Science
Translational
Medicine
AAAS

www.sciencetranslationalmedicine.org/cgi/content/full/9/413/eaan5325/DC1

Supplementary Materials for

Evolution-informed forecasting of seasonal influenza A (H3N2)

Xiangjun Du, Aaron A. King, Robert J. Woods, Mercedes Pascual*

*Corresponding author. Email: pascualmm@uchicago.edu

Published 25 October 2017, *Sci. Transl. Med.* **9**, eaan5325 (2017)
DOI: 10.1126/scitranslmed.aan5325

This PDF file includes:

Table S4. Model comparison based on data from the U.S. HHS region 3.
Table S5. H3N2 risk level forecasts for the U.S. HHS region 3.

ADDITIONAL DESCRIPTION OF MATERIALS AND METHODS

Data

Outpatient illness surveillance data include information on patient visits to health care providers for ILI, which is collected through the US Outpatient Influenza-like Illness Surveillance Network (ILINet). The percent of patients presenting with ILI among all patient visits each week were used as indication of ILI incidence rate in the US population. Viral surveillance data, including weekly proportion of ILI samples test positive for influenza and subtype specific percentage data, were both from the US World Health Organization (WHO) Collaborating Laboratories and National Respiratory and Enteric Virus Surveillance System (NREVSS) laboratories.

Epidemiological model: seasonality based on humidity and regional formulation

For the humidity-forced model, $\beta(t)$ is given by the following expression based on (*21, 23, 24*):

$$\beta(t) = \frac{e^{-180H(t)}(R_{0\,max} - R_{0\,min}) + R_{0\,min}}{\gamma} \langle \frac{d\Gamma}{dt} \rangle, \tag{S1}$$

where $H(t)$ is the specific humidity at time t, and $R_{0\,max}$ and $R_{0\,min}$ denote the maximum and minimum basic reproductive numbers and the basic reproductive number is here given by (*21, 23, 24*):

$$R_0(t) = e^{(-180H(t)(R_{0\,max} - R_{0\,min}) + R_{0\,min})}. \tag{S2}$$

We recall that the basic reproductive number is given by:

$$R_0(t) = \frac{\beta(t)}{\frac{1}{\gamma} + \mu}. \tag{S3}$$

For the regional analysis, we revised our model slightly to make its fitting less sensitive to the data during the low seasons, which are mostly interpolated. We did this by adding a constant (200) to the reporting error through ρ in equations 10 and S4 so that likelihoods calculated based on the low seasons vary less and contribute less in differentiating model performance. We also increased the importation rate from 0.1/day to 10/day to allow for the more frequent movement of people between regions within US (*53, 54*). Additionally, we used the national sequence data for the evolutionary covariate, under the assumption that from the perspective of evolutionary change the whole country would be largely synchronized (*5, 53*). The choice of spatial scale could be examined further in the future, although more limited sequences are available for the regional level.

Measurement model

Reported cases were sampled from a normal distribution such that

$$\tilde{C}(t) = normal(\varphi I, \rho I), \tag{S4}$$

where φ is the reporting rate and the coefficient ρ defines the standard deviation as proportional to the size of the infected population. In addition, we impose the condition:

$$C(t) = \begin{cases} [\tilde{C}(t)], if\ \tilde{C}(t) \geq 0 \\ 0, otherwise \end{cases} \tag{S5}$$

We note that the parameters of the measurement error model are fitted as part of the inference process. This is important since the value of the reporting rate can often be confounded with the degree of population immunity. As a result, we constrained its value to remain under 1, but also obtained a profile likelihood for this key parameter (Fig. S11).

Continuous Evolutionary index

The biology behind $E(t)$ relates to the immune memory or protection existing at a given time in the human population for a new virus: the more similar this variant is to viruses in the past, the less likely it will be to infect people, since a higher probability exists that antibodies induced by viruses from previous infections will bind to it and stop the infection. Thus, the idea of a sum of weighted distance in sequence space is that of a quantity reflecting the movement of the virus away from variants the human population has been exposed to in the past. We include a decay function (controlled by a parameter θ that needs to be estimated) so that distances to more recent viruses have a higher weight when computing this average, given a time decay of antibody-mediated immunity. In other words, movement away from more recently observed antigenic variants would result in a higher evolutionary index and reflect a virus that is more novel from the perspective of the immunity landscape in the current human population.

We recognize that the weights can implicitly reflect additional processes that are not explicitly represented in the model, including the complex interaction of the age structure of the infected population (and contacts) and the effects on immune memory of age of exposure. The proposed quantity is intended to measure with a simple expression how much the virus has moved away from its recent predecessors in the sequence space that to the best of our knowledge reflects changes in the phenotype of interest. We note that since the rate of the decay backwards in time is one of the parameters inferred as part of fitting the overall model to the incidence time series, a possible outcome is for the decay to be negligible. In that sense, the inference process (and not an a priori assumption) determines the relevant time extent over which to evaluate the change in the virus.

In the formula for $E(t)$, distances were summed over time after weighting them back in time with an exponential decaying factor whose time scale is defined by the parameter θ. Only a total of twenty years were considered to make sure there were no data missing for calculating $E(t)$ of each month starting from October 2002. Distances $\widetilde{d(s,t)}$ for a given month t were calculated relative to previous seasons (and not individual months) in the past because fewer viruses are typically sequenced during the summer season due to lower levels of incidence and associated weaker surveillance efforts. Also, early previous years exhibit multiple months without any reported sequence due to weaker sampling and sequencing efforts. Months and previous years were assigned based on date information that is at least monthly for sequences after 1992. For the earlier period between 1982 and 1992, although there are enough sequences for our calculations (described below), most of them lack detailed monthly information. As a result, approximate previous season assignments were made based on the reported calendar year (with calendar year 1990 assigned to season 1989/1990 and so on).

First, an average distance $D(s,t)$ was calculated based on 1000 distances ($d_{m,st}$) among 1000 random pairs of sequences sampled from month t and previous season s. The actual value of $d_{m,st}$ is calculated as the number of amino acid differences for epitope regions of

HA1 (*30*). For distances to the current season ($s = 0$), distances were calculated based only on comparisons to sequences from earlier months. To avoid geographical and temporal sampling biases, we followed the practice of subsampling each random pair of sequences from sequences in both month t and previous season s, respectively, with equal probability from different states and from different months (or different previous seasons for earlier period before 1992) (*13, 14*). This subsampling process was repeated 1000 times to get a mean value:

$$D(s,t) = \frac{\sum_{m=1}^{1000} d_{m,st}}{1000}. \tag{S6}$$

We note that $D(s,t)$ is a matrix whose rows are previous seasons (from $s = 0$ in row 1, to $s = 19$ in the last row) and whose columns correspond to the months of interest in the time series of incidence to be analyzed (starting with October 2002 in column 1). We now proceed to normalize the entries of this matrix for each row, to correct for the effect of the passing of time within each season, which introduces an artificial trend in the un-normalized metric. Specifically, we normalize each term of the matrix by a mean value as follow:

$$D_{av}(s,t) = \sum_{t' \mid mod(t',12)=mod(t,12)} D(s,t') \,/\, \sum_{t' \mid mod(t',12)=mod(t,12)} 1, \tag{S7}$$

where the numerator sum is over all entries of the given row s that fall in the same month (as specified by the condition using the modulo operation mod), and the denominator sum simply counts the number of corresponding months. We then normalize the distances to obtain $d(s,t)$:

$$d(s,t) = \frac{D(s,t)}{D_{av}(s,t)}. \tag{S8}$$

Finally, $d(s,t)$ was interpolated (for months without sequences after October 2002, including March/May/July in 2004, May/July/September in 2005, May/June/July in 2006, December in 2009, February in 2010 and June in 2011) and smoothed by a cubic smoothing spline at a monthly scale (using the smooth.spline function in R package stats which uses a leave-one-out test to determine the smoothing parameter) to calculate $\widetilde{d(s,t)}$ in equation 8.

Forecasts based on the continuous model and risk level prediction

For the continuous model, we also need to know $C_{H1}(t)$ and $E(t)$. For $C_{H1}(t)$, the average monthly values of H1N1 incidence were used. $E(t)$ was linearly extrapolated up to September using the data available until June of the current season, then kept constant.

When predicting the risk level (high or low) of a target season, we first define an epidemic relative to a reference threshold, defined initially as the median of the seasonal total incidence in the training dataset. We then calculated the percentage of simulations above this threshold among 1000 simulations for the given target season. This percentage provides a probability of exceeding the given epidemic level. We can again use ROC curve and the training data set to establish which probability should be exceeded to predict a high risk. If the percentage is above a cutoff (upper bound of best accuracy), the target season is predicted with high risk; otherwise, it is a low risk season (see Fig. S5 for the US national data based on the cluster model, Fig. S12 for the US national data based on the continuous model, and Fig. S13 for the HHS region 3 data). Although a natural choice might be 50%

of the simulations, the ROC curve can indicate that a lower percentage is more appropriate to signal high risk, given the tendency of the model to under-predict the size of the peaks (with the mean or median of the simulation ensemble).

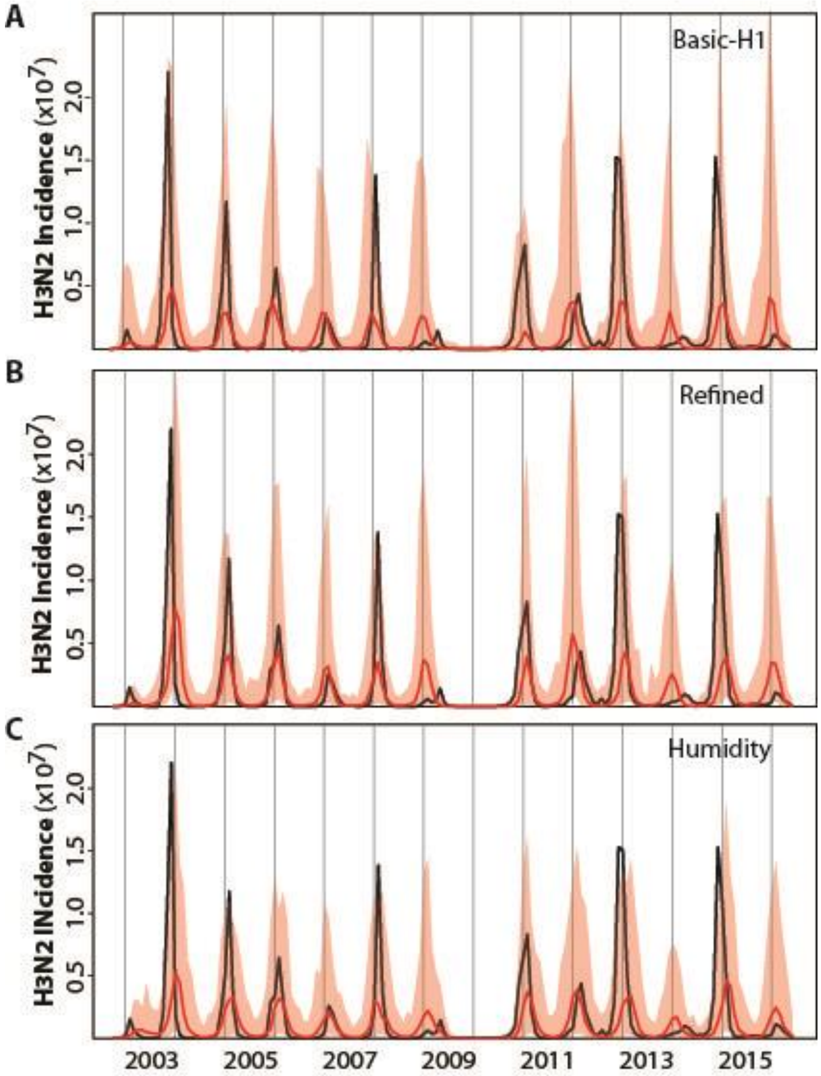


Figure S1. Illustration of the best model fits for alternative models. **(A)** The basic H1, **(B)** the refined and **(C)** the humidity models. See Table 1 for the specification and statistical comparison of the different models. Here, simulations of the respective models with the MLE (Maximum Likelihood Estimates) parameters are shown for the median (in red) and 95% uncertainty intervals (shaded red) of 1000 simulations starting from estimated initial conditions in October 2002. For comparison, the data are shown in black. The basic-H1 model includes an additional class to rely on the observed incidence of H1N1 as a covariate. The refined model additionally considers a different reporting error for the summer and winter seasons, and allows for the sub-exponential growth of the epidemic curve. The humidity model uses a humidity-based function for the seasonal forcing. All three models only partially capture the interannual variation of H3N2 incidence.

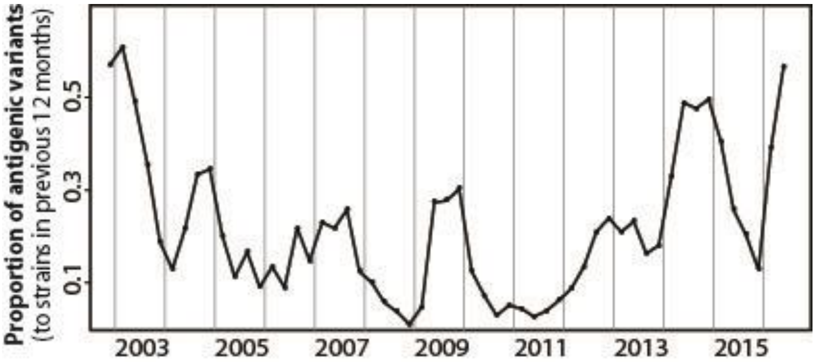


Figure S2. Proportion of antigenic variants (PAV). Antigenic variants were identified using the way described in the Materials and Methods. For strains in a given quarter, the number of antigenic variants was identified relative to the strains in the time window given by the previous 12 months. This number was then normalized by the number of all possible pairs. See the Materials and Methods for more details.

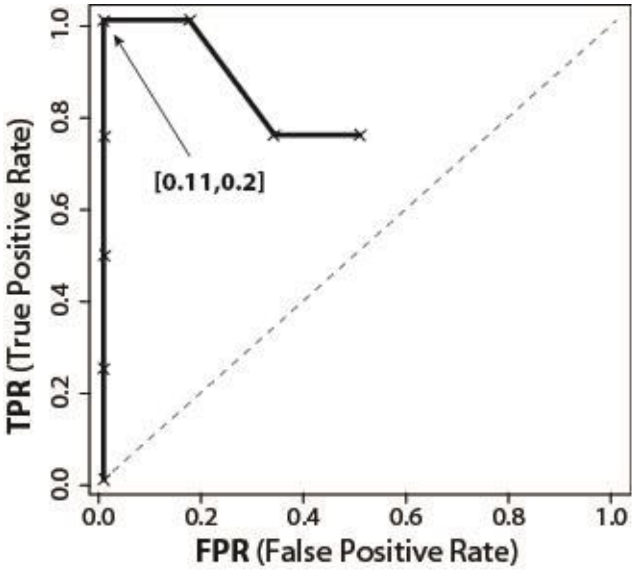


Figure S3. ROC curve for predicting antigenic cluster transitions based on the training dataset covering the period from 2002 to 2011 in US. Observed antigenic cluster transitions were identified based on CDC reports (see Materials and Methods for more details). Predictions were made based on the pattern of PAV change as explained in the Materials and Methods. When PAV for the third quarter crosses a cutoff, an antigenic cluster transition is anticipated for the next influenza season if PAV increases from the first quarter to the second quarter (or decreases but no antigenic cluster transition was assigned for the current influenza season; see Materials and Methods for more details). The range of the PAV cutoff with the best accuracy when predicting an antigenic cluster transition is also shown.

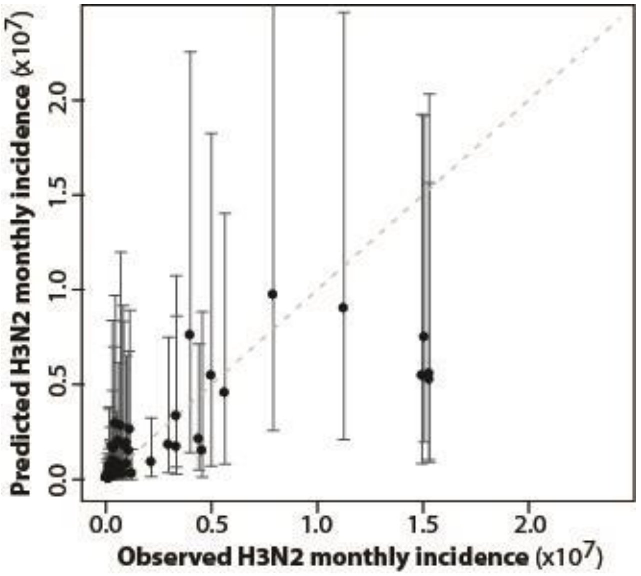


Figure S4. Comparison of monthly observations and forecasts generated with the cluster model for the out-of-fit period covering the period from 2011 to 2017. 95% uncertainty intervals from 1000 random simulations with the best models are shown. The observed incidence data for the 2016/2017 influenza season, which were not available when this study was conducted, are based on data from the weekly US influenza surveillance report until week 14 ending on April 8, 2017.

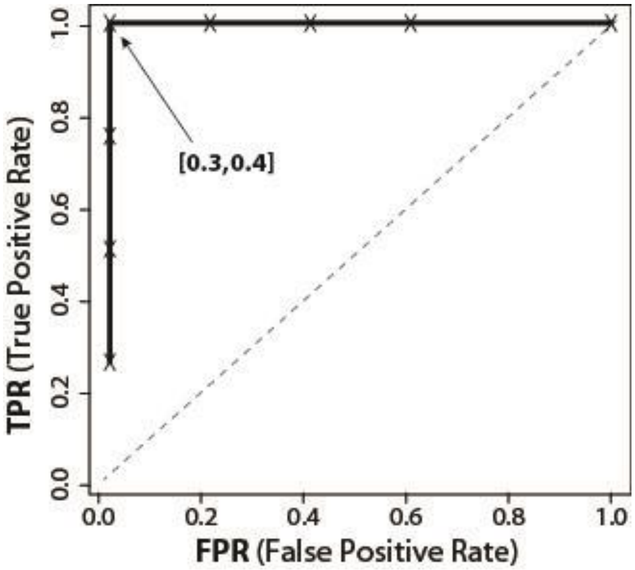


Figure S5. ROC curve for choosing the percentage cutoff applied to risk level prediction for the cluster model (Table 2 and Table S1). This choice is based only on the training dataset covering the period from 2002 to 2011 for the US. An observed high risk season is defined as a season whose total incidence is above a chosen threshold, here the median of the seasonal totals from the whole training dataset. A high risk level is predicted for a season if the percentage of simulations that are classified as high risk among 1000 simulations is above a cutoff; otherwise, a low risk level is predicted for that season. The range of the percentage cutoff with the best accuracy when predicting seasonal risk level is also indicated.

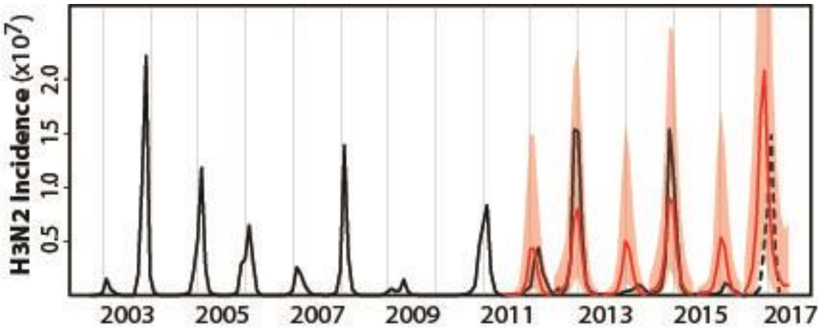


Figure S6. Forecasts based on the continuous model for the US. Both retrospective forecasts (for each influenza season from 2011/2012 to 2015/2016) and a real-time forecast for the 2016/2017 influenza season are presented. These forecasts are simulated on a seasonal basis from estimated initial conditions starting in June and based on parameters estimated with all the data up to that point in time. The black curve is the monthly observed H3N2 incidence; the red curve is the predicted monthly median incidence with shaded 95% uncertainty intervals from 1000 random simulations with the best models. The observed incidence data for the 2016/2017 influenza season, which were not available when this study was conducted, are shown with the dotted line, and are based on data from the weekly US influenza surveillance report until week 14 ending on April 8, 2017. Correlations between observed incidence and median incidence from the forecasts are $r = 0.51$ and $r^2 = 0.26$ for the monthly data, $r = 0.55$ and $r^2 = 0.30$ for the seasonal data. All correlations are statistically significant (with p value < 0.05).

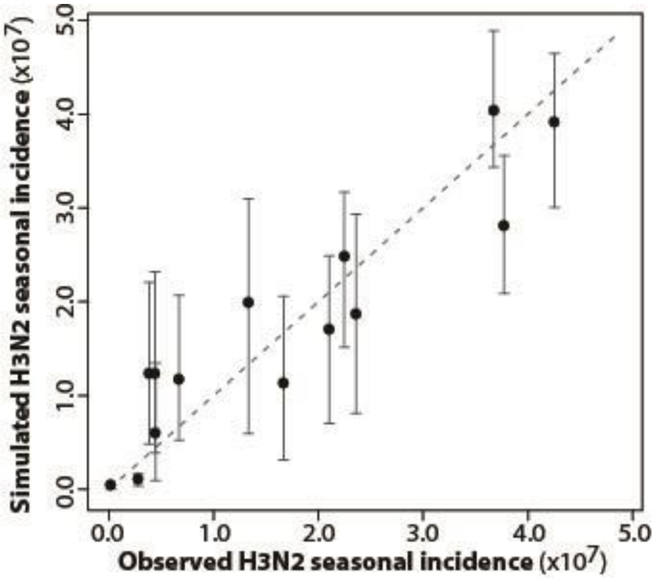


Figure S7. Scatter diagram for seasonal simulations and observations for the US dataset covering the period from 2002 to 2016 based on the cluster model. The correlation between observed H3N2 incidence and median incidence from the simulations is $r = 0.83$ and $r^2 = 0.69$, which is statistically significant (p value < 0.05). 95% uncertainty intervals are also shown based on 1000 random simulations with the best model. The seasonal incidence is the sum of the monthly incidence for a specific influenza season.

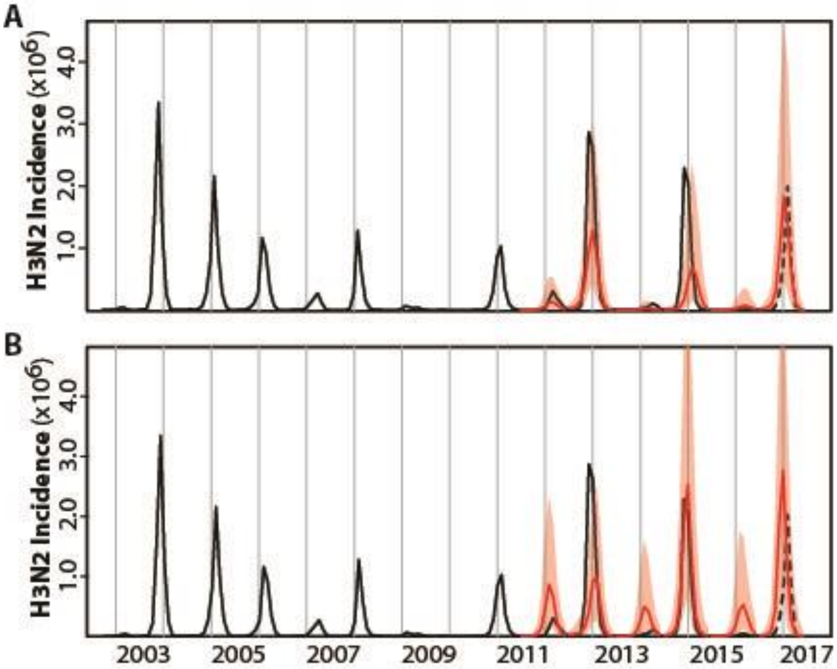


Figure S8. H3N2 incidence forecasts for the HHS region 3. **(A)** The cluster model and **(B)** the continuous model. Both retrospective forecasts (for each influenza season from 2011/2012 to 2015/2016) and a real-time forecast for the 2016/2017 influenza season are presented. These forecasts are simulated on a seasonal basis from estimated initial conditions starting in June, and are based on parameters estimated with all the data up to that point in time. The black curve is the monthly observed H3N2 incidence; the red curve is the predicted monthly median incidence with shaded 95% uncertainty intervals from 1000 random simulations with the best models. The observed incidence data for the 2016/2017 influenza season, which were not available when this study was conducted, are shown with the dotted line, and are based on data from the weekly US influenza surveillance report until week 14 ending on April 8, 2017. For the cluster model, correlations between observed incidence and median incidence from forecasts are $r = 0.68$ and $r^2 = 0.47$ for the monthly data, $r = 0.76$ and $r^2 = 0.57$ for the seasonal data (all correlations are significant, with p value < 0.05). For the continuous model, correlation between observed incidence and median incidence from forecasts are $r = 0.62$ and $r^2 = 0.38$ for the monthly data, $r = 0.60$ and $r^2 = 0.36$ for the seasonal data. All correlations are statistically significant, with p value < 0.05).

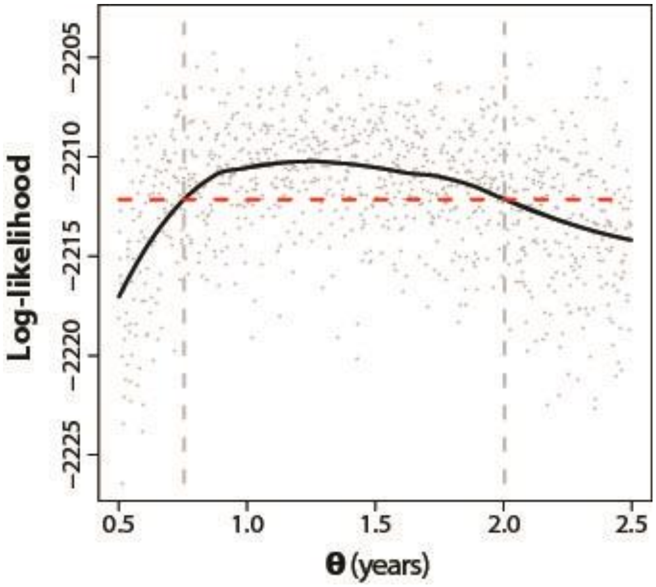


Figure S9. Log-likelihood profiling of the average effective time θ. The best likelihood among 10 random repeats for each θ fixed at a given value was obtained by allowing other parameters to be freely optimized (grey dots). The black curve is the fitting based on local polynomial regression. The intersection between the red dashed line and the vertical dashed lines indicates the 95% confidence intervals.

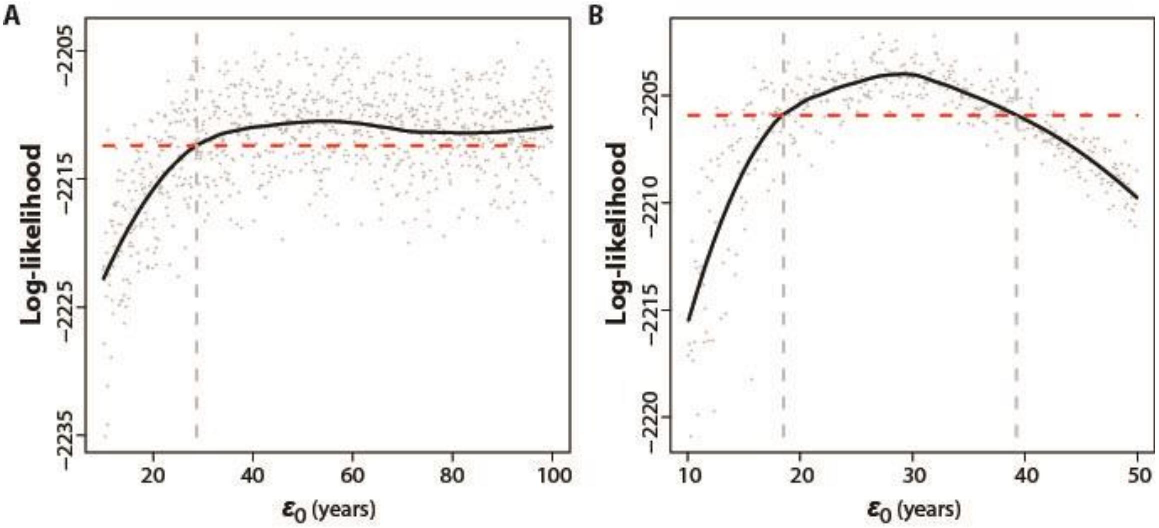


Figure S10. Log-likelihood profiling of the basic average latent time ε_0. **(A)** The best fitness model and **(B)** the cluster model. The intersection between the red dashed line and the vertical dashed lines indicates the 95% confidence intervals.

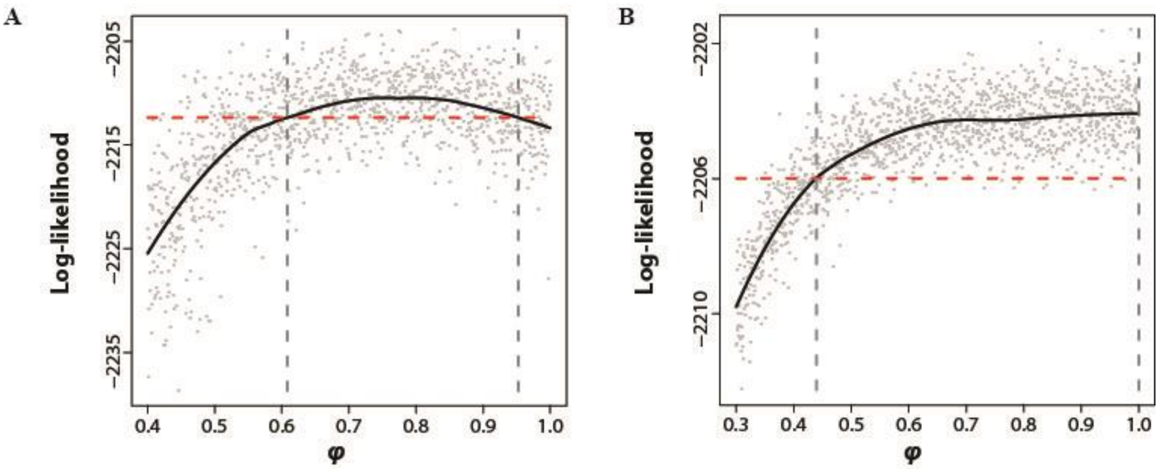


Figure S11. Log-likelihood profiling of the reporting rate φ. **(A)** The continuous model and **(B)** the cluster model. The intersection between the red dashed line and the vertical dashed lines indicates the 95% confidence intervals.

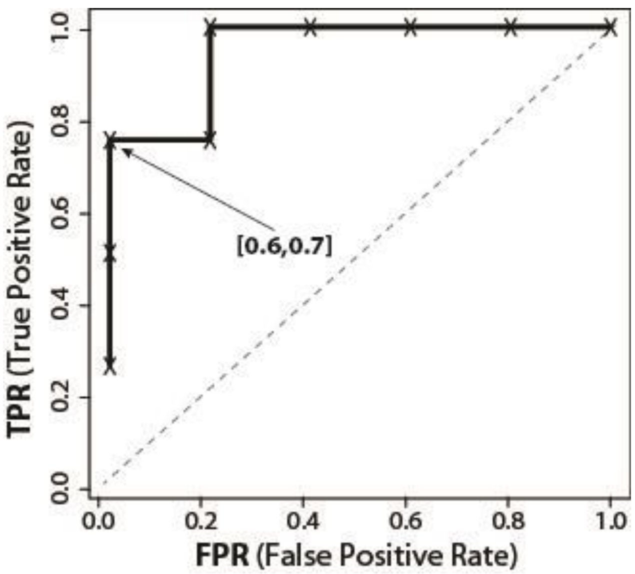


Figure S12. ROC curve for choosing the percentage cutoff applied to risk level prediction for the continuous model (Table S3). This choice is based on the training dataset covering the period from 2002 to 2011 in US. An observed high risk season is defined as a season with total incidence above a threshold, defined here as the median of the seasonal totals from the whole training dataset. A high risk level is predicted for a given season when the percentage of simulations classified as high risk among 1000 simulations is larger than a cutoff; otherwise a low risk level is predicted for that season. The range of the percentage cutoffs with the best accuracy when predicting seasonal risk level is also indicated.

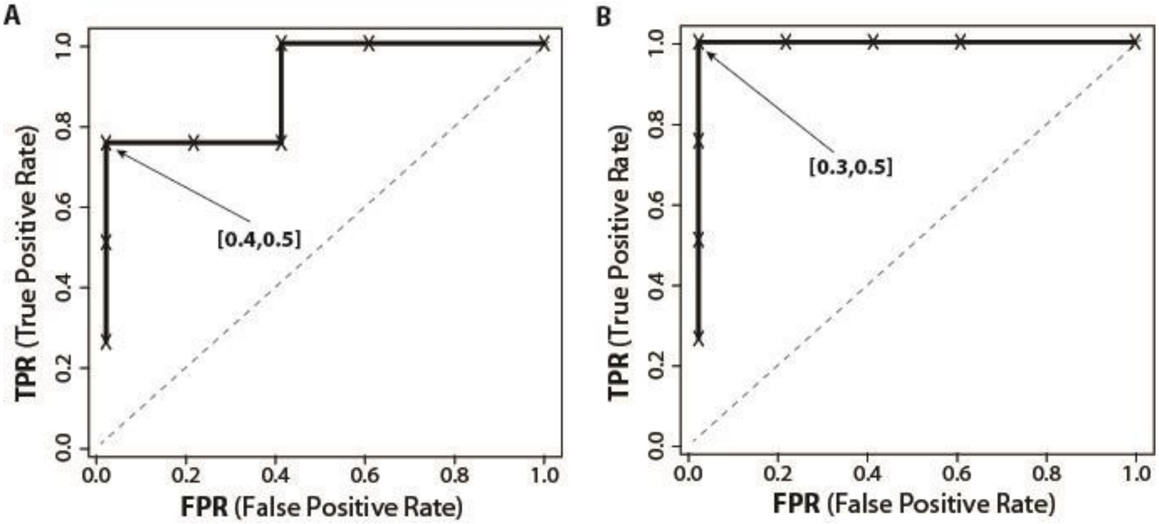
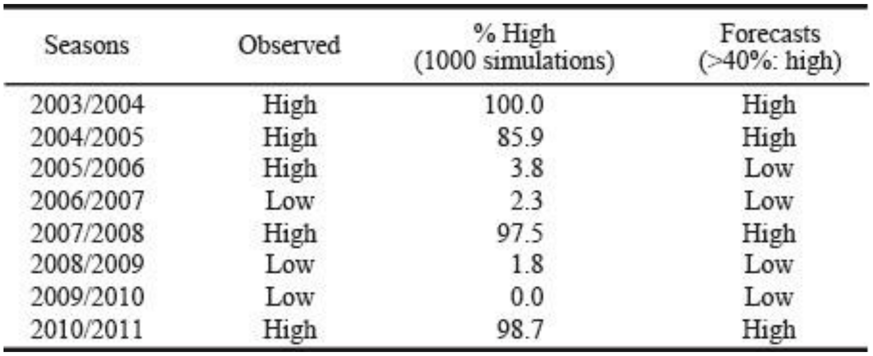


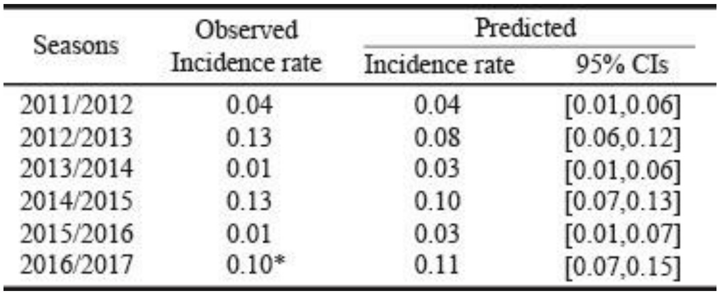
Figure S13. ROC curves for choosing the percentage cutoff applied to risk level prediction for the HHS region 3 (Table S5). **(A)** The cluster model and **(B)** the continuous model. This choice is based on the training dataset that covers the period from 2002 to 2011. An observed high risk season is a season whose total incidence surpasses a chosen threshold, here the median of the seasonal totals from the whole training dataset. A high risk level is predicted for a season if the percentage of simulations classified as high risk among 1000 simulations is above a given cutoff; otherwise, a low risk level is predicted for that season. The ranges of the percentage cutoffs with the best accuracy when predicting seasonal risk level are also indicated.

Table S1. H3N2 risk level forecasts based on leave-one-out cross validation using the cluster model for the period between 2003 and 2011. For a target season, incidence risk level is defined as high or low compared to a reference level chosen here as the median of the seasonal total incidence over all seasons excluding the target season. We defined an observed season as high risk, when the observed total incidence surpass the reference level; and an observed season as low risk, otherwise. For the forecasts, the percentage of 1000 simulations that exhibit a high risk was obtained. When this percentage exceeded 40% (chosen based on Fig. S5), we forecasted a high risk season. Otherwise, a low-risk season was predicted.



Seasons	Observed	% High (1000 simulations)	Forecasts (>40%: high)
2003/2004	High	100.0	High
2004/2005	High	85.9	High
2005/2006	High	3.8	Low
2006/2007	Low	2.3	Low
2007/2008	High	97.5	High
2008/2009	Low	1.8	Low
2009/2010	Low	0.0	Low
2010/2011	High	98.7	High

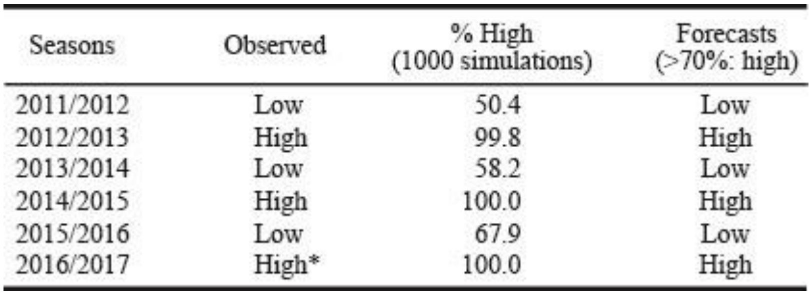
Table S2. Observed and predicted H3N2 seasonal incidence rate for the US based on the cluster model for the period between 2011 and 2017. Seasonal incidence rate is defined as the seasonal total incidence normalized by the average population size in that season. The predicted incidence rate is defined as the median from 1000 simulation. 95% confidence intervals (CIs) for the predicted incidence rate from 1000 simulations are also shown.



Seasons	Observed	Predicted	
	Incidence rate	Incidence rate	95% CIs
2011/2012	0.04	0.04	[0.01,0.06]
2012/2013	0.13	0.08	[0.06,0.12]
2013/2014	0.01	0.03	[0.01,0.06]
2014/2015	0.13	0.10	[0.07,0.13]
2015/2016	0.01	0.03	[0.01,0.07]
2016/2017	0.10*	0.11	[0.07,0.15]

* Based on the updated data from the weekly US influenza surveillance report until week 14 ending on April 8, 2017

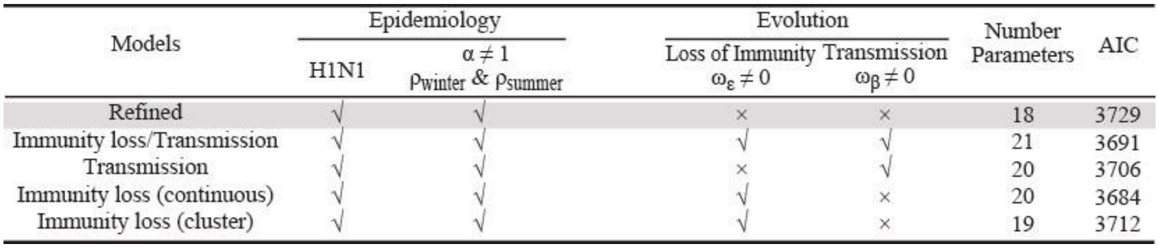
Table S3. H3N2 risk level forecasts based on the continuous model for the US. Seasonal risk level is defined as high or low for each influenza season between 2011 and 2017 , by comparison to a reference level defined here as the median of the seasonal total incidence in the training dataset. We defined an observed season as H3N2 high risk, when the observed total H3N2 incidence surpasses the reference level; and an H3N2 low risk season, otherwise. For the forecasts, the percentage of 1000 simulations that exhibit a H3N2 high risk was obtained. When this percentage exceeded 75% (chosen based on Fig. S12), we forecasted an H3N2 high risk season. Otherwise, an H3N2 low risk season was predicted.



Seasons	Observed	% High (1000 simulations)	Forecasts (>70%: high)
2011/2012	Low	50.4	Low
2012/2013	High	99.8	High
2013/2014	Low	58.2	Low
2014/2015	High	100.0	High
2015/2016	Low	67.9	Low
2016/2017	High*	100.0	High

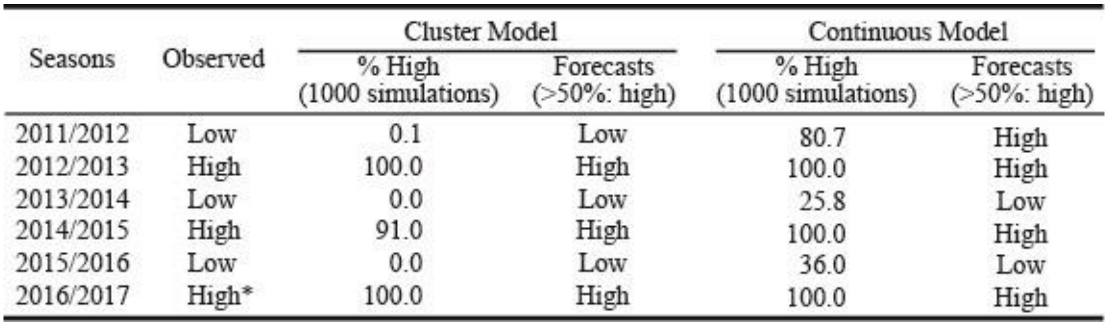
* Based on the updated data from the weekly US influenza surveillance report until week 14 ending on April 8, 2017

Table S4. Model comparison based on data from the HHS region 3. See Table 1 for the different models and for how to compare them. Based on the likelihood ratio test, models with evolutionary change, including the cluster model, are significantly better than those without it (the refined model here, shaded). The continuous model is significantly better than other models with continuous evolutionary change (the transmission model and the immunity loss/transmission model), and the models with continuous evolutionary change are significantly better than the cluster model.



Models	Epidemiology		Evolution		Number Parameters	AIC
	H1N1	$\alpha \neq 1$ ρ_{winter} & ρ_{summer}	Loss of Immunity $\omega_\varepsilon \neq 0$	Transmission $\omega_\beta \neq 0$		
Refined	√	√	×	×	18	3729
Immunity loss/Transmission	√	√	√	√	21	3691
Transmission	√	√	×	√	20	3706
Immunity loss (continuous)	√	√	√	×	20	3684
Immunity loss (cluster)	√	√	√	×	19	3712

Table S5. H3N2 risk level forecasts for the HHS region 3. Seasonal risk level is defined as high or low for each season of the out-of-fit period (2011-2017) compared to a reference level, defined here as the median of the seasonal total incidence in the corresponding training dataset. We defined an observed season as H3N2 high risk, when the observed total H3N2 incidence surpasses the reference level; and an H3N2 low risk season, otherwise. For the forecasts, the percentage of 1000 simulations that exhibit a H3N2 high risk was obtained. When this percentage exceeded 50% (chosen based on Fig. S13), we forecasted a H3N2 high risk season. Otherwise, a H3N2 low risk season was predicted.



Seasons	Observed	Cluster Model		Continuous Model	
		% High (1000 simulations)	Forecasts (>50%: high)	% High (1000 simulations)	Forecasts (>50%: high)
2011/2012	Low	0.1	Low	80.7	High
2012/2013	High	100.0	High	100.0	High
2013/2014	Low	0.0	Low	25.8	Low
2014/2015	High	91.0	High	100.0	High
2015/2016	Low	0.0	Low	36.0	Low
2016/2017	High*	100.0	High	100.0	High

* Based on the updated data from the weekly US influenza surveillance report until week 14 ending on April 8, 2017