planning for additional staffing/supplies at hospitals, and ramping up public health messaging.

Real-time, or near real-time, observations are critical for the generation of real-time forecasts. The primary data source for ILI forecasts in the US is provider-reported outpatient ILI visit rates collected through the ILINet (Centers for Disease Control and Prevention, 2018a). Several methods for supplementing these surveillance data with alternate estimates of ILI inferred from public non-surveillance proxies have also been proposed (Wang et al., 2015; Farrow, 2016; Kandula et al., 2017; Santillana et al., 2016, 2015; Lampos et al., 2015; Paul et al., 2014; Yang et al., 2015). In addition to capturing information that traditional surveillance systems are not designed to capture, these approaches can more directly address delays in ILINet reporting and dissemination. Some of these alternative estimates are generated using autoregression, a common time series modelling approach in which the response variable is modeled wholly or partly as a linear combination of past values of the response variable. Indeed, reasonably accurate estimates of current ILI have been developed using a combination a of autoregressive terms and real time data from Google search trends (Kandula et al., 2017; Lampos et al., 2015; Yang et al., 2015), twitter messages (Paul et al., 2014), or a combination of sources (Wang et al., 2015).

Some of these methods extend ILI forecast further into the future: Paul et al (Paul et al., 2014) used an autoregressive model with three lag variables to forecast incidence up to 10 weeks ahead; (Ray et al., 2017; Ray and Reich (2018)) implemented a fixed order Seasonal Autoregressive Integrated Moving Average (SARIMA) model to forecast incidence for all remaining weeks of the season; and, Wang et al(Wang et al., 2015) proposed multiple dynamic autoregressive models with exogenous variables for forecasting both seasonal and short-term targets. While the SARIMA model is better equipped to capture the known seasonality of influenza, the use of a fixed order for the duration of the season by the first two models potentially constrains them, as different phases of the season have different characteristics and hence may require different optimal orders. The dynamic models studied by Wang et al are more adaptive but are reliant on several sources, and their performance without these sources and in relation to non-autoregressive models has not been reported.

In this study, we propose three autoregressive methods for near-term ILI forecast (1–4 weeks into the future) that share some of the techniques of these earlier models and try to address their deficiencies. In limiting to these approaches, our intention is to explore the utility of simple, easy-to-implement forecast methods that are supported in standard statistical libraries and hence lend themselves to operational deployment in diverse operational settings. Furthermore, the chosen methods could serve as non-naïve baselines for comparing near-term ILI forecasts from other competing more complex approaches.

The first method is a simple seasonal ARIMA model of varying order; the second method decomposes the time series before fitting a dynamic ARIMA model; and the last attempts autoregression with an artificial neural network. Using these methods, we generated retrospective near-term forecasts during five seasons at the US national and 10 HHS regional levels in the US, and report the accuracy of these methods relative to each other and against two large validation sets that each include forecasts from more than 20 diverse models. Our results suggest that the methods proposed here are reasonably accurate in forecasting near-term ILI incidence and can match or outperform most of the methods in the two validation sets.

2. Materials and methods

In this section, we describe the two data sources used to generate retrospective forecasts. One data source is a CDC surveillance system while the other is based on a method we implemented to estimate ILI from search trend activity. This is followed by a description of the three autoregressive methods and implementation details pertaining to how they were used to generate point and probabilistic forecasts for near-term targets. The section ends with information on two validation sets and the evaluation measures used to compare the forecasts.

2.1. Data sources

2.1.1. U.S. Outpatient influenza-like illness surveillance network (ILINet)

Through the ILINet system, the CDC collects data from approximately 2800 healthcare providers on outpatient visits for ILI, which is defined as fever (temperature greater than 100°F) co-occurring with cough and/or sore throat. Providers voluntarily submit to the system weekly counts of patients seen for ILI and for all causes. These count data are aggregated to US national, HHS regional- and state-levels and are used to calculate percentage of outpatient visits that are for ILI, often referred to as ILI rates. Both population-weighted and unweighted estimates of aggregated ILI rates are available, and henceforth in this study by ILI rate we mean weighted ILI rates.

A surveillance week runs from Sunday thru Saturday and provider reports are due by Tuesday of the following week. The aggregated ILI rates are publicly released through the FluView website (Centers for Disease Control and Prevention, 2018b) on Friday. A majority of the providers report data on time, but the system allows for delayed reporting with the delayed data included in subsequent weekly releases. Hence, the ILI estimates for a week can change for multiple weeks following the initial release and while referring to the ILI weekly rate it is necessary to identify not only the week but also the revision version, i.e. the week when the revision was released. In generating retrospective forecasts, we only used the version of the data that would have been available if the forecasts had been generated in real-time. Snapshots of the revisions to ILI data are not available through FluView, but an archive of revisions from 2009/10 season onwards exists (Epidemic Prediction Initiative 2015-2016). Furthermore, revisions for seasons 2013/14 and later are programmatically accessible through the DELPHI group's *epidata* API (Delphi Research Group, 2019).

2.1.2. Google extended trends (GET) API

As evident from the above description, there is a lag of 5 days between the end of a surveillance week and the week's data being first available on FluView. Previously, we explored the generation of alternative proxy estimates of ILI for the current or ongoing week using Google search trends (Kandula et al., 2017). To discriminate an ILI estimate for a future week – a forecast – from an estimate for the current week, we refer to these search trends based estimates as nowcasts.

Search trends were downloaded from the GET API (Google Trends Team, 2019; Ginsberg et al., 2009), which provides timeline data of the probability that a specified term is queried during a search session. Additional parameters allow for the specification of geographical and temporal granularities and time period of interest. From (Google Correlate (2011); Mohebbi et al., 2011) and other sources we identified a set of query terms whose search activity is historically well-correlated with ILI rates. Random forest regression models were then built with the search frequencies of these terms as the feature set and ILI rates as response. Separate nowcast models were fit for the 11 locations at each week.

For the US national level we retrieved search trends from the API at the 'country' resolution, but as trends at regional resolution are not directly available, we approximate these as population weighted means of state trends. Unlike ILINet data, search trends data are not revised, but as the response of the random forest models are ILINet rates, only versions of the rates that would have been available in real-time were used to generate nowcasts. See supplementary information for a more detailed description of the nowcast method.

2.2. Methods

2.2.1. Autoregressive integrated moving average (ARIMA)

ARIMA models estimate future observations of a time series as a function of past observations and forecast errors (also known as moving average). A non-seasonal ARIMA model is specified by three parameters: *p*, the order of the autoregressive component; *q*, the order of the moving average component; and *d*, the degree of differencing required to make the given time series stationary. For a time series, *Y*, let *y* denote the time series obtained by *d* degree differencing. If $d = 1$, $y_t = Y_t - Y_{t-1}$ and if $d = 2$, $y_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2})$, etc. A Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test (Kwiatkowski et al., 1992) can be used to determine if differencing is required for a given *Y* and repeated KPSS tests can be used to identify an appropriate degree of differencing.

Thus, an *ARIMA(p, d, q)* is a model of the form:

$$y_t = c + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_q \varepsilon_{t-q}$$

where the elements, ε_i, represent the forecast errors at the i^{th} time step.

Although values for parameters c, ϕ_1, ..., ϕ_p, θ_1, ...θ_q can be estimated through maximum likelihood estimation, determining appropriate values for *p* and *q* is non-trivial. In this study we used an implementation of an iterative method proposed by (Hyndman and Khandakar (2008)) from the R (R Core Team, 2013) *forecast* (Hyndman, 2015) package to find *p*, *q* and to fit ARIMA models. Briefly, the method is initiated with the model that has the lowest Akaike's Information Criterion (AIC) (Akaike, 1974) amongst 4 models, *[(p, q): (0, 0), (0,1), (1, 0), (2, 2)]*. This is followed by iteratively varying *p* and/or *q* by ± 1, and picking the variant with the lowest AIC. The process is terminated when the pre-specified parameter space for *p, q* is exhausted or when all variants of the selected model result in a higher AIC.

A seasonal time series model is specified with additional terms *P, D, Q* where *D* is seasonal differencing and *P, Q* are analogous to *p, q* described above. Hyndman and Khandakar's method can also estimate appropriate values for these seasonal parameters. In the current study, we have not mandated use of a seasonal model and a non-seasonal model was used when it was found to be better than a seasonal model. Additionally, to inform the model of known holiday effects (Brooks et al., 2015) (anomalous changes in ILI rates during the weeks of Thanksgiving, Christmas and New Year's), week numbers were provided as external regressors.

2.2.2. ARIMA with seasonal and trend decomposition (ARIMA-STL)

The seasonality of influenza outbreaks in the US is well established and we were interested in examining the change in forecast quality if ARIMA models are fit to a seasonally adjusted time series instead of the raw ILI rates. Seasonal and Trend Decomposition using Loess (STL) (Cleveland et al., 1990) is one of the more robust methods to decompose a time series into its seasonal, trend and remainder components. Using STL, we decomposed the raw ILI rates to remove the seasonal component before fitting an ARIMA model. As the time series is weekly, and the seasonality is annual with 52 or 53 weeks per year, we used a periodicity of 52.18 (Hyndman, 2014). Additionally, since STL only handles additive decomposition and ILI rates are more likely multiplicative, the rates were log transformed prior to applying STL decomposition (Hyndman and Athanasopoulos, 2014). As in the ARIMA models, week numbers were used as external regressors.

To generate a final forecast, the seasonal component is added back to the model predictions and the summand is then reverse log transformed (i.e. exponentiated) to obtain ILI rate forecasts. Note that the prediction intervals estimated by these models do not capture the uncertainty of the seasonal component and can be expected to be narrower than the prediction intervals of ARIMA models fit on non-seasonally adjusted time series.

2.2.3. Autoregression with neural network (AR-NN)

A simple neural network (McCulloch and Pitts, 1943; Rosenblatt, 1962; Rumelhart et al., 1985; Werbos, 1974) that connects input nodes (explanatory variables) to a single output node (response) can mimic linear regression, with the regression coefficients given by the weights of the connectors. When a layer of nodes is added between the input and output layers, the regression becomes two-step as the network can extract linear combinations of the inputs as derived features, and through the use of an appropriate activation function (such as the sigmoid) can model the response as a nonlinear function of these features. As the intermediate layers are unobserved they are commonly referred to as hidden layers (Hyndman and Athanasopoulos, 2014; Hastie et al., 2009). Additional intermediate layers would allow the network to model more complex non-linear functions, but would also make finding a solution computationally more expensive. By using lagged values of the response as the inputs, the neural network can be implemented as an autoregressive model. To model a non-linear autoregressive function, as is required here, a neural network with a single hidden layer will suffice. Such a network is specified by three parameters - *p*, the number of lagged inputs; *k*, the number of nodes in the hidden layer; and for seasonal data, *P*, the number of previous seasons to consider - and can be denoted by *NN(p, P, k)*. For example, a specification of *NN* (6, 2, 3) to denote values of p, P and k respectively; for a monthly time series with annual seasonality, implies that the model has 3 nodes in the hidden layer and uses observations of the previous 6 months and of the same month in the previous 2 years as input, $y_t = f(y_{t-1}, y_{t-2}, \ldots, y_{t-6}, y_{t-12}, y_{t-24},)$

As in the above two methods, we used *forecast* package's implementation of an autoregressive neural network with the following parameters: $P = 1$; p is chosen so as to minimize AIC; and $k = (p + P + 1)/2$. To avoid overfitting through assignment of excessive weights on some of the connectors a decay parameter of 0.5 was used.

2.3. Forecast generation and validation

Using the above methods, retrospective forecasts were generated at US National and the 10 HHS regions during the 2012/13 to 2016/17 influenza seasons beginning from the Morbidity and Mortality Weekly Report (MMWR) (Centers for Disease Control and Prevention, 2018c) week 40. At each week, the ILI estimates for all weeks beginning week 40 of the 2010/11 season and the nowcast estimate for the next week were used to fit the models. As a log transformation was required for ARIMA-STL, this was applied to all three models for consistency. This has an added advantage of stabilizing the observational variance prior to model fitting.

At a given week *t*, the models are fit using the time series $X_1, \ldots, X_t, Z_{t+1}$, where *X* and *Z* denote the transformed rates and nowcast, respectively, and are used to forecast rates for weeks $t + 2, \ldots, t + 4$. Let $\hat{X}_{t+k}$ denote the point forecast from the model for ILI rate *k*-weeks ahead and $[\hat{X}^{.025}_{t+k}, \hat{X}^{.975}_{t+k}]$ the corresponding 95% prediction interval. The probabilistic forecast *k*-weeks ahead is calculated on 1000 random draws from a truncated (minimum = 0) normal distribution defined by,

$$N(\hat{X}_{t+k},\ (\hat{X}^{.975}_{t+k} - \hat{X}^{.025}_{t+k})/(1.96*2))$$

The nowcast Z_{t+1} doubles as the 1-week ahead point forecast and a bootstrap distribution of the random forest model that generated the nowcast is used as the 1-week ahead probabilistic forecast.

The point and probabilistic forecasts thus obtained are sufficient to compare the three autoregressive methods. To evaluate the quality of these forecasts relative to other statistical and mechanistic models in use by the ILI modelling community, we used two validation sets.

2.3.1. Set 1: FluSightNetwork component model forecasts

Multiple groups of infectious disease modelers have been generating real-time forecasts of influenza over the past several years, either independently or as part of CDC coordinated Epidemic Prediction Initiative (EPI) (Biggerstaff et al., 2016, 2018; Epidemic Prediction Initiative, 2019a). The (FluSightNetwork (2019)) was formed in 2017 to bring together forecasts from these different models in a standardized template to systematically compare forecast accuracy and develop ensemble methods that build on the strengths of individual models. Participants were required to submit retrospective forecasts for 32 weeks during each of the 2010/11 through 2016/17 influenza seasons, at US National and 10 HHS regions using only data that would have been available if the forecasts had been generated in real-time. Detailed guidelines for contributing forecasts are published elsewhere (Lab PR., 2019a). At the time of the start of this study, forecasts from 21 models have been submitted and publicly available (FluSightNetwork (2019)). Forecasts from models submitted since have not been included in this analysis.

Included in the FluSightNetwork common template are the four near-term targets that are the focus of this study. In generating forecasts from the three autoregressive models, we adhered to the common template guidelines, thus allowing for a comparison of the forecast quality of these methods against the components of the FluSightNetwork.

2.3.2. Set 2: epidemic prediction initiative (EPI) 2016/17 season challenge

EPI's 'Forecast the 2016/17 Influenza Season Collaborative Challenge' (FluSight, 2017) was an effort coordinated by the Influenza division of CDC to solicit and compare real-time forecasts of ILI at US National and 10 HHS regions during the 2016/17 season. This was the fourth consecutive annual iteration of the challenge and 29 teams participated. The challenge was held between November 7, 2016 and May 15, 2017 and participating teams were required to submit weekly forecasts per a standard template that is identical to the FluSightNetwork template. We used an archive (Epidemic Prediction Initiative, 2019b) of submitted forecasts as a second validation set.

2.3.3. Evaluation measures

The primary evaluation measure used in this study is a variation of the log-score, which is a 'proper' scoring rule (Gneiting and Raftery, 2007). This variant is defined for probabilistic forecasts and is identical to the measure used to compare the component models of the FluSightNetwork and by the EPI 2016/17 challenge to rank participant models (Niemi, 2019).

The probabilistic forecast for region *r's* near-term target *g* using ILI rates through week *w,* denoted by the tuple *(r, g, w)*, is a set of probabilities for the *possible* intensity outcomes. The log-score is calculated as, $ln\left(\sum_{i\in O_g^r} p_i\right)$, where O_g^r is the set of *acceptable* outcomes, a subset of the possible outcomes, and p_i is the probability assigned by the model to outcome *i*. Following EPI's guidelines, intensity intervals [0, 0.1), [0.1, 0.2), …, [12.9, 13), and [13, 100] were used as possible outcomes and interval bins within 0.5% of the observed intensity were considered acceptable. If $\sum_{i\in O_g^r} p_i = 0$, the log-score was set to -10.

In addition, we define the point forecast of *(r, g, w)*, a single scalar value, as the predicted intensity for *g*. The accuracy of the point forecast was measured by the absolute proportional error, i.e. error as a proportion of the observed intensity.

We use simple means of scores/errors to aggregate across targets, weeks, seasons and regions. Note that when comparing forecasts of the autoregressive methods with each other, we ignore the 1-week ahead target (i.e. the identical nowcasts) and only use the 2–4 week ahead forecasts, but when comparing these with the two validation sets described above, we use forecasts for all 1–4 weeks.

3. Results

3.1. Comparison of the three autoregressive methods

For the probabilistic forecasts, the forecast quality of the near-term forecasts of AR-NN were better than that of ARIMA and ARIMA-STL overall, and at a majority of seasons and regions (see Table 1, *Overall*). For ARIMA and ARIMA-STL, there is no discernable difference in forecast quality at each of the three weekly horizons, but noticeable variability exists across seasons and locations. As longer time series are available to train the models during later seasons, we expected to observe a corresponding improvement in forecast accuracy; however, no such trend is evident - 2012/13 had the lowest log score on average, but three of the remaining four seasons have similar log-scores. Among locations, HHS regions 1, 8 and US National scored high whereas HHS region 6 had the lowest score.

In instances when the mean log-scores of ARIMA-STL and ARIMA are similar, the standard deviation of the former tends to be smaller. As anticipated in the *Methods* section, this is a result of the ARIMA-STL model omitting the uncertainty of the seasonal component. Figure S1, which plots the median log-score and the inter-quartile ranges for the three methods, shows that AR-NN often also has a better median skill score than ARIMA and ARIMA-STL.

Table 2 shows the corresponding mean absolute proportional error of the point forecasts. As with probabilistic forecasts, AR-NN outperforms the other two methods, overall, at each of the three time horizons and a majority of the locations. ARIMA has lower error than ARIMA-STL at most disaggregation criteria examined.

Ignoring the magnitude of the difference, the forecast quality of the three methods was ranked from best (rank = 1) to worst (rank = 3) for each combination of region-season-week-target. As seen in Fig. 1, for about 47% (2358 of 4983[1]) of the instances by log-score and 58% (2908 of 4983) by proportional error, AR-NN ranked best. ARIMA more often ranked second than ARIMA-STL. Taken together, Tables 1 and 2 and Fig. 1, S1 indicate that AR-NN most often yields the best quality forecasts (both point and probabilistic) but also has a number of instances of large errors relative to ARIMA.

3.2. Forecast quality relative to component models of FluSightNetwork

Fig. 2 shows the overall skill (exponentiation of the log-score) for the three autoregressive methods introduced in this paper relative to the 21 component models of the FluSightNetwork. All three methods presented here have a higher skill than all FluSightNetwork component models, perhaps largely benefiting from their performance in the 2015/16 season ('+' data points). When disaggregated by target (Fig. 3), we see that the forecasts of the autoregressive methods are better than the forecasts of the component models at all horizons except 4-week. All 3 autoregressive methods and models prefixed *CU_* have identical 1-week ahead scores as they all used identical nowcasts.

The forecast quality decreases with increasing time horizon for all component models, but this deterioration is larger in the proposed models relative to the best performing FluSightNetwork component models, with the 3-week ahead forecast of ARIMA and ARIMA-STL matching the best component model but noticeably underperforming for the 4-week ahead forecast. Even at 4-weeks the forecasts are better than a majority of the component models including all *CU_* models. Figure S2 is a counterpart of Fig. 1, and shows a distribution of ranks of the three methods relative to all models of the FluSightNetwork. All

[1] Forecasts were generated for 3 targets (2- to 4-week ahead) at 11 regions over 151 weeks (across 5 seasons). Hence a total of 4983 forecasts from each method were compared (3*11*151). Of these, in 2358 cases, the probabilistic scores of the AR-NN forecasts were better than both ARIMA and ARIMA-STL. 2908 is the analogous count for the point forecasts.

Table 1
Mean (std. dev) of log-score of the 3 methods - overall, disaggregated by target, season and region. In each row, the best score is underlined.

		ARIMA	AR w/ NeuralNet	ARIMA Deseasoned
Overall		−0.82 (1.22)	−0.8 (1.22)	−0.83 (.96)
Horizon	**2 week ahead**	−0.59 (0.97)	−0.59 (1.04)	−0.61 (0.8)
	3 week ahead	−0.84 (1.27)	−0.82 (1.25)	−0.85 (0.96)
	4 week ahead	−1.03 (1.35)	−0.99 (1.32)	−1.03 (1.06)
Season	**2012/13**	−1.06 (1.43)	−0.97 (1.47)	−1.08 (0.94)
	2013/14	−0.81 (1.31)	−0.79 (1.3)	−0.82 (0.88)
	2014/15	−0.8 (1.21)	−0.78 (1.2)	−0.83 (1.06)
	2015/16	−0.61 (0.81)	−0.58 (0.66)	−0.62 (0.75)
	2016/17	−0.78 (1.19)	−0.87 (1.27)	−0.76 (1.09)
Region	**US National**	−0.65 (1.12)	−0.57 (0.93)	−0.68 (0.86)
	HHS Region 1	−0.52 (1.22)	−0.54 (1.2)	−0.5 (0.87)
	HHS Region 2	−0.94 (0.79)	−1.01 (1.02)	−0.99 (0.68)
	HHS Region 3	−1.01 (1.71)	−0.99 (1.64)	−0.97 (1.3)
	HHS Region 4	−0.94 (1.19)	−0.93 (1.31)	−0.96 (1)
	HHS Region 5	−0.73 (1.26)	−0.73 (1.27)	−0.74 (1)
	HHS Region 6	−1.22 (1.42)	−1.2 (1.36)	−1.28 (1.27)
	HHS Region 7	−0.86 (0.88)	−0.82 (0.9)	−0.91 (0.76)
	HHS Region 8	−0.64 (1.42)	−0.58 (1.35)	−0.57 (0.85)
	HHS Region 9	−0.82 (0.99)	−0.82 (0.91)	−0.84 (0.76)
	HHS Region 10	−0.66 (0.98)	−0.6 (1.14)	−0.67 (0.75)

Table 2
Mean (std. dev) absolute proportional error of the 3 methods - overall, disaggregated by target, season and region. In each row, the lowest errors is underlined.

		ARIMA	AR w/ NeuralNet	ARIMA Deseasoned
Overall		0.234 (0.23)	0.221 (0.21)	0.256 (0.26)
Horizon	**2 week ahead**	0.18 (0.17)	0.172 (0.16)	0.19 (0.18)
	3 week ahead	0.233 (0.23)	0.222 (0.21)	0.255 (0.24)
	4 week ahead	0.289 (0.27)	0.269 (0.24)	0.323 (0.32)
Season	**2012/13**	0.303 (0.29)	0.249 (0.22)	0.368 (0.36)
	2013/14	0.254 (0.24)	0.242 (0.23)	0.275 (0.24)
	2014/15	0.203 (0.2)	0.191 (0.18)	0.216 (0.2)
	2015/16	0.222 (0.23)	0.221 (0.22)	0.228 (0.23)
	2016/17	0.184 (0.17)	0.201 (0.18)	0.185 (0.17)
Region	**US National**	0.156 (0.13)	0.139 (0.11)	0.189 (0.17)
	HHS Region 1	0.182 (0.16)	0.198 (0.17)	0.201 (0.17)
	HHS Region 2	0.224 (0.2)	0.216 (0.19)	0.237 (0.21)
	HHS Region 3	0.217 (0.19)	0.208 (0.19)	0.241 (0.2)
	HHS Region 4	0.276 (0.21)	0.262 (0.2)	0.309 (0.27)
	HHS Region 5	0.215 (0.17)	0.2 (0.17)	0.234 (0.19)
	HHS Region 6	0.178 (0.15)	0.169 (0.13)	0.211 (0.18)
	HHS Region 7	0.334 (0.32)	0.309 (0.29)	0.343 (0.35)
	HHS Region 8	0.229 (0.21)	0.22 (0.18)	0.252 (0.21)
	HHS Region 9	0.176 (0.13)	0.184 (0.13)	0.19 (0.15)
	HHS Region 10	0.388 (0.41)	0.324 (0.33)	0.408 (0.45)

three methods are unlikely to be the worse ranked among the component models, very likely to score among the top 10, and overall have a good rank distribution.

3.3. Forecast quality relative to participants of 2016/17 EPI challenge

Fig. 4 shows the overall skill score of the three autoregressive methods during the 2016/17 season relative to the 29 participant models of the 2016/17 EPI challenge. ARIMA-STL, ARIMA, AR-NN are the top three ranked models with average skill of .468, .458 and .419 respectively. As can be seen in Table 1, AR-NN had greater difficulty forecasting 2016/17 season over three of the five seasons studied, yet was found to perform better than the best performing participating model.

3.4. Sensitivity analysis

The log-score computation described in an earlier section used two ad hoc values: the size of the acceptance window i.e. the margin around the observed truth, and a lower bound on the log-score. The results reported so far are with a window size of 0.5 and a lower bound of -10. Although this is consistent with the scoring scheme recommended and used by EPI challenges, to assess the sensitivity of the results to these two values, we performed additional analysis using alternative window sizes of 0.2, 0.7 and 1, and alternative lower bounds of -5 and -7. Both alternative lower bounds assign smaller penalty to missed forecasts and hence the log-scores are expected to improve over those seen with a lower bound of -10. Similarly log-scores should improve with the more lenient window sizes of 0.7 and 1. Table S1 shows that the changes in log-score are in line with these expectations and the relative accuracy of the three methods remains largely unchanged.

Extending the sensitivity analysis to the component models of FluSightNetwork (Table S2) shows that the superior performance of the three methods is independent of the acceptable window and the lower bound used, and they are consistently well ranked in all attempted scenarios.

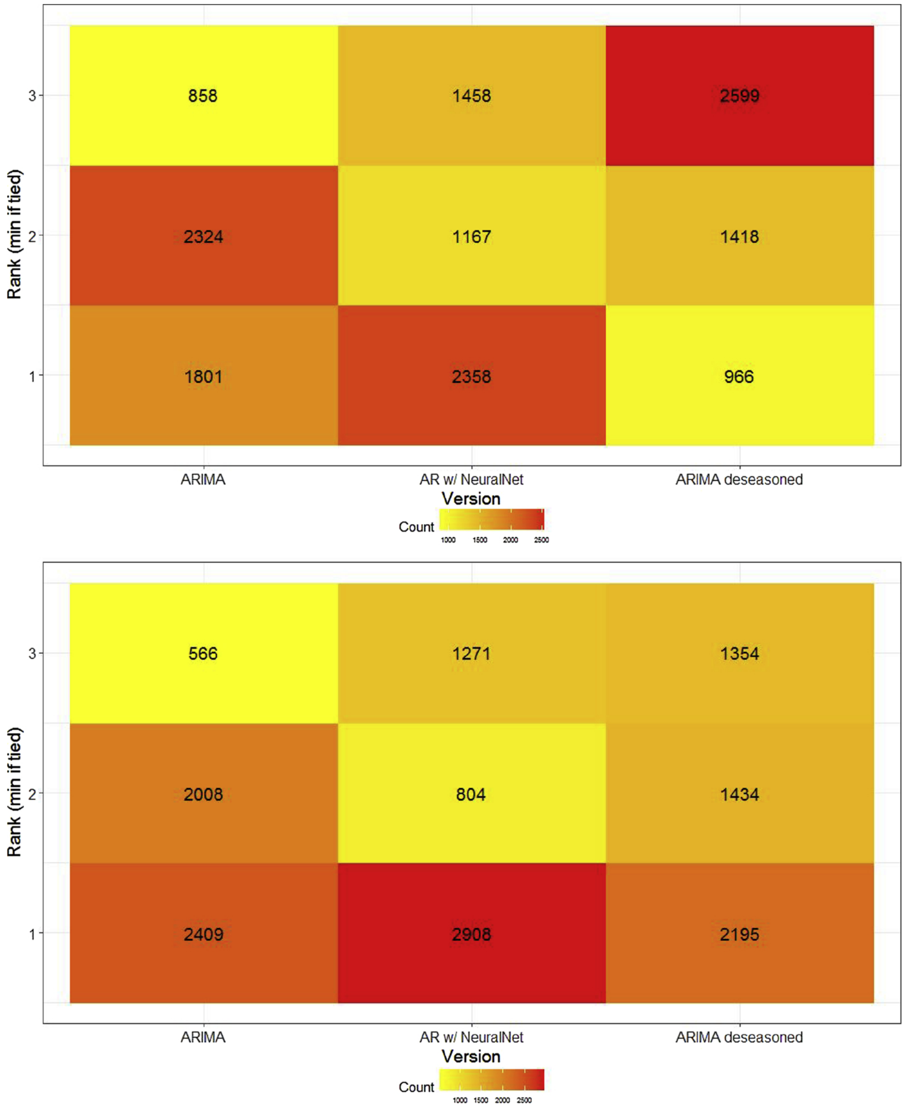

Fig. 1. Heat map showing the distribution of ranks among the 3 methods for a) log-score; b) absolute proportional error. For each [region, season, week, target] combination the scores/errors of the 3 methods were ordered from best (rank = 1) to worst (rank = 3). Tied methods were assigned the same minimum rank; a lot more ties occur with proportional error.

3.5. *Tests for statistical significance*

We performed a Friedman rank sum test (Friedman, 1937) followed by a (Nemenyi (1962)) test to assess whether the differences in forecast quality of the three approaches are statistically significant. The Friedman rank sum test is a non-parametric test that ranks the score/error of each group (here, method) at each region-week-target combination. The Nemenyi test extends the comparison to each pair of methods. We tested for significance with all targets taken together and for each target individually.

The differences in probabilistic forecasts (Table S3) and errors (Table S4) of ARIMA/ARIMA-STL and AR-NN/ ARIMA-STL were found to be statistically significant ($p < .001$), but not between ARIMA and AR-NN. Extending the analysis to component models of the FluSightNetwork (Table S5), we found that log-scores of ARIMA and AR-NN are statistically different from all component models; ARIMA-STL was not dissimilar from a few of the component models, mostly models with shared nowcasts.

Similarly, Table S6 shows results from Friedman-Nemenyi applied to the 2016/17 participating models. ARIMA and ARIMA-STL were found to differ from all of the participatory models at statistically significant levels; AR-NN, the lowest ranked of the three for this season, was not significantly different from a few of the higher scoring participants.

3.6. *Forecasts without nowcasts*

For the models presented here to be portable to scenarios where nowcasts are not readily available, it is pertinent to quantify their dependence on nowcasts. In order to measure this, we regenerated all forecasts using ILI alone i.e. at a given week t, the models were fit using the time series $X_1, \ldots, X_t$, where X denotes the log transformed ILI rates, and forecasts are obtained for weeks $t + 1, \ldots, t + 4$. Tables S7 and S8 show that there was a decrease of 14% in the overall skill score of the models when nowcasts were not used. The difference between 1-week ahead score and nowcasts is 6–9%, and there is considerable knock-on effect on the 2–4 week forecasts. There is noticeable variability among seasons (2014/15 had the largest change and 2016/17 the smallest) and locations, with HHS Region 9 showing a loss of more than a third of the skill when nowcasts were dropped. Similar results were seen with the proportional error of point forecasts (Table S9).

Relative to the component models of FluSightNetwork, these variant

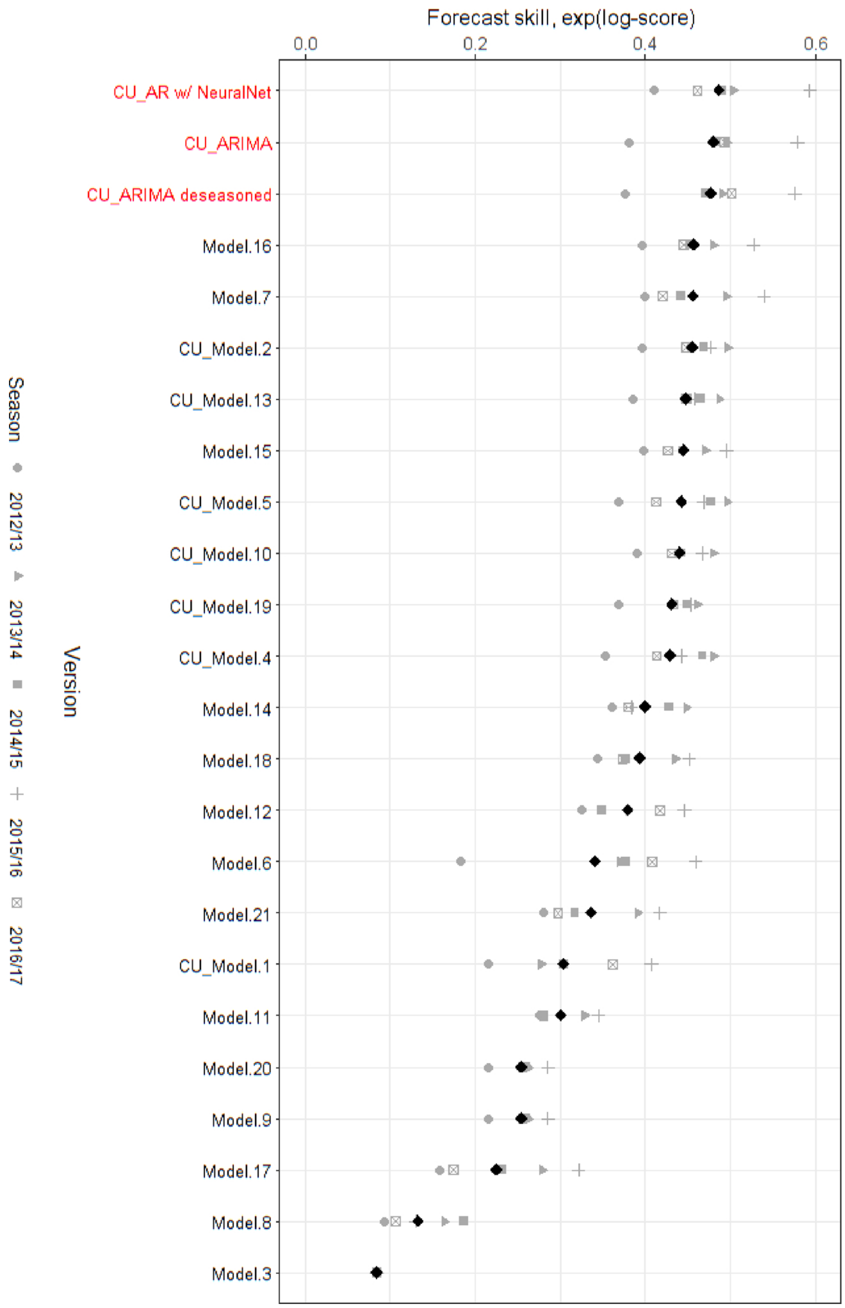

Fig. 2. Mean forecast skill of the 3 methods discussed in this paper (labeled in red), plotted against the 21 component models of the FluSightNetwork. The mean skill across 5 seasons (2012/13–2016/17) is show in black and the other data points denote skills during specific seasons. The 'Version' axis is ordered by the overall mean score. *Model.11* is based on historical outbreaks and can be considered to be a naïve baseline (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

forecasts of *ARIMA, ARIMA-STL* and *AR-NN* were ranked 10–12 respectively (Figure S3). The ranks were largely invariant by target (Figure S4), with a slightly better accuracy at 1-week. Note that at least 6 of the better ranked methods used nowcasts. With respect to participant models of EPI 2016/17 (Figure S5), the three models remain the top ranked as perhaps foreshadowed by an earlier finding - 2016/17 had only a small improvement from nowcasts (Tables S7/S8).

4. Discussion

The results presented here indicate that forecasting near-term ILI with time series approaches is a promising alternative to more sophisticated statistic and mechanistic models. Combined with their computational efficiency and support in standard open-source statistical software, these approaches can provide suitable baselines for comparing existing, as well as newly developed methods. Beginning with the 2017/18 season, ILINet data are being released at the state-level and we do not foresee barriers to using the methods presented here with these, or other, more geographically resolved datasets. Application to different data streams, such as ILI hospitalization rates and virologic data or incidence data from other respiratory and vector-borne diseases may be possible but should be preceded by analysis such as presented here.

This study is to our knowledge the first application of a neural network to forecasting influenza incidence at national scales and the results are promising. Although the improvement over ARIMA models is slim, it would add to the diversity of components of superensembles, such as FluSightNetwork, and potentially improve their accuracy. As

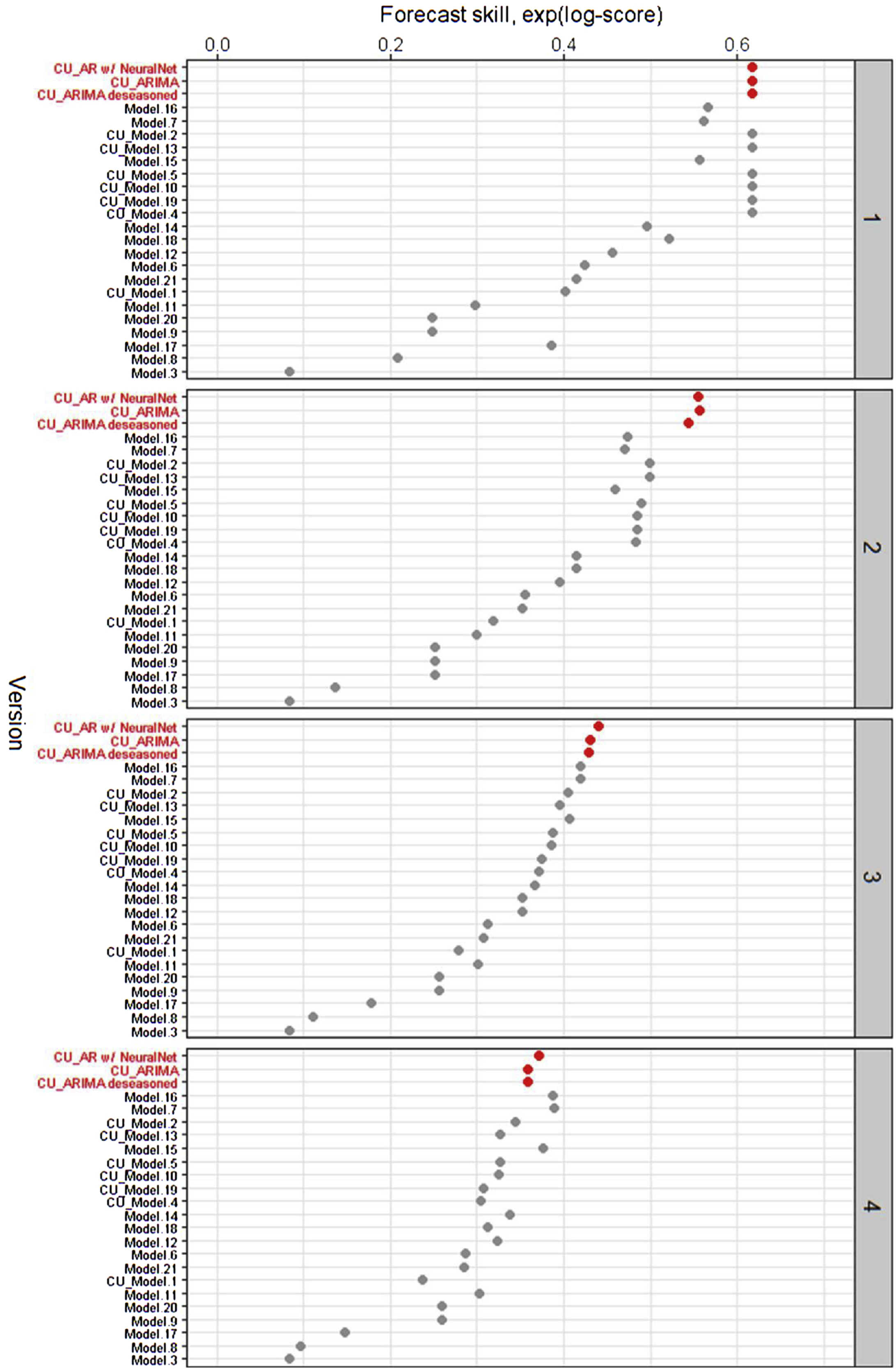

Fig. 3. Mean forecast skill of the 3 methods discussed in this paper (labeled in red) at each of the four time horizons, plotted against the 21 component models of the FluSightNetwork. For instance, panel labeled '1' plots scores for the 1 week ahead forecasts. The 'Version' axis continues to be ordered by the overall (1–4 week ahead) mean score. *Model.11* is based on historical outbreaks and can be considered to be a naïve baseline (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

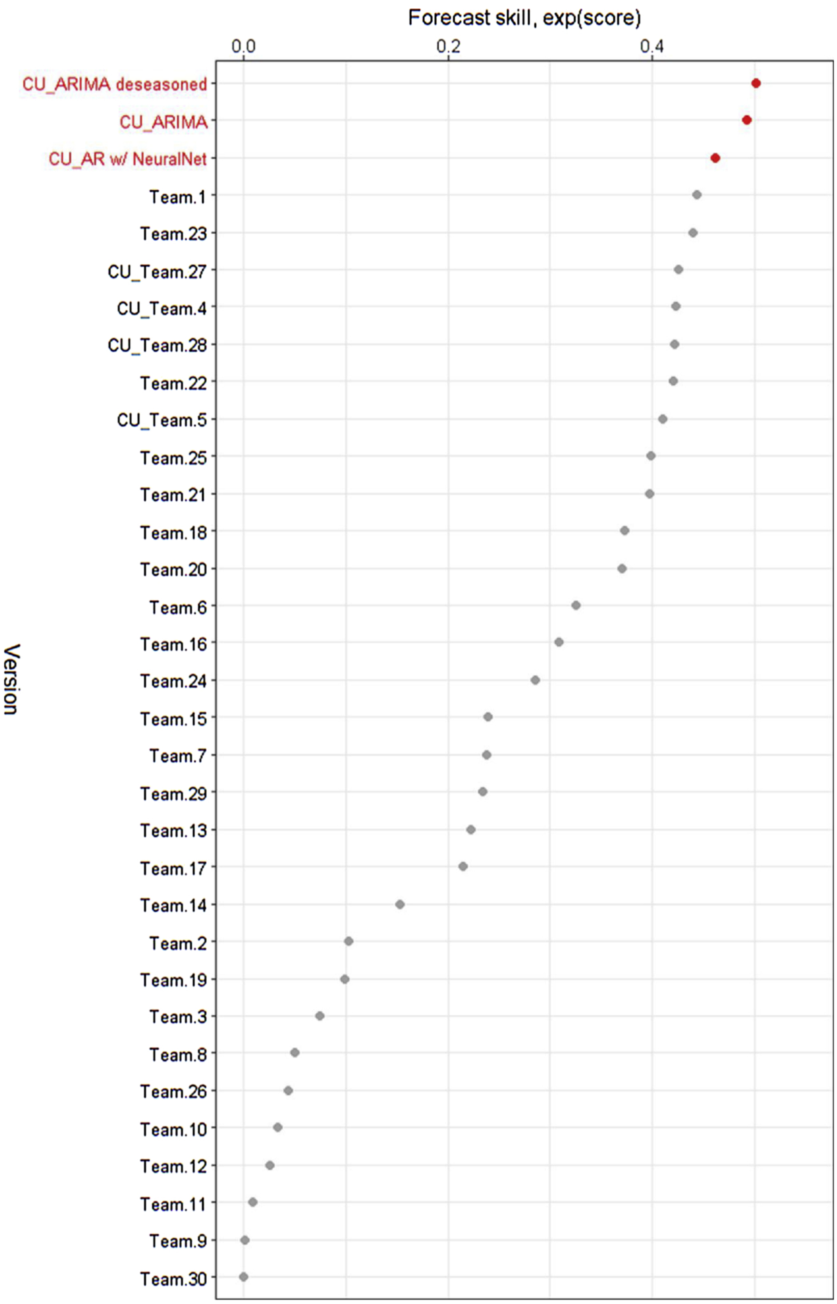

Fig. 4. Mean forecast skill of the 3 methods discussed in this paper (labeled in red), plotted against the 29 models that participated during the 2016/17 CDC EPI challenge. The 'Version' axis is ordered by the overall (1–4 week ahead) mean score. *Team.24* can be considered as a naïve baseline whose forecasts are an average of historical observations (the forecast for a week is the average of ILI rates observed at that week in previous seasons) (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

noted in an earlier section, the underperformance of AR-NN's probabilistic forecasts in a sizeable number of instances (Fig. 1a) is quite likely a result of incorrectly calibrated prediction intervals, and can potentially be remedied. Use of an ad hoc scaling factor to artificially extend the intervals, was found to improve scores (not presented) but more systematic alternatives need to be explored. Parameter choices such as the number of nodes in the hidden layer and the number of hidden layers will need to be revisited.

The decrease in forecast quality when nowcasts were not used to train the models underscores the need for more up-to-date outbreak information and makes a case for the adoption of nowcasts in operational settings whenever possible. In addition to providing observation for an additional week, the nowcasts could help counteract the errors present in the initial release of surveillance data. Our results indicate that even when nowcasts are not available, the forecasts from these methods improve over the quality of most of the existing methods.

4.1. Limitations and future work

We have not attempted to analyze the effect of the magnitude of revisions to the initial release of ILINet data on the forecast quality and it would be interesting to see if there are significant differences in the sensitivity of the three models to these errors. A related sub-analysis would be to compare model performance at different phases of a season – weeks of low activity at the beginning and end of season versus weeks around the peak week. These analyses would have operational implications, as users may need them to pick the model most appropriate to their observational data stream or outbreak.

The log-score presented here uses the same acceptable window for all regions and seasons, and we chose it in order to be consistent with the EPI challenges and the FluSightNetwork. However, as outbreak size varies by HHS region and season, a fixed window size does not score all regions/seasons equitably. We believe a scoring scheme that defines acceptable margins relative to observed ground truth rather than a common fixed window would be more appropriate.

Our original motivation to propose simple, easy-to-implement methods limited exploration of potentially promising albeit more complex alternative approaches such as artificial neural network with multiple hidden layers and recurrent neural networks. Future efforts should explore the benefits and costs of these more advanced methodologies. Furthermore, as the three methods presented here do not consider underlying disease transmission dynamics, seasons that deviate considerably from a 'typical' season may see larger errors. This is a characteristic shared with other statistical ILI forecasting methods and it is recommended that whenever possible these should be used in conjunction with process-based models, such as simple compartmental models, that are better equipped to capture ILI transmission dynamics in populations.

Competing financial interests

JS and Columbia University declare partial ownership of SK Analytics. SK was a consultant for SK Analytics.

Acknowledgments

This work was supported by grants from US National Institutes of Health [GM110748]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We would like to thank coordinators of the Epidemic Prediction Initiative's FluSight Collaborative for access to historical incidence rates and scores for the 2016/17 season, and to all contributors of the FluSightNetwork and FluSight challenges for sharing their model forecasts.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.epidem.2019.01.002.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Automat. Contr. 19 (6), 716–723.

Balcan, D., Colizza, V., Gonçalves, B., Hu, H., Ramasco, J.J., Vespignani, A., 2009. Multiscale mobility networks and the spatial spreading of infectious diseases. Proc. Natl. Acad. Sci. 106 (51), 21484–21489.

Biggerstaff, M., Alper, D., Dredze, M., Fox, S., IC-H, Fung, 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 (1), 357.

Biggerstaff, M., Johansson, M., Alper, D., Brooks, L.C., Chakraborty, P., Farrow, D.C., et al., 2018. Results from the second year of a collaborative effort to forecast influenza seasons in the United States. Epidemics.

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, 2018a. Overview of Influenza Surveillance in the United States. Available from:. http://www.cdc.gov/flu/weekly/overview.htm.

Centers for Disease Control and Prevention, 2018b. FluView Interactive. Available from:. http://gis.cdc.gov/grasp/fluview/fluportaldashboard.html.

Centers for Disease Control and Prevention, 2018c. National Notifiable Diseases Surveillance System. Available from:. MMWR Weeks. https://wwwn.cdc.gov/nndss/document/MMWR_week_overview.pdf.

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.

Cleveland, R.B., Cleveland, W.S., Terpenning, I., 1990. STL: a seasonal-trend decomposition procedure based on loess. J. Off. Stat. 6 (1), 3.

Delphi Research Group, 2019. Epidemiological Data API. Available from:. https://github.com/cmu-delphi/delphi-epidata.

Epidemic Prediction Initiative: FluSight 2015-2016, Data archive Available from: https://predict.phiresearchlab.org/post/5a6232f8da94b605acafdca8.

Epidemic Prediction Initiative. Available from:. https://predict.cdc.gov/.

Epidemic Prediction Initiative. Available from:. FluSight Forecast Submissions Archive. https://github.com/cdcepi/FluSight-forecasts.

Farrow, D., 2016. Modeling the Past, Present, and Future of Influenza. [Doctoral dissertation]. Carnegie Mellon University.

Farrow, D.C., Brooks, L.C., Hyun, S., Tibshirani, R.J., Burke, D.S., Rosenfeld, R., 2017. A human judgment approach to epidemiological forecasting. PLoS Comput. Biol. 13 (3) e1005248.

FluSight, 2017, Available from: https://predict.cdc.gov/post/57f3f440123b0f563ece2576.

FluSightNetwork Guidelines and Forecasts for a Collaborative U.S. Influenza Forecasting Project. https://zenodo.org/record/1255023#.W3GwfthKjOY.

Friedman, M., 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J. Am. Stat. Assoc. 32 (200), 675–701.

Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L., 2009. Detecting influenza epidemics using search engine query data. Nature. 457 (7232), 1012–1014.

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. J. Am. Stat. Assoc. 102 (477), 359–378.

Google Correlate Available from: https://www.google.com/trends/correlate, 2011.

Google Trends Team, 2019. Extended Health Trends API. Available from:. https://research.googleblog.com/2015/08/the-next-chapter-for-flu-trends.html.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning, 2nd edition. Springer, New York.

Hyder, A., Buckeridge, D.L., Leung, B., 2013. Predictive validation of an influenza spread model. PLoS One 8 (6) e65459.

Hyndman, R., 2015. Forecasting Functions for Time Series and Linear Models. R package version 6.1.

Hyndman, R.J., 2014. Forecasting Weekly Data. Available from:. https://robjhyndman.com/hyndsight/forecasting-weekly-data/.

Hyndman, R.J., Athanasopoulos, G., 2014. Forecasting: Principles and Practice. OTexts.

Hyndman, R., Khandakar, Y., 2008. Automatic Time Series Forecasting: The Forecast Package for R 7. 2007.

Niemi, J., 2019. FluSight: An R Package Containing Utility Functions for the CDC Flu Forecasting Competition. Available from:. https://github.com/jarad/FluSight.

Kandula, S., Hsu, D., Shaman, J., 2017. Subregional nowcasts of seasonal influenza using search trends. J. Med. Internet Res. 19 (11).

Kwiatkowski, D., Phillips, P.C., Schmidt, P., Shin, Y., 1992. Testing the null hypothesis of stationarity against the alternative of a unit root: how sure are we that economic time series have a unit root? J. Econom. 54 (1-3), 159–178.

Lampos, V., Miller, A.C., Crossan, S., Stefansen, C., 2015. Advances in nowcasting influenza-like illness rates using search query logs. Sci. Rep. 5, 12760.

McCulloch, W.S., Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. Bull. Math. Biophys. 5 (4), 115–133.

Mohebbi, M., Vanderkam, D., Kodysh, J., Schonberger, R., Choi, H., Kumar, S., 2011. Google Correlate Whitepaper.Google Correlate Whitepaper.

Molinari, N.A., Ortega-Sanchez, I.R., Messonnier, M.L., Thompson, W.W., Wortley, P.M., Weintraub, E., et al., 2007. The annual impact of seasonal influenza in the US: measuring disease burden and costs. Vaccine. 25 (27), 5086–5096.

Nemenyi, P. (Ed.), 1962. Distribution-Free Multiple Comparisons. Biometrics.

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 Other Respir. Viruses 8 (3), 309–316.

Osthus, D., Hickmann, K.S., Caragea, P.C., Higdon, D., Del Valle, S.Y., 2017. Forecasting seasonal influenza with a state-space SIR model. Ann. Appl. Stat. 11 (1), 202–224.

Paul, M.J., Dredze, M., Broniatowski, D., 2014. Twitter improves influenza forecasting. PLOS Currents Outbreaks.

Ray, E.L., Reich, N.G., 2018. Prediction of infectious disease epidemics via weighted density ensembles. PLoS Comput. Biol. 14 (2) e1005910.

Ray, E.L., Sakrejda, K., Lauer, S.A., Johansson, M.A., Reich, N.G., 2017. Infectious disease prediction with kernel conditional density estimation. Stat. Med.

Rolfes, M., Foppa, I., Garg, S., Flannery, B., Brammer, L., Singleton, J., 2016. Estimated Influenza Illnesses, Medical Visits, Hospitalizations, and Deaths Averted by Vaccination in the United States. Centers for Disease Control and Prevention.

Rosenblatt, F., 1962. Principles of Neurodynamics.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1985. Learning Internal Representations by Error Propagation. California Univ San Diego La Jolla Inst for Cognitive Science.

Santillana, M., Nguyen, A.T., Dredze, M., Paul, M.J., Nsoesie, E.O., Brownstein, J.S., 2015. Combining search, social media, and traditional data sources to improve influenza surveillance. PLoS Comput. Biol. 11 (10) e1004513.

Santillana, M., Nguyen, A., Louie, T., Zink, A., Gray, J., Sung, I., et al., 2016. Cloud-based electronic health records for real-time, region-specific influenza surveillance. Sci. Rep. 6.

Shaman, J., Karspeck, A., 2012. Forecasting seasonal outbreaks of influenza. Proc. Natl. Acad. Sci. 109 (50), 20425–20430.

R Core Team, 2013. R: a Language and Environment for Statistical Computing.

U.S. Department of Health & Human Services Regional Offices. Available from: https://www.hhs.gov/about/agencies/regional-offices/index.html.

Viboud, C., Boëlle, P.-Y., Carrat, F., Valleron, A.-J., Flahault, A., 2003. Prediction of the spread of influenza epidemics by the method of analogues. Am. J. Epidemiol. 158 (10), 996–1006.

Dynamic poisson autoregression for influenza-like-illness case count prediction. Wang, Z., Chakraborty, P., Mekaru, S.R., Brownstein, J.S., Ye, J., Ramakrishnan, N. (Eds.), Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Werbos, P., 1974. Beyond Regression: New Fools for Prediction and Analysis in the Behavioral Sciences. PhD thesis. Harvard University.

Yang, S., Santillana, M., Kou, S.C., 2015. Accurate estimation of influenza epidemics using Google search data via ARGO. Proc. Natl. Acad. Sci. 112 (47), 14473–14478.