RESEARCH ARTICLE

A stacked ensemble method for forecasting influenza-like illness visit volumes at emergency departments

Arthur Novaes de Amorim[1]*, Rob Deardon[2], Vineet Saini[3,4]

1 Department of Economics, University of Calgary, Calgary, Alberta, Canada, **2** Department of Mathematics and Statistics and Faculty of Veterinary Medicine, University of Calgary, Calgary, Alberta, Canada, **3** Research and Innovation, Provincial Population and Public Health, Alberta Health Services, Calgary, Alberta, Canada, **4** Department of Community Health Sciences and O'Brien Institute for Public Health, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada

* arthur.novaesdeamori@ucalgary.ca

OPEN ACCESS

Citation: Novaes de Amorim A, Deardon R, Saini V (2021) A stacked ensemble method for forecasting influenza-like illness visit volumes at emergency departments. PLoS ONE 16(3): e0241725. https://doi.org/10.1371/journal.pone.0241725

Editor: Shinya Tsuzuki, National Center for Global Health and Medicine, JAPAN

Received: October 14, 2020

Accepted: February 27, 2021

Published: March 22, 2021

Data Availability Statement: All relevant data are within the manuscript and its Supporting information files.

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Accurate and reliable short-term forecasts of influenza-like illness (ILI) visit volumes at emergency departments can improve staffing and resource allocation decisions within hospitals. In this paper, we developed a stacked ensemble model that averages the predictions from various competing methodologies in the current frontier for ILI-related forecasts. We also constructed a back-of-the-envelope prediction interval for the stacked ensemble, which provides a conservative characterization of the uncertainty in the stacked ensemble predictions. We assessed the accuracy and reliability of our model with 1 to 4 weeks ahead forecast targets using real-time hospital-level data on weekly ILI visit volumes during the 2012-2018 flu seasons in the Alberta Children's Hospital, located in Calgary, Alberta, Canada. Our results suggest the forecasting performance of the stacked ensemble meets or exceeds the performance of the individual models over all forecast targets.

Introduction

Influenza-like illness (ILI) causes significant burden to healthcare systems [1, 2]. To address this burden, a growing body of literature focuses on developing accurate and reliable forecasts of ILI to help inform public health decisions and resource planning [3–5]. These research efforts led to multiple candidate models which may be suitable for forecasting ILI activity within a country or region.

While forecasts at a regional level may improve public health decisions, substantial heterogeneity within regions imply a one size fits all forecast at such a large level of spatial aggregation may be of little use for hospital-level staffing and resource allocation decisions [6]. Here, we bridge this literature gap by assessing the forecasting performance of various models from the recent literature using Emergency Department (ED)-level data from the Alberta Real-Time Syndromic Surveillance Net (ARTSSN). Our goal is to equip decision makers in hospitals with near-term forecasts (1 to 4 weeks ahead) of the number of weekly ILI visits to the ED.

We used an ensembling method to leverage the strength of different modeling approaches. Evidence from weather forecasting research [7] and infectious disease forecasts [8] suggest that predictions from ensembling are at least as good as the best performing individual models that they build upon, making the ensemble a potential candidate for predicting ILI positive visit volumes at EDs. We used stacking—a data-driven approach—when leveraging the contribution of each model to the ensemble. In our retrospective analysis, we found that the stacked ensemble model performed as well as—when not better than—the best individual models using various quantitative performance measures on each of the 1 to 4 weeks ahead forecast targets.

Methods

We performed a retrospective analysis with the objective of forecasting weekly ILI positive visits within an ED. The data and modelling decisions are described below.

Data

We used data from the ARTSSN, which offers daily records of patients screened ILI positive in Calgary, Alberta. Our sample featured two EDs: the Alberta Children's Hospital (ACH)—with a high volume of ILI visits—and the Foothills Medical Center (FMC)—with a low volume of ILI visits. The data covered flu seasons from 2012 to 2018, with each flu season starting roughly on the last week of August and lasting 52 or 53 weeks depending on leap year status. The ARTSSN records additionally include patient characteristics such as age, sex, and postal code of last known residence. We supplemented ARTSSN with the following data sets: (1) patients' recent flu immunization status and dates, obtained from Alberta Health; (2) laboratory diagnostics for ILI-causing viral pathogen confirmation (e.g. influenza A H1N1 and H3, influenza B, rhino/enterovirus), obtained from the Alberta Public Health Laboratory; (3) annual mid-year population estimates in Calgary, obtained from Calgary's Civic Census; and, (4) weather information, obtained from Environment and Climate Change Canada.

The unit of observation and analysis in the final linked data set was the weekly-aggregated count of ILI visits at the high-volume ED. We used the low-volume ED data (i.e. FMC) for sensitivity analysis of how the magnitude of visits affects model performance. In some richer model specifications, we further disaggregated the ILI positive visit counts within age brackets and distance buffers around the ED and assessed the impact of these alternative models on forecast performance. Predictors available in a daily format (e.g. minimum temperature) were transformed into weekly averages for the analysis.

Statistical models

Our objective was to estimate at week t, a function $\hat{f}$ such that the forecast of ILI ℓ weeks into the future is given by:

$$\hat{y}_{t+\ell} = \hat{f}(y_t, y_{t-1}, y_{t-2} \ldots, \mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2} \ldots), \tag{1}$$

where y_t is the observed count of ILI visits to the ED at week t and $\mathbf{x}_t$ are other predictors of ILI which are available at time t. Each model drew information from different predictors and specified different relationships between predictors and the outcome of interest. Model implementation is described below.

Empirical Bayes. The Empirical Bayes (EB) model we draw upon is a framework developed by Brooks *et al* for predicting epidemics by relying on slightly modified versions of past epidemics to form possibilities for the current season [9, 10]. With this approach, we first modelled previous flu seasons using non-parametric smoothed piecewise quadratic curves generated with the *cv.trendfilter* method of the *genlasso* R package [11]. This gave us a set of smoothed trajectories $\{f^s\}$ for each flu season s in the training data. Next, we drew at random with equal probability one trajectory from $\{f^s\}$ and applied a series of transformations to this curve, resulting in a curve $\tilde{f}^s$. The transformations shift the epidemic curve's peak height, peak location, and pacing toward the peak, respectively. For these transformations, we randomly sampled a peak height, peak location, and pacing parameter, where candidate peak heights and locations came from the smoothed trajectories f^s and the pacing parameters came from a uniform distribution $U[0.75, 1.25]$. We then assigned a likelihood weight to the transformed curve based on how closely the curve approximated observed ILI up to the current week of the season. We lastly injected noise to the transformed curve, where the noise terms were drawn from a normal distribution $\mathcal{N}(0, \sigma_s^2)$ with σ_s being a noise level derived from the trajectory f^s.

The EB prediction of ILI visits ℓ weeks from week t, $\hat{y}^{EB}_{t+\ell}$, was the weighted median of $N = 10^5$ random samples of transformed trajectories, with the likelihood weights w described above. A prediction interval was also constructed from percentiles of the N weighted samples. For the weighted median point prediction, we ranked the forecast ILI at period $t+\ell$ from the N samples from smallest to largest, $\tilde{f}_{1,t+\ell} < \tilde{f}_{2,t+\ell} \ldots \tilde{f}_{N,t+\ell}$ and computed:

$$\hat{y}^{EB}_{t+\ell} = \tilde{f}_{k,t+\ell}, \quad \text{where } k \text{ is such that}$$
$$\sum_{i=1}^{k-1} w_i <= 0.5 \quad \text{and} \quad \sum_{i=k+1}^{N} w_i <= 0.5. \tag{2}$$

Autoregressive integrated moving average. The Autoregressive integrated moving average (ARIMA) model estimates future periods' ILI positive visit counts as a function of previous observations and forecast errors. An ARIMA(p, d, q) model is specified by three parameters: an autoregressive order term p, a degree of differencing d for making the time series stationary, and, the order of the moving average q. We used seasonal and trend decomposition using locally estimated scatterplot smoothing (STL) to remove seasonality from the raw weekly ILI counts before fitting an ARIMA(2,0,1) model. The analysis was implemented via the *stlm* method of the *forecast* R package [12]. Determination of the number of time lags $p = 2$, degree of differencing $d = 0$, and order of moving average $q = 1$ was based on optimization of the Akaike Information Criterion (AIC) [13].

Upon seasonally adjusting the ILI data with STL, the ARIMA(2,0,1) prediction of ILI positive visits ℓ weeks from week t, $\hat{y}^{ARIMA}_{t+\ell|t}$, is given by the forecasting equation:

$$\hat{y}^{ARIMA}_{t+\ell|t} = \beta + \hat{\phi}_1 y_{t-1} + \hat{\phi}_2 y_{t-2} + \hat{\theta}_1 e_{t-1} + S_{t+\ell}, \tag{3}$$

where $\hat{\phi}_1, \hat{\phi}_2, \hat{\theta}_1$ are the estimated model parameters and $S_{t+\ell}$ is the seasonal component which was computed using STL and removed before fitting the model. The prediction interval is given by $\hat{y}_{t+\ell|t} \pm c\hat{\sigma}_\ell$ where $\hat{\sigma}_\ell$ is an estimate of the standard error of the ℓ weeks ahead prediction distribution and c comes from the interval coverage probability assuming a normal distribution of forecast errors. Note that the prediction intervals are estimated from the seasonally-adjusted data but may be too narrow as they ignore uncertainty associated with the STL estimation of the seasonal components.

Quantile regression forest. Random forests are collections of bagged regression trees [14]. Upon sampling a random set of predictors, each tree generates one prediction for the next period's count of ILI positive visits. A random forest forecast consists of the average of the predictions of all trees.

We implemented an extension to random forests, called quantile regression forest (QRF), using the *quantregForest* R package [15]. QRF generalizes the usual random forest prediction by estimating conditional quantiles, useful when constructing prediction intervals. The predictors included in the QRF model were: lags of weekly ILI counts; current epi-week (for epi-week definition, see [16]); lags of weekly minimum and maximum temperature; lags of weekly flu immunization rate; and city population. Other predictors were considered but were excluded by recursive feature elimination using the rfe method of the R *caret* package [17]. These included weekly counts of diagnosed cases of the most common strain types, which are influenza A H1N1 and H3, influenza B, rhino/enterovirus, and respiratory syncytial virus. Information on patient age, sex, and location within city were also excluded since they rendered the model computationally expensive and yielded no performance gains. To illustrate the impact of these additional covariates on performance, S1 Fig shows the effect of adding patient sex, age, sex and age together, location within city, or strain types on the performance of the "baseline" QRF. The RMSE performance metric in this S1 and S2 Figs is described in the analysis section of the paper.

Our implemented QRF used $n = 2000$ decision trees, sampling $m = 4$ of the available predictors each time. The parameter m was chosen using the conventional heuristic $\sqrt{P}$, where P represents the number of predictors. The QRF prediction of ILI positive visits ℓ weeks from week t, $\hat{y}^{QRF}_{t+\ell}$, is given by:

$$\hat{y}^{QRF}_{t+\ell}(\mathbf{x}_t) = \frac{1}{n}\sum_{b=1}^{n} T_b(\mathbf{x}_t), \tag{4}$$

where each T_b is one of the n decision trees resulting from a bootstrapped sample of the training data with m randomly selected predictors and $\mathbf{x}_t$ are the values of the model predictors at week t. The prediction interval of the QRF stems from conditional quantiles computed by the *quantregForest* method in R. [18].

Linear regression. We fitted a standard linear regression (LR) model with the following predictor variables: current week's ILI positive visit count, weekly-averaged minimum and maximum temperature, city population, immunization rate, year trend, epi-week fixed effects, slope of the ILI curve at current period, and a categorical variable counting upward movements in weekly ILI positive visits over the preceding three weeks. The ILI slope and count of upward movements helped on detecting sharp increases in recent ILI positive visits and improved the prediction near the peak. The model's prediction of ILI positive visits ℓ weeks from week t is given by:

$$\hat{y}^{LR}_{t+\ell} = \hat{\alpha} + \mathbf{X_t}\hat{\boldsymbol{\beta}}, \tag{5}$$

where $\mathbf{X_t}$ is a row vector of the values for the predictors at week t, $\hat{\boldsymbol{\beta}}$ is a column vector with the estimated marginal effect of each predictor on the ℓ weeks ahead count of ILI positive visits and $\hat{\alpha}$ is the intercept estimate. The prediction interval is $\hat{y}_{t+\ell} \pm c\hat{\sigma}(\mathbf{X_t})$ where $\hat{\sigma}(\mathbf{X_t})$ is the standard error of the prediction given observed values $\mathbf{X_t}$ and c once again comes from the amount of coverage for the prediction interval under the assumption that errors are normally distributed.

Stacked ensemble. The stacked ensemble (SE) computes a weighted-average prediction $\hat{y}^{SE}_{t+\ell}$ based on the predictions of M contributing models $\{EB, QRF, ARIMA, LR\}$ described above:

$$\hat{y}^{SE}_{t+\ell} = \sum_{m=1}^{M} w_m \hat{y}^{m}_{t+\ell}, \text{ with } \sum_{m=1}^{M} w_m = 1 \text{ and } w_m >= 0 \; \forall \; m. \tag{6}$$

We used stacking, a data-driven approach, to find a set of weights w_m for combining the predictions of individual models in a manner that minimized prediction error in a held-out data set. Specifically, we found the weights:

$$\begin{gathered} \underset{\{w_m\}_{m=1}^{M}}{\arg\min} \sqrt{\frac{1}{H}\sum_{t=1}^{H}\left(y_t - \sum_{m=1}^{M} w_m \hat{f}^{m}(X_t \mid \hat{\theta}_{-t})\right)^2}, \\ s.t. \sum_{m=1}^{M} w_m = 1 \;, \; w_m \geq 0 \; \forall \; m, \end{gathered} \tag{7}$$

where H is the size of the hold-out data and $\hat{f}^{m}(X_t \mid \hat{\theta}_{-t})$ is the prediction at period t of model m trained without data from the flu season encompassing period t. In practice, the algorithm for deriving the SE weights is as follows:

1. Train each individual model to the dataset holding out all weeks comprising an flu season;
2. Obtain fitted values for the weeks in the hold-out data;
3. Repeat steps 1 and 2 using every other flu season as hold-out data;
4. Compute the weights using Eq 7 with the H weeks of predictions obtained in steps above.

The rationale for ensembling is its potential to reduce prediction error by reducing prediction variance and, in some cases, bias [19]. The benefits of ensembling are typically smaller as the predictions of individual models become more positively correlated. Furthermore, as shown by Claeskens *et al* [20], estimation of averaging weights introduces additional randomness to the ensembled prediction. As such, we also show results for a "naïve" ensemble using equal weights on each individual model, i.e. the case where $w_m = \frac{1}{M}$ for each of the M models contributing to the ensembled prediction.

Deriving the sampling distribution of $\hat{y}^{SE}_{t+\ell}$ is nontrivial as ensembling mixes the distributions of each contributing model and some individual models are non-parametric to begin with. The literature on model averaged prediction intervals is scarce, and even in simpler contexts with parametric models the constructed intervals perform poorly in terms of coverage rate on validation exercises [21]. We report a weighted average of quantiles of the distribution of each individual model prediction as the ensembles' own distributions. In our analysis, we show the coverage rate of this back-of-the-envelope ensemble prediction interval, as well as the coverage rate of intervals constructed from each individual model.

Analysis

Each model was trained on a partition of the available data and subsequently evaluated against held-out data. The model evaluation used a leave-one-out approach: each flu season was held out once and used as a test set, with the model being trained on the remaining seasons. In each instance, the test set was removed before executing the algorithm for the ensemble methods to ensure that test data did not contribute to the construction of the ensemble weights. We considered four commonly used metrics for comparing model performance: mean absolute error

(MAE); root mean square error (RMSE); mean absolute percentage error (MAPE); and log scoring. These are described below.

Mean absolute error & root mean square error. Our forecasting targets were the $\ell \in \{1, 2, 3, 4\}$ weeks ahead ILI positive visit counts. We summarized model performance for each target using standard measures of prediction error: MAE and RMSE. These are given by:

$$\mathrm{MAE(m)} = \frac{1}{T-\ell}\sum_{t=1}^{T-\ell} |y_{t+\ell} - \hat{y}^{m}_{t+\ell}|, \text{ and} \tag{8}$$

$$\mathrm{RMSE(m)} = \sqrt{\frac{1}{T-\ell}\sum_{t=1}^{T-\ell} (y_{t+\ell} - \hat{y}^{m}_{t+\ell})^2}, \tag{9}$$

where $y_{t+\ell}$ is the realized weekly ILI positive count at period $t + \ell$ and $\hat{y}^{m}_{t+\ell}$ is model m's generated ℓ weeks ahead prediction made at period t. MAE and RMSE express an *average* prediction error over the $T - \ell$ point forecasts performed in the sample. Both metrics are increasing with the average error, meaning the models with best accuracy in the test set are the ones with the lowest MAE and RMSE. Squaring of the error implies the RMSE imposes a greater penalty for larger deviations between predicted and observed values relative to the MAE. As such, a decision maker assigning higher importance to prediction accuracy near the flu season peak may prefer a model with lower RMSE.

Mean absolute percentage error. For the sake of interpretability, we also present each model's MAPE, which expresses the average percentage deviation between forecast and realized outcomes. The MAPE formula is given by:

$$\mathrm{MAPE(m)} = \frac{1}{T-\ell}\sum_{t=1}^{T-\ell} \left|\frac{y_{t+\ell} - \hat{y}^{m}_{t+\ell}}{y_{t+\ell}}\right|. \tag{10}$$

MAPE is known to produce infinite or undefined values when the denominator of any summation term in Eq 10 approaches zero [22]. This disadvantage did not apply to our analysis since we had a minimum of 52 weekly ILI visits to the ED during the sample period.

Log score. The last performance metric in our analysis is a variant of the log score measure used to rank participants on the United States Centers for Disease Control (US CDC) flu forecasting challenge [23]. The log score of a forecast measures how much probability our model assigns to an "acceptable" prediction range. While MAE, RMSE, and MAPE relate to a model's prediction accuracy, log scoring assesses the confidence that the model's probabilistic forecast falls within a tolerance level.

We defined a prediction as acceptable if it fell within ± 25 visits of the realized weekly ILI positive weekly count. Since we compare a mix of parametric and non-parametric models, we approximate the probability assigned to the acceptable range by the count of centiles of the forecast inside the range. That is, the log score of a ℓ weeks ahead prediction from model m at week t is computed as follows:

$$\mathrm{log\_score}(\hat{y}^{m}_{t+\ell}) = \log\left(\frac{1}{98}\sum_{c=1}^{98} \mathbb{1}(y_{t+\ell} + 25 \geq \hat{y}^{m}_{c,t+\ell} \geq y_{t+\ell} - 25)\right), \tag{11}$$

where $\hat{y}^{m}_{c,t+\ell}$ is the c centile of the (probabilistic) forecast generated by model m. In the event the observed ILI visits fell below the first centile or above the 99th centile, we assigned the value -5 to the log score, which is slightly lower than $\log(0.01) \approx -4.6$.

Descriptive statistics

Table 1 summarizes key variables used in the analysis. The high visit volume ED at ACH received on average 208 ILI positive visits per week during the sample period. This average increased almost 60% to 332 ILI positive visit counts during high visit volume weeks—i.e. those on the top 25% of ILI visit counts—within each season. This increased pattern was also visible for lab confirmed cases of different virus strains, except for rhino/enteroviruses which varied less predictably throughout the flu season. The last column of Table 1 shows the t statistic for the difference in means between high and low visit volume weeks—i.e. weeks on the bottom 75% of ILI visit counts—within each season. In addition to the mechanical increase in ILI during high volume weeks, we see as expected lower temperatures during high volume weeks, but little evidence of differences in patient demographic characteristics and flu immunization rates by visit volume.

Results

Stacked ensemble weights

Table 2 shows the optimal weights derived from the data through Eq 7. In all flu seasons and all forecast targets, the ensemble draws information from at least three of the individual models for its prediction. While the principle of ensembling is combining rather than selecting models, the weights obtained suggest ARIMA offers little contribution to the SE forecast.

Overall comparison

We compared each model's performance metrics over all prediction targets. Figs 1 and 2 show the MAE and RMSE of all models for the 1 to 4 weeks ahead predictions. These figures

Table 1. Descriptive statistics.

Sample → **↓ Variable**	**All weeks (N = 348)**		**High volume (N = 89)**		**Low volume (N = 259)**		**High-Low Volume**
	Mean	**Sd**	**Mean**	**Sd**	**Mean**	**Sd**	**t stat**
ILI-related variables							
ILI count	208.42	107.06	332.90	103.02	165.65	67.94	17.37***
Influenza A H1N1 count	9.36	28.68	28.71	50.85	2.71	7.12	8.02***
Influenza A H3 count	13.74	34.71	42.09	58.51	4.00	8.95	10.16***
Influenza B count	8.07	17.05	18.10	27.34	4.63	9.47	6.84***
Rhino/Enterovirus count	23.82	12.01	20.65	8.54	24.90	12.83	-2.91***
RSV count	13.44	20.05	33.65	25.40	6.49	11.47	13.65***
Environmental variables							
Minimum temperature	-5.86	11.24	-15.56	8.59	-2.53	10.05	-10.94***
Maximum temperature	17.01	9.94	9.20	7.08	19.70	9.35	-9.68***
Female rate	0.43	0.04	0.44	0.03	0.43	0.04	2.18**
Age 0-1 rate	0.42	0.06	0.41	0.07	0.43	0.06	-1.76*
Age 2-4 rate	0.32	0.04	0.33	0.03	0.32	0.05	3.55***
Age 5-8 rate	0.16	0.04	0.17	0.05	0.16	0.04	1.98**
Age 9-17 rate	0.09	0.03	0.08	0.03	0.10	0.03	-4.30***
Immunization rate	0.23	0.05	0.22	0.05	0.23	0.05	-0.96

Notes. (1) * p <.1, ** p <.05, *** p <.001. (2) Except for temperature information, all data is derived from visits screened ILI positive at the ACH ED. (3) The label "high volume" applies to weeks on the top 25% of ILI counts within each season. (4) Weekly counts of virus strains/subtypes represent the subset of ILI positive visits with lab confirmed diagnostics. (5) Each minimum/maximum temperature observation is a weekly average of daily temperature records. (6) Immunization rates represent the fraction of ILI positive visits of patients who received flu immunization within the last 365 days.

https://doi.org/10.1371/journal.pone.0241725.t001

Table 2. Stacked ensemble weights.

Flu Season → **↓ Model**	**2012-2013**	**2013-2014**	**2014-2015**	**2015-2016**	**2016-2017**	**2017-2018**	**2018-2019**
1-wk forecast							
Empirical Bayes	0.47	0.25	0.20	0.14	0.54	0.33	0.29
ARIMA	0.08	0.00	0.00	0.00	0.08	0.01	0.00
Quantile Regression Forest	0.16	0.28	0.49	0.64	0.15	0.23	0.39
Linear Regression	0.29	0.47	0.32	0.22	0.23	0.42	0.32
2-wks forecast							
Empirical Bayes	0.22	0.48	0.28	0.22	0.15	0.38	0.18
ARIMA	0.00	0.01	0.00	0.00	0.00	0.10	0.04
Quantile Regression Forest	0.56	0.17	0.25	0.48	0.64	0.11	0.25
Linear Regression	0.23	0.34	0.47	0.30	0.21	0.40	0.53
3-wks forecast							
Empirical Bayes	0.17	0.14	0.47	0.27	0.20	0.15	0.48
ARIMA	0.03	0.01	0.09	0.04	0.02	0.00	0.10
Quantile Regression Forest	0.46	0.62	0.15	0.25	0.44	0.57	0.13
Linear Regression	0.34	0.24	0.29	0.44	0.34	0.28	0.30
4-wks forecast							
Empirical Bayes	0.28	0.18	0.08	0.38	0.25	0.23	0.19
ARIMA	0.04	0.01	0.00	0.25	0.20	0.16	0.10
Quantile Regression Forest	0.24	0.50	0.68	0.25	0.22	0.34	0.44
Linear Regression	0.43	0.32	0.24	0.12	0.33	0.27	0.27

Notes. Reported values are rounded to two decimal places.

https://doi.org/10.1371/journal.pone.0241725.t002

illustrate how predictions worsen as we look further into the future. Both MAE and RMSE yield the same ordering of model performance, except for a higher RMSE on the 3 and 4 weeks forecast of the EB model. This suggests EB predictions are farther from the truth on the longer forecast horizons. Figs 1 and 2 also highlight the fact that predictions which draw information solely on the trajectory of ILI (EB and ARIMA forecasts) perform considerably worse than predictions which additionally model the relationship between ILI and environmental factors such as temperature and immunization rates. Fig 3 shows the best performing model's (SE) predictions deviate from the observed ILI count by an average of 12% for 1 week ahead and 19% for 4 weeks ahead predictions, in absolute terms.

This overall comparison of models suggests the SE was competitive across all performance metrics and all prediction targets. Predicted counts generated by this approach deviated, on average, by 30 and 50 weekly ILI visits in the nearest 1 week and farthest 4 weeks ahead targets, respectively. In fact, forecasts generated by combining predictions from various models are gaining traction as they tend to be on average more accurate than the individual models they combine.

Flu season breakdowns

The results presented in the overall comparison subsection paint a general picture about each model's ability to predict future ILI weekly counts. Table 3 provides a more granular perspective by examining model performance separately for each season. Table columns alternately report the models' mean log score for each season. The SE outperforms individual models for the 1 and 2 weeks ahead forecast, falling slightly behind the QRF on the more distant forecasts.

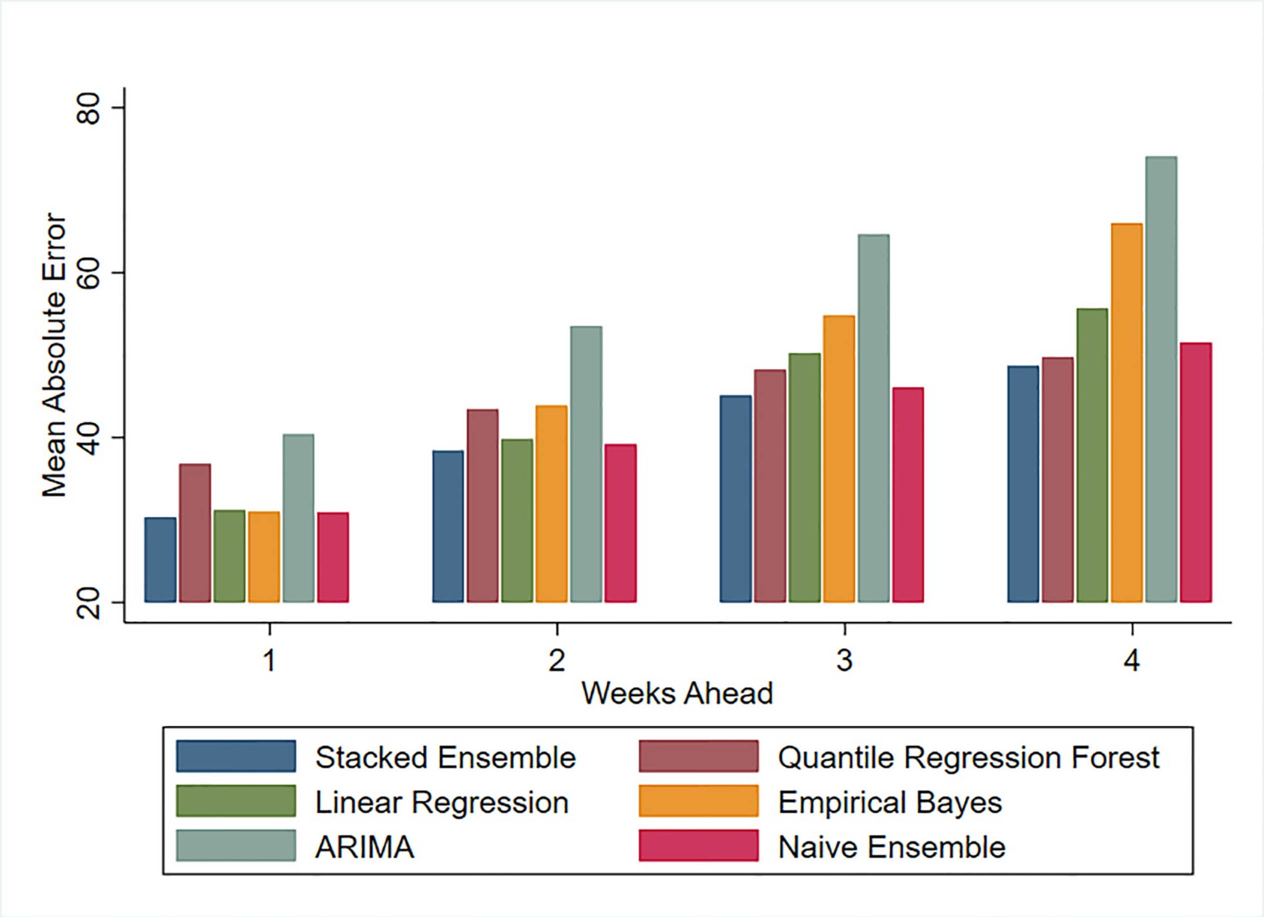

Fig 1. MAE comparison for test seasons 2012-2018.

https://doi.org/10.1371/journal.pone.0241725.g001

Table 4 reports for each model and flu season the percentage of predictions that fall within 20% of the observed weekly counts on short-term, 1 week ahead, forecasts. Each column represents a prediction target for a specific flu season. The SE method produced the highest percentage of predictions within 20% of observed visit counts in more seasons than any other model, also ranking highest on the last two columns aggregating predictions over all seasons.

Prediction interval coverage rate

Each model uses a different approach when characterizing the uncertainty of the predictions. Constructing prediction intervals for the ARIMA and LR models is straightforward and follows from estimates of the standard error of the prediction. Meanwhile the EB estimates have a prediction interval computed from quantiles of the posterior distribution and the QRF prediction interval stems from quantiles of the output from individual regression trees. Lastly the interval around the SE estimate (as well as naive ensemble) is a back-of-the-envelope calculation from weighting the quantiles of the predictions from individual models.

We assessed the coverage of each model's prediction interval on the complete test seasons from 2012-13 to 2017-18. Fig 4 shows, for each model and flu season, the 1 week ahead prediction against the realized ILI visit counts in that week. The shaded region corresponds to a 90% prediction interval generated by the model and the coverage rate refers to the percentage of

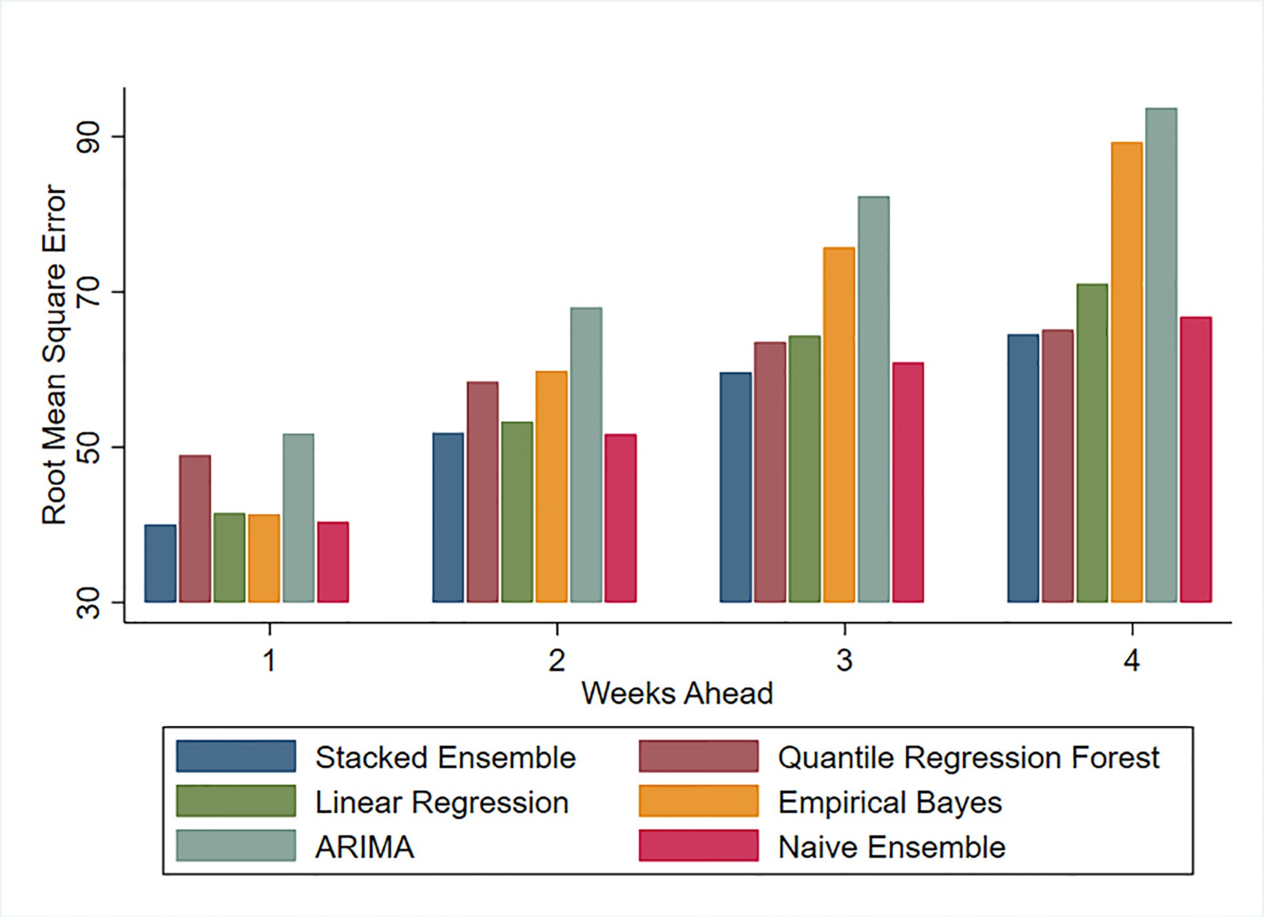

Fig 2. RMSE comparison for test seasons 2012-2018.

https://doi.org/10.1371/journal.pone.0241725.g002

observed weekly ILI visits that falls inside the model's prediction interval. The figures show ARIMA prediction intervals had the lowest coverage, potentially since uncertainty associated with the STL decomposition was ignored. Except for ARIMA and the EB model, the coverage rates of models were slightly higher than 90% and as such the intervals on our most competitive models can be viewed as conservative estimates.

Alternative data set

We evaluated whether our findings were robust to using data from the lower visit volume ED at FMC hospital, located in the vicinity of ACH in Calgary, Alberta. S2 Fig shows the RMSE of each model when trained and evaluated on data from the FMC ED. Results were qualitatively consistent with those in the main dataset, with the SE still outperforming the individual models. In addition to lower visit volumes, the FMC ED caters to an adult population, in contrast to the children's hospital used in the main analysis.

Discussion

We tested the performance of multiple models for predicting ILI visits at the ED level. Our findings promote stacked ensembling—a data-driven model averaging method—as a viable

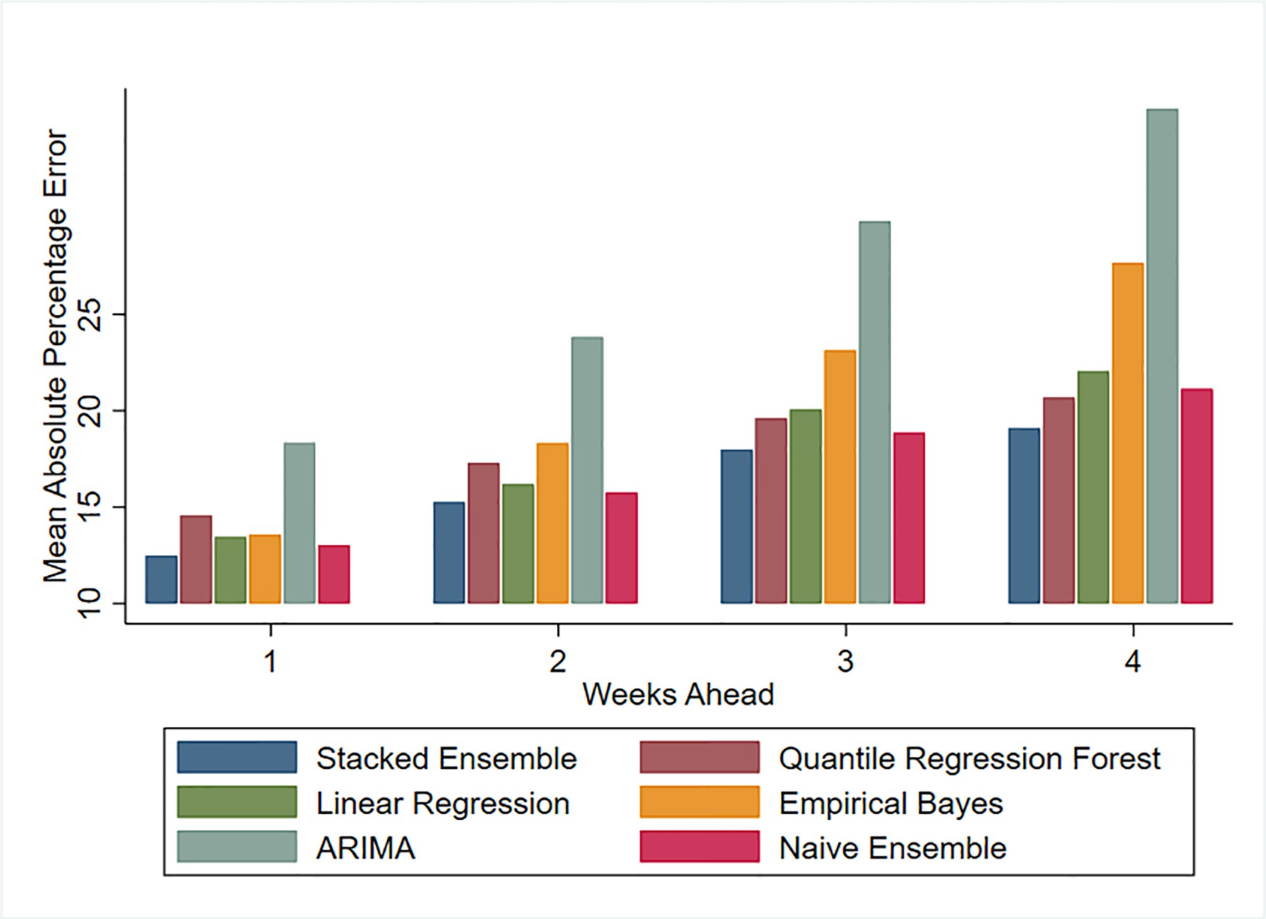

Fig 3. MAPE comparison for test seasons 2012-2018.

https://doi.org/10.1371/journal.pone.0241725.g003

approach for short term forecasts of weekly ILI visits. We show, through various exercises comparing model performance, that our SE prediction leverages the strength of individual models and therefore provides a robust estimate on both short-term (1 week) and longer-term (4 weeks) forecast horizons.

Previous research on forecasting ILI have discussed potential improvements of incorporating demographic and spatial information into the statistical models. We found no evidence of such improvements in our research design, possibly due to small sample size and the potential for overfitting when including additional predictors. Instead, our preferred model required only ED-level weekly data on ILI positive patient counts, the rate of these patients who were immunized in the past and current season, and temperature and population data which are typically available in urban areas. In addition, the tested models are easily implementable with free software for statistical computing.

We found that the predictions from our SE outperformed those of individual models in two distinct EDs located in Calgary, Alberta. Future work should seek refinements in the ensembling methodology. For instance, as the relative performance of each model may vary within a flu season, the SE can be refined by allowing weights to vary throughout the weeks. Other improvements may be achieved by including a larger sample of hospitals in the province and attempting to model cross effects of ILI visits in an ED on the prediction of ILI for neighboring

Table 3. Mean log score of each model, by flu season and forecast horizon.

Flu Season → **↓ Model**	**2012-13**	**2013-14**	**2014-15**	**2015-16**	**2016-17**	**2017-18**	**2018-19**	**All Seasons**
1 week ahead forecast								
Empirical Bayes	-1.13	-1.01	-1.31	-1.54	-1.11	-1.12	-2.47	-1.27
ARIMA	-2.10	-1.81	-1.77	-1.59	-1.66	-1.46	-4.42	-1.88
Quantile Regression Forest	-1.10	-0.94	-1.14	-1.11	-1.21	-1.26	-2.87	-1.22
Linear Regression	-1.17	-0.95	-1.18	-1.53	-1.08	-1.06	-2.63	-1.24
Stacked Ensemble	**-0.96**	**-0.89**	-1.06	**-1.10**	**-1.02**	-0.99	**-2.24**	**-1.07**
Naive Ensemble	-1.02	-0.92	**-1.05**	**-1.10**	-1.03	**-0.98**	-2.58	-1.10
2 weeks ahead forecast								
Empirical Bayes	-1.39	-1.18	-1.39	-1.58	-1.52	-1.37	**-2.77**	-1.48
ARIMA	-2.66	-2.62	-1.53	-1.97	-2.27	-1.56	-5.84	-2.31
Quantile Regression Forest	-1.20	**-1.10**	-1.27	-1.26	-1.32	-1.41	-3.51	-1.38
Linear Regression	-1.24	-1.20	-1.41	-1.63	-1.43	-1.20	-3.83	-1.49
Stacked Ensemble	**-1.15**	**-1.10**	-1.18	-1.26	-1.22	-1.12	-3.35	**-1.29**
Naive Ensemble	-1.21	-1.18	**-1.14**	**-1.21**	**-1.21**	**-1.08**	-3.39	**-1.29**
3 weeks ahead forecast								
Empirical Bayes	-1.60	-1.42	-1.95	-1.83	-1.72	-1.35	-3.42	-1.74
ARIMA	-3.28	-3.26	-1.93	-2.32	-2.86	-1.57	-5.71	-2.71
Quantile Regression Forest	**-1.19**	-1.14	**-1.34**	**-1.36**	**-1.32**	-1.47	**-3.19**	**-1.41**
Linear Regression	-1.41	-1.38	-1.38	-1.80	-1.67	-1.33	-5.07	-1.69
Stacked Ensemble	-1.24	**-1.11**	-1.38	-1.46	-1.38	-1.26	-3.57	-1.43
Naive Ensemble	-1.37	-1.32	-1.27	-1.39	-1.39	**-1.19**	-4.00	-1.47
4 weeks ahead forecast								
Empirical Bayes	-1.78	-1.58	-1.84	-1.97	-1.84	-1.52	-3.62	-1.86
ARIMA	-3.95	-3.82	-1.94	-2.55	-3.05	-1.81	-6.66	-3.06
Quantile Regression Forest	**-1.24**	**-1.14**	-1.29	**-1.44**	**-1.34**	-1.41	**-3.02**	**-1.41**
Linear Regression	-1.54	-1.45	-1.43	-1.77	-1.78	-1.43	-5.04	-1.76
Stacked Ensemble	-1.40	-1.21	**-1.25**	-1.53	-1.53	**-1.27**	-3.59	-1.49
Naive Ensemble	-1.45	-1.38	-1.27	-1.51	-1.53	**-1.27**	-4.18	-1.55

Notes. Due to data availability, the 2018 season uses only 14 weeks of data, starting from epiweek 37 or approximately mid September.

https://doi.org/10.1371/journal.pone.0241725.t003

Table 4. Percentage of 1 week ahead predicted ILI visits within 20% of observed visits, by flu season.

Flu Season → **↓ Model**	**2012-13**	**2013-14**	**2014-15**	**2015-16**	**2016-17**	**2017-18**	**2018-19**	**All Seasons**
All weeks in sample								
Empirical Bayes	76.7	76.7	69.8	67.4	79.1	88.4	**80.0**	76.6
ARIMA	48.8	58.1	74.4	**79.1**	76.7	86.0	66.7	70.3
Quantile Regression Forest	60.5	**88.4**	69.8	**79.1**	**86.0**	90.7	60.0	78.0
Linear Regression	67.4	83.7	74.4	74.4	81.4	93.0	66.7	78.4
Stacked Ensemble	**79.1**	83.7	**79.1**	72.1	**86.0**	**95.3**	**80.0**	**82.4**
Naive Ensemble	72.1	83.7	74.4	69.8	83.7	**95.3**	73.3	79.5
High volume weeks								
Empirical Bayes	63.6	83.3	**90.9**	81.8	81.8	90.9	**100.0**	83.1
ARIMA	54.5	58.3	81.8	81.8	72.7	90.9	75.0	73.2
Quantile Regression Forest	72.7	83.3	**90.9**	**90.9**	81.8	**100.0**	25.0	83.1
Linear Regression	63.6	**100.0**	72.7	81.8	**90.9**	**100.0**	75.0	84.5
Stacked Ensemble	**81.8**	83.3	**90.9**	**90.9**	**90.9**	**100.0**	75.0	**88.7**
Naive Ensemble	**81.8**	91.7	**90.9**	81.8	81.8	**100.0**	75.0	87.3

Notes. (1) Due to data availability, the 2018 season uses only 14 weeks of data, starting from epiweek 37 or approximately mid September.

https://doi.org/10.1371/journal.pone.0241725.t004

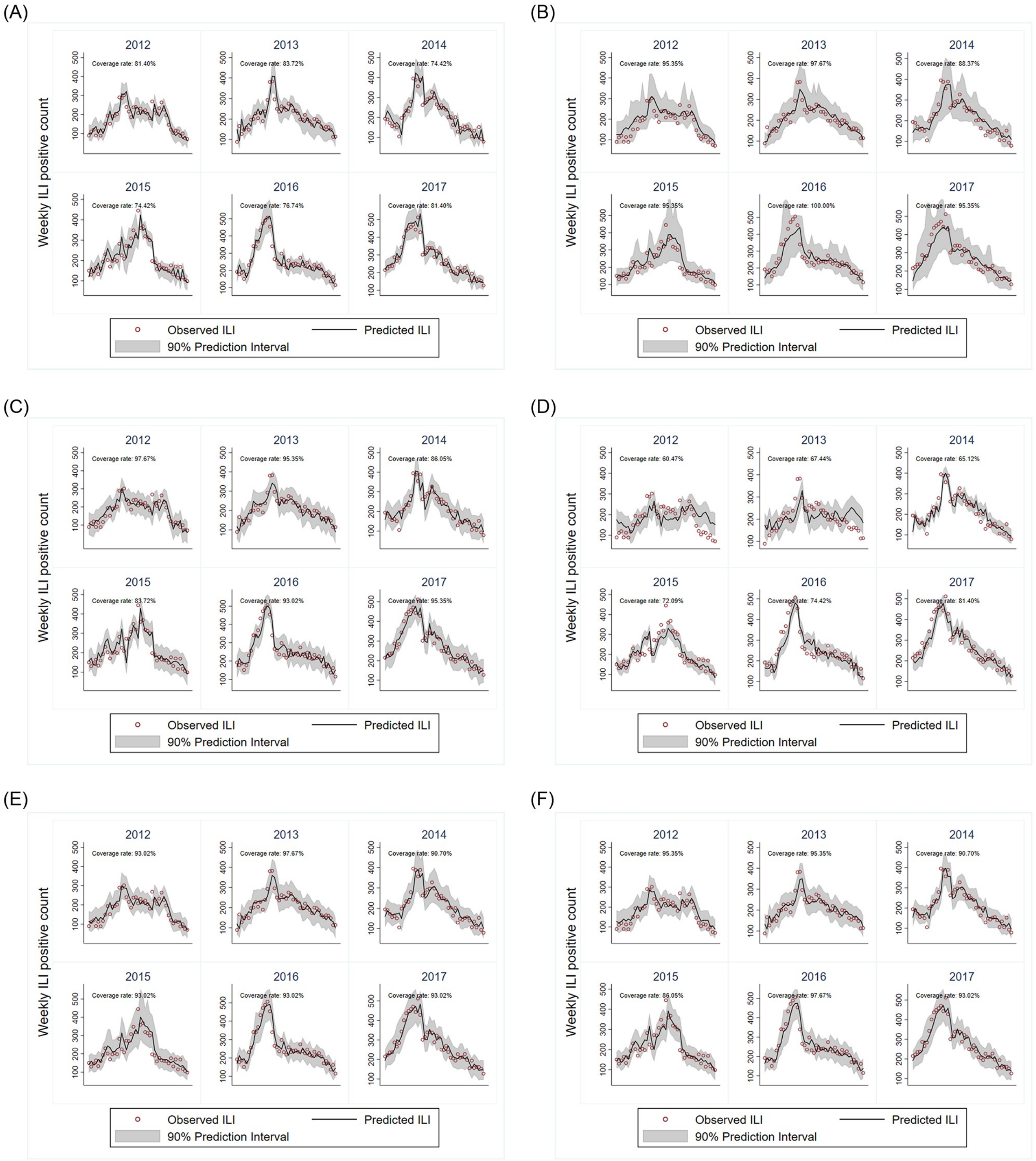

Fig 4. Predicted vs observed ILI visits for 1 week forecast. A: Empirical Bayes. B: Quantile Regression Forest. C: Linear Regression. D: ARIMA. E: Stacked Ensemble. F: Naive Ensemble.

https://doi.org/10.1371/journal.pone.0241725.g004

EDs. This could potentially improve predictive power at lower visit volume EDs where outbreak detection could be of relevance.

Supporting information

S1 Fig. RMSE comparison for quantile regression forests with added features. Notes. (1) Data from weekly ILI visits at the ACH. (2) The Baseline QRF is described in section 2.2; the other QRF specifications build on the Baseline. E.g., Sex means we disaggregate weekly ILI counts by male and female patients. (3) Calculation of RMSE uses 42 weeks on each test season of the ACH data, starting from epiweek 37 or approximately mid September. (4) Due to data availability, the 2018 season uses only 14 weeks of data, also starting from epiweek 37.
(TIF)

S2 Fig. RMSE comparison for test seasons 2012-2018 (FMC hospital).
(TIF)

S1 Data.
(ZIP)

Acknowledgments

We thank members of Public Health Surveillance and Infrastructure at Alberta Health Services. Hussain Usman, Adrienne Macdonald and Jason Cabaj provided helpful input and feedback during the model development stage. We also thank Sherry Trithart and Shaun Malo for data provision from ARTSSN.

Author Contributions

Data curation: Vineet Saini.

Formal analysis: Arthur Novaes de Amorim.

Methodology: Rob Deardon.

Project administration: Vineet Saini.

Supervision: Rob Deardon.

Visualization: Arthur Novaes de Amorim.

Writing – original draft: Arthur Novaes de Amorim, Rob Deardon, Vineet Saini.

References

1. Menec V. The Impact of Influenza-Like Illness on the Winnipeg Health Care System: Is an Early Warning System Possible? Manitoba Centre for Health Policy and Evaluation; 2001.
2. Molinari NAM, Ortega-Sanchez IR, Messonnier ML, Thompson WW, Wortley PM, Weintraub E, et al. The annual impact of seasonal influenza in the US: measuring disease burden and costs. Vaccine. 2007; 25(27):5086–5096. https://doi.org/10.1016/j.vaccine.2007.03.046 PMID: 17544181
3. Viboud C, Boëlle PY, Carrat F, Valleron AJ, Flahault A. Prediction of the spread of influenza epidemics by the method of analogues. American Journal of Epidemiology. 2003; 158(10):996–1006. https://doi.org/10.1093/aje/kwg239
4. Kandula S, Yamana T, Pei S, Yang W, Morita H, Shaman J. Evaluation of mechanistic and statistical methods in forecasting influenza-like illness. Journal of The Royal Society Interface. 2018; 15 (144):20180174. https://doi.org/10.1098/rsif.2018.0174
5. Ward MA, Stanley A, Deeth LE, Deardon R, Feng Z, Trotz-Williams LA. Methods for detecting seasonal influenza epidemics using a school absenteeism surveillance system. BMC public health. 2019; 19 (1):1232. https://doi.org/10.1186/s12889-019-7521-7

6. Yang W, Olson DR, Shaman J. Forecasting influenza outbreaks in boroughs and neighborhoods of New York City. PLoS computational biology. 2016; 12(11). https://doi.org/10.1371/journal.pcbi.1005201 PMID: 27855155

7. Krishnamurti TN, Kishtawal CM, Zhang Z, LaRow T, Bachiochi D, Williford E, et al. Multimodel ensemble forecasts for weather and seasonal climate. Journal of Climate. 2000; 13(23):4196–4216. https://doi.org/10.1175/1520-0442(2000)013%3C4196:MEFFWA%3E2.0.CO;2

8. Ray EL, Reich NG. Prediction of infectious disease epidemics via weighted density ensembles. PLoS computational biology. 2018; 14(2):e1005910. https://doi.org/10.1371/journal.pcbi.1005910

9. Brooks LC, Farrow DC, Hyun S, Tibshirani RJ, Rosenfeld R. Flexible modeling of epidemics with an empirical Bayes framework. PLoS computational biology. 2015; 11(8):e1004382. https://doi.org/10.1371/journal.pcbi.1004382

10. Farrow D. Modeling the past, present, and future of influenza. Phd thesis. 2016;.

11. Arnold TB, Tibshirani RJ. Path algorithm for generalized lasso problems; 2020.

12. Hyndman RJ, Athanasopoulos G, Bergmeir C, Caceres G, Chhay L, O'Hara-Wild M, et al. Forecasting functions for time series and linear models; 2020.

13. Hyndman RJ, Khandakar Y. Automatic time series for forecasting: the forecast package for R. Monash University, Department of Econometrics and Business Statistics; 2007.

14. Breiman L. Random forests. Machine learning. 2001; 45(1):5–32. https://doi.org/10.1023/A:1010933404324

15. Meinshausen N. Quantile regression forests; 2016.

16. Grolemund G, Wickham H. Dates and times made easy with lubridate. Journal of statistical software. 2011; 40(3):1–25. https://doi.org/10.18637/jss.v040.i03

17. Kuhn M. Building predictive models in R using the caret package. Journal of statistical software. 2008; 28(5):1–26. https://doi.org/10.18637/jss.v028.i05

18. Meinshausen N. Quantile regression forests. Journal of Machine Learning Research. 2006; 7 (Jun):983–999.

19. Dormann CF, Calabrese JM, Guillera-Arroita G, Matechou E, Bahn V, Bartoń K, et al. Model averaging in ecology: A review of Bayesian, information-theoretic, and tactical approaches for predictive inference. Ecological Monographs. 2018; 88(4):485–504. https://doi.org/10.1002/ecm.1309

20. Claeskens G, Magnus JR, Vasnev AL, Wang W. The forecast combination puzzle: A simple theoretical explanation. International Journal of Forecasting. 2016; 32(3):754–762. https://doi.org/10.1016/j.ijforecast.2015.12.005

21. Claeskens G, Hjort NL. Model selection and model averaging. Cambridge Books. 2008;.

22. Kim S, Kim H. A new metric of absolute percentage error for intermittent demand forecasts. International Journal of Forecasting. 2016; 32(3):669–679. https://doi.org/10.1016/j.ijforecast.2015.12.003

23. Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. Journal of the American statistical Association. 2007; 102(477):359–378. https://doi.org/10.1198/016214506000001437