STATISTICS IN MEDICINE
Statist. Med. 2005; **24**:3447–3461
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2196

Modelling antigenic drift in weekly flu incidence

B. F. Finkenstädt[1,*,†], A. Morton[1,2] and D. A. Rand[2]

[1] *Department of Statistics, University of Warwick, U.K.*
[2] *Mathematics Institute, University of Warwick, Coventry CV4 7AL, U.K.*

SUMMARY

Since influenza in humans is a major public health threat, the understanding of its dynamics and evolution, and improved prediction of its epidemics are important aims. Underlying its multi-strain structure is the evolutionary process of antigenic drift whereby epitope mutations give mutant virions a selective advantage. While there is substantial understanding of the molecular mechanisms of antigenic drift, until now there has been no quantitative analysis of this process at the population level. The aim of this study is to develop a predictive model that is of a modest-enough structure to be fitted to time series data on weekly flu incidence. We observe that the rate of antigenic drift is highly non-uniform and identify several years where there have been antigenic surges where a new strain substantially increases infective pressure. The SIR-S approach adopted here can also be shown to improve forecasting in comparison to conventional methods. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: influenza; SIR-S models; antigenic drift; flu incidence time series

1. INTRODUCTION

Influenza continues to exhibit high levels of incidence in human populations through its capacity to undergo antigenic transformation leading to partial escape from immune response [1–4]. Moreover, flu is a major contributor to mortality and morbidity throughout the world [4, 5] posing enormous medical and public health challenges. The flu virus is unusual in that it is continuously undergoing immunologically significant evolution that allows it to evade the host immune surveillance through the genetic processes termed *antigenic drift* and *shift*. A flu infection is thought to entail lifelong immunity with respect to the infecting strain and the host immune response changes continuously in the sense that immunity to one strain of influenza confers partial immunity against related strains [6]. However, genetic variation

*Correspondence to: Bärbel F. Finkenstädt, Department of Statistics, University of Warwick, Coventry CV4 7AL, U.K.
†E-mail: b.f.finkenstadt@warwick.ac.uk

Contract/grant sponsor: Health Predict Ltd.

Received 30 July 2004
Accepted 19 January 2005

produces antigenic novel strains at such a high rate that most people who have had flu are susceptible to a new circulating strain of flu within a few years of infection [5].

The influenza virus is divided into three main types (A, B and C) according to differences in two important internal proteins [7]. Both influenza type A, which is found in a large variety of birds and mammals, and type B, which is largely confined to humans, are subject to *drift* dynamics, the evolutionary process whereby mutations in the appropriate epitope of the hemagglutinin molecule give the mutant virions a selective advantage [1, 2]. Selection is applied by the host immune system as the mutations change the properties of the antigenic sites (epitopes) that the human system must recognize in order to suppress flu. In this way influenza has the potential to recur like an endemic disease in populations that can have substantial immunity from previous exposures to the disease [4]. In addition, there is the possibility of *antigenic shift* through reassortment events whereby infection of a single cell, often of an avian host, by two different flu viruses of type A, can result in the production of a new virus whose genome contains mixtures of the parental RNA segments. Although antigenic shifts occur only every few decades they are potentially of far greater concern as the more dramatic changes allow for sub-types to occur to which there exist no or little immunity and therefore entail unimpeded spread through the host population [4]. Two of the three pandemics that occurred during the 20th century were reassortants and all three, the 1918, 1957 and 1968 pandemic viruses, contain gene segments from reservoirs of the influenza A virus in birds [4].

The strain stability of some other diseases such as measles simplifies the mathematical modelling of the natural history of the disease and this facilitates the construction of predictive models for them (see, for example, References [8–10]). Household data studies, where the serological statuses of all individuals in a household before and after an epidemic period are determined, have also received attention in the study of influenza transmission. The studies in References [11, 12] use a probabilistic model to estimate the probability of becoming infected through transmission within a household or within a community. This modelling approach is extended in References [13, 14] to allow for heterogeneity in infectiousness and susceptibility. In Reference [15] a Bayesian MCMC approach is proposed to estimate the duration of the infectious period and instantaneous risk of infection using household data that was obtained from the 1999 to 2000 'Epigrippe' study in France [16] where daily data was obtained on the presence or absence of influenza symptoms in 334 households for 15 days after an index case was observed in the households.

The literature contains a number of deterministic mathematical models for multi-strain diseases (see, for example, References [17–19], and [20] for an overview) which is currently a highly active area of research [21–23]. In principle, one could use these to develop a predictive model. However, they require detailed information about both the cross-reactivity of the different strains and the corresponding antibodies and, additionally, the infective pressure of the strains. The authors of a recent review in Reference [5] point out the overwhelming complexity of the models as even for a modest number of strains modelling becomes analytically and computationally intractable, although References [18, 24, 25] have used assumptions to reduce dimensionality. Another problem is that the size of populations in the different compartments of a multi-strain model may become very small so that stochastic rather than deterministic modelling is required [5].

Given the nature and quality of the data for flu it is inconceivable that the parameters of a multi-strain compartmental model could be estimated or even be identified from non-compartmental data. Moreover, the time-series of case numbers fail to distinguish between influenza A, influenza B and flu-related symptoms. Our study aims to develop a model that is able to take important aspects of strain structure into account but has a modest enough complexity to allow parameter identification and estimation from time series data on weekly flu incidence. We suggest a discrete-time susceptible-infected-recovered-susceptible (SIR-S) approach to model the process by which previously infected individuals become susceptible either through the waning of their immunity or through antigenic drift in a very simple way, essentially in terms of a function u_t that measures the effective rate at which previously infected and immune individuals re-enter the susceptible class.

2. DATA

Several sources of data are available for studying the population dynamics of flu. Monthly mortatility from pneumonia and influenza since the beginning of the 20th century is available for the U.S.A. and weekly records of influenza-like illnesses (ILI) on a discrete scale from 0 to 4 for various countries have been compiled since 1995 by the World Health Organization. For the purpose of this study we use weekly incidence rates of ILI per 100 000 inhabitants compiled by the French Sentinel surveillance and publically available on their webpage. We use the weekly time series from week 44, 1984 to week 21, 2002 for model fitting giving a sample size of 914 observations.

The time series plot (see Figure 1) shows that the data exhibit pronounced yearly cyclicity where the intensity of the epidemics varies from year to year. All of the cycles are sharply peaked: the climb from a low trough value to the peak occurs within less than 4 weeks. There is only one peak per epidemic year which is located between week 45 and week 5, usually around week 50, in any year. Epidemic activity gradually decreases after week 5. There is little epidemic activity in summer and the lowest values are observed from week 20 to week 35 of the calendar year. This yearly cyclicity may be the result of an endogenous population cycle due to the non-linear predator–prey interaction excited by an exogenous yearly seasonality. However, the reasons for the yearly seasonality are not yet understood. Temperature, humidity, crowding indoors in winter and/or the school season may all be explanatory factors [5, 26]. In common with other infectious disease incidence data, the marginal distribution of the incidence time series is strongly asymmetric which can, for example, be inferred from the simple five-number summary statistics given by $(\text{minimum, lower quartile, median, upper quartile, maximum}) = (0, 11, 34, 100, 1793)$.

We also considered weekly ILI data gathered by surveillance networks from other countries such as the Netherlands, England and Wales and Portugal and found that there are large discrepancies between data from different countries. The main reason lies in both the case definition of clinical influenza used in these networks and the behaviour of the individual suffering from ILI. For example, the proportion of patients who will consult a GP with a particular combination of symptoms will differ from one country to another, as will the diagnosis by a GP in the different surveillance networks. Whilst such information was hardly available for the other countries the data from the French Sentinelle surveillance system are relatively well understood and their availability has facilitated active research on influenza in France (see, for example, References [27, 28]).

The French Sentinelle data are consultations for ILI in general practice and it is estimated that on average over the epidemic year 40 per cent of patients consulting the GP with ILI

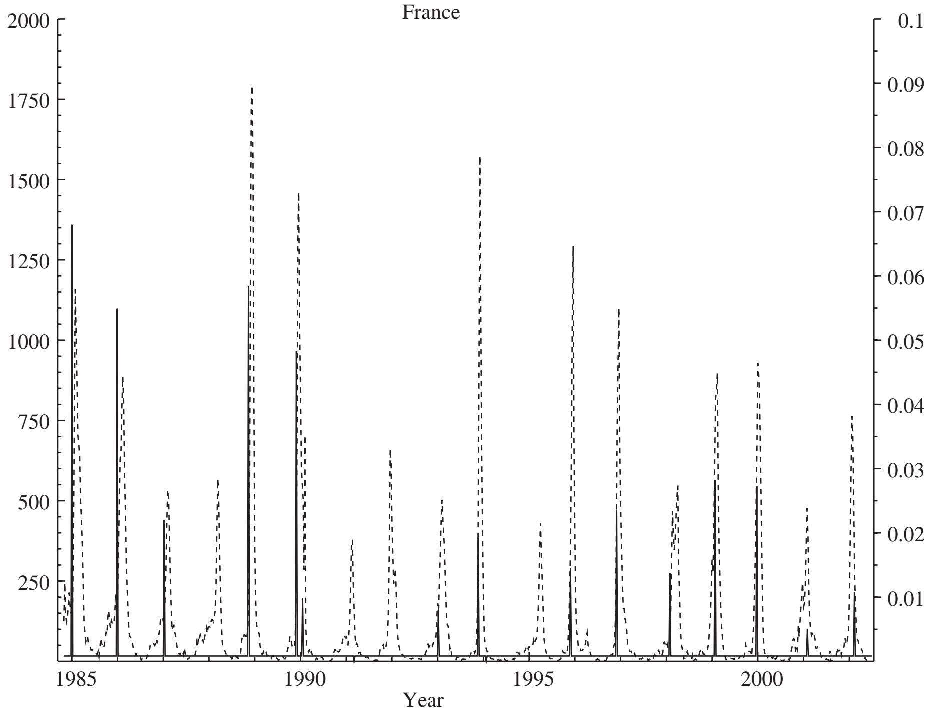

Figure 1. Estimates of u_t (solid line) which indicate the proportion of recovered individuals who lose immunity in each week for comparison with reported incidence (dashed line). The number of influenza cases per 100 000 is indicated on the left axis and the values of u_t on the right axis.

are actually infected by an influenza virus [29]. On the other hand, it is estimated that in the French Sentinelle network 40–50 per cent of patients that suffer from ILI will not consult a GP. Experimental challenging of humans with influenza and serological survey showed that the rate of latent influenza infection is between 25 and 40 per cent and with all these rules the true rate of influenza in France would be around 12–15 per cent of the population per epidemic year [29]. We found that the incidence reported in the observed time series over the epidemic year is in agreement with these rates taken with respect to a population size of 100 000. This information gives us some confidence that the reported level of the French data corresponds approximately to the true level.

3. MODEL AND INFERENCE

3.1. Model

Let S_t, I_t and R_t denote the number of susceptible, newly infected and recovered (immune) individuals at week t in a population of constant size P so that $P = S_t + I_t + R_t$. We assume that, given the numbers of infecteds and susceptibles I_{t-1} and S_{t-1} at week $t-1$, then the number of infecteds I_t in week t is drawn from a gamma distribution with mean $E(I_t|I_{t-1}, S_{t-1})$

given by $\lambda_t = \beta_t I_{t-1}^{\alpha} S_{t-1}$ and shape parameter $\gamma_t = cI_{t-1}$ so that $\mathrm{Var}(I_t|I_{t-1}, S_{t-1}) = \lambda_t^2/\gamma_t$. Here β_t represents a factor of proportionality which is allowed to be seasonally changing with a period of 1 year. We refer to β_t as contact rate noting that, since the data are aggregated over heterogeneous populations as well as over time, the parameters estimated for β_t will not represent the contact rates of a pure mathematical multi-strain model but will adjust to accommodate for the effects of such an aggregation. The approach outlined here has proved successful in approximating measles epidemics [10]. Note that both the conditional mean and the conditional variance are allowed to change over the epidemic time. The parameter α is around one if the mixing between the infecteds and susceptibles is homogeneous following the mass action paradigm. For a further detailed interpretation of these parameters in a model for childhood diseases see also References [8, 9].

In the week following infection each individual joins the recovered class. To model the antigenic drift and the decline in immunity of recovered individuals we assume that a proportion $0 \leqslant u_t \leqslant 1$ leaves the recovered population every week and re-enters the susceptible population. Thus the prescription $S_{t+1} = S_t - I_{t+1} + u_t R_t$, $R_{t+1} = (1-u_t)R_t + I_{t-1}$ completes the definition of the SIR-S model. Since the surveillance data is normalized as a rate per 100 000 inhabitants we always set $P = 100\,000$.

3.2. *Methods of inference*

We assume that the rate u_t at which immune individuals re-enter the susceptible class is constant and equal to $\bar{u}$ except for at most 1 week in each flu season. The corresponding time can be any in that season. We can interpret this as a new strain arriving each year potentially causing a larger proportion of the immune class to become susceptible again. Thus we estimate the value $\bar{u}$, the times t when the changes in u_t occur and the values of u_t at these times. We also simultaneously estimate the following epidemiological parameters: the shape parameter c, the mass-action correction α, and the weekly contact rates β_s, $s = 1, \ldots, 52$. To enable comparisons we refer to the above model as the *full model* and compare it with one where $u_t = u$ is forced to be constant for all time. We call this restricted SIR-S model the *null model*.

Likelihood inference for the null model is straightforward. Let $\boldsymbol{\theta} = (c, u, \alpha, R_1, \boldsymbol{\beta})$ denote the vector of all unknown parameters where $\boldsymbol{\beta}$ is a vector containing 52 values of the contact rate and let $\mathbf{I}$ denote the vector of data $I_1, \ldots, I_N$. The log of the likelihood, conditional on the initial value I_1, for the model above is

$$\log L(\boldsymbol{\theta}|\mathbf{I}) = \sum_{t=2}^{N} \log f(I_t|I_{t-1}, S_{t-1}) \tag{1}$$

The gamma distribution f of the transmission function is

$$f(I_t|I_{t-1}, S_{t-1}) = \frac{I_t^{(\gamma_t - 1)} (\gamma_t/\lambda_t)^{\gamma_t} \exp(-I_t \gamma_t/\lambda_t)}{\Gamma(\gamma_t)} \tag{2}$$

with λ_t and γ_t as defined in the previous section. The ML estimator $\hat{\boldsymbol{\theta}}$ is the vector of parameter values which maximizes (1).

Our likelihood optimization procedure uses a mixture between non-linear and linear optimization since, for given values of u, α and R_1, the likelihood with respect to the contact rate parameters is maximized by $\hat{\beta}_s = \sum_{t=2}^{N} J_{t-1}D_t / \sum_{t=2}^{N} I_{t-1}D_t$ where $J_{t-1} = I_{t-1}I_t/I_{t-1}^{\alpha}S_{t-1}$, $D_t = 1$ if t is a multiple of s and $D_t = 0$ otherwise, $s = 1, \ldots, 52$.

In the null model a time constant u_t implies a geometrically declining lag structure as past numbers of infecteds leave the susceptible class and thus S_{t-1} can be expressed as

$$S_{t-1} = P - (1-u)^{t-2}R_1 - I_{t-1} - \sum_{i=1}^{t-2}(1-u)^{t-2-i}I_i$$

In order to adapt the estimation procedure to the full model we assume that the usual rate at which immune individuals re-enter the susceptible class is constant (and equal to $u_t \equiv \bar{u}$) except for $N^* \ll N$ occasions when u_t is free and does not necessarily equal $\bar{u}$. We associate with the parameter vector $\boldsymbol{u}^* = (u_1^* \; u_2^* \cdots u_{N^*}^*)'$ a vector of time points $\mathbf{t}^*$ such that in week $t = t_j^*$, $u_t = u_j^*$. Therefore the equation representing changes in the immune population is

$$R_t = (1-u_t)R_{t-1} + I_{t-1} \text{ where } \begin{cases} u_t = \bar{u} & \text{if } t \notin \mathbf{t}^* \\ u_t = u_j^* & \text{if } t = t_j^*, j = 1, \ldots, N^* \end{cases}$$

The question is how to allocate the time points t_j^*, $j = 1, \ldots, N^*$, throughout the course of the data and how to choose N^* itself. We decided to take one element of $\mathbf{t}^*$ to be in each flu season so that N^* is (approximately) the number of years covered by the data. The corresponding time can be any in that season. We can interpret this as a new strain arriving each year and instantly causing a proportion of the immune class to become susceptible again. The parameter vector is now $\boldsymbol{\theta} = (c, \bar{u}, \alpha, R_1, \boldsymbol{\beta}, \mathbf{t}^*, \mathbf{u}^*)$. Although the likelihood can easily be evaluated for given parameter values it cannot be directly maximized with respect to $\mathbf{t}^*$ and $\boldsymbol{u}^*$. Therefore we search for the maximum likelihood point by using the Monte-Carlo sampling method of simulated annealing [30]. This technique is shown in Reference [31] to be particularly suited to maximization of likelihood functions with a multitude of local maxima, as is the case for our model. From a statistical viewpoint the method is most easily understood as an adaptation of a regular Monte Carlo Markov chain algorithm.

We begin by simulating a Markov chain with stationary density proportional to the likelihood. This is achieved by using a single-component Metropolis algorithm (see, for example, Reference [32]) evaluating the one-dimensional conditional distributions as shown in Reference [33]. If we denote the current state of the ith parameter x and the value of the conditional density at this point $\pi(x)$, we first sample a candidate point y from some proposal distribution $q(y|x)$ which is symmetric around x. Then we accept the candidate point y with probability $\delta(x, y) = \min\{1, (\pi(y)/\pi(x))^m\}$ where the starting value of m is 1. Thus there is less chance of y being accepted the smaller its conditional density in comparison to that of x. Using this Metropolis algorithm to sample each parameter in turn is done repeatedly to provide a dependent sample from the required density. To locate the maximum the constant m in the Metropolis algorithm is slowly increased so that candidates are only likely to be accepted when they increase the likelihood. As m increases the corresponding stationary density for the chain moves closer and closer to the point mass at the maximum of $L(\boldsymbol{\theta})$. For each new set of parameter values we evaluate the likelihood and store the parameters which correspond to the highest observed value of $L(\boldsymbol{\theta})$.

One important issue is the specification of a 'flu season', i.e. how to choose the sets $\mathbf{T}_j$, $j = 1, \ldots, N^*$, such that $t_j \in \mathbf{T}_j$. So far we have taken each $\mathbf{T}_j$ to have a period of 1 year and our first strategy was to take each $\mathbf{T}_j$ to be a calendar year. However, our imposed 'flu season' then dissects the actual one and so we also tried taking each $\mathbf{T}_j$ to begin on week 45 of the calendar year. It is straightforward to estimate the optimal week of the year in which to commence the flu season within the current framework. However the season should be chosen so that the estimated shifts in susceptibility can most effectively explain the observed epidemics from a biological viewpoint.

4. RESULTS

4.1. Model parameters and immunity decay

The parameter estimates for both models are summarized in Table I. The coefficient of determination R^2 shows a good fit for both the null model ($R^2=0.92$) and the full model ($R^2=0.95$). The increase in the likelihood between the full model and the null model is substantial despite a maximum of two extra parameters for each year. The Akaike information criterion (AIC) which penalizes model complexity introduced through extra parameters clearly shows the significance of the changes in susceptibility. One conclusion is therefore that the data shows significant evidence of considerable heterogeneity in the seasonal level of antigenic shifts and, in particular, the presence of antigenic surges where a jump that is considerably larger than usual occurs. The location and magnitude of the jumps in immunity loss are shown in Figure 1 for comparison with the time-series data. It is intriguing that the jumps tend to occur at the beginning of the yearly epidemic. Their sizes are within a range so that each year different proportions but not all of the recovered individuals return to the susceptible class and therefore the size of the following epidemic is positively correlated with the previous immunity decay.

Because the time intervals cover only 1 week the numerical values of $\bar{u}$ are very small and thus the values of the weekly immunity parameter $(1-\bar{u})$ are very close to one. For

Table I. Maximum likelihood estimates with their approximate standard error (in parentheses), the negative log-likelihood (NLLik), Akaike's information criterion, and coefficient of determination for the full (left) and the null (right) model fitted to French ILI incidence rates (sample size $N=914$).

	Full model	Null model
c	0.153 (0.007)	0.096 (0.004)
$(1-\bar{u})^{52}$	0.958 (1.2e-3)	0.926 (2.5e-3)
α	0.845 (0.007)	0.890 (0.009)
R_1	94,527 (480)	63,713 (2553)
NLLik	3859	4101
AIC	7726	8211
R^2	0.949	0.918

For the full model the immunity loss rate $\bar{u}$ in the table corresponds to the regular rate of loss in immunity when t does not correspond to one of the occasions when variable jumps are allowed ($t \notin \mathbf{t}^*$) and so excludes the size of one jump per year (see Figure 1). It is therefore smaller than the value reported for the null model where $(1-\bar{u})^{52}$ is equal to $(1-u)^{52}$ giving the total immunity decay over the year.

convenience we present the factor $U_{\text{ann}} = (1 - \bar{u})^{52}$ in Table I. This gives us, approximately, the proportion of immunes that remain immune over a year. The estimated likelihood decreases sharply if $(1 - \bar{u})$ is approaching one and constructing confidence intervals from the estimated standard errors show that $(1 - \bar{u})$ is significantly different from one. This is the case for both the null and the full model. It therefore seems that the data clearly supports the presence of immunity loss and the use of an SIR-S model in contrast to a simpler SIR model.

Stochastic simulations of the models suggest that very small changes in $\bar{u}$ (or in u for the null model) have a substantial impact on the dynamics of the system. The estimated value of around 0.926 would suggest that approximately 7.4 per cent of the recovered individuals have re-entered the susceptible class after 1 year. Although this is somewhat comparable with Reference [34] where an immunity loss of about 5 per cent per year is concluded from a linear fit of loss of immunity against time of strain isolation using data from experiments with human volunteers [35] an interpretation with regard to the precise estimated size of the parameters would be overambitious given that the data are a collection of ILI and also parameter estimates will adjust to the reporting level. On the other hand, predictions from this model can still be useful for planning the distribution of vaccine if one can assume that the number of cases consulting a GP and reported as ILI would be proportional to the number of cases demanding medical care.

4.2. Modelling contact parameters

Although the seasonal pattern of the contact rates for the two models are almost identical their levels are not with the estimated contact rates for the main model being larger. This is compensated for in this model by the smaller estimated values of α. It is interesting to investigate the seasonal pattern of the 'residual' contact rate series that results from taking the ratio $I_t/I_{t-1}^{\alpha}S_{t-1}$ using the estimated parameters for the null model as reported in Table I[‡] The empirical autocorrelation in Figure 2 clearly shows that this residual contact rate series has a yearly cycle albeit dramatically reduced in comparison to the yearly autocorrelation of the incidence series I_t implying that most of the annual cyclicity observed in I_t is already explained by the fitted predator–prey interaction term $I_{t-1}^{\alpha}S_{t-1}$.

We started by fitting a seasonal indicator model to capture this cycle. Estimation is reasonably unproblematic as the model is linear in these parameters and we have at least 18 years of data. However, a further reduction in the number of parameters could, for example, be achieved by fitting a sinusoidal curve of the form

$$\beta_t = c + \sum_{j=1}^{p} (a_j \cos(t\mathrm{j}\omega) + b_j \sin(t\mathrm{j}\omega)) \tag{3}$$

to the contact rates, where $\omega = 2\pi/52$ is the fundamental frequency and $p \leqslant 26$. Note that the seasonal indicator model arises as a linear combination of the sinusoidal model (3) when $p = 26$ so that in this case the two models for the contact rate are identical. The AIC chooses $p = 7$ and the estimated curve is plotted on top of the estimated contact rates for $p = 26$ in Figure 2. A further conclusion is thus that the French data support the conjecture that contact rates are subject to seasonal variation. Although we find that a smooth curve can achieve some

[‡] Alternatively, one could use the parameters of the full model but as the contact rates were virtually the same it is hard to believe that it would make any difference to our conclusions.

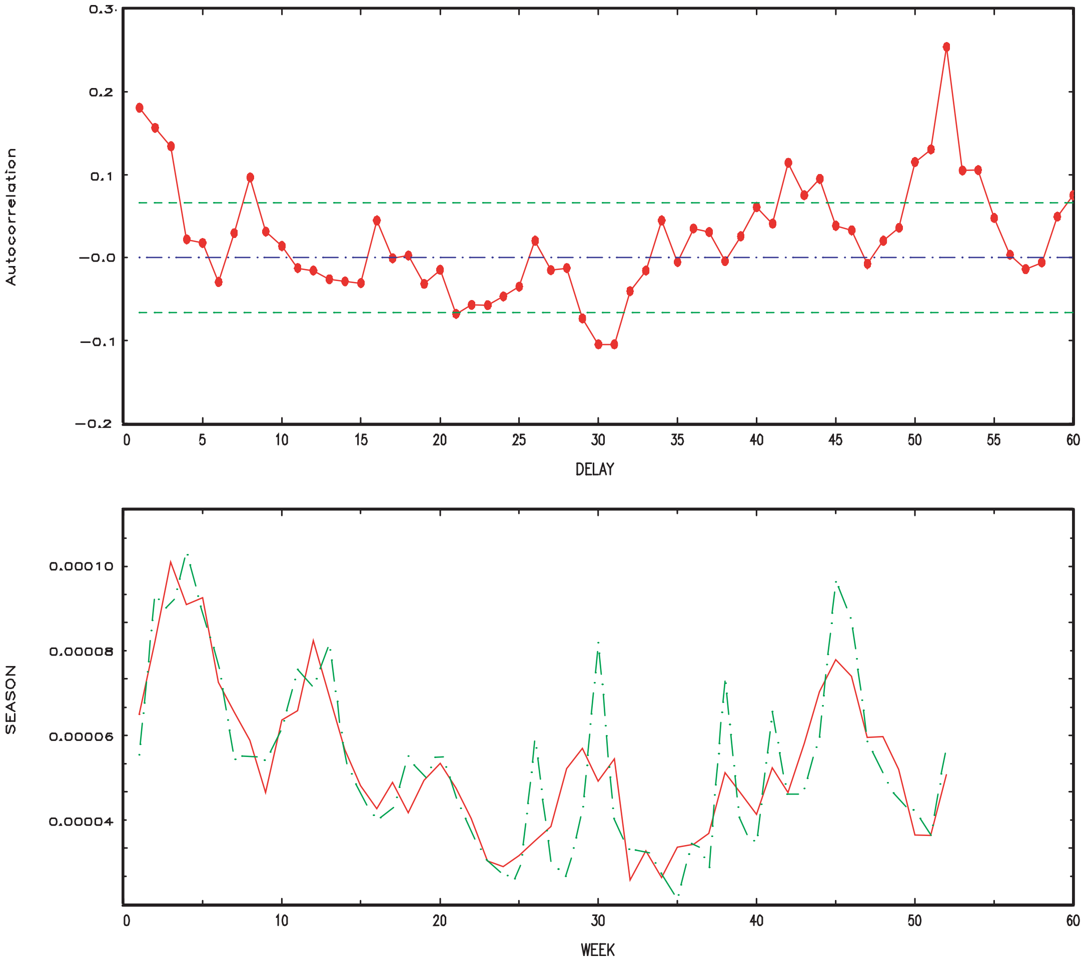

Figure 2. Top: Empirical autocorrelation for the ‘residual’ series that is obtained by taking $I_t/I_{t-1}^{\alpha}S_{t-1}$. It clearly indicates the existence of a yearly cycle in the contact rates. Bottom: Estimated contact rates over the calendar year. The bold lines are the rates for the smoother sinusoidal model with $p=7$ and the dash–dotted lines for the full seasonal indicator model with $p=26$.

reduction in model complexity given the data it is yet to be explored whether real contact rates do vary smoothly over the year, as for instance, contact processes may change abruptly (when school starts) whilst temperature variations usually happen more gradually.

4.3. *Forecasting influenza incidence*

One of our aims was to forecast influenza incidence. In Figure 3 we plot deterministic forecasts and stochastic simulations for the French data from $t=1$ (week 44 of 1984) for comparison with the actual series. For the null model, once the initial conditions are ‘forgotten’, the model trajectory settles down to a steady limit cycle and any variations in the heights of the

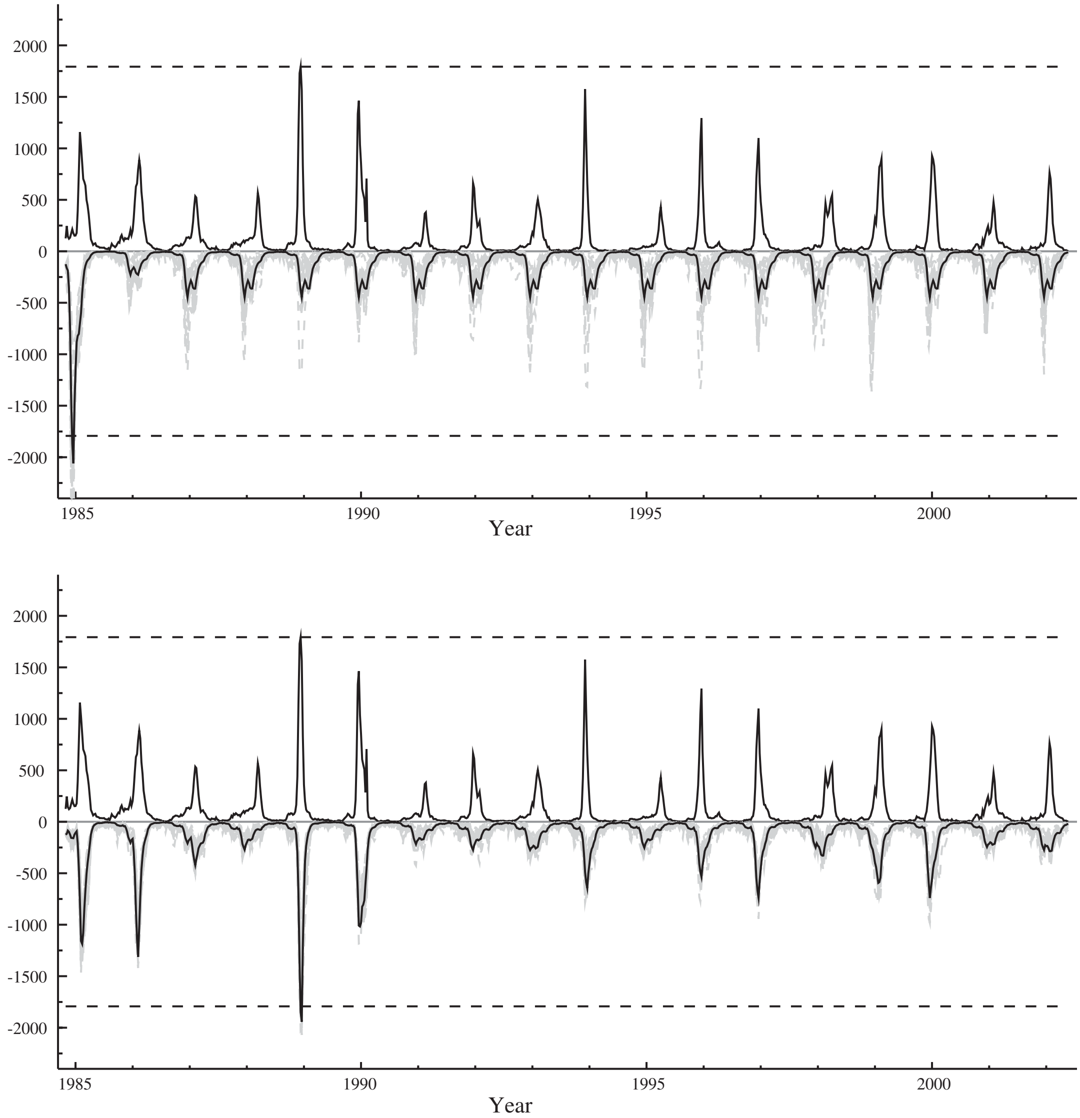

Figure 3. Deterministic and stochastic simulations for the French series for the null model (top graph) and the full model (bottom graph). In both graphs the deterministic model forecasts conditional on I_1 are plotted below the horizontal axis as a mirror image to the observed data which is plotted above the horizontal axis. The dotted lines below the horizontal axis show 20 stochastic simulations from each model. The dashed horizontal line indicates the value of the highest peak (observed in winter 1989/1990).

peaks are merely a result of stochasticity. This is in contrast to the full model where the estimated systematic shifts in susceptibility explain the observed variability in the intensity of the influenza epidemics. Although for in-sample prediction the full model outperforms the

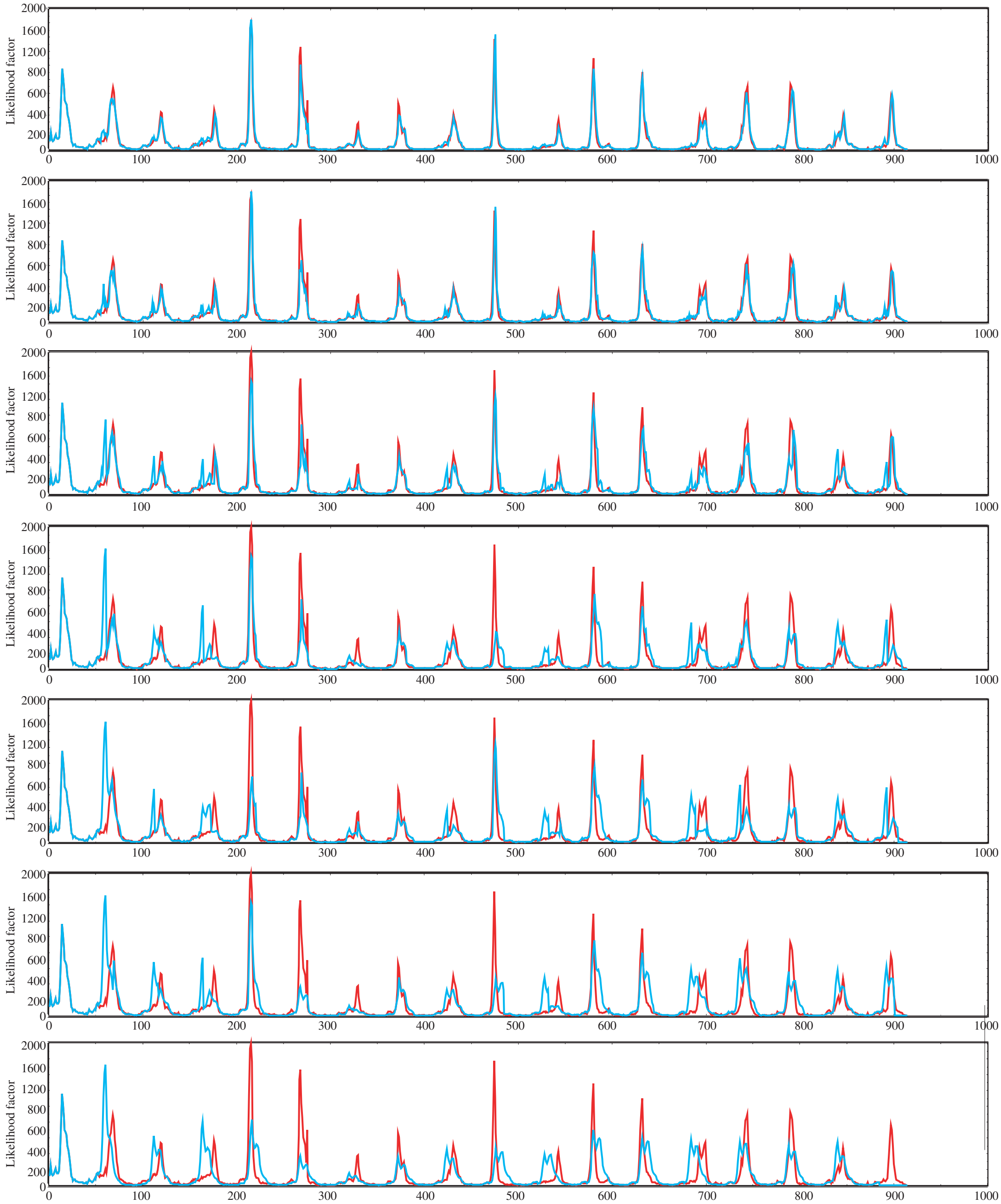

Figure 4. In-sample prediction (light solid line) from the null model for French incidence rates (dark solid line). Here we used the first year as a transient and then started prediction updating every 1, 2, 4, 8, 12, 16, 52 weeks (from top to bottom panel).

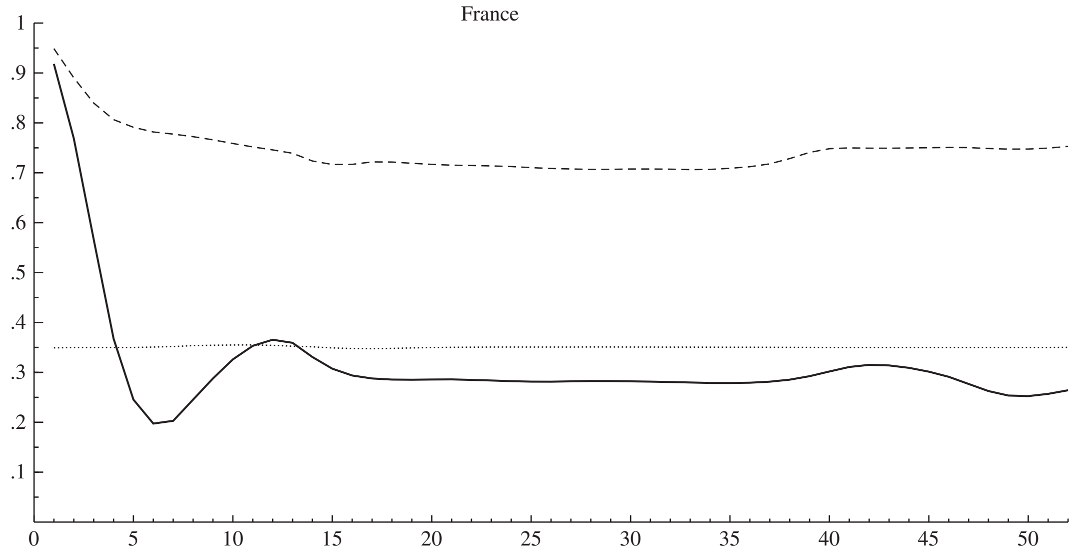

Figure 5. The proportion of variability explained by model forecasts using prediction steps of between 1 and 52 weeks. The bold line indicates the null model, the dashed line the full u_t model, and the dotted line indicates the performance from simply using the mean case count for each week of the year.

null model, currently it cannot be used for out-of-sample prediction unless it is extended to incorporate explanatory factors of these shifts in susceptibility. This could be, for example, genetic information about the virus evolution that quantifies how much the genetic material of a new strain is different from the previous one.

Figure 4 demonstrates the prediction performance of the SIR-S null model with parameter estimates as reported in Table I and a seasonal indicator model for the contact rates (see Figure 2, bottom panel). Each panel in Figure 4 shows the observed time series against the predicted time series when the information about incidence was updated every $1, 2, 4, 8, 12, 16, 52$ weeks. The top panel thus shows the model fit (one-step-ahead predictions) whilst the bottom panel demonstrates the prediction performance if updates are only taken every 52 weeks, that is at each time point we tried to predict the incidence for the whole next year. Obviously, the higher the frequency of updating, the closer are the predictions to the true curve. For France, the one-step-ahead error grows by a factor of 1.4 if prediction is updated every 2 weeks and has doubled if updating takes place every 5 weeks. We also assessed out-of-sample prediction performance of the null model, where we estimated parameters from 1 year less of data and attempted to predict the omitted last epidemic year assuming different step lengths for the updating process. As the estimated parameters for the restricted data set were virtually identical to the estimates reported in Table I the conclusions about the out-of sample prediction are the same as for the in-sample prediction.

The fact that the null model has significant potential for providing better short-term forecasts than the conventional method of using just the average yearly cycle is demonstrated in Figure 5 where we show the proportion of variability explained by model forecasts for a range of lead times from 1 week to 1 year. The dotted line indicates the quality of the forecasts obtained by taking the predicted number of cases in week i of the calendar year to be the mean of the number in week i across all years of the data. Thus for predicting up to 1 month into the future this mechanistic SIR-S-based-model is significantly superior to the usual predictor of using an average cycle.

5. SUMMARY AND CONCLUSIONS

This analysis has used time-series data on weekly flu and flu-like incidence in France to identify and quantify antigenic shifts in terms of the infective pressure induced by them at the population level. It shows that the size is highly variable with large surges occurring at certain times. It has also shown the potential of the restricted null model for providing short-term forecasts. The fact that the parameter estimates hardly changed at all when two more years of data are added suggests the model can be used with confidence for forecasting influenza incidence up to a month in advance. We also demonstrated that the full model has potential for influenza prediction. However, a predictor variable of the future jump size of u_t is needed to make this effective and we believe that some limited information will enable the model to provide superior longer term forecasts. Such an approach would be particularly interesting as information about the genetic variation of new strains and sub-types which usually appear firstly in the southern hemisphere would be available before the following epidemic year of the northern hemisphere and could be exploited for prediction. Other developments include using the model to investigate the effect of vaccination strategies and to study the role of other physical and biomedical covariates (e.g. temperature and death rates). For example, a preliminary analysis of temperature time series indicates that the comparative harshness of the winter weather only plays a minor role in influencing the epidemics.

It should be noted that the model suggested here is designed to reflect the dynamics of the observed data and our study shows that it does so adequately. One could in principle extend the current modelling approach to allow for a seasonally varying stochastic reporting process. However, not enough information is currently available to attempt this. If there does exist such a seasonally recurring pattern in the reporting rate, then this is likely to be reflected in the seasonally changing contact rates.

Although ideally we would want information about the population dynamics of variants within each subtype and better understanding of the cross-reactivity among strains, it is likely to be a long time until enough data can be collected to estimate such processes. The aim of this paper was to suggest a simpler modelling approach that can avoid some of the current severe data limitations but that allows some insight into the dynamics of flu at the population level and can be used for forecasting. However, we have only been able to test this out for the French surveillance data which is reasonably well understood and has been collected for a long time. To fully make use of this approach it would be extremely desirable to have weekly surveillance data of higher quality with confirmary testing of cases from several countries from both hemispheres.

ACKNOWLEDGEMENTS

This research was supported by Health Predict Ltd., Oxford, U.K. We are grateful to Andrew Roddam and Fabrice Carrat for some useful discussions that have influenced this work.

REFERENCES

1. Fitch WM, Leiter JME, Palese P. Positive Darwinian evolution in human influenza A viruses. *Proceedings of the National Academy of Sciences of the United States of America* 1991; **88**:4270–4274.
2. Fitch WM, Bush RM, Bender CA, Cox NJ. Long term trends in the evolution of H(3) HA1 human influenza type A. *Proceedings of the National Academy of Sciences of the United States of America* 1997; **94**:7712–7718.
3. Pease CM. An evolutionary epidemiological mechanism, with applications to type A influenza. *Theoretical Population Biology* 1987; **31**:422–452.
4. Webby RJ, Webster RG. Emergence of influenza A viruses. *Philosophical Transactions of the Royal Society* 2001; **B**(356):1817–1828.
5. Earn DJD, Dushoff J, Levin SA. Ecology and evolution of the flu. *Trends in Ecology and Evolution* 2002; **17**(7):334–340.
6. de Jong JC, Rimmelzwaan GF, Fouchier RAM, Osterhaus ADME. Influenza virus: a master of metamorphosis. *Journal of Infection* 2000; **40**:218–228.
7. Webster RG, Bean WJ, Gorman OT, Chambers TM, Kawaoka Y. Evolution and ecology of influenza A viruses. *Microbiological Reviews* 1992; **56**:152–179.
8. Finkenstädt BF, Grenfell BT. Time series modelling of childhood diseases: a dynamical systems approach. *Journal of the Royal Statistical Society* 2000; **C**(49):187–205.
9. Bjørnstad ON, Finkenstädt BF, Grenfell BT. Dynamics of measles epidemics: estimating scaling of transmission rates using a time series SIR model. *Ecological Monographs* 2002; **72**:169–184.
10. Morton A, Finkenstädt BF. Discrete-time modelling of disease incidence time series using MCMC. *Journal of the Royal Statistical Society* 2005, in press. doi:10.111/j.1467-9876.2005.05366.x
11. Longini IM, Koopman JS. Household and community transmission parameters from final distributions of infections in households. *Biometrics* 1982; **38**(1):115–126.
12. Longini IM, Koopman JS, Monto AS, Fox JP. Estimating household and community transmission parameters for influenza. *American Journal of Epidemiology* 1982; **115**(5):736–751.
13. Haber M, Longini IM, Cotsonis GA. Models for the statistical analysis of infectious disease data. *Biometrics* 1988; **44**(1):163–173.
14. Addy CL, Longini IM, Haber M. A generalized stochastic model for the analysis of infectious disease final size data. *Biometrics* 1991; **47**(3):961–974.
15. Cauchemez S, Carrat F, Viboud C, Valleron AJ, Boelle PY. A Bayesian MCMC approach to study transmission of influenza: application to household longitudinal data. *Statistics in Medicine* 2004; **23**(22):3469–3487.
16. Carrat F, Sahler C, Leruez M, Roger S, Freymuth F, Le Gales C, Bungener M, Housset B, Nicolas M, Rouzioux C. Influenza burden of illness: estimates from a national prospective survey of household contacts in France. *Archives of Internal Medicine* 2002; **162**(16):1842–1848.
17. Andreasen V, Levin SA, Lin J. A model of influenza A drift evolution. *Zeitschrift fur Angewandte Mathematik und Mechanik* 1996; **76**(S2):421–424.
18. Andreasen V, Lin J, Levin SA. The dynamics of cocirculating influenza strains conferring partial cross-immunity. *Journal of Mathematical Biology* 1997; **35**:825–842.
19. Haraguchi Y, Sasaki A. Evolutionary pattern of intra-host pathogen antigenic drift: effect of cross-reactivity in immune response. *Philosophical Transactions of the Royal Society* 1997; **B**(352):11–20.
20. Gomes MGM, Medley GF. Dynamics of multiple strains of infectious agents coupled by cross-immunity: a comparison of models. In *Mathematical Approaches for Emerging and Reemerging Infections: Models, Methods and Theory*, Blower S, Castillo-Chavez C, Cooke KL, Kirschner D, Van der Driessche P (eds). Springer: New York, 2001:171–191.
21. Andreasen V. Dynamics of annual influenza A epidemics with immuno-selection. *Journal of Mathematical Biology* 2003; **46**:504–536.
22. Ferguson NM, Galvani AP, Bush RM. Ecological and immunological determinants of influenza evolution. *Nature* 2003; **422**:428–433.
23. Boni MF, Gog JR, Andreasen V, Christiansen FB. Influenza drift and epidemic size: the race between generating and escaping immunity. *Theoretical Population Biology* 2004; **65**:179–191.
24. Gupta S, Ferguson N, Anderson RM. Chaos, persistence, and evolution of strain structure in antigenically diverse infectious agents. *Science* 1998; **280**:912–915.
25. Lin J, Andreasen V, Levin SA. Dynamics of influenza A drift: the linear three-strain model. *Mathematical Biosciences* 1999; **162**:33–51.
26. Gog J. Multiple strains and influenza. *Ph.D. Thesis*, 2003, University of Cambridge, unpublished.
27. Flahault A, Letrait S, Blin P, Hazout S, Mnars J, Valleron AJ. Modelling the 1985 influenza epidemic in France. *Statistics in Medicine* 1988; **7**:1147–1155.
28. Carrat F, Tachet A, Rouzioux C, Housset B, Valleron AJ. Evaluation of clinical case definitions of influenza: detailed investigation of patients during the 1995–1996 epidemic in France. *Clinical Infectious Diseases* 1999; **28**:283–290.

29. Personal communication with Fabrice Carrat, INSERM Paris.
30. Ripley BD. *Stochastic Simulation*. Wiley: New York 1987.
31. Brooks SP, Morgan BTJ. Optimisation using simulated annealing. *Journal of the Royal Statistical Society* 1995; **D**(44):241–257.
32. Gilks WR, Richardson S, Spiegelhalter DJ. Introducing Markov chain Monte Carlo. In *Markov chain Monte Carlo in Practice*, Gilks WR, Richardson S, Spiegelhalter DJ (eds). Chapman & Hall: London, 1996: 1–20.
33. Gilks WR. Full conditional distributions. In *Markov chain Monte Carlo in Practice*, Gilks WR, Richardson S, Spiegelhalter DJ (eds). Chapman & Hall: London, 1996:75–88.
34. Pease CM. An evolutionary epidemiological mechanism, with applications to type A influenza. *Theoretical Population Biology* 1987; **31**:422–452.
35. Potter CW, Jennings R, Nicholson K, Tyrell DAJ, Dickinson KG. Immunity to attenuated influenza virus WRL 105 infection induced by heterologous, inactivated influenza A virus vaccines. *Journal of Hygiene* 1977; **79**: 321–332.