Predicting Seasonal Influenza Hospitalizations

Kevin W. McConeghy, Jason R. Gantenberg, Andrew R. Zullo, Chanelle J. Howe, . . . (author list tentative)

Compiled: 2019-08-15

Purpose

Question 1

Which of a set of candidate statistical models most closely fits target features in a distribution of hypothetical influenza hospitalization curves?

Secondary question

Stratified by the Centers for Disease Control and Prevention's (CDC) season severity designation (Centers for Disease Control and Prevention, 2016; Biggerstaff *et al.*, 2018), which of a set of candidate statistical models most closely fits target features in a distribution of hypothetical influenza hospitalization curves?

Background

MMWR Weeks

All weeks will be specified based on the *MMWR* Week convention (Centers for Disease Control and Prevention, n.d.). Referred to as "epiweeks", *MMWR* weeks are integer values assigned to each week of the year, ranging from 1–53 (Centers for Disease Control and Prevention, n.d.). A typical flu season begins in epiweek 40. For the purposes of our analysis, we will renumber epiweeks to begin at 1 in epiweek 40 and end at 52 (or 53, in leap years) in epiweek 39.

Season severity

Recently, the CDC developed a methodology to designate the severity of each flu season based on hospitalizations and mortality (Biggerstaff *et al.*, 2018; Centers for Disease Control and Prevention, 2018). Severity designations are assigned by age group (child, adult, over 65) and overall (Biggerstaff *et al.*, 2018). For the purposes of the current study, we will match historical seasons with their corresponding overall severity designations.

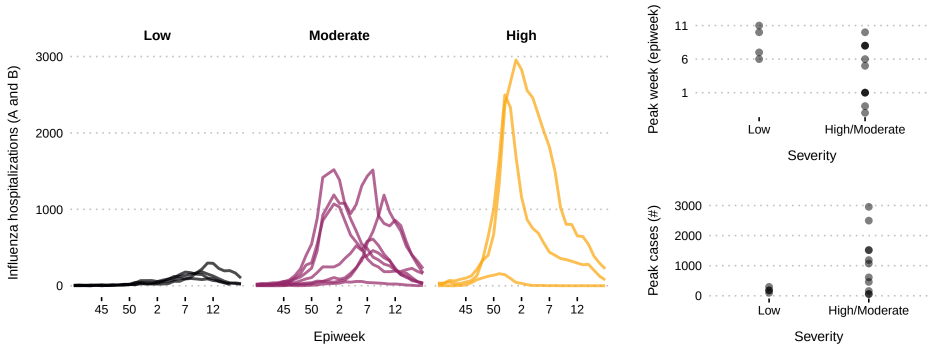

Figure 1: Empirical hospitalization curves, peak weeks, and peak number of cases — 2003–2017 (Source: FluSurv-NET, CDC). Data excludes 2009–2010 pandemic influenza season and 2018–2019 due to no official severity designation.

General approach

1. Fit a quadratic trend filter model (Kim *et al.*, 2009; Tibshirani, 2014; Brooks *et al.*, 2015) to empirical FluSurv-NET hospitalization data for the seasons 2003–2004 through 2017–2018, based in part on the approach described by Brooks et al. (Brooks *et al.*, 2015).

2. Simulate N hypothetical influenza hospitalization curves using the empirical Bayes approach described in *Curve fitting and simulation*.

3. For each candidate statistical model, find the set of variables that results in best fit to the distribution of hypothetical influenza hospitalization curves. Compare the best model for each model type against pre-specified target parameters to determine which candidate statistical model performs best across these targets.

Targets

For each of the following target parameters, we will calculate the root mean squared errors (RMSE), Bayesian information criterion (BIC), and relative bias for each of the best-fitting candidate statistical model types:

1. *Peak height* - largest number of hospitalizations in a single week
2. *Peak week* - the week in which this peak height occurred
3. *Total hospitalizations* - the cumulative number of hospitalizations in the season

These targets follow in part from (Brooks *et al.*, 2015), who focused on influenza-like illness rather than hospitalizations.

Curve fitting and simulation

We will simulate hypothetical influenza hospitalization curves using a modified version of the curve-fitting approach described by Brooks et al. (Brooks *et al.*, 2015).

First, using the `glmgen` (Brooks *et al.*, 2015) package in R, we will fit a quadratic trend filter (Kim *et al.*, 2009; Tibshirani, 2014; Brooks *et al.*, 2015) to each empirical hospitalization curve released by the CDC (Centers for Disease Control and Prevention, 2016), beginning with the 2003–2004 season.

For each season s and week i, Brooks et al. conceptualize a seasonal influenza curve as some function plus noise (Brooks *et al.*, 2015)[i]:

$$y_i^s = f^s(i) + \epsilon_i^s, \epsilon \sim N(0, \tau^s),$$

where

$$f^s(i) = \frac{\theta}{\max_j f(j)} \left[f\left(\frac{i-\mu}{v} + \overset{\arg\max j}{f}(j) \right) \right].$$

For each empirical season s, we fit a quadratic trend filter and average the residuals over i to estimate τ^s:

$$(\hat{\tau}^s)^2 = \underset{i}{\text{avg}} \left[y_i^s - \hat{f}^s(i) \right]^2.$$

We use the fitted trend filters in the curve simulation procedure as $f(j))$, where j represents the indicator for a given week in the trend filter model for the corresponding season.[ii] introducing noise for each simulated weekly hospitalization count based this estimate of τ^2.

We alter the original curve formula to impose a lower bound of 0 via the following transformation of $\hat{y}_i^s$, denoted below as $\hat{z}_i^s$:

$$\hat{z}_i^s = 0.5 \bigg(|\hat{y}_i^s| + \hat{y}_i^s \bigg)$$

Parameters in quadratic trend filtering model

The hypothetical influenza hospitalization curves are simulated using the following sampling scheme for each parameter used to simulate the hospitalization rate in week i, using the shape of season s as a starting point (y_i^s). Note that all notation included below either is adapted from or appears in (Brooks *et al.*, 2015).

$$\big\langle f, o, \nu, \theta, \mu \big\rangle$$

[i] Brook et al. use their method to forecast a current flu season, where b represents the current season's epidemic threshold of weighted influenza-like illness percent. Because hospitalization curves have a lower bound of 0, we drop b from the original equation.

[ii] In the Brooks paper, what j stood for was not defined explicitly. After studying the paper in detail and simulating my own curves, I believe the only sensible interpretation is the one provided in this analysis plan.

f : Shape

$f \sim U\{\hat{f} : \text{historical season } s\}$

σ : Noise

$\sigma \sim U\{\hat{\tau}^s : \text{historical season } s\}$

θ : Peak height

$\theta \sim U\left[\theta_m, \theta_M\right]$[iii]

Results in the following curve:

$f_3(i) = f_2(i - \mu + \text{arg max}_j f_2(j))$

ν : Pacing

Curve-stretching around peak week:

$\nu \sim U[0.75, 1.25]$

Results in following curve:

$f_4(i) = f_3\left(\frac{i - \text{arg max}_j\ f_3(j)}{\nu} + \text{arg max}_j\ f_3(j)\right)$

Candidate models

Serfling model (least squares)

Modified from (McConeghy *et al.*, n.d.):

$$Y = \beta_0\alpha + \beta_1 t + \beta_r X_r + ... + \beta_p cos\left(\frac{2\pi t}{52}\right) + \beta_q sin\left(\frac{2\pi t}{52}\right)$$

Where t = time (week), subscript r denotes a vector of β coefficients and variables, and subscripts p and q take on particular numbers based on the length of r.

Modified Serfling model

Modified from (McConeghy *et al.*, n.d.):

$$y = \alpha_0 + \beta_1 t + \beta_2 Flu_t + \beta_p X_p + ... + sin\left(\frac{2\pi t}{period}\right) + cos\left(\frac{2\pi t}{period}\right) + u$$

Where t = time (week) and subscript p denotes a vector of β coefficients and variables.

[iii] Brooks et al. say they get “unbiased estimators” for the minimum (θ_m) and maximum (θ_M) peak heights, respectively. However, given their notation does not seem to indicate that these parameters are estimates, I am using the observed minimum and maximum heights from the CDC data.

Generalized additive model (Prophet)

This model will be implemented using the R package *prophet* (Taylor & Letham, 2018), which implements a generalized additive modeling approach developed by engineers at Facebook, Inc.

The general form of the equation that will be fit to the hypothetical influenza hospitalization curves:

$$y(t) = g(t) + ... + h(t) + \epsilon_t,$$

where $g(t)$ models nonperiod trends and $h(t)$ stands for a vector of holidays or other events that are known to correlate with flu hospitalization (Brooks *et al.*, 2015). As with the Serfling and modified Serfling models, the Prophet model will include additional model terms included to improve fit.

Model terms

The following model terms will be entered as predictor variables to be considered for each model:

a) Cyclical terms (Serfling, Fourier, etc.)

b) Historical (empirical) flu hospitalizations

c) Historical data on viral activity (NREVSS), outpatient surveillance (ILI-Net)

d) Average weekly temperature

e) Climate factors (e.g., prior summer temperatures)

f) Lags and leads of c) or d)

g) Indicators for weeks of Thanksgiving and/or Christmas (Brooks *et al.*, 2015; Taylor & Letham, 2018)

Automated variable selection be used to select the best-fitting model for each model type using 2, 3, 4, and 5 prior seasons to predict the n^{th} season's curve.

In each case, prior seasons will be drawn from the distribution of hypothetical hospitalization curves and concatenated.

Goodness of fit

- Root mean squared error (RMSE)
- Relative bias

Measures of fit will be calculated for each model against each of the target parameters (*i.e.*, peak height, peak week, cumulative hospitalizations).[iv]

[iv] Question for Dr. Naimi: Should we be doing cross-validation or sample-splitting given the aims of this analysis? We can simulate an arbitrary number of hypothetical hospitalization curves, so the limited number of historical flu season available may not be an issue.

Sensitivity analysis

Challenge

The composition of institutions reflected in the FluSurv-NET has changed over time (Kandula *et al.*, 2019; Centers for Disease Control and Prevention, n.d.), meaning the FluSurv-NET estimates for influenza-related hospitalizations may not be comparable across time.

Response

Redo the analysis three times: one for each set of years in which the same participating institutions reported flu data to CDC. See (Centers for Disease Control and Prevention, n.d.) for more information.

References

Biggerstaff, M., Kniss, K., Jernigan, D.B., Brammer, L., Bresee, J., Garg, S., Burns, E., & Reed, C. (2018) Systematic assessment of multiple routine and near Real-Time indicators to classify the severity of influenza seasons and pandemics in the united states, 2003-2004 through 2015-2016. *Am. J. Epidemiol.*, **187**, 1040–1050.

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**, e1004382.

Centers for Disease Control and Prevention (2016) Laboratory-confirmed influenza hospitalizations.

Centers for Disease Control and Prevention (2018) How CDC classifies flu severity.

Centers for Disease Control and Prevention (n.d.) MMWR week overview.

Centers for Disease Control and Prevention (n.d.) Flu view phase 3 quick reference guide.

Kandula, S., Pei, S., & Shaman, J. (2019) Improved forecasts of influenza-associated hospitalization rates with google search trends. *J. R. Soc. Interface*, **16**, 20190080.

Kim, S., Koh, K., Boyd, S., & Gorinevsky, D. (2009) ℓ_1 Trend filtering. *SIAM Rev.*, **51**, 339–360.

McConeghy, K.W., Aalst, R. van, Zullo, A.R., & Joyce, N. (n.d.) An R package for estimating attributable influenza morbidity and mortality.

Taylor, S.J. & Letham, B. (2018) Forecasting at scale. *Am. Stat.*, **72**, 37–45.

Tibshirani, R.J. (2014) Adaptive piecewise polynomial estimation via trend filtering. *Ann. Stat.*, **42**, 285–323.