A Model for Characterizing Annual Flu Cases

Miriam Nuño and Marcello Pagano

Department of Biostatistics,
Harvard School of Public Health, Boston, MA 02115, USA
mnuno@hsph.harvard.edu, pagano@biostat.harvard.edu

Abstract. Influenza outbreaks occur seasonally and peak during winter season in temperate zones of the Northern and Southern hemisphere. The occurrence and recurrence of flu epidemics has been alluded to variability in mechanisms such temperature, climate, host contact and traveling patterns [4]. This work promotes a Gaussian–type regression model to study flu outbreak trends and predict new cases based on influenza–like–illness data for France (1985–2005). We show that the proposed models are appropriate descriptors of these outbreaks and can improve the surveillance of diseases such as flu. Our results show that limited data reduces our ability to predict unobserved cases. Based on laboratory surveillance data, we prototype each season according to the dominating virus (H3N2, H1N1, B) and show that high intensity outbreaks are correlated with early peak times. These findings are in accordance with the dynamics observed for influenza outbreaks in the US.

1 Background

Seasonal variation of infectious diseases is associated with several factors that include, environmental mechanisms, host–specific behavior and pathogen's ability to continuously invade the host population. The influenza virus, a well studied pathogen, is known for its ability to continuously invade the human host by constantly mutating and successfully evading a host's immune system. Due to constant minor (antigenic drift) and major changes (antigenic shift) in the virus surface proteins, flu vaccines are updated yearly to enhance protection against new infections. Influenza seasons occur during winter in temperate zones of the Northern and Southern hemisphere. It is estimated that some 50 million people get infected, more than 200,000 people are hospitalized from flu complications, and about 47,200 people die from flu each year.

A goal of research in bio–surveillance of communicable diseases such as influenza involves the development and implementation of reliable methods for early outbreak detection [1]. Effective surveillance methods enhance the preparedness and facilitate immediate response from health officials in the event of epidemic and pandemic events [3, 5, 8, 9]. The recent SARS outbreak (2002–2003) event showed that control and containment of the outbreak was mainly based on rapid diagnosis coupled with effective patient isolation according to the modeling in [2]. A commonly, and widely used method for automatic detection of

D. Zeng et al. (Eds.): BioSurveillance 2007, LNCS 4506, pp. 37–46, 2007.

flu epidemics from time–series data was proposed by Serfling in 1963 [6]. Since then, the Center for Disease Control and Prevention has implemented the Serfling methodology to parameterize a baseline model based on statistical expectations (95% confidence interval of the baseline) by training data from non–epidemic years. A surveillance system based on the Serfling methodology signals an epidemic whenever the observed time–series data exceeds a threshold. The model assumes an average mortality (β_0), linear trend ($\beta_1 t$), and a 52–week cyclical period denoted by $\beta_2 cos(2\pi t/52)+\beta_3 sin(2\pi t/52)$. This model

$$Y(t) = \beta_0 + \beta_1 t + \beta_2 cos(2\pi t/52) + \beta_3 sin(2\pi t/52)$$

assumes that flu outbreaks are unimodal, cyclical and symmetric between the peak and troughs.

One of the aims of flu surveillance is the early detection of outbreaks, however, understanding the underlying mechanisms driving the observed fluctuations can be instrumental in developing effective monitoring systems. We study a time series regression model that summarizes outbreak trends and describes the observed seasonality. We apply the model to influenza–like–illness (ILI) weekly data for France reported during 1985–2005. We estimate the model parameters through least squares and validated the model numerically (adjusted R^2) and graphically (residual analysis). We assess the impact of timely reporting in predicting new flu cases. Finally, we study the correlation of outbreak peak times, intensity for virus-specified seasons and discuss the relevance of the proposed model for surveillance and prediction of flu outbreaks.

2 Methods

Weekly data of influenza–like–illness is illustrated in Figure 1 and demonstrates that flu outbreaks are highly seasonal with irregular intervals. The model propose incorporates several features that are characteristic of flu outbreaks. In particular, the model adjusts for variable peak times, intensity, and duration of outbreaks. Although the majority of flu outbreaks exhibit single peaks, Figure 1 shows that multiple peaks are also possible (Figure 1: the seasons 91–92, 97–98, 00-01 exhibit such behavior).

2.1 Statistical Model

We apply a Gaussian–type regression model to weekly influenza–like–illness (ILI) epidemic data to study outbreak trends for France during 1985–2005. For each year, we estimate the intensity, time of peak and duration of these outbreaks using least squares. The time series $Y_i(t)$ denotes the number of ILI cases observed in year i at the predictor time t. The value of j denotes the number of peaks considered in the model. That is, j=1 represents a single peak outbreak while j=1,2 assumes multiple peaks. The general form of the Gaussian model is as follows:

$$Y_i(t) = a_{i_0} e^{-b_{i_0} t} + \sum_{j=1}^{2} a_{ij} e^{-\left(\frac{t-b_{ij}}{c_{ij}}\right)^2}. \qquad (1)$$

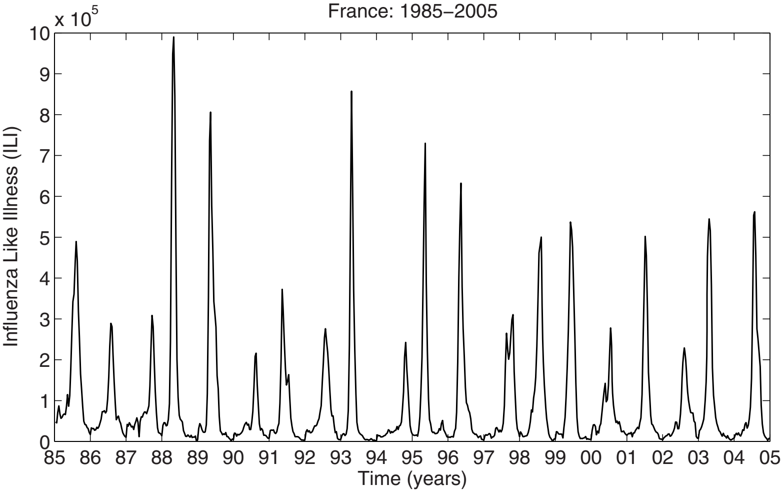


Fig. 1. Influenza–Like–Illness (ILI) weekly data for France (1985–2005)

Parameter a_{i_0} estimates the baseline (at time zero) of i–th year outbreak, a_{i_1} and a_{i_2} estimate the intensity of each peak in the i–th year, b_{i_0} estimate the decay rate for the background model, b_{i_1} and b_{i_2} evaluate the time of each peak, and c_{i_1} and c_{i_2} describe the duration of these outbreaks.

We identify outbreak trends for all seasons together, as well as, for outbreaks that are dominated by specific virus subtypes (H3N2, H1N1, B) by fitting the Gaussian model with single and multiple peaks. For each model, we estimate the mean, median and standard deviation of each of the parameters fitted in these models (e.g. outbreak peak time, intensity and duration). We summarize and compare the goodness of fit for each model numerically and graphically.

2.2 Dominating Virus Subtypes in Epidemic Seasons

Using laboratory surveillance data, influenza seasons (1985–2000) were summarized according to the prototype strain responsible each year. From the 20 seasons studied, virus A (H3N2) predominated 13 seasons, A (H1N1) dominated 3 seasons and the remaining 4 seasons are dominated by B type viruses. Figure 2 gives a box–plot description of the data for each season. The prototype seasons are distinguished by solid (H3N2), dotted (H1N1) and boxed (B) notches in order to illustrate the frequency and magnitude of these outbreaks and the corresponding dominating viruses.

2.3 Measuring the Uncertainty of Predicting New Cases

Based on the bimodal regression model, we estimate the likelihood of predicting ILI cases for unobserved times. We calculated prediction bounds for a new

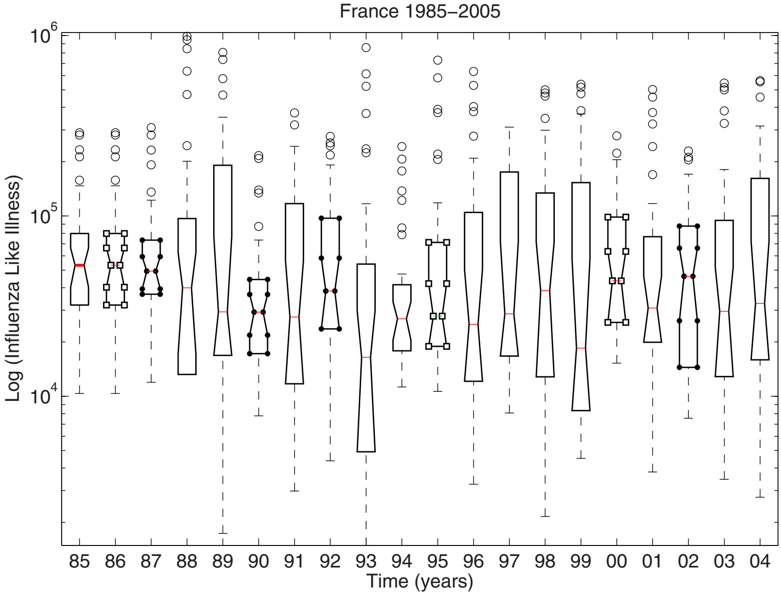


Fig. 2. French ILI data reported weekly. Solid, dotted and boxed notches denote outbreaks dominated by H3N2, H1N1 and B virus subtypes, respectively.

Table 1. Parameter estimation for single and bimodal peak models based on French ILI data from 1985 to 2005. We calculate the mean, median and standard deviation for each of the parameters estimated. Note that a_1 describes the model baseline, a_2 and a_3 the intensity of peaks 1 and 2, with peak times at b_2 and b_3, and corresponding duration given by c_1 and c_2. Parameter b_1 denotes the decay rate of each season. Non applicable findings are denoted by na.

$Y_i(t)$	a_1	a_2	b_1	b_2	c_1	a_3	b_3	c_2	R^2	adj R^2	RMSE
1985–2005											
Model 1											
mean	30464	507823	0.20	16.2	3.3	na	na	na	0.9550	0.9485	29092
median	26640	497000	0.03	14.9	3.2	na	na	na	0.9728	0.9689	25990
STD	20419	208105	0.52	5.4	0.9	na	na	na	0.0471	0.0538	17561
Model 2											
mean	29864	383466	0.01	13.6	3.3	240189	17.8	3.8	0.9919	0.9899	13133
median	27370	413500	0.04	11.5	2.0	240300	17.1	3.5	0.9934	0.9915	11780
STD	16188	262955	0.08	4.8	1.7	161857	4.9	1.9	0.0062	0.0077	4077

observation assuming that data was not available (extrapolation). The prediction bounds are calculated simultaneously and measure the confidence that a new observation lies within the interval regardless of time.

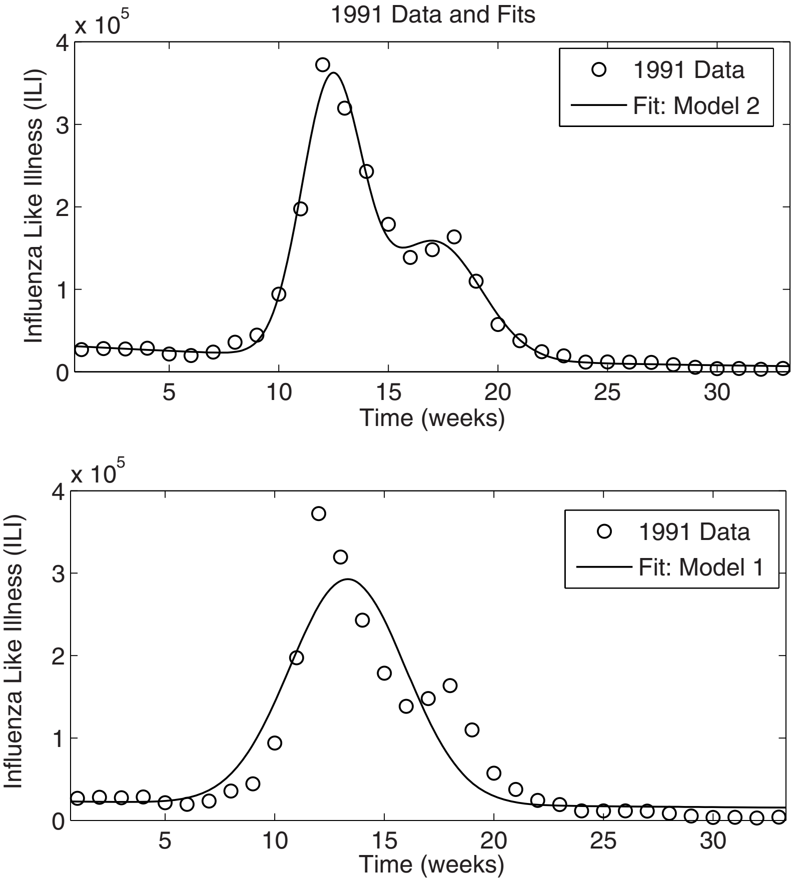
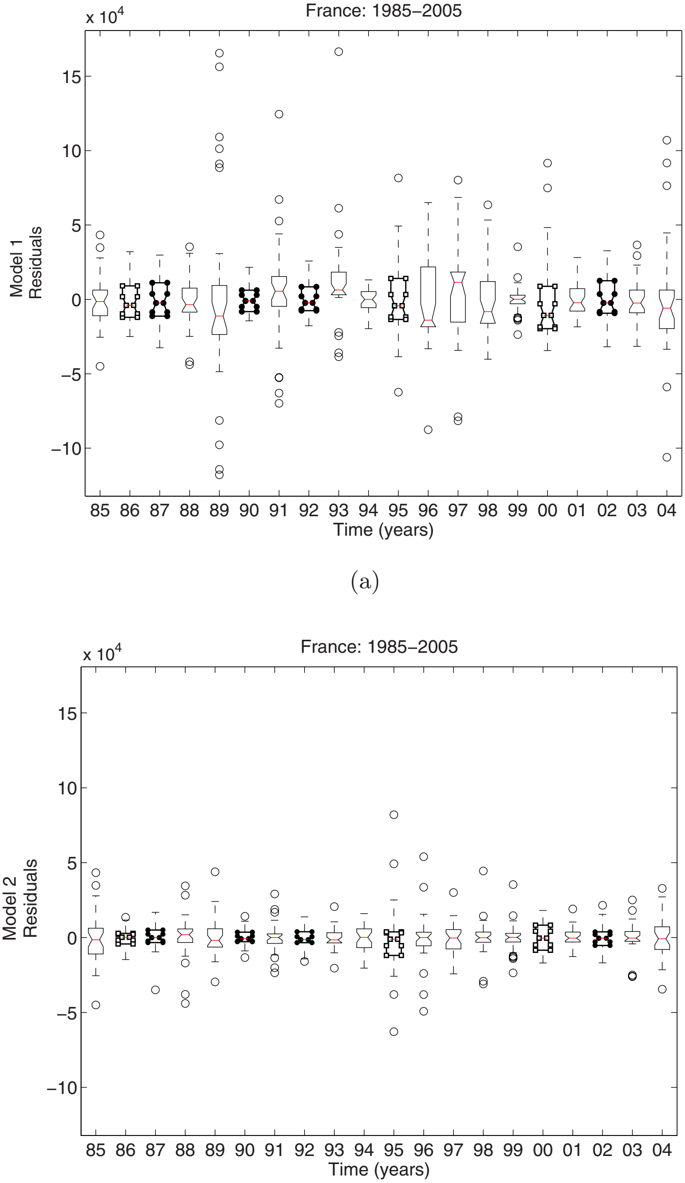


Fig. 3. Fits obtained with regression Models 1 and 2

3 Results

To assess the outbreak trends for all seasons of data available, we fit Models 1 and 2 to each year of data. We estimate the parameters of the single peak model (denoted by Model 1) and multiple peak model (denoted by Model 2) through least squares for each of the 20 years of data available (Table 1). After estimating the best fits for all 20 years, we calculate the mean, median and standard deviation (STD) of the parameters in these models. Fitting results of Model 1 yield a mean of 4.3×10^5, 4.4×10^5 median and 2.2×10^5 STD for the parameter estimating the intensity of the outbreaks. Similarly, we estimate the mean, median and standard deviation for the peak times and obtain a 18.4 weeks, 17.2 weeks, and 5 weeks, respectively. Table 1 shows that Model 2 describes the data better than Model 1 (adjusted R^2). We illustrate the actual fit of both Model for the 1991 season. These fits illustrate that the single peak model is ill specified to capture the bimodality of the data.



(a)

(b)

Fig. 4. Residual plots for Model 1 (a) and Model 2 (b) for French ILI data from 1985-2005. Solid, dotted and boxed notches denote the residuals corresponding to H3N2, H1N1 and B type strains.

We find strong evidence that Model 2 fits the data better than Model 1. The goodness of fit of each of these models was assessed by analyzing their residuals for each of the years fitted. We illustrate the goodness of fit of these model for the 1990-1991 flu season. Figure 3 illustrates the fits obtained with Model 2 (top–panel), Model 1 (bottom–panel) for these fits.

In order to investigate the goodness of fit of Model 1, we calculated these statistics for the adjusted R^2 values of this data. The adjusted R^2 values estimated were 0.9423 (mean), 0.9606 (median) and 0.0588 (STD). In order to compare the goodness of fit of Models 1 and 2, we carry out a similar analysis for Model 2 (see Table 1). Our results show that Model 2 improves the data fit. That is, the adjusted R^2 estimates for the mean is improved from 0.9423 to 0.987, median from 0.9609 to 0.988 with corresponding standard deviation reduced by approximately 85% (from 0.0588 to 0.009). We further assess these fits graphically by calculating the residuals for each model. Figure 4 (top–panel) illustrates the residuals for Model 1 and Figure 4 (bottom–panel) for Model 2.

3.1 Correlating Trends with Subtype–Specific Outbreaks

We further assessed the trends of these outbreaks by analyzing them according to dominating virus in each season. Our results show that intensity of H3N2 outbreaks were significantly higher than H1N1 and B subtypes combined ($p = 0.0226$, Kruskal-Wallis test, two-tailed). Table 2 also shows that H3N2 outbreaks tend to peak sooner than H1N1 and B. We carried out a similar test to assess any significant difference among the peak times for H3N2 and those for H1N1 and B and find that peak times for H3N2 subtype outbreaks occur earlier than H1N1 and B. Note that these results are supported by the parameter estimates obtained from fitting Model 1 and Model 2. For each subtype dominant season, our goodness of fit results (adjusted R^2) showed that the latter model improves the fit.

Although the data in this study was available weekly, we assumed several scenarios with limited data and assessed the likelihood of predicting new ILI cases for unspecified times. We assumed that data is available weekly (as in the current study), biweekly and monthly. Figure 5 illustrates the prediction bounds (dashed–dotted) obtained assuming weekly (left–panel), biweekly (middle–panel) and monthly (left–panel) data. Evaluating the prediction bounds for each of the scenarios shows that our ability to predict new ILI cases decreases as less data becomes available. That is, we show that for weekly available data we can predict all data points with high certainty since all data lies within the 95% confidence interval of prediction. As the data becomes more scarce (biweekly), we show that we are no longer able to predict the highest intensity data of the outbreak (Figure 5: middle–panel) . Finally, for monthly available data, we show that we are no longer able to predict the bimodality of our data in addition to the highest intensity peak.

4 Discussion

The findings in this study promote two Gaussian–type of regression models that assess influenza outbreak trends. Unlike the well–known cyclical Serfling model,

Table 2. Fit results implementing Model 1 and Model 2 for ILI data that is grouped according to the subtype–specific strains (H1N1, H3N2, and B) dominating in each season. Note that a_1 describes the model baseline, a_2 and a_3 the intensity of peaks 1 and 2, with peak times at b_2 and b_3, and corresponding duration given by c_1 and c_2. Parameter b_1 denotes the decay rate of each season. Non applicable findings are denoted by na.

$Y_i(t)$	a_1	a_2	b_1	b_2	c_1	a_3	b_3	c_2	R^2	adj R^2	RMSE
H3N2											
Model 1											
mean	30464	507823	0.20	16.2	3.3	na	na	na	0.9550	0.9485	29092
median	26640	497000	0.03	14.9	3.2	na	na	na	0.9728	0.9689	25990
STD	20419	208105	0.52	5.4	0.9	na	na	na	0.0471	0.0538	17561
Model 2											
mean	29864	383466	0.01	13.6	3.3	240189	17.8	3.8	0.9919	0.9899	13133
median	27370	413500	0.04	11.5	2.0	240300	17.1	3.5	0.9934	0.9915	11780
STD	16188	262955	0.08	4.8	1.7	161857	4.9	1.9	0.0062	0.0077	4077
H1N1											
Model 1											
mean	42843	364367	0.01	16.3	2.5	na	na	na	0.9133	0.9011	24683
median	45400	246100	0.01	17.9	2.5	na	na	na	0.9559	0.9497	26710
STD	12404	240319	0.01	4.1	0.4	na	na	na	0.0945	0.1080	7091
Model 2											
mean	33040	318300	0.11	14.0	2.3	115903	21.7	4.9	0.9825	0.9775	15410
median	33600	212800	0.01	12.6	2.4	84370	19.5	2.1	0.9802	0.9746	11870
STD	2644	285561	0.19	3.4	0.1	104945	5.2	5.64	0.0101	0.0130	10785
B											
Model 1											
mean	31985	221450	0.004	21.1	3.1	na	na	na	0.9588	0.9528	14030
median	29930	216950	0.01	20.5	2.9	na	na	na	0.9582	0.9521	13950
STD	13091	35613	0.02	2.5	0.9	na	na	na	0.0205	0.0234	3692
Model 2											
mean	28325	78820	0.04	18.0	5.7	177523	21.7	3.8	0.9867	0.9830	8578
median	24795	72175	0.05	19.6	5.6	187100	21.4	2.7	0.9865	0.9827	8776
STD	13293	49331	0.03	3.9	3.2	66625	2.14	2.6	0.0050	0.0064	1764

these models adjust for variability in time of peaks, intensity and duration of outbreaks. We show that these models are highly effective in describing outbreak trends, and thereby facilitate the assessment of flu patterns. The data presented in this study illustrates that flu outbreaks depict multiple peaks and therefore appropriate models are needed to regard for these dynamics. A residual evaluation of the fits of these models shows that these models are highly effective in describing the data presented here. These models show that they are highly effective in describing the data presented here, particularly, the bimodal model.

The results of this study support previous observations of the correlation in the time of peak and intensity of influenza epidemics for A (H3N2) virus outbreaks for the US [7]. That is, high intensity outbreaks tend to occur early–on the season, while lower intensity outbreaks (H1N1 and B) occur later. Moreover, our study

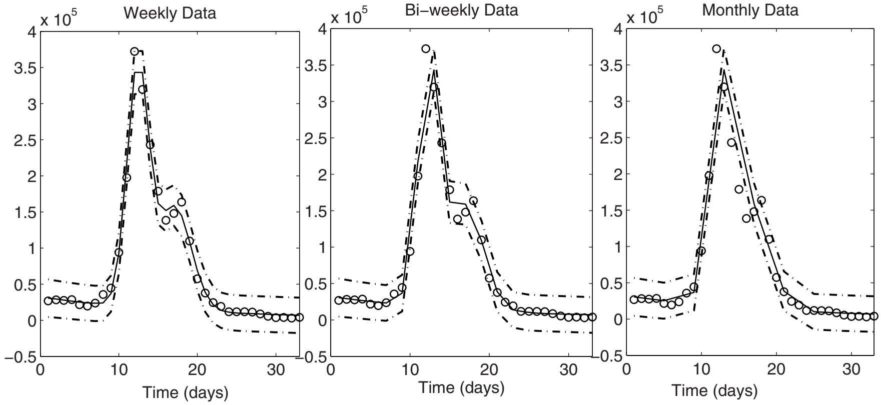


Fig. 5. Predictions based on weekly (left–panel), biweekly (middle–panel) and monthly data available (right–panel)

suggests that low–intensity peaks are more likely to be followed by high–intensity peaks since it is possible that frequently dominating viruses benefit from the recruitment of susceptibles during for almost two years. However, less dominating virus subtypes (H1N1 and B) do not rely on this recruiting advantage, and therefore face a continuous challenge to become established in the population as the dominating virus.

Effective surveillance of infectious diseases involves the timely assessment and implementation of monitoring systems aimed to reduce morbidity and mortality while minimizing societal disruption. Influenza surveillance is a combined effort of virological identification, clinical and epidemiological monitoring of multiple source data such as influenza–like–illness, pneumonia and influenza related mortality, hospitalization, to name a few. However, in this study we show that characterizing outbreak trends improves our understanding of the underlying mechanisms driving influenza epidemics, and therefore, is key for developing effective surveillance systems.

A critical limitation of this work lies in the prediction of unobserved cases. Our work evaluates the likelihood of predicting ILI cases during a particular season for unobserved times (retrospectively), however, we do not assess prediction of cases in future seasons. It is evident that effective surveillance systems should include a clear understanding of outbreak trends and retrospective assessment of unobserved cases. Moreover, the prediction of future cases based on current and historical data is an essential component of effective surveillance. To this end, our current research efforts are placed in predicting flu cases for future seasons based on the descriptive models proposed herein.

Acknowledgments. Miriam Nuño was funded by the Yerby Postdoctoral Fellowship and the National Institutes of Health grant T32AI07358. The research of Marcello Pagano was funded by the National Institutes of Health grant R01EB006195. The authors are thankful to Martine Valette of NIC–Lyon

Laboratoire de Virologie for providing prototype strain isolate information for the data used in this study.

References

[1] Buffington, J., Chapman, L. E., Schmeltz, L. M., Kendal, A, P.: Do family physician make good sentinels for influenza?, Arch. Fam. Med. **2:8** (1993) 859–864.

[2] Chowell, G., Fenimore, P, W., Castillo–Garsow, M, A., Castillo–Chavez, C.: SARS outbreaks in Ontario, Hong Kong and Singapore: the role of diagnosis and isolation as a control mechanism, J. Theor. Biol. **24** (2003) 1-8.

[3] Lazarus, R., Kleinman, K, P., Dashevsky, I., DeMaria, A., Platt, R.: Using automated medical records for rapid identification of illness syndromes (syndromic surveillance): the example of lower respiratory infection, BMC Public Health **1:9** (2001).

[4] Lofgre, E., Fefferman, N., Naumov, Y, N., Gorski, J., Naumova, E, N.: Infuenza Seasonality: Underlying Causes and Modeling Theories, J. Virol. (2006) 1680–1686.

[5] Reis, B,Y., Mandl, K,D.: Integrating syndromic surveillance data accross multiple locations: effects on outbreak detection performance, AMIA Annu. Symp. Proc. (2003) 549–53.

[6] Serfling, R, E.: Methods for current statistical analysis of excess Pneumonia-Influenza deaths, Public Health Reports **78:6** (1963) 494–506.

[7] Simonsen, L., Reichert, T, A., Viboud, C., Blackwelder, W, C., Taylor, R, J., Miller, M. A.: Impact of Influenza Vaccination on Seasonal Mortality in the US Elderly Population, Arch. Intern. Med. **165** (2005) 265–272.

[8] Toubiana, L., Flahault, A.: A space-time criterion for early detection of epidemics of influenza-like-illness, Eur. J. Epidemiol. **14:5** (1998) 465–70.

[9] Upshur, R, E., Knight, K., Goel, V.: Time-series analysis of the relation between influenza virus and hospital admissions of the elderly in Ontario, Canada, for pneumonia, chonic lung disease, and congestive heart failure, Am. J. Epidemiol. **149:1** (1999) 85–92.