Forecasting Influenza in the U.S. with a Collaborative Ensemble from the FluSight Network

Nicholas G Reich, Logan Brooks, Sasikiran Kandula, Craig McGowan, Dave Osthus, Evan Ray, Abhinav Tushar, Teresa Yamana, Jeff Shaman, Roni Rosenfeld

November 2017

Abstract

Problem: Our aim is to combine forecasting models for seasonal influenza in the US to create a single ensemble forecast. The central question is **can we provide better information to decision makers by combining forecasting models**, and specifically by using past performance of the models to inform the ensemble approach.

Materials and Methods: The FluSight Network is a multi-institution and multi-disciplinary consortium of teams that have participated in past CDC FluSight challenges. Prior to the start of the 2017/2018 influenza season in the US, we assembled 21 distinct forecasting models for influenza, each with forecasts from the last seven influenza seasons in the US. Subsequently, we conducted a cross-validation study to compare five different methods for combining these models into a single ensemble forecast.

Conclusions: In our study, across the past seven seasons, four of our collaborative ensemble methods had higher average scores than any of the individual forecasting model. We chose the best performing ensemble model and are submitting forecasts from this model each week to the CDC 2017/2018 FluSight Challenge.

Executive Summary

In the 2016/2017 influenza season, the CDC ran the 4th annual FluSight competition and received 28 submissions from 19 teams. During the 2016/2017 season, analysts at the CDC built an ensemble model that combined all of the submitted models by taking the average forecast for each influenza target. This model was one of the top performing models for the entire season.

In March 2017, a group of influenza forecasters who have participated in this challenge in past seasons established the FluSight Network, a multi-institution and multi-disciplinary consortium of forecasting teams. Out of the previous participants, 4 groups decided to participate and contributed 21 models using a diverse array of methodologies. This group worked throughout 2017 to create a set of guidelines and an experimental design that would enable submission of a publicly available, multi-team, real-time submission of a collaborative ensemble model with validated and performance-based weights for each model (i.e. not a simple average of models).

This document provides a high-level overview of that effort, highlighting the results and documenting the chosen model that was designated for real-time submission during the 2017/2018 U.S. influenza season.

FluSight Network Participants for 2017/2018 season

Institution	No. of models	Team leaders
Delphi team at Carnegie Mellon	9	Logan Brooks, Roni Rosenfeld
Columbia University	7	Teresa Yamana, Sasikiran Kandula, Jeff Shaman
Los Alamos National Laboratory	1	Dave Osthus
Reich Lab at UMass-Amherst	4	Nicholas Reich, Abhinav Tushar, Evan Ray

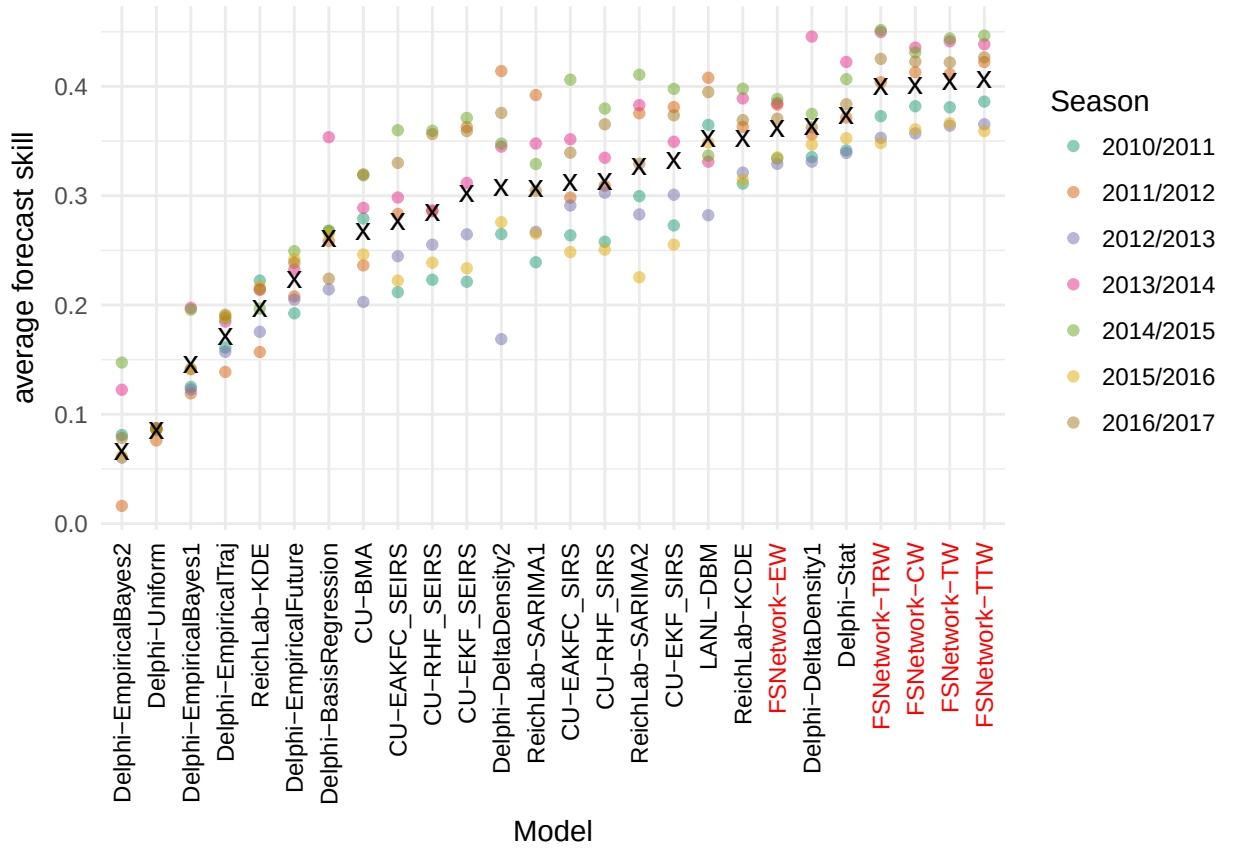

Figure 1: Average performance for all models, by season. The average forecast skill for each model within a season is plotted with a colored dot. The average across all seasons is shown with a black 'x'. Higher values indicate more accurate predictive performance, as they are a measure of how much probability on average a forecast from the given model assigned to the eventually observed value. The FluSightNetwork ensemble models are highlighted in red text. Models are sorted left to right in order of increasing accuracy.

Choosing an Ensemble Model for Real-time Influenza Forecasting

In early November 2017, the FluSight Network team chose a single model to submit to the CDC throughout the 2017/2018 influenza season. This model had the highest overall score among all individual component models and ensemble models examined (Figure 1). This model is called the "target-type weight" (TTW) model because it assigns each model an individual weight for each target-type (see next section for details). Forecasts from this model have been, and continue to be, submitted to the CDC in real-time, starting on November 6, 2017. They may be viewed at a public website by visiting: https://flusightnetwork.io.

We use an algorithm to estimate an optimal distribution of weights for the 21 different component models (Figure 2). For example, the `DeltaDensity1` model from the Delphi team is given 42% of the weight when creating ensemble predictions for incidence in the next four weeks and is given 12% of the weight for the seasonal targets (onset week, peak week, peak incidence). For the week-ahead forecasts, 8 models are given at least 1% weight. For the seasonal forecasts, 6 models are given at least 11% weight and no other models are given more than 1% weight. In total, 11 of the models are weighted by at least 1% for at least one of the target-types.

While the weights are optimized to choose the best combination, they should not be interpreted as a ranking of models. For example, if two models make very similar forecasts but one is always a little bit better, it is possible that the slightly worse model will receive very little weight since most of the information it has to contribute is already contained in the better model. Typically, the ensemble will choose a set of models that contribute different information to the forecast, as this "diversity of opinion" will improve the forecast.

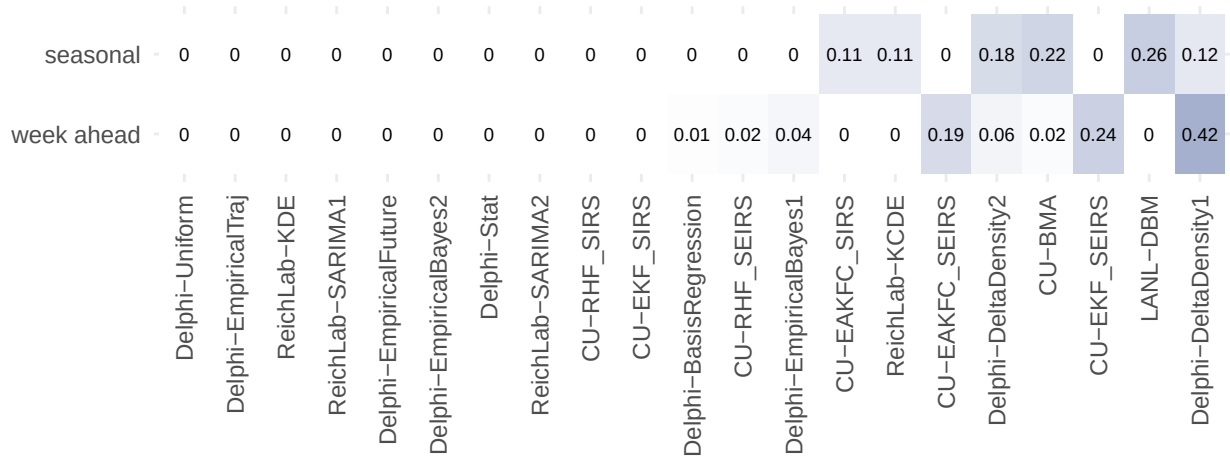

Figure 2: Estimated model weights by target type. The number in each cell corresponds to the weight assigned to the model in each column and the target in each row. The weights in each row sum to 1. These weights are used to create the weighted average ensemble model for the 2017/2018 season.

Technical details

Targets: For every week in a season, each component model submission contains forecasts for seven targets of public health interest specified by the CDC for each of the 11 HHS regions. The region-level targets are: weighted influenza-like-illness (wILI) in each of the next four weeks of the season, the week of season onset, the week in which the peak wILI occurs, and the level of the peak wILI. Forecasts within 0.5 percentage points of the target wILI and within 1 week of the weekly targets are given full credit for having been "correct".

Ensemble specifications: All of our ensemble models are built by taking weighted averages of the component models. We examined the performance of five different possible ensemble specifications (see table below). The "equal weights" model takes a simple average of all of the models, with no consideration of past performance. The other four approaches estimated weights for models based on past performance.

Model	No. of weights	description
Equal weights (EW)	1	Every model gets same weight.
Constant weights (CW)	21	Every model gets a single weight, not necessarily the same.
Target-type-based weights (TTW)	42	Two sets of weights, one for seasonal targets and one for weekly wILI targets.
Target-based weights (TW)	147	Seven sets of weights, one for each target separately.
Target-and-region-based weights (TRW)	1,617	Target-based weights estimated separately for each region.

Forecast Evaluation: We measured performance by (1) comparing the average score across all targets and all relevant weeks in the last seven seasons and (2) comparing the variability in average score. The variability is important because a model can achieve good average performance by having a model that captures typical trends fairly well, but in a season showing unusual timing or dynamics it might fail. We want to ensure that we choose a model that shows good average performance but also is consistently good, especially in unusual seasons.

For submitting in real-time in 2017-2018, we selected the ensemble model that achieved the best overall score in a cross-validation experiment over the last seven seasons. This was the target-type-based model that assigned one set of weights to each component model for the weekly incidence targets and another set of weights for the seasonal targets (onset timing, peak timing, and peak incidence).