



## Prediction of the Spread of Influenza Epidemics by the Method of Analogues

Cécile Viboud<sup>1,2</sup>, Pierre-Yves Boëlle<sup>1,3</sup>, Fabrice Carrat<sup>1,3</sup>, Alain-Jacques Valleron<sup>1,3</sup>, and Antoine Flahault<sup>1,2,3</sup>

<sup>1</sup> Epidemiology and Information Sciences, INSERM Unit 444, Université Pierre et Marie Curie, Paris, France.

<sup>2</sup> WHO Collaborating Center for Electronic Diseases Surveillance, Paris, France.

<sup>3</sup> Assistance Publique–Hôpitaux de Paris, Paris, France.

Received for publication January 3, 2002; accepted for publication May 7, 2003.

This study was designed to examine the performance of a nonparametric forecasting method first developed in meteorology, the “method of analogues,” in predicting influenza activity. This method uses vectors selected from historical influenza time series that match current activity. The authors applied it to forecasting the incidences of influenza-like illnesses (ILI) in France and in the country’s 21 administrative regions, using a series of data for 938 consecutive weeks of ILI surveillance between 1984 and 2002, and compared the results with those for autoregressive models. For 1- to 10-week-ahead predictions, the correlation coefficients between the observed and forecasted regional incidences ranged from 0.81 to 0.66 for the method of analogues and from 0.73 to –0.09 for the autoregressive models ( $p < 0.001$ ). Similar results were obtained for national incidence forecasts. From the results of this method, maps of influenza epidemic forecasts can be made in countries in which national and regional data are available.

communicable disease control; diffusion; epidemiologic methods; forecasting; influenza; statistics, nonparametric

Abbreviations: CV, cross-validation; ILI, influenza-like illnesses.

Over the past 20 years, surveillance networks for real-time monitoring of influenza activity have been developed worldwide to detect epidemics rapidly and to estimate the annual impact of influenza (1–10). However, these networks provide information on present, not future, activity. There is a lack of epidemiologic forecasts designed for public health authorities, which would allow for the organization of health care facilities during winter outbreaks (11). For instance, because the 1999–2000 influenza outbreak was both unpredictable and severe, record levels of hospital admissions for influenza put exceptional pressure on health care facilities in the United Kingdom during that winter (12). Also lacking are epidemiologic forecasts designed for the general population, as opposed to weather forecasts, which are part of daily news bulletins. While meteorologic forecasting methods have improved greatly, prediction of the occurrence or spread of infectious diseases, especially influenza, is not yet operational (11).

A large body of work has been devoted to the real-time detection of influenza outbreaks, defined as some increase above a historical baseline threshold (5, 8, 13, 14). A limited range of approaches has been developed to predict the spread of the epidemic process (15–19). These approaches fall into two categories: those that model the diffusion mechanisms and those that model the epidemic curve.

Large-scale mathematical deterministic models, also known as susceptible infected recovered models, are based on the mechanism of serial person-to-person transmission (20) and describe the time and geographic spread of influenza (15–18). Application of these models to retrospective data in the former Soviet Union, in Western countries (16, 18), and on a global scale (15, 17) provided useful insight into the diffusion mechanisms but did not prove efficient for prospective forecasting. More flexible models such as chain binomial models allowed for simulation of the spread of influenza epidemics in structured micropopulations, that is, families or small cities (19). Although these models were

Correspondence to Cécile Viboud, Fogarty International Center, National Institutes of Health, 16 Center Drive, Bethesda, MD 20892 (e-mail: viboudc@mail.nih.gov).

potentially excellent exploratory tools, they were not designed primarily for larger communities because of their computational complexity (21).

A second approach is based on time-series modeling of the epidemic curve. Autoregressive seasonal linear models (22) have previously been applied to influenza surveillance data (5, 23). However, these models did not take into account the geographic correlations relating to the diffusion process and could not be adjusted to sudden changes in dynamics (21). Nevertheless, although these models were not used in the past for operational forecasts of influenza epidemics on a national or regional geographic scale, they can be considered reference models against which new methods should be tested.

The method of analogues is a nonparametric approach first developed by Lorenz in 1969 to forecast meteorologic time series (24). It was then applied for prediction purposes in other fields, including physics (25), finance (25), hydrology (26), and geophysics (27). A similar approach has also been used to detect chaotic behavior in epidemiologic time series (28, 29). The aim of the present study was to evaluate application of this approach to the prediction of surveillance data. First, we applied the method of analogues to prediction of weekly national incidences of influenza-like illnesses (ILI) in France. We then extended this method to forecasting of the regional spread of ILI epidemics.

## MATERIALS AND METHODS

### Source of data

Data were obtained from the French Sentinel Network, a computerized public health surveillance system (4, 30) that, since November 1984, has been collecting reports from 1,790 general practitioners. These Sentinel general practitioners are voluntary unpaid participants and are located all over France. They are requested to log onto the system as often as possible, which is available around the clock, but they have to connect to the network at least once a week. If the interval between two successive reports from a Sentinel general practitioner exceeds 12 days, the last report is not taken into account in the surveillance, and the general practitioner is considered silent for the corresponding period. In accordance with a standardized surveillance protocol, the participating general practitioners enter an individual report for each visit for ILI that meets the following criteria: sudden fever of more than 39°C, myalgia, and respiratory symptoms. From the raw reports, a set of routine procedures produces ILI incidences for different time units (week, month, or year) and geographic areas (department, region, whole of France).

For this study, we used the times series of weekly ILI incidences for France and its 21 administrative regions. The series covered the 938 weeks spanning the period from November 1984 to October 2002, during which 18 epidemic seasons occurred. National and regional ongoing ILI incidence estimates are published on the following Web site (in French): <http://www.u444.jussieu.fr/sentiweb>. Epidemic weeks were defined according to a periodic seasonal regression model (13, 31), used routinely in the French Sentinel

Network. A similar model is applied to pneumonia and influenza mortality surveillance data by the Centers for Disease Control and Prevention (Atlanta, Georgia) (32). Epidemic onset is defined as the first week during which the national ILI incidence exceeds a baseline nonepidemic threshold given by the upper limit of the 95 percent confidence interval of the periodic model, provided the incidence remains above this threshold for at least 2 consecutive weeks. During the epidemic periods spanning 1984–2002, national ILI incidence estimates ranged from 105 to 1,793 cases per 100,000. For each epidemic, we also defined a preepidemic period as the 4 weeks immediately preceding onset of the epidemic. The preepidemic period captured periods of early increase in influenza activity prior to epidemic onset. The number of cases rises abruptly at the start of an influenza epidemic (21), and, in the fourth week before epidemic onset, the incidence was only 12 percent (range, 8–16 percent) of that observed during the epidemic. Prior to the fourth week before epidemic onset, influenza activity was even lower and hence was considered negligible. The total duration of the epidemic and preepidemic periods was 241 weeks.

### Principle of the method of analogues

Time-series prediction by using the method of analogues has been described in detail elsewhere (25, 33, 34). Here, we first briefly introduce the method of analogues for predicting national ILI incidences and then describe its extension to multivariate time series for predicting regional incidences.

*National ILI incidence forecasts.* Let  $I(t)$  denote the observed national ILI incidence at week  $t$ ,  $1 \leq t \leq N$ , where  $N = 938$  denotes the total number of weeks in the surveillance record. Suppose we wish to forecast the future from current week  $T$ . The principle of the method of analogues is to select historical sections of the time series that most closely match the observations at  $T$ . Prediction of future observations is based on the values that follow these closely matching sections (table 1 and figure 1).

We define influenza activity at week  $T$  by using the vector of incidence measures over the past  $l$  weeks,  $\mathbf{X}(T) = (I(T), I(T-1), \dots, I(T-l))$ . We compare  $\mathbf{X}(T)$  with historical vectors of  $(l+1)$  consecutive incidences and select the best matches on the basis of a distance criterion expressed as

$$\text{dist}(\mathbf{X}(T), \mathbf{X}(t)) = \sum_{j=0}^l (I(T-j) - I(t-j))^2,$$

$t < T$ . The best matches will be referred to as the “nearest neighbors” of  $\mathbf{X}(T)$ . We write  $\{\mathbf{X}(T^1), \dots, \mathbf{X}(T^v), \dots, \mathbf{X}(T^v)\}$ ,  $1 \leq i \leq v$  for the  $v$  nearest neighbors of  $\mathbf{X}(T)$  and  $F(T+h)$  for the  $h$ -week-ahead forecast from week  $T$  ( $h \geq 1$ ,  $l+1 \leq T \leq N-h$ ).  $F(T+h)$  is computed as the weighted mean of the incidences that follow the nearest neighbors; therefore,

$$F(T+h) = \sum_{1 \leq i \leq v} w^i \times I(T^i + h),$$

where  $w^i$  is the weight assigned to neighbor  $i$  (we used previously published weights (34); details below).

*Regional ILI incidence forecasts.* In the setting of the geographic spread of an epidemic disease, we now consider

**TABLE 1. Prediction of incidence + 1 based on one to four historical vectors of measures of incidence\* of influenza activity over the past 2 weeks, France, 1984–2002†**

Vector	Week no. in the series of 938 weeks	$\left\{ \begin{array}{l} \text{Incidence} \\ \text{Incidence} - 1 \end{array} \right\}$	Distance to $\mathbf{X}(T)$ (incidence <sup>2</sup> )	Rank	Incidence + 1
<i>Current vector</i>					
$\mathbf{X}(T)$	$T = 289$	$\left\{ \begin{array}{l} I(T) = 800 \\ I(T - 1) = 450 \end{array} \right\}$			
<i>Historical vectors</i>					
$\mathbf{X}(D)$	$D = 217$	$\left\{ \begin{array}{l} I(D) = 925 \\ I(D - 1) = 423 \end{array} \right\}$	16,354	1	$I(D + 1) = 1,575$
$\mathbf{X}(B)$	$B = 85$	$\left\{ \begin{array}{l} I(B) = 663 \\ I(B - 1) = 348 \end{array} \right\}$	29,173	2	$I(B + 1) = 513$
$\mathbf{X}(C)$	$C = 153$	$\left\{ \begin{array}{l} I(C) = 503 \\ I(C - 1) = 339 \end{array} \right\}$	100,530	3	$I(C + 1) = 436$
$\mathbf{X}(A)$	$A = 35$	$\left\{ \begin{array}{l} I(A) = 379 \\ I(A - 1) = 236 \end{array} \right\}$	223,037	4	$I(A + 1) = 244$
<i>Forecast based on</i>					
One neighbor ( $\mathbf{X}(D)$ , $w_D = 1$ )					$F(T + 1) = 1,575$
Two neighbors ( $\mathbf{X}(D)$ , $w_D = 0.64$ ; $\mathbf{X}(B)$ , $w_B = 0.36$ )					$F(T + 1) = 1,193$
Three neighbors ( $\mathbf{X}(D)$ , $w_D = 0.58$ ; $\mathbf{X}(B)$ , $w_B = 0.33$ ; $\mathbf{X}(C)$ , $w_C = 0.09$ )					$F(T + 1) = 1,122$
Four neighbors ( $\mathbf{X}(D)$ , $w_D = 0.56$ ; $\mathbf{X}(B)$ , $w_B = 0.31$ ; $\mathbf{X}(C)$ , $w_C = 0.09$ ; $\mathbf{X}(A)$ , $w_A = 0.04$ )					$F(T + 1) = 1,090$

\* Influenza-like illnesses cases/100,000.

†  $\mathbf{X}(T)$  denotes the vector of current activity at week  $T$ . From the historical series of incidences, four vectors, or “nearest neighbors” ( $\mathbf{X}(A)$ – $\mathbf{X}(D)$ ), are selected as the best matching  $\mathbf{X}(T)$  and are ranked according to their euclidian distance to  $\mathbf{X}(T)$ . The incidence + 1 forecast is calculated as the mean of the incidences + 1 of the nearest neighbors, weighted by the inverse of their distance to  $\mathbf{X}(T)$  ( $w_A$ – $w_D$ ). Refer to figure 1 for a graphic representation.

multivariate time series, that is, the set of  $\{I_k(t)\}$  where  $I_k(t)$  denotes the observed ILI incidence at week  $t$ ,  $1 \leq t \leq N$ , in region  $k$ ,  $1 \leq k \leq 21$ . We define influenza activity at current week  $T$  by the matrix  $\mathbf{X}(T) = (I_k(T - r))$ ,  $0 \leq r \leq l$ ,  $1 \leq k \leq 21$  and use a distance criterion to compare it with historical matrices expressed as

$$\text{dist}(\mathbf{X}(T), \mathbf{X}(t)) = \sum_{k=1}^{21} \sum_{j=0}^l (I_k(T - j) - I_k(t - j))^2.$$

Computation of the  $h$ -week-ahead forecast in region  $k$ , similar to that described in the preceding section, follows as

$$F_k(T + h) = \sum_{1 \leq i \leq v} w^i \times I_k(T^i + h),$$

where the  $v$  nearest neighbors  $\{\mathbf{X}(T^1), \dots, \mathbf{X}(T^i), \dots, \mathbf{X}(T^v)\}$ ,  $1 \leq i \leq v$  are selected by minimization of the distance criterion, and  $w^i$  is the weight assigned to neighbor  $i$ . Note that for a given week, identical  $w^i$  weights in all regions are chosen.

**Parameter estimation.** We used a cross-validation (CV) criterion based on the root mean square error (35) to select the combination of parameters ( $l$ ,  $v$ , and  $w^i$ ) that yielded the most accurate forecasts in the retrospective series. The CV criterion is defined as the error that occurs when predicting the future from week  $T$ , excluding the incidences surrounding  $I(T)$  from the library of historical observations. Specifically, we exclude the  $l$  consecutive incidences

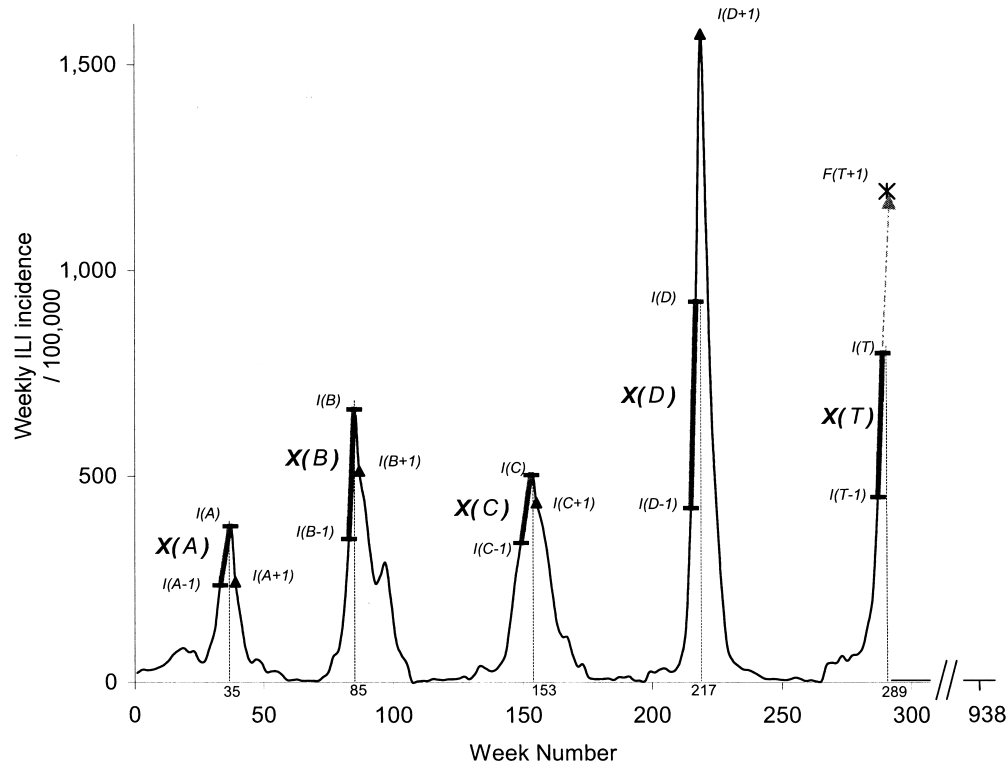
preceding  $I(T)$  and the  $h$  consecutive incidences following  $I(T)$ . This algorithm avoids redundancy between the forecasts and the model (28, 29). The CV criterion for an  $h$ -week-ahead national prediction was expressed as

$$CV_h = \frac{\sum_{l+1 \leq T \leq N-h} (I(T+h) - F(T+h))^2}{\sum_{l+1 \leq T \leq N-h} (I(T+h) - \bar{I})^2},$$

where  $I(T + h)$  is the observed national incidence at week  $T + h$ ,  $F(T + h)$  is the forecasted national incidence at week  $T + h$ , and  $\bar{I}$  is the average national incidence. At a regional level, the CV criterion for an  $h$ -week-ahead prediction in region  $k$  was expressed as

$$CV_{k,h} = \frac{\sum_{l+1 \leq T \leq N-h} (I_k(T+h) - F_k(T+h))^2}{\sum_{l+1 \leq T \leq N-h} (I_k(T+h) - \bar{I}_k)^2},$$

where  $I_k(T + h)$  is the observed incidence at week  $T + h$  in region  $k$ ,  $F_k(T + h)$  is the forecasted incidence at week  $T + h$  in region  $k$ , and  $\bar{I}_k$  is the average incidence in region  $k$ . The



**FIGURE 1.** Illustration of the principle of the method of analogues for time-series prediction of the incidence of influenza-like illnesses (ILI) 1 week ahead from week  $T$  (week no. 289), France, 1984–2002. The time series of observed incidences is represented by a thin solid line; vectors are thick, solid lines. The vector of incidence measured over the past 2 weeks,  $\mathbf{X}(T) = (I(T), I(T - 1))$ , represents current influenza activity. The four historical vectors of incidence measures shown ( $\mathbf{X}(A)$ – $\mathbf{X}(D)$ ) best match  $\mathbf{X}(T)$ . The details of the distances between  $\mathbf{X}(T)$  and the historical vectors (the so-called nearest neighbors) are given in table 1. The forecast  $F(T + 1)$  (star) is calculated as a weighted mean of the observed incidences following the historical vectors (triangles). In this figure, the two nearest neighbors,  $\mathbf{X}(B)$  and  $\mathbf{X}(D)$ , are used, and the forecast is based on  $I(B + 1)$  and  $I(D + 1)$ .

mean error for the  $h$ -week-ahead prediction of ILI incidences in the 21 regions was estimated as

$$CV_h = \frac{\sum_{1 \leq k \leq 21} CV_{k,h}}{21}.$$

A grid search was conducted in the parameter space (for  $h = 1$ –10 weeks, with  $v = 2$ –16 nearest neighbors,  $l = 0$ –10 weeks,  $w^i =$  equally distributed weights or  $w^i =$  weights proportional to the inverse of the distance criterion (34)) to select the combination of values that minimized  $CV_h$  over the 241 epidemic and preepidemic weeks.

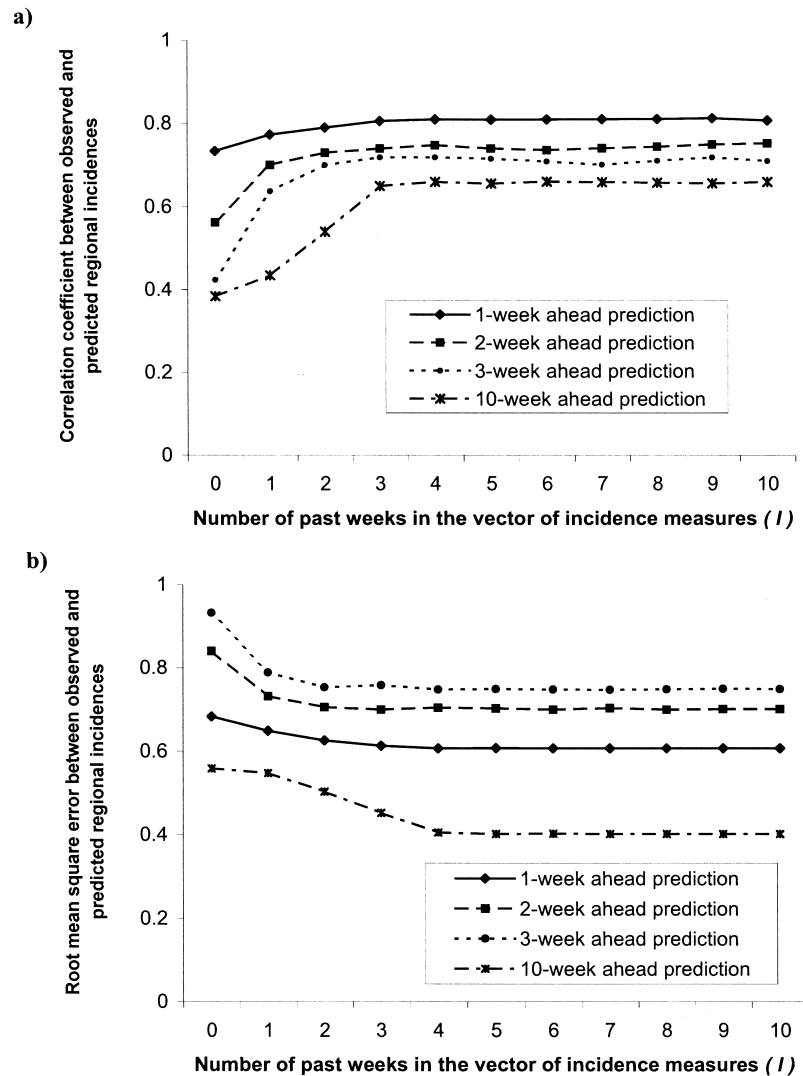
**Overall accuracy of predictions.** We used two measures of prediction accuracy to evaluate the method of analogues: the correlation coefficient and the root mean square error between observed and forecasted incidences, denoted by  $CV_h$  in the text (25, 29, 33, 34, 36). We defined the prediction horizon (denoted by  $h$  in the text) as the number of weeks in advance that the prediction is made. Both measures of prediction accuracy were plotted against the prediction horizon.

**Comparisons with linear methods.** The forecasts estimated by using the method of analogues were compared with

those estimated by using the “naïve” method of prediction and by using regional linear autoregressive models (29, 36). The naïve method constitutes a bottom line in the comparison. It is defined as the method in which predicted incidences are equal to the current incidence, hence  $F(T + h) = I(T)$  for national predictions and  $F_k(T + h) = I_k(T)$  for regional predictions,  $h \geq 1$ .

An autoregressive model of order  $p$  can be written as  $I^*(T + 1) = \phi_0 + \phi_1 I^*(T) + \phi_2 I^*(T - 1) + \dots + \phi_p I^*(T - p + 1) + \epsilon(T)$ , where  $I^*(\cdot)$  is the detrended series of incidence measures,  $\phi_i$ ,  $0 \leq i \leq p$  are autoregressive parameters to be estimated from the sample data, and  $\epsilon(\cdot)$  are independent random normal deviates. For the 21 regions separately, the series of incidence measures were detrended, and the autocovariances were computed. The autoregressive parameters were calculated from the autocovariances in a Yule-Walker framework, and the model was built with a backward selection procedure (Forecast, SAS software, version 8; SAS Institute, Inc., Cary, North Carolina). To account for the seasonality of the disease, we allowed for annual terms to be included in the model.

The performances of the three methods (naïve, autoregressive, and analogues) were assessed by computing the corre-



**FIGURE 2.** Parameter estimation procedure for the method of analogues for prediction of the incidence of influenza-like illnesses, France, 1984–2002: a) selection of the number of past weeks in the vector of incidence measures ( $l$ ; refer to the text) by optimizing the correlation coefficient or b) the cross-validation criterion based on the root mean square error.  $v$ , the number of neighbors, is fixed at 3; optimum  $l$  is 4.

lation coefficient and the root mean square error between predicted and observed incidences (29, 36). The prediction horizon ranged from 1 to 10 weeks for the 241 epidemic and preepidemic weeks.

**Mapping of the 1999–2000 influenza epidemic.** Observed versus predicted regional incidences were mapped for the 1999–2000 influenza epidemic. Spearman correlation coefficients were used to compare the two series of incidences.

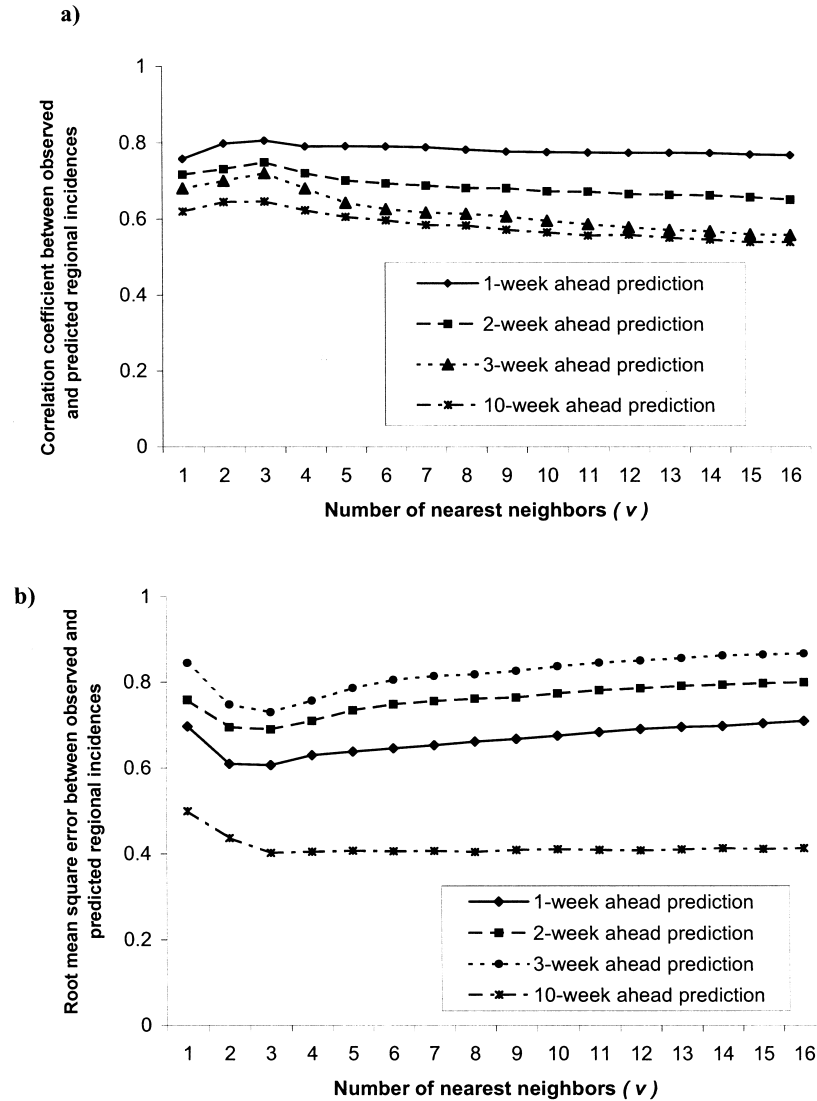
## RESULTS

### Parameter estimation

The optimal predictions in our influenza time series for the period 1984–2002 were reached (i.e., the CV criterion was

minimal) when 5 consecutive weeks comprising the current week and 4 previous weeks were used to select the three nearest neighbors in the past series (i.e.,  $l = 4$  weeks,  $v = 3$ ) (figures 2 and 3). The CV criterion and the correlation statistics yielded very similar results for the parameter estimation procedure (compare figure 2a with 2b and 3a with 3b).

The quality of the predictions was enhanced by the use of weights ( $w^i$ ) proportional to the inverse of the distance between the current activity and the selected nearest neighbors. For a given prediction horizon, the CV criterion based on the root mean square error was systematically reduced with this weighting, the other parameters ( $l$  and  $v$ ) being held fixed. It was an average of 1.0 percent lower with this weighting than with the equal weighting (range, 0.5–1.5 percent).



**FIGURE 3.** Parameter estimation procedure for the method of analogues for prediction of the incidence of influenza-like illnesses, France, 1984–2002: a) selection of the number of nearest neighbors ( $v$ ; refer to the text) by optimizing the correlation coefficient or b) the cross-validation criterion based on the root mean square error.  $l$ , the number of past weeks in the vector of incidence measures, is fixed at 4; optimum  $v$  is 3.

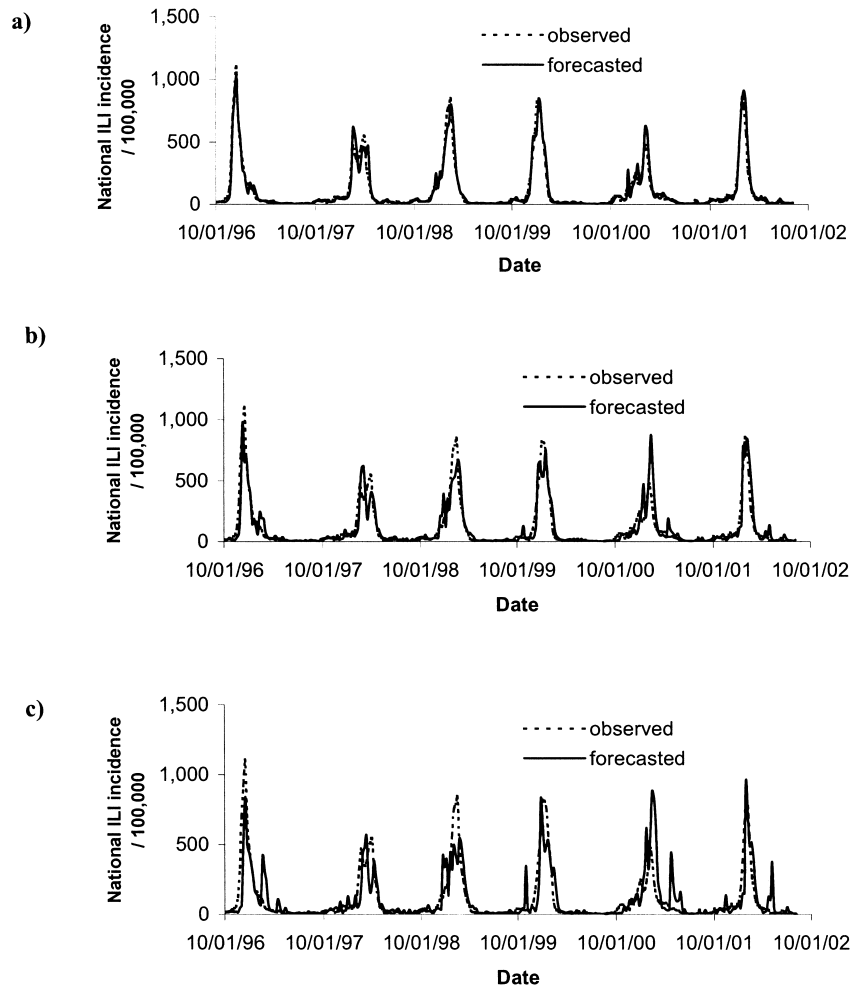
**National ILI incidence forecasts**

The correlation coefficients between 1-, 2-, and 3-week-ahead predicted and observed national incidences were, respectively, 0.90, 0.76, and 0.63 over the 241 epidemic and preepidemic weeks (figure 4). Over the same period, the corresponding coefficients obtained with the “naïve” method were, respectively, 0.84, 0.53, and 0.21. Although the correlation decreased with the prediction horizon, the degree of accuracy of the 3-week-ahead forecasts by the method of analogues was high. For a prediction horizon of 10 weeks, the correlation coefficient was still 0.58 for the method of analogues, compared with  $-0.19$  for the “naïve” method.

**Regional ILI incidence forecasts**

The distribution of the 1-week-ahead regional forecasts was similar to that of the observed incidences, except for extreme values. The lowest values were overestimated, and the highest were underestimated. On the whole, the forecasts were closer than the observed incidences to the mean of the time series, so that the range of predicted values was narrower.

The autoregressive method gave higher correlation coefficients (similarly, a lower root mean square error) than the naïve method did, but the difference was not large (figure 5). In the same way as for national predictions, the quality of the predictions decreased with the prediction horizon. For 1- to



**FIGURE 4.** Observed incidences of influenza-like illnesses (ILI) in France in 1996–2002 vs. a) 1-week-ahead, b) 2-week-ahead, and c) 3-week-ahead forecasted incidences. Forecasts are made with the method of analogues ( $v = 3$ ,  $l = 4$ , and  $w^i \propto 1/\text{distance}$  criterion) by using all of the 1984–2002 data and a cross-validation algorithm. Dates abbreviated as follows, for example: 10/01/96, October 1, 1996.

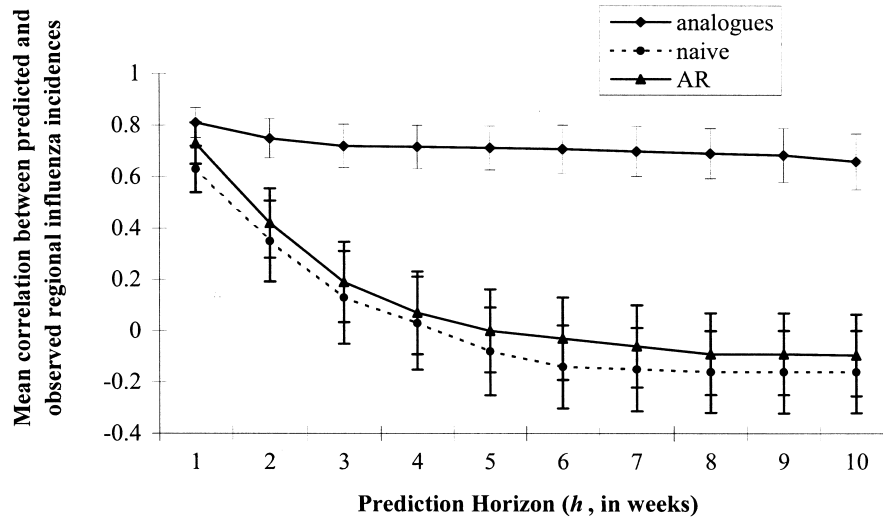
10-week-ahead predictions, the correlation coefficients between the observed and forecasted regional incidences ranged from 0.81 to 0.66 for the method of analogues and from 0.73 to  $-0.09$  for the autoregressive models ( $p < 0.001$ ). For the method of analogues, there was a relative decrease of 21 percent in the correlation coefficients between the 1- and 10-week-ahead predictions. The corresponding decreases were 112 percent and 125 percent, respectively, for the autoregressive and naive methods. The root mean square error was 10 percent lower with the method of analogues for 1-week-ahead predictions and 105 percent lower for 10-week-ahead predictions. Hence, for up to 10-week-ahead predictions, the method of analogues provided better predictions than the autoregressive and naive methods.

Note that for the method of analogues, the prediction accuracy before epidemic onset was different from that at the time of peak incidence—which occurred about 4 weeks after onset. The correlation coefficients were, respectively, 0.55, 0.42, and 0.27 for the 1-, 2-, and 3-week-ahead forecasts conducted 4

weeks before epidemic onset, and they were 0.85, 0.81, and 0.78 for the corresponding forecasts conducted 4 weeks after this onset. However, on average for 1-, 2-, and 3-week-ahead forecasts 4 weeks before the epidemic occurred, the root mean square error was lower by 30.3 percent than for forecasts conducted 4 weeks after onset. This difference likely was due to a low level of influenza activity 4 weeks before onset of the epidemic. When incidence measures fluctuate around zero, the magnitude of the prediction bias is small, yielding on average small errors, while the correlation statistics reflect noise around the estimates of zero.

#### Mapping of the 1999–2000 influenza epidemic

Extension of the method of analogues to regional ILI incidence forecasts is illustrated in figure 6 by a set of maps corresponding to the 1999–2000 influenza season. The first set of predictions was based on the influenza activity recorded during the first week of December 1999 (i.e., the



**FIGURE 5.** Prediction accuracy (y-axis) vs. the number of weeks in advance that the prediction is made (prediction horizon, x-axis) for the method of analogues, the “naive” method, and the autoregressive (AR) method. Prediction accuracy is defined as the correlation coefficient between observed and predicted regional incidences averaged over the 21 regions of France and calculated for the total duration of the epidemic and preepidemic periods for 1984–2002 (241 weeks). Vertical bars, the 95 percent confidence interval limits of the correlation coefficients.

week of epidemic onset, week 49 of 1999, started on December 6 (figure 6a)). The method of analogues forecasted the regional incidence levels for the last 3 weeks of December 1999, with a correlation coefficient between the observed and forecasted incidences of 0.68 ( $p < 0.001$ ). A second set of predictions was based on the influenza activity recorded during the last week of December (i.e., 1 week before peak incidence, week 52 of 1999, started on December 27 (figure 6b)). The method of analogues forecasted the regional incidence levels for the first 3 weeks of the year 2000, with a correlation coefficient between the observed and forecasted incidences of 0.78 ( $p < 0.001$ ). Because information about the incidence of influenza is often delivered dichotomously, that is, as being above or below the epidemic threshold, we also examined the accuracy of the forecasts in this discrete way. Thus, for the first 3 weeks of 2000, the numbers of regions for which the observed incidence exceeded the threshold were, respectively, 20, 19, and 19. For 18 of the 21 regions, the method of analogues forecasted an epidemic status that matched the observed one over the whole 3 weeks.

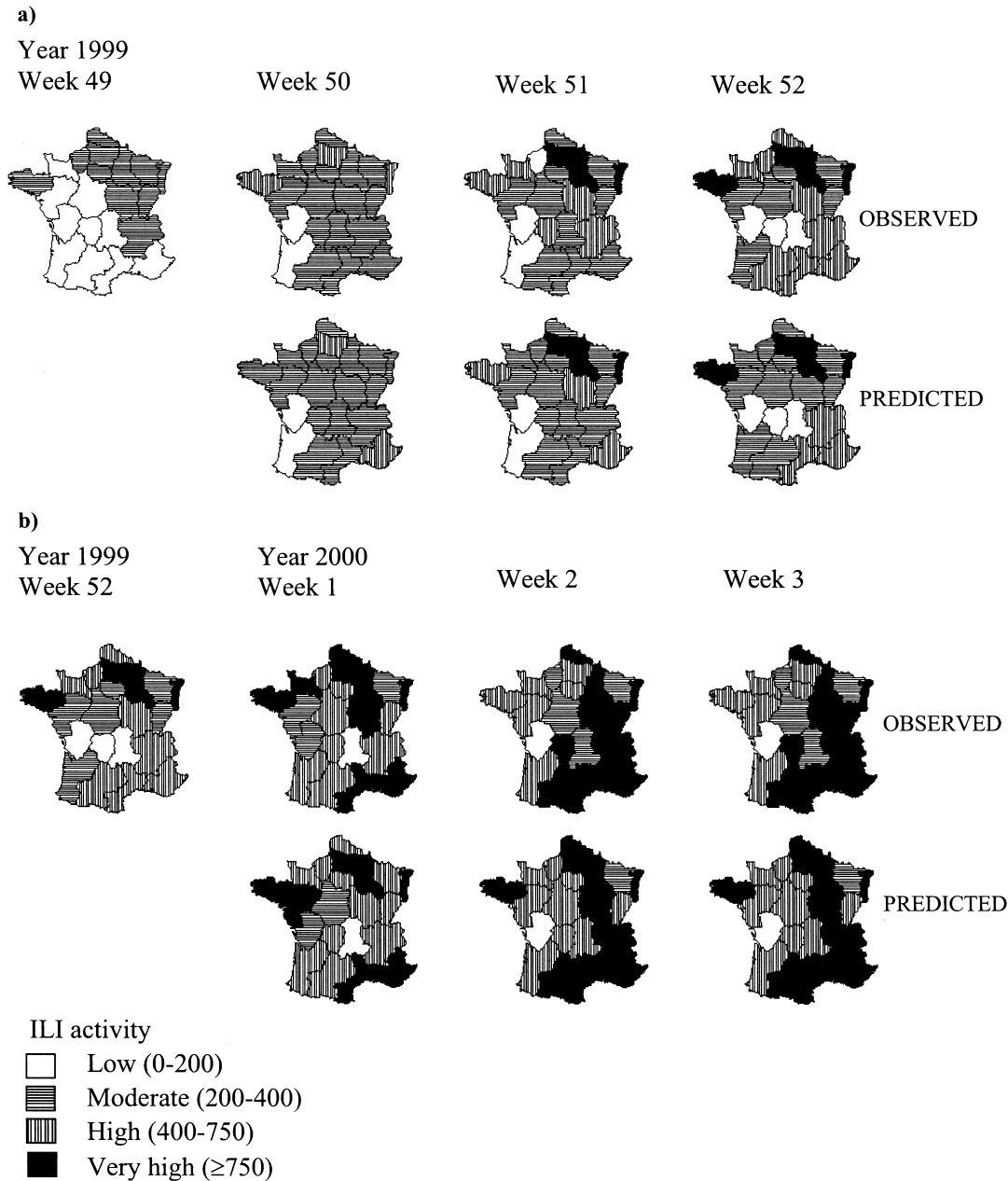
## DISCUSSION

In this study, we applied the method of analogues to forecasting the time and geographic spread of ILI epidemics in France from 1984 to 2002. This approach proved appropriate for forecasting both national and regional ILI incidences up to 10 weeks in advance during the epidemic and preepidemic periods. The correlations between the observed and predicted national incidences were above 0.58 (10-week-ahead forecasts) and up to 0.90 (1-week-ahead forecasts). The corresponding correlations between the observed and predicted regional incidences were above 0.66 and up to

0.81. This method yielded forecasts that were more accurate than those provided by linear autoregressive models.

A reason that might explain why the method of analogues outperforms autoregressive-like models in influenza series is the absence of exact periodicity regarding this disease. Although influenza epidemics occur in wintertime, the time of year of epidemic onset varies between November and March in France as well as in other temperate areas of the Northern Hemisphere (20). For this reason, models including seasonal terms, as autoregressive-like models do, do not fit influenza series well. The method of analogues is nonparametric and makes no assumption about the distribution or seasonality of the disease, which may explain the improvement in fit. In addition, several authors have suggested that local models (such as the method of analogues) outperformed global models (such as autoregressive-like models), especially when the system under study was complex (33, 37). In particular in epidemiology, Sugihara and May (36) reported a similar result in their study of measles time series.

Various distance criteria have been used for past applications of the method of analogues (24–27, 33, 34, 36, 37), but, to our knowledge, the issue of whether the forecasts are sensitive to the choice of distance criteria has not been studied. In our study, we used the euclidian distance criterion defined by the squared differences of raw incidence rates. ILI incidences are likely to be Poisson distributed, and the latter distance criterion is more appropriate for normal random variables. We also studied three other distance criteria: a euclidian distance applied to the log-transformed incidences, an exponentially weighted euclidian distance, and a chi-square-type distance. In particular, we used the log transformation as a variance-stabilizing transformation to obtain normal-like variables. Similarly, the chi-square-type



**FIGURE 6.** Prediction of the geographic spread of the 1999–2000 influenza epidemic in France vs. the observed spread by using the method of analogues. a) Upper panels, the regional incidences of influenza-like illnesses (ILI) observed at week 49 (starting on December 6, the week of epidemic onset) and weeks 50, 51, and 52 of 1999; lower panels, the 1-, 2-, and 3-week-ahead predictions obtained from influenza activity at week 49 of 1999. b) Upper panels, the regional ILI incidences observed at week 52 of 1999 (starting on December 27, 1 week before the peak incidence) and weeks 1, 2, and 3 of 2000; lower panels, the 1-, 2- and 3-week-ahead predictions obtained from influenza activity at week 52 of 1999. The parameters used for the predictions were those that minimized the error criterion ( $v = 3$ ,  $l = 4$ , and  $w' \propto 1/\text{distance}$  criterion). Numbers in parentheses, ILI incidence/100,000.

distance is more appropriate for count data characterized by increasing variance with increasing mean. This analysis revealed that the selection of neighbors was almost identical for all of the distance criteria; hence, the forecasts were not substantially different (data not shown). The choice of the distance criteria had very little impact on this study.

The method of analogues was applied here to surveillance data collected during interpandemic periods. An influenza pandemic is a major epidemic due to a completely novel or reemerging influenza virus spreading on a global scale. The viruses isolated and reported in France during the 18 influenza epidemics that we studied were restricted to the types

and subtypes isolated worldwide in humans in recent years—B, A/H1N1, and A/H3N2. In the case of a pandemic, the method of analogues would probably yield worse performances than those reported here. During a pandemic, incidences are expected to be much higher than those recorded during interpandemic periods. For a pandemic, other types of models, such as susceptible infected recovered mechanistic models similar to those used in the retrospective forecast of the global spread of the 1968–1969 Hong Kong influenza pandemic (15), would probably be more appropriate.

Exogenous covariates such as meteorologic or demographic factors may alter the diffusion process of influenza (21), but the determinants of the spread of this disease are still controversial (20). The method of analogues could be refined so as to include these covariates in the process of selecting the nearest neighbors, that is, in the definition of influenza activity. In previous applications of this method, a weighting algorithm was implemented to assign relative importance to covariates on the basis of the magnitude of their association with the dynamics under study (27). However, for influenza, these covariates and their corresponding weights are not yet known.

In conclusion, the method of analogues described here constitutes a nonparametric approach that, at least during interpandemic periods, is available for forecasting the diffusion of influenza epidemics. This method makes extensive use of past observed epidemic patterns to estimate the temporal and geographic dynamics of the diffusion process. The method of analogues is probably suitable for predicting other communicable diseases such as acute diarrhea, as long as historical observations are numerous enough to provide an extensive description of likely patterns and the disease displays recurring cycles. Like any other prediction method, it also requires real-time collection of data to make prediction worthwhile. Lastly, the method of analogues is a self-learning process, which allows for accuracy to improve with the length of the time series (24, 33).

## ACKNOWLEDGMENTS

The first author is supported by a grant from the French Ministère de l'Éducation Nationale et de la Recherche.

## REFERENCES

1. Flahault A, Dias-Ferrao V, Chaberty P, et al. FluNet as a tool for global monitoring of influenza on the Web. *JAMA* 1998; 280:1330–2.
2. Simonsen L, Clarke MJ, Williamson GD, et al. The impact of influenza epidemics on mortality: introducing a severity index. *Am J Public Health* 1997;87:1944–50.
3. Snacken R, Manuguerra JC, Taylor P. European influenza surveillance scheme on the Internet. *Methods Inf Med* 1998;37: 266–70.
4. Fourquet F, Drucker J. Communicable diseases surveillance: the Sentinel Network. *Lancet* 1997;349:794–5.
5. Quenel P, Dab W. Influenza A and B epidemic criteria based on time-series analysis of health services surveillance data. *Eur J Epidemiol* 1998;14:275–85.
6. Fleming DM, Zambon M, Bartelds AI, et al. The duration and magnitude of influenza epidemics: a study of surveillance data from sentinel general practices in England, Wales and the Netherlands. *Eur J Epidemiol* 1999;15:467–73.
7. Goddard NL, Joseph CA, Zambon M, et al. Influenza surveillance in England and Wales: October 1999 to May 2000. *Commun Dis Public Health* 2000;3:261–6.
8. Hashimoto S, Murakami Y, Taniguchi K, et al. Detection of epidemics in their early stage through infectious disease surveillance. *Int J Epidemiol* 2000;29:905–10.
9. Snacken R. Weekly monitoring of influenza impact in Belgium (1993–1995). *Pharmacoeconomics* 1996;9:34–7, 50–3.
10. Fleming DM, Cohen JM. Experience of European collaboration in influenza surveillance in the winter of 1993–1994. *J Public Health Med* 1996;18:133–42.
11. Laporte RE. How to improve monitoring and forecasting of disease patterns. *BMJ* 1993;307:1573–4.
12. Woodman R. Doctors and politicians clash over size of flu problem. *BMJ* 2000;320:138.
13. Serfling R. Methods for current statistical analysis of excess pneumonia-influenza deaths. *Public Health Rep* 1963;78:494–506.
14. Le Strat Y, Carrat F. Monitoring epidemiologic surveillance data using hidden Markov models. *Stat Med* 1999;18:3463–78.
15. Longini IM Jr, Fine PE, Thacker SB. Predicting the global spread of new infectious agents. *Am J Epidemiol* 1986;123: 383–91.
16. Flahault A, Letrait S, Blin P, et al. Modelling the 1985 influenza epidemic in France. *Stat Med* 1988;7:1147–55.
17. Flahault A, Deguen S, Valleron AJ. A mathematical model for the European spread of influenza. *Eur J Epidemiol* 1994;10: 471–4.
18. Baroyan OV, Rvachev LA, Basilevsky UV, et al. Computer modelling of influenza epidemics for the whole country (USSR). *Adv Appl Probabil* 1971;3:224–6.
19. Elveback L, Ackerman E, Gatewood L, et al. Stochastic two-agent epidemic simulation models for a community of families. *Am J Epidemiol* 1971;93:267–80.
20. Cox NJ, Subbarao K. Global epidemiology of influenza: past and present. *Annu Rev Med* 2000;51:407–21.
21. Cliff AD, Haggett P. Statistical modelling of measles and influenza outbreaks. *Stat Methods Med Res* 1993;2:43–73.
22. Box GEP, Jenkins GM. Time series analysis: forecasting and control. San Francisco, CA: Holden Days, 1976.
23. Stroup DF, Thacker SB, Herndon JL. Application of multiple time series analysis to the estimation of pneumonia and influenza mortality by age 1962–1983. *Stat Med* 1988;7:1045–59.
24. Lorenz EN. Atmospheric predictability as revealed by naturally occurring analogies. *J Atmosphere Sci* 1969;26:636–46.
25. Weigend AS, Gershenfeld NA. Time series prediction: forecasting the future and understanding the past. Reading, MA: Perseus Books Publishing, 1994.
26. Solomatine DP, Rojas CJ, Velickov S, et al. Chaos theory in predicting surge water levels in the North Sea. Presented at the 4th International Conference on Hydroinformatics, Iowa City, Iowa, July 23–27, 2000.
27. Brabec B, Meister R. A nearest neighbor model for regional avalanche forecasting. *Ann Glaciol* 2001;32:130–4.
28. Sugihara G, Grenfell B, May RM. Distinguishing error from chaos in ecological time series. *Philos Trans R Soc Lond B Biol Sci* 1990;330:235–51.
29. Stone L. Coloured noise or low-dimensional chaos? *Proc R Soc Lond B Biol Sci* 1992;250:77–81.
30. Garnerin P, Saidi Y, Valleron AJ. The French Communicable Diseases Computer Network. A seven-year experiment. *Ann N*

- Y Acad Sci 1992;670:29–42.
31. Costagliola D, Flahault A, Galinec D, et al. A routine tool for detection and assessment of epidemics of influenza-like syndromes in France. *Am J Public Health* 1991;81:97–9.
  32. Simonsen L, Clarke MJ, Stroup DF, et al. A method for timely assessment of influenza-associated mortality in the United States. *Epidemiology* 1997;8:390–5.
  33. Casdagli M. Chaos and deterministic versus stochastic nonlinear modelling. *J R Stat Soc B* 1992;54:303–28.
  34. Grassberger P, Schreiber T, Schaffrath C. Nonlinear time sequence analysis. *Int J Bifurcat Chaos* 1991;1:521–47.
  35. Hastie T, Friedman J, Tibshirani R. The elements of statistical learning: data mining, inference, and prediction. New York, NY: Springer Verlag, 2001.
  36. Sugihara G, May RM. Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. *Nature* 1990;344:734–41.
  37. Farmer JD, Sidorowich JJ. Predicting chaotic time series. *Phys Rev Lett* 1987;59:845–8.