


Research Report
Statistical Research Unit
Department of Economics
Göteborg University
Sweden

Modeling influenza incidence
for the purpose of on-line monitoring

Eva Andersson, David Bock & Marianne Frisén

Research Report 2007:5
ISSN 0349-8034

Mailing address:	Fax	Phone	Home Page:
Statistical Research Unit	Nat: 031-786 12 74	Nat: 031-786 00 00	http://www.statistics.gu.se/
P.O. Box 640	Int: +46 31 786 12 74	Int: +46 31 786 00 00	
SE 405 30 Göteborg			
Sweden			

Modeling influenza incidence for the purpose of on-line monitoring

Eva Andersson[1, 2] David Bock[1] and Marianne Frisén[1]
1: Statistical Research Unit, Department of Economics, Göteborg University, Göteborg, Sweden
2: Occupational and environmental medicine, The Sahlgrenska University hospital, Göteborg, Sweden

We describe and discuss statistical models of Swedish influenza data, with special focus on aspects which are important in on-line monitoring. Earlier suggested statistical models are reviewed and the possibility of using them to describe the variation in influenza-like illness (ILI) and laboratory diagnoses (LDI) is discussed. Exponential functions were found to work better than earlier suggested models for describing the influenza incidence. However, the parameters of the estimated functions varied considerably between years. For monitoring purposes we need models which focus on stable indicators of the change at the outbreak and at the peak.

For outbreak detection we focus on ILI data. Instead of a parametric estimate of the baseline (which could be very uncertain,), we suggest a model utilizing the monotonicity property of a rise in the incidence. For ILI data at the outbreak, Poisson distributions can be used as a first approximation.

To confirm that the peak has occurred and the decline has started, we focus on LDI data. A Gaussian distribution is a reasonable approximation near the peak. In view of the variability of the shape of the peak, we suggest that a detection system use the monotonicity properties of a peak.

Adress for correspondence: Marianne Frisén, Statistical research unit, Goteborg University, PO Box 640, SE 405 30 Goteborg, Sweden.
E-mail: marianne.frisen@statistics.gu.se
Grant sponsor: Swedish Emergency Management Agency; grant number: 0622/2004

1 Introduction

Influenza epidemics impose huge costs on society due to, for example, high levels of work absence and heavy demands on the health-care system, see [1]. Hannoun and Tumova [2] investigated diagnostic procedures and surveillance systems in the European countries and found differences which hamper comparability. Information on the Swedish influenza incidence is found in the publications by the Swedish Institute for Infectious Disease Control (SMI). Spatial issues (see for example [3], [4], [5] and [6]) are of interest in modeling the spread of the influenza. Regional effects within Sweden are analyzed in [7] while aggregated data for Sweden is used here.

Different models are useful for different purposes. Probabilistic models for the transmission of infection are important for the causal understanding of the variation in influenza incidence, for example the effect of vaccination. A review of probabilistic models based on epidemiological theory of measles and influenza is given in [8]. The classical susceptible-infectious-recovered (SIR) paradigm, with its many variants, is important for the causal understanding of what factors influence the incidence of infectious diseases. An attempt to explain the large seasonal variation in influenza incidence is made in [9], where it

is demonstrated that the large oscillations in incidence may be caused by small, otherwise undetectable seasonal changes in the influenza transmission rate. A disease is transmitted within and between communities when infected and susceptible individuals interact. However, the parameters in an advanced causal model are not identifiable by means of the data available here (ILI and LDI). For surveillance purposes, simpler models which capture the most important features are useful. The models used in [10] provide links between theoretical epidemic probabilistic modeling and simple statistical models. In [11] it is stated that since small-scale movements and contacts between people are generally not recorded, available data regarding infectious disease are often aggregations in space and time. Thus, in [11] a spatially descriptive temporally dynamic model is used, where the intensity depends on space and time.

Prediction is an important aim. In [12] we examined some simple prediction rules based on multiple regression and there we found that an algorithm based only on a measure of the time of the start of the epidemic phase gave a good prediction of the height of the peak. An early start is a warning for a high peak. Prediction can also be used as a component in a surveillance system. The authors of [13] use 2-weeks-ahead predictions as a monitoring tool.

Modelling for monitoring is in focus here. For reviews and discussions of prospective statistical surveillance in public health, see [14] and [15] and Section 0. The aim of this article is to demonstrate aspects of modelling, when the purpose is monitoring of influenza. We make an exploratory analysis of influenza-like illness (ILI) and laboratory diagnoses (LDI) in Sweden in order to study the statistical properties of these variables. We try to find reasonable stochastic models and find out which characteristics are stable between years and which are not. Thus we want to find models for the Swedish influenza data that could work in a future surveillance system. Important issues here are to examine the data quality and to examine if it is possible to find a model to describe the process before the change. It is important to be realistic and not base the decision on too many assumptions. This report focuses on those issues of modeling that are of concern in the construction of a surveillance system. Some general aspects on monitoring will be discussed, whereas the construction of the surveillance system itself will be presented in a forthcoming article.

The outline is the following: In Section 2 we describe the Swedish data available for our analysis. Least squares estimations of how the incidence depends on the time of year are presented in Section 3, for both non-parametric and parametric models. The distribution around the curves is analyzed in Section 4, both at the outbreak and at the peak. Some aspects on monitoring systems are given in Section 0. Conclusions are given in Section 6.

2 The Swedish influenza data

Two different types of data are analyzed: weekly ILI and LDI data. Their respective reporting systems are described in e.g [16] and in the annual reports from the National Influenza Reference at SMI.

2.1 Reports on influenza-like illness (ILI)

Weekly data on ILI are collected from a number of sentinel physicians, who report the number of patient visits. For patients showing ILI, the date of the visit and the patient's age and sex are recorded, see [17]. The data are collected during the weeks when an influenza epidemic can be expected, according to SMI (week 40 up to week 20). This will also be those

weeks during which the surveillance is applied. Around 2% of the Swedish general practitioners participate. Their involvement is voluntary and their representativeness can therefore be questioned. In [18] the representativeness of the physicians participating in a French sentinel system for ILI is discussed.

The percentages of patients showing ILI (%ILI) were available for the years 1999–2005. The time from autumn one year to spring next year (e.g. autumn 1999 to spring 2000) is hereafter denoted as a period (e.g. period 99_00). For the last five of these periods (00_01 to 04_05), we also had access to the total number of patients (#PAT) as well as the number of patients showing ILI (#ILI). As is seen in Figure 1, the total number of patients varies considerably between weeks. The number of influenza patients contributes only marginally to this variation. The variances of the estimates of %ILI will also vary. The varying number of patients might reflect physicians' inclination to send reports. After the peak of the influenza, some of the physicians might refrain from doing so. This could explain the decrease in the total number of patients, since aggregated data are used. The ILI data should consequently be interpreted with care since the data might be influenced by time-dependent effects, such as an interest in an expected outbreak or a lack of interest after the peak has been reached. In [10] a time-dependent underreporting (of measles) is modelled, but the possibility of identifying this when also the transmission intensity is time-dependent can be questioned.

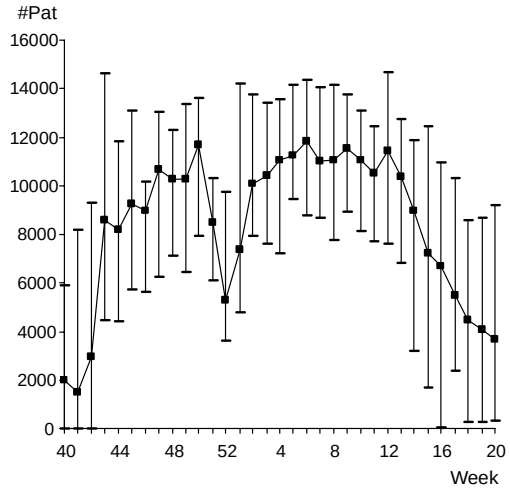

Figure 1. The number of patients per week, reported to be seen by the sentinel physicians for the six influenza periods 00_01 to 05_06. The average and the range are illustrated

2.2 Reports on laboratory diagnosed cases (LDI)

The LDI data consist of weekly reports (during the time from week 40 to week 20) from five virus laboratories (at university hospitals and at SMI) and a number of microbiological laboratories (usually between 15 and 20, see [19] and [17]). The laboratory reporting mainly concerns patients who are severely ill and in need of hospital care. In the laboratories the influenza is typed as either A, B or C, which all belong to the group ortomyxovirus. It is mainly A and B that give rise to the typical influenza infection, see [20]. In this report, LDI consists of the sum of A and B cases. We had access to the number of laboratory diagnosed cases of influenza for seven influenza periods (1998–2005). In the weekly influenza reports

from SMI, the number of laboratory confirmed cases of influenza are given for each week during the monitoring period (in our data from week 40 in the fall to week 20 in the spring).

2.3 Covariation between ILI and LDI

The covariation between ILI and LDI has been studied. Unfortunately, the conclusion was that it is hard to use the ILI data as an early indicator since the relation between the processes is different before and after the peak. Similar differences between correlations of variables before and after the peak could be seen in a figure in [13]. The two series, ILI and LDI, do not have the same peak times or even the same relation between the peak times for ILI and LDI for different years. This is a further indication that ILI is not a simple leading indicator to LDI. Also, there is no clear-cut relation between %ILI(t) and LDI(t+j).

LDI data will be used for the peak and decline. We will only use ILI data for the outbreak period where no other variable is early enough. We will thus model for univariate surveillance, i.e. surveillance of only one process and thus we investigate the properties of each process separately.

3 Least squares estimation of the incidence over time

Monitoring of influenza concerns a change in the incidence. Therefore it is important to investigate the incidence. We consider two types of changes that are of interest, namely the increase of the incidence at the outbreak (start) of the epidemic phase and the decrease of the incidence after the peak of the influenza. The reason for the second surveillance is for example to be able to detect other contagious diseases that might show themselves in an increased number of cases of influenza-like illnesses.

In [13], the expected value of each of four processes, is modeled. For each process, the expected value at time t is a function of the previous observed value of that process and, for some processes, also a function of previous values of some of the other processes.

3.1 Parametric models

Different regression functions for the variation of the incidence with time have been suggested. A cyclical pattern over the year with a peak during the winter is natural.

In an early model by [21], the proportion of deaths due to pneumonia or influenza is modeled using trigonometric regression

$$X(t)=\mu_0+\alpha_0\cdot t+\sum_{i=1}^{q}\alpha_i\cdot\cos\left(\left(2\cdot\pi\cdot t\right)/q\right)+\sum_{i=1}^{q}\beta_i\cdot\sin\left(\left(2\cdot\pi\cdot t\right)/q\right)+\varepsilon(t)\,,$$

where q is the periodicity and $\varepsilon\sim$iid N(0, σ^2). Trigonometric regression models have later been used for different variables and both for epidemic and non-epidemic phases. Usually, a 52-week periodicity is assumed, i.e. q=52 as in, for example, [22] for non-epidemic ILI in France.

In [18], a trigonometric regression function (with periodicity of 52) is used for modelling the weekly French ILI incidence. For the surveillance they use a Hidden Markov Model (HMM), which allows for switching between epidemic and non-epidemic states. The transitions are according to a Markov chain. In many papers, there are attempts separately to estimate the non-epidemic seasonal effect and the epidemic effect. In [18], the model

included both seasonality and switching between epidemic and non-epidemic phases. A general problem with the modeling of seasonality and the duration of non-epidemic phases is that the separation depends heavily on the assumptions made. Since the seasonality and the non-epidemic phase usually are found to be the same, or nearly the same, it is also hard to separate them. In [4] a trigonometric regression is also used for infectious disease data but with an additional autocorrelation effect (a "parameter- and data-driven" model). In [13], the seasonal variation in the expected value during non-epidemic phases is also captured by trigonometric regression. Deviations from a trigonometric regression with a constant cycle length and a constant peak height might be used to detect unusually severe epidemics. However it is not suitable in surveillance for detecting the outbreak or the peak of an ordinary influenza since the characteristics of the influenza curve are not the same from one year to the next. A reference curve (e.g. a trigonometric curve that captures the expected value during non-epidemic phases) would be needed at the start of the season. But at the start, the characteristics of the coming season (peak time, peak height, shape of peak) are unknown. Therefore a reference curve must be modelled using data from previous seasons, which would result in a average curve. Thus a deviation from this curve will only tell if a particual year was very different. In Section 5 we suggest that a surveillance system for the start or peak of the influenza is based on monotonicity properties instead of an average curve.

We fitted a trigonometric regression to the Swedish LDI data from the seven periods 98_99 to 04_05 in hope that the peak would be well represented. However the fit was poor (R^2=0.2) due to a much more pointed peak than by the trigonometric regression and also due to the parametric restriction of constant amplitude and cycle length for all the periods.

In [23], the incidence during the non-epidemic phase was estimated as constant level. A very simple model for the incidence near the peak is to assume that the curve is linear on each side of the peak. Such a model fits the Swedish influenza incidence rather well, see [12]. If a piecewise linear model for the incidence curve holds, then the successive differences have a constant expected value on each side of the peak. This is also the case when X is a random walk with drift, $X(t)=X(t-1)+\beta+\varepsilon(t)$, where ε is iid and the drift β changes from negative to positive. However, the stochastic properties are different for the two models, which will be of importance in the surveillance. If X can be described as $X(t)=\mu(t)+\varepsilon(t)$, where μ is piecewise linear and ε is iid, then the error term of the first difference $X(t)-X(t-1)$ would be an MA(1) process. If, instead, X can be described as a random walk with drift, then it is reasonable to differentiate the data and monitor the observed successive differences, which will be independent. The implications of dependencies are discussed further in Section 0.

An exponential curve is a natural choice for the expected value of the incidence because of the biological process. Also, it does not allow the expected incidence to be negative, as some other models do. We tried different approaches for fitting the piecewise exponential curve

$$\mu(t|j)=\begin{cases}\beta_0\cdot\exp(\beta_1\cdot t), & t\le j\\ \beta_0\cdot\exp(\beta_1\cdot j+\beta_2\cdot(t-j)), & t>j\end{cases},$$

where j is the time of the peak and $\beta_1>0$ and $\beta_2<0$.

One way of fitting the exponential model is to linearize the curve by the logarithmic transform. This transformation implies that the deviations from the curve are multiplicative. In the case of our data, heteroscedasticity was a consequence. Another drawback was that the data contains zeros for which the logarithm cannot be computed and which cannot be deleted since they are important. Instead, we used non-linear least squares estimation.

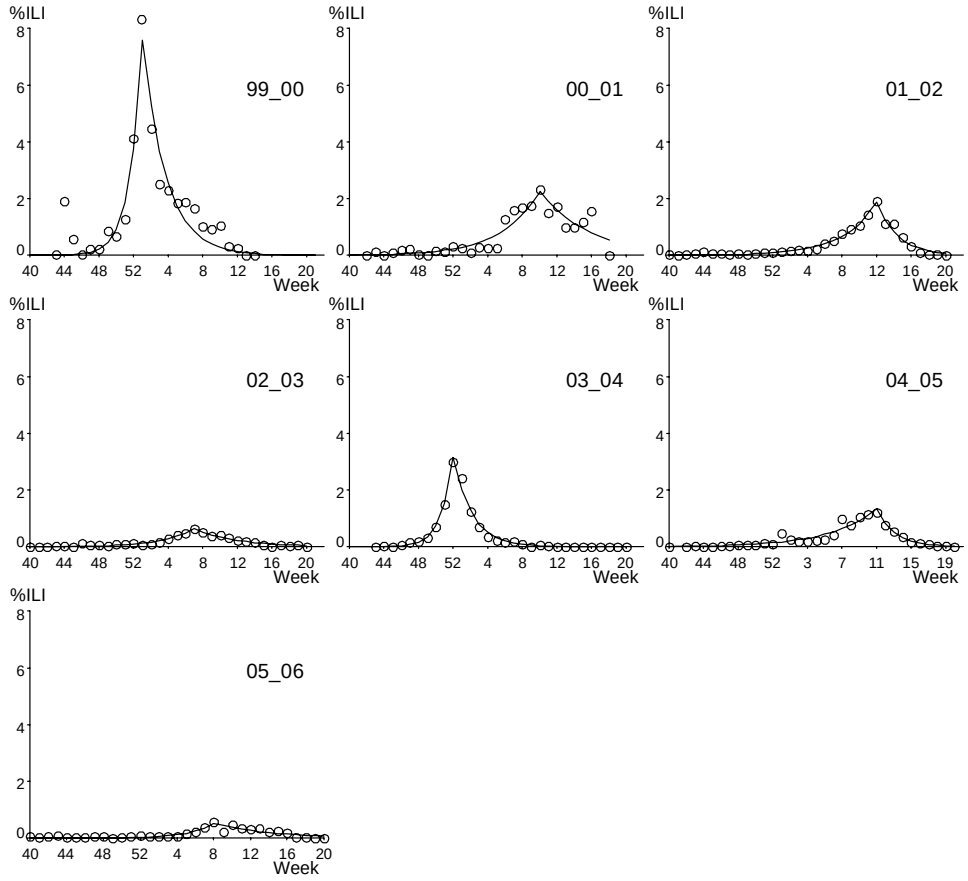

Figure 2. Observed values of %ILI (circle) and the exponential regression curve (solid line).

From Figure 2 and Figure 3 we conclude that the six (seven for LDI) influenza periods are very different. The curves differ much in both height and growth. The curves are un-symmetric for all the influenza periods, but there is no consistency in whether the up-phase or down-phase has the largest slope. The possibility of using a parametric surveillance method (. basing a surveillance system on models for μ with known parameters β_1 and β_2) is hampered by this lack of consistency, also discussed in Section 3.1.

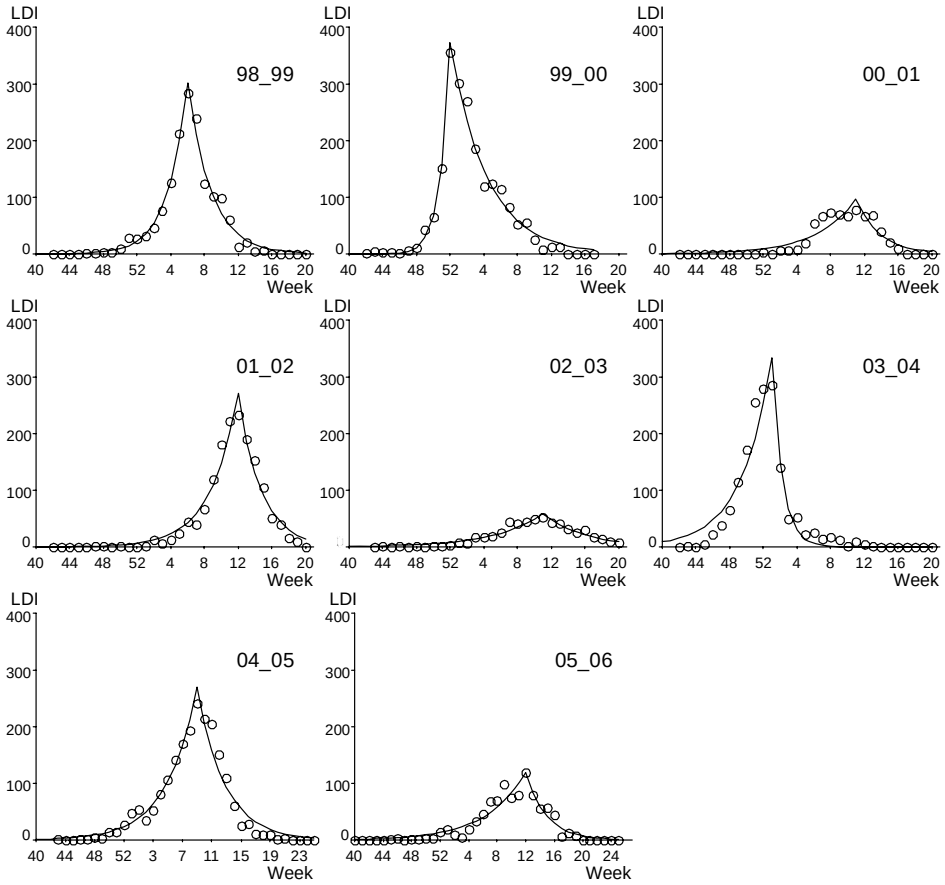

Figure 3. Observed values of LDI (circle) and the exponential regression curve (solid line).

3.2 Nonparametric models

Unimodal regression can be used to estimate a function without making any assumptions about the parametric shape but using only order restrictions. Consider the model

$$X(t) = \mu(t) + \varepsilon(t),$$

where $E[\varepsilon(t)]=0$, X is a process measuring the incidence and t is the time (in weeks). We want an estimate of the expected incidence, $E[X(t)]=\mu(t)$, and, based on data from the latest seven periods, μ has one peak every influenza period. However, the height and the time of the peak varies from one period to the next, making it difficult to use a single parametric model for prediction and surveillance in future periods. Therefore, in the estimation of μ, we use only the information of unimodality:

$$\mu(1)\leq \mu(2)\leq \ldots \leq \mu(j-1) \leq \mu_{max} \text{ and } \mu_{max}\geq \mu(j)\geq \mu(j+1) \geq \ldots \geq \mu(t).$$

μ_{max} is the peak value (not necessarily observable). The unimodal regression is consisting of an up-phase, where $E[X(t)]$ is monotonically increasing with t up to an unknown time, and a down-phase, where the regression is monotonically decreasing with t. The solution technique [24] is based on the "Pool adjacent" procedure and gives a least square estimate where the sum of squares is minimized under the unimodal restriction above. A free computer program

is available from the corresponding author. When ε is iid N(0; σ^2), this least square estimate is also the maximum likelihood estimate.

The values estimated by unimodal regression and the raw data are shown in Figure 4 and 5. The cycle length and the height of the peaks are seen to change considerably from one period to the next.

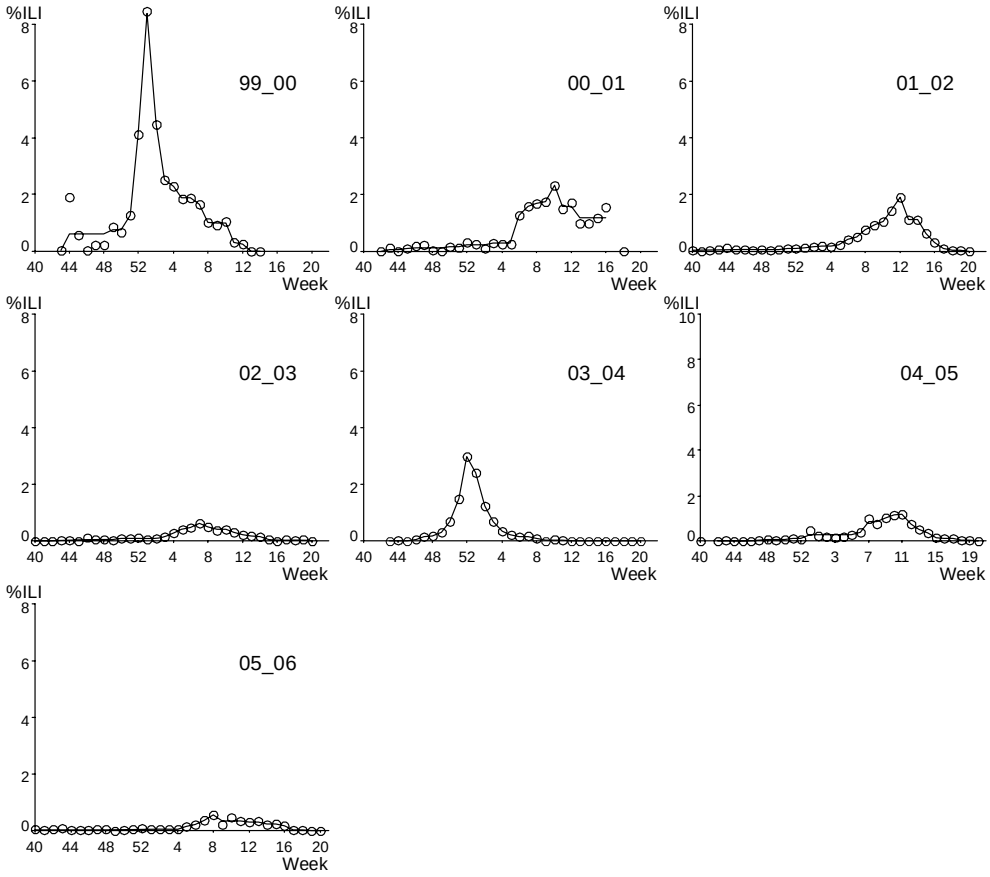

Figure 4. Observed values of %ILI (circle) and the unimodal regression estimates (conected by solid lines).

When the estimator of the incidence is strongly consistent for each time, it follows that the unimodal regression technique also gives strongly consistent estimates of the time of the peak, and the height of the peak, see [24].

Often, when considering non-parametric methods for trend estimation, moving averages and kernel smoothing are considered (see e.g [25]). However, they have the disadvantage of not preserving the peak location (see [24]) and this preservation is necessary here, since one goal is to detect the peak.

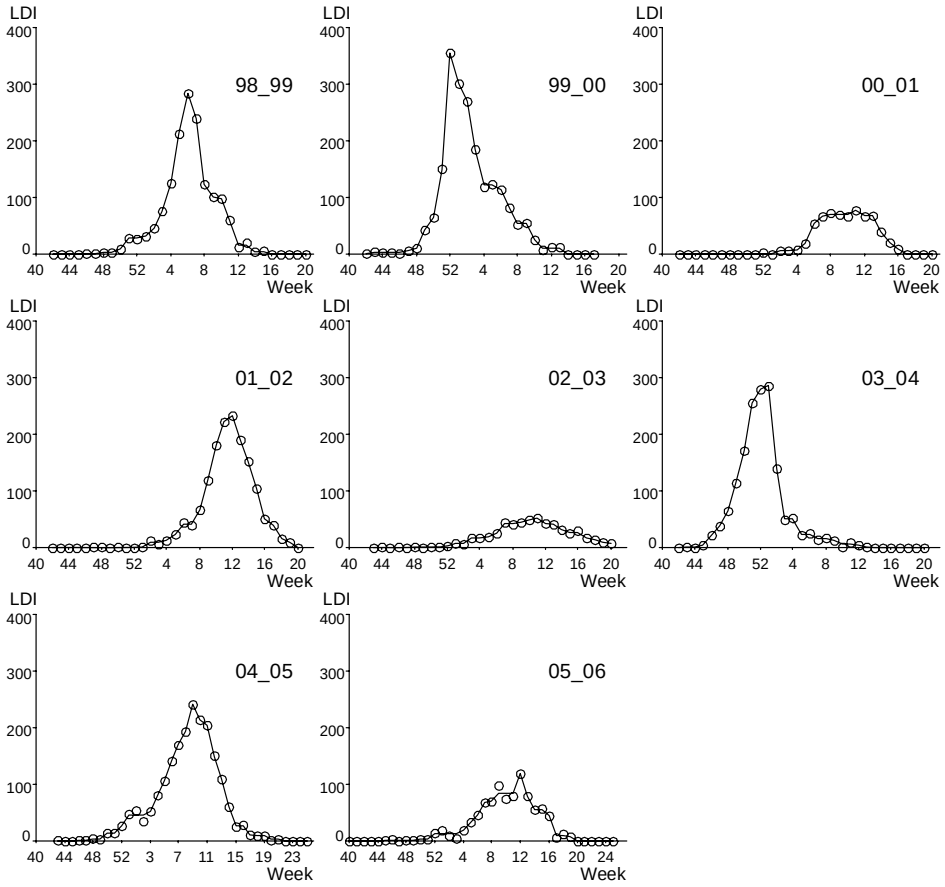

Figure 5. Observed values of LDI (circle) and the unimodal regression estimates (connected by solid lines).

4 Stochastic distributions

In order to construct a surveillance system to detect changes in influenza incidence, we need models that describe the stochastic properties of the process at the start of as well as during the epidemics. In public health monitoring, outbreak detection is probably the most interesting aspect. For the outbreak detection we have chosen to analyze the ILI variable and not the LDI.

In many situations it is also important to detect the peak. One example is when surveillance is carried out in order to detect new infectious diseases. If we have information that the "ordinary" influenza has reached its peak and we then detect an increase in the ILI incidence, this indicates an unusual situation which requires extra caution, since it could be an indication of another infectious disease. For this reason and for planning purposes we also need models for the peak. For peak detection it is reasonable to use the variable LDI, since it contains data of high quality including the confirmed cases of those who visited a physician. For peak detection, knowledge of the stochastic properties at the peak is needed.

When the curve is not known, it is difficult to make conclusions about the distributions of the deviations from the expected value at each time. If a very flexible nonparametric curve is used, then the deviation will be underestimated. If a mis-specified parametric curve is used, then the deviations will be overestimated. We use the residuals from the estimated

exponential models in Figures 2 and 3. The problem of determining the distribution for LDI at the peak appeared to be the easiest one to solve and will be described first.

4.1 The distribution at the peak

In most papers, an iid Gaussian distribution is used to describe the variation around the curve representing influenza incidence. In [21], a Gaussian distribution is used it for death rates, in [18] it is used for ILI and in [13] it is used for hospital cases.

In [26], [27] and [28], different Gaussian models allowing for autocorrelations were used. In our data for LDI near the peak (where LDI≥30), the autocorrelation was found to be small and within a distance of two standard errors from zero at nearly all time lags. Linearizing the exponential curve by the logarithm transformation resulted in residuals near the peak that were larger than the rest and a variance that was clearly not constant (not shown here). This is an indication that the stochastic term is not multiplicative but additive. No evidence for serious heteroscedasticity in the additive model was found. For details, see [12].

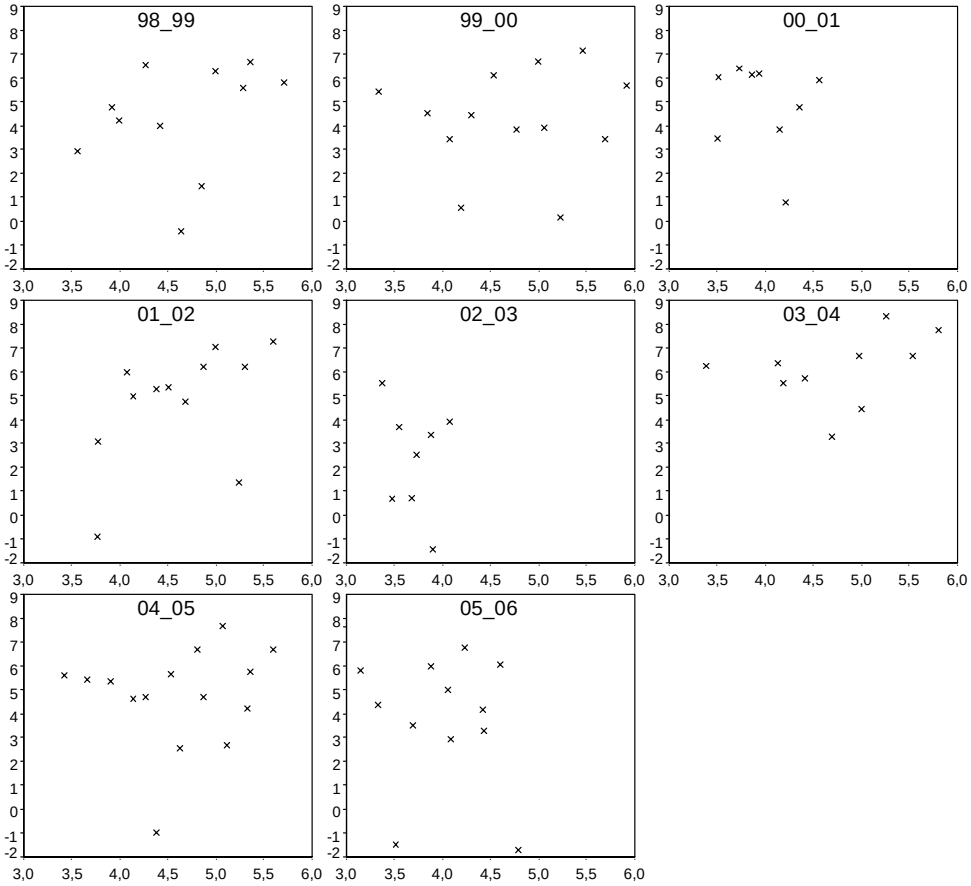

Figure 6. Residuals from the exponential, non-symmetric curves, fitted to LDI data. The graphs show $\ln(\hat{\varepsilon}^2(t))$ (vertical axis) as a function of $\ln(\hat{\mu}(t))$ (horizontal axis).

The residual variances differ between the periods. The periods 00_01 and 02_03 differ from the others in that both the average level of the incidence and the residual variance are lower. We analyze the variance near the peak in Figure 6, where the squared residuals $\hat{\varepsilon}^2(t)$

are plotted against the value of the estimated exponential curve $\hat{\mu}(t)$. If the disturbances of the LDI process are consistent with an iid Gaussian assumption, we would expect $\hat{\varepsilon}^2(t)$ to be constant. A natural alternative for count data is a discrete distribution such as a Poisson or negative binomial distribution. If LDI follows a Poisson distribution, we would expect $\hat{\varepsilon}^2(t)$ to be equal to $\hat{\mu}(t)$. A negative binomial distribution has one more parameter and allows a model with over dispersion. The relation is illustrated in Figure 6.

The conclusion from the analysis of the residuals is that for high values of $\hat{\mu}$ (near the peak) $\hat{\varepsilon}^2(t)$ is fairly constant and there is no evidence for a specific alternative. A reasonable approximation for the stochastic term is that it is additive with a constant variance. It is of course not possible to give evidence for this being the correct distribution but it has earlier been assumed and this investigation does not contradict it.

4.2 Distributions at the outbreak

Also for ILI at the outbreak an iid Gaussian distribution is used in most papers. Gaussian distributions were, however, criticized in [18] and [29] because they assign a non-zero probability for negative incidences. The suggestion in [29] is a Gaussian distribution for the epidemic state but an exponential distribution for the non-epidemic state. In [4] the Poisson distribution is discussed for infectious disease surveillance data. The negative binomial distribution is suggested here and also in [30] when there is an over-dispersion compared to the Poisson distribution. In [13], a log-normal distribution is used for ILI data. We tried linearizing the exponential curve by the logarithm transformation. However, this resulted in residuals near the peak that were larger than the rest and the variance that was clearly not constant for epidemic data. In [11] a Bayesian hierarchical model with Poisson distribution is used for the incidence counts, while normal distributions are used for several of the parameters. In [13], a Poisson log-linear model was used for death rates.

In our data of ILI at the outbreak, the autocorrelations were found to be small (within a distance of two standard errors from zero). Neither a Gaussian distribution nor a constant variance is realistic as regards ILI at the outbreak. First, the variance is clearly dependent on the incidence, which is quite different before and after the outbreak. Second, the total number of patients (#PAT) varies considerably, and this will influence the variance of %ILI as well as the variance of #ILI conditionally on #PAT. The binomial or Poisson distributions are simple and natural choices for the distribution of #ILI. Since the incidence is low, these distributions will be similar. We have chosen to examine the fit to the Poisson distribution. We have to consider #PAT when we examine the Poisson assumption for #ILI. If #ILI follows a Poisson distribution, then, conditional on #PAT(t), we would have

$$E[\#ILI(t)] = Var[\#ILI(t)].$$

We have no good direct estimate of Var[#ILI(t)]. When we study the residuals from the expected value for each week we have to use the curve for %ILI, since we do not have a natural curve for #ILI. Var[%ILI(t)] can be estimated by $\hat{\varepsilon}^2(t)$, where $\hat{\varepsilon}(t)=\%ILI(t)-\hat{\mu}(t)$ is the residual from the estimated exponential curve in Section 3.1. We express #ILI as a function of %ILI, so that

$$E[\#ILI(t)] = (\#PAT(t)/100) \cdot E[\%ILI(t)] \text{ and}$$
$$Var[\#ILI] = (\#PAT/100)^2 \cdot Var[\%ILI], \text{ which implies}$$
$$E[\%ILI] = (\#PAT/100) \cdot Var[\%ILI]$$

The Poisson assumption can thus be examined by relating

$$g(\hat{\varepsilon}(t)) = (\#PAT/100)\cdot \hat{\varepsilon}^2(t) / \hat{\mu}(t)$$

to $\hat{\mu}(t)$. The comparison is shown in Figure 7 for values of $\hat{\mu}(t)$ below or equal to 0.5% in the start of the outbreak. The figure does not clearly demonstrate that $g(\hat{\varepsilon}(t))$ is independent of $\hat{\mu}(t)$ as it should be for a Poisson distribution but gives no strong evidence for an alternative natural hypothesis. The stochastic variation is large. Since no direct evidence against the Poisson distribution is found (e.g. obvious overdispersion), the conclusion is that the Poisson distribution can be used as a first approximation. Even though no hard evidence can be given that this is the true distribution, neither other investigations in the literature nor our data point towards a serious disagreement. Since we have very little data, a model with more parameters might be overfitted. With more data, it might be useful to estimate a model with more flexibility (e.g. a negative binomial distribution).

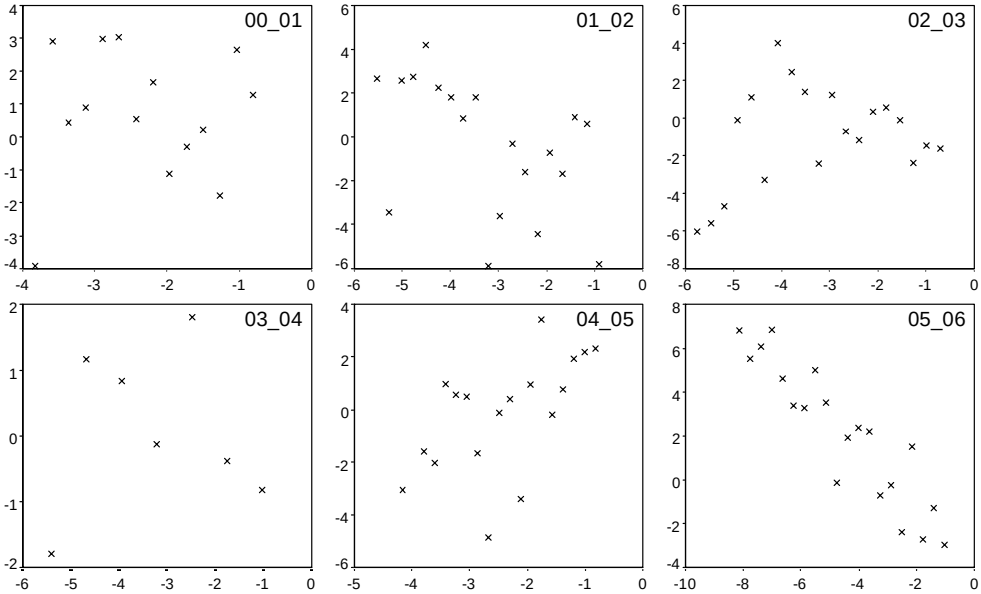

Figure 7. Residuals from the exponential curves, fitted to data on %ILI. The graphs show $\ln(g(\hat{\varepsilon}(t)))$ (vertical axis) as a function of $\ln(\hat{\mu}(t))$ (horizontal axis).

5 Surveillance systems

The aim in a surveillance system is to detect a change in the underlying process, for example the change from a non-epidemic to an epidemic period. The starting time of the influenza differs from one year to the next, as was demonstrated in Figures 4 and 5. In Sweden, data are collected weekly during the period from week 40 to week 20. Thus a surveillance system would have its first observation at week 40, and after that once a week until week 20 in the spring. The start of the influenza period (the outbreak), however, is of course not identical to week 40, but varies.

The most commonly used surveillance systems are based on the Shewhart method (see [31] and [32]). In its original form, we compare the current observation (of e.g. ILI) to a reference model, which shows the expected value that week, given that the epidemic phase

has not started yet. The difference between the observed and the expected value is calculated and compared to an alarm limit, which is usually three standard deviations. If the observed difference for the current week is below the alarm limit, we conclude that the epidemic phase has not started yet. The next week we get a new observation, calculate the difference to the expected value of that week and compare the difference to the alarm limit. In a more advanced setting we use a derived statistic. In [13], the 2-weeks-ahead predictions are used and here an alarm (for an influenza incidence that is higher than expected) is given when the predicted value is large.

In the Shewhart method, only the latest (derived) observation is used as a basis for inferring whether the epidemic phase has started or not. The Shewhart approach is optimal for detecting a large change immediately. In contrast, other methods use all the observations and are better to detect smaller changes where an accumulation of the information improves the performance, see [33]. Some commonly used methods which use all the observations since the start of the surveillance are, CUSUM ([34]), EWMA ([35]) and the full likelihood ratio method ([36], [37]). Since all the methods are based on the likelihood ratio, we need information about the distribution of the process, with and without the change. In order for a surveillance system to be efficient, it is important to find the stable characteristics of the change. If this model is seriously misspecified, the surveillance system will be of little use (see e.g. [38]). In order to choose the correct alarm limit, the distributional properties such as the variance of the process and whether the observations can be considered to be independent over time are especially important. Thus, for constructing the surveillance system, the investigation of the distributional properties of the process is important.

An important question is whether the surveillance system would be improved by using several processes in the monitoring. Potentially this is so and in [13], the relation between four series is successfully utilized.

Below we briefly exemplify surveillance using the Shewhart method using a parametric model. The situation when we want to detect a peak is used in the example. We want to detect (confirm) the peak, i.e. we want to detect a deviation from the model where the incidence is increasing. We monitor X(t), which is modeled as

$$X(t) = \mu(t) + \varepsilon(t), \text{ where } \varepsilon \text{ is iid } N(0;\sigma).$$

If the peaks were relatively alike, from one year to the next, we could use parametric functions for μ. Then, at decision time s, we would expect an incidence of

$$\mu(t) = \exp(\beta 0 + \beta 1 * t).$$

The alarm statistic, according to the Shewhart method, is the difference between the observed value and the expected value, given that the incidence is still increasing. Thus, the alarm statistic at time s would be

$$x(s) - \exp(\beta_0 + \beta_1 t).$$

The alarm limit at time s is often constructed as

$$g(s) = k * \sigma,$$

where k is a constant (often the value 3). Depending on the desired false alarm probability, the constant k can be set to different values.

Another possibility, than the Shewhart approach, could be to use the full likelihood method using the theory in [36], where we could have the following model for the expected incidence

$\mu(s) = \exp(\beta 0 + \beta 1 s)$, $s<\tau$
$\mu(s) = \exp(\beta 0 + \beta 1(\tau-1) - \beta 2(s-\tau+1))$, $s\geq\tau$

and the alarm statistic would be the ratio between the two likelihoods for X (conditional on $s<\tau$ and conditional on $s\geq\tau$, respectively).

Alternatively, a maximum likelihood approach could be used. Since the peaks are very different from one year to the next, a non-parametric approach might be useful. We suggest a nonparametric maximum likelihood approach based on the comparison between the likelihood conditional on $s<\tau$ and conditional on $s\geq\tau$, respectively. We suggest that the expected incidence is estimated under monotonicity restrictions (see e.g. [24] and [39]). For peak detection [40] the relevant conditions to be compared are

μ: $\mu(1)\leq \mu(2)\leq \ldots \leq\mu(s)$, $s<\tau$
μ: $\mu(1)\leq \ldots \leq \mu(\tau-1)$ and $\mu(\tau)\geq \ldots \geq\mu(s)$, $s\geq\tau$.

For outbreak detection the relevant conditions to be compared are

μ: $\mu(1)= \mu(2)= \ldots =\mu(s)$, $s<\tau$
μ: $\mu(1)= \ldots=\mu(\tau-1)$ and $\mu(\tau)\leq\ldots\leq\mu(s)$, $s\geq\tau$.

Outbreak detection using these monotonicity restrictions is a topic for further research.

6 Conclusions

To construct a system for on-line detection of changes in the influenza incidence, by monitoring ILI and LDI, we need first find reasonable models for these processes. Generally, the modeling of processes which are subject to many disturbing factors implies many challenges. For monitoring purposes, however, it is especially important to decide which assumptions are valid approximations and which would result in bad conclusions. Robust simple models may be better than models with many estimated parameters which are not stable over the years.

In this paper we have analyzed the statistical properties of two variables related to the evolution of influenza in Sweden, namely influenza-like illness (ILI) and laboratory diagnoses (LDI). This analysis forms a basis for constructing a system for on-line surveillance.

For outbreak detection we have to use ILI since no observations on LDI are available until the influenza is manifest. As regards non-epidemic phases, very sparse data are available in the present data set. No data are available prior to week 40. This means that there is a lack of good baseline data. A very simple model with a constant incidence during non-epidemic phases could be satisfactory as a first approximation. From the present data, however, such an incidence can be estimated only with large uncertainty. The lack of good baseline data is a serious problem for the detection of a change from a baseline. A surveillance system which relies on an estimate of the baseline level is very vulnerable to errors in the estimate. Instead, we suggest that the fact that there is a level rise at the start of an epidemic be utilized. Thus, a

detection system which utilizes the monotonicity property at a level rise, rather than an estimate of the non-epidemic level, is suggested.

For the detection of the peak and decline of the influenza, we consider LDI. Several suggestions of the shape of the curve at the peak of the influenza incidence have been discussed and compared with the present data set. Both parametric and non-parametric models can describe the data near the peak. Trigonometric curves were shown to fit the Swedish data poorly. Several other suggestions of the shape of the curve at the peak of the influenza incidence agreed rather well with the present data set. We found an exponential model with additive stochastic term to be preferred. A problem in the construction of a surveillance system based on estimates of a parametric curve is that the parameters vary a lot from one influenza period to the next. Since the data are accruing in time conclusions for the present time cannot be based on earlier information unless this information it stable over the periods. Thus, a non-parametric model for the shape is an important alternative. Surveillance methods based only on monotonicity properties (first increase and then decline) are more robust.

The available data are sparse and do not give any evidence of complicated stochastic models. A general problem with the modeling of both heteroscedasticity and autocorrelation is that the estimated parameters depend heavily on the assumptions made about the time-dependent expected value of the process. Based on the data available to us, we conclude that a Poisson distribution can be used as a first approximation for the outbreak phase and that Gaussian distributions with a constant variance can be used near the peak.

Information about the data for different geographical areas is valuable since it is possible that the influenza reaches different areas at different times. In this article, however, the data are collected on a national level and, consequently, only the temporal aspects are studied.

Acknowledgement

The data were made available to us by the Swedish Institute for Infectious Disease Control, and we are grateful for discussions about the aims and the data quality.

References

[1] Szucs T. The socio-economic burden of influenza. Journal of Antimicrobial Chemotherapy. 1999;4:11-5.
[2] Hannoun C, Tumova B. Survey on influenza laboratory diagnostic and surveillance methods in Europe. European Journal of Epidemiology. 2000;16(3):217-22.
[3] Nakaya T, Fotheringham AS, Brunsdon C, Charlton M. Geographically weighted Poisson regression for disease association mapping. Statistics in Medicine. 2005;24(17):2695-717.
[4] Held L, Höhle M, Hofmann M. A statistical framework for the analysis of multivariate infectious disease surveillance data. Statistical Modelling. 2005;5(3):187-99.
[5] Diggle P. Spatio-temporal point processes, partial likelihood, foot and mouth disease. Statist Meth Med Res. 2006;15(4):325-36.
[6] Clark AB, Lawson AB. Surveillance of individual level disease maps. Statist Meth Med Res. 2006;15(4):353-62.
[7] Bock D, Pettersson K. Exploratory analysis of spatial aspects on the Swedish influenza data. Stockholm: Report from the Swedish Institute for Infectious Disease Control; 2006. Report No.: 3:2006.
[8] Cliff AD, Haggett P. Statistical modelling of measles and influenza outbreaks. Statistical Methods in Medical Research. 1993;2(1):43-73.
[9] Dushoff J, Plotkin JB, Levin SA, Earn DJD. Dynamical resonance can account for seasonality of influenza epidemics. Proceedings of the National Academy of Sciences of the United States of America. 2004 Nov 30;101(48):16915-6.
[10] Morton A, Finkenstädt BF. Discrete time modelling of disease incidence time series by using Markov chain Monte Carlo methods. J of the Royal Statistical Society C. 2005;54(3):575-94.
[11] Mugglin AS, Cressie N, Gemmell I. Hierarchical statistical modelling of influenza epidemic dynamics in space and time. Statistics in Medicine. 2002;21(18):2703-21.
[12] Andersson E, Bock D, Frisén M. Exploratory analysis of Swedish influenza data: Report from the Swedish Institute for Infectious Disease Control; 2006. Report No.: 1:2006.
[13] Sebastiani P, Mandl KD, Szolovits P, Kohane IS, Ramoni MF. A Bayesian dynamic model for influenza surveillance. Statistics in Medicine. 2006;25(11):1803-16.
[14] Sonesson C, Bock D. A review and discussion of prospective statistical surveillance in public health. Journal of the Royal Statistical Society A. 2003;166(1):5-21.
[15] Lawson AB, Kleinman K, eds. Spatial and Syndromic Surveillance for Public Health: Wiley 2005.
[16] Ganestam F, Lundborg CS, Grabowska K, Cars O, Linde A. Weekly antibiotic prescribing and influenza activity in Sweden: a study throughout five influenza seasons. Scandinavian Journal of Infectious Diseases. 2003;35(11-12):836-42.
[17] Linde A, Brytting M, Johansson M, Wiman Å, Högberg L, Ekdahl K. Annual Report September 2003 - August 2004: The National Influenza Reference Center. Stockholm: Swedish Institute for Infectious Disease Control; 2004.
[18] Le Strat Y, Carrat F. Monitoring epidemiologic surveillance data using hidden Markov models. Statistics in Medicine. 1999;18(24):3463-78.
[19] Linde A, Brytting M, Petersson P, Mittelholzer C, Penttinen P, Ekdahl K. Annual Report September 2001 - August 2002: The National Influenza Reference Center: Swedish Institute for Infectious Disease Control; 2002.

[20] Andersson Y, Normann B, Tideström L. Fakta om smittsamma sjukdomar. Stockholm: Swedish institute of infectious disease control; 1999.
[21] Serfling R. Methods for current statistical analysis of excess pneumonia-influenza deaths. Public Health Reports. 1963(78):494-506.
[22] Costagliola D, Flahault A, Galinec D, Garnerin P, Menares J, Valleron AJ. A Routine Tool for Detection and Assessment of Epidemics of Influenza-like Syndromes in France. American Journal of Public Health. 1991;81(1):97-100.
[23] Quenel P, Dab W, Hannoun C, Cohen JM. Sensitivity, Specificity and Predictive Values of Health- Service Based Indicators For the Surveillance of Influenza-a Epidemics. International Journal of Epidemiology. 1994;23(4):849-55.
[24] Frisén M. Unimodal regression. The Statistician. 1986;35(4):479-85.
[25] Andersson M, Ekdahl K, Mölstad S, Persson K, Hansson HB, Giesecke J. Modelling the spread of penicillin-resistant <I>Streptococcus pneumoniae</I> in day-care and evaluation of intervention. Statistics in Medicine. 2005;24(23):3593-607.
[26] Stroup DF, Thacker SB, Herndon JL. Application of Multiple Time-Series Analysis to the Estimation of Pneumonia and Influenza Mortality By Age 1962-1983. Statistics in Medicine. 1988;7(10):1045-59.
[27] Quenel P, Dab W. Influenza A and B epidemic criteria based on time-series analysis of health services surveillance data. European Journal of Epidemiology. 1998;14(3):275-85.
[28] Hussain S, Harrison R, Ayres J, Walter S, Hawker J, Wilson R, et al. Estimation and forecasting hospital admissions due to Influenza: Planning for winter pressure. The case of the West Midlands, UK. Journal of Applied Statistics. 2005;32(3):191-206.
[29] Rath TM, Carreras M, Sebastiani P. Automated Detection of Influenza Epidemics with Hidden Markov Models. *Advances in Intelligent Data Analysis V*. Berlin, Germany: Springer-Verlag 2003:521-31.
[30] Grabowska K. Occurrence of invasive pneumococcal disease and number of excess cases due to influenza. Examensarbete: Mathematical Statistics, Stockholm University; 2005.
[31] Shewhart WA. Economic Control of Quality of Manufactured Product. London: MacMillan and Co. 1931.
[32] Frisén M. Properties and Use of the Shewhart Method and Followers. Sequential Analysis. 2006:to appear.
[33] Frisén M. Statistical surveillance. Optimality and methods. International Statistical Review. 2003;71(2):403-34.
[34] Page ES. Continuous inspection schemes. Biometrika. 1954;41:100-14.
[35] Roberts SW. Control Chart Tests Based on Geometric Moving Averages. Technometrics. 1959;1:239-50.
[36] Shiryaev AN. On optimum methods in quickest detection problems. Theory of Probability and its Applications. 1963;8:22-46.
[37] Frisén M, de Maré J. Optimal Surveillance. Biometrika. 1991;78:271-80.
[38] Andersson E, Bock D, Frisén M. Statistical surveillance of cyclical processes. Detection of turning points in business cycles. Journal of Forecasting. 2005;24:465-90.
[39] Robertson T, Wright FT, Dykstra RL. Order Restricted Statistical Inference: John Wiley & Sons Ltd 1988.
[40] Bock D, Andersson E, Frisén M. On statistical surveillance of Swedish influenza incidence. Peak detection.: Smittskyddsinstitutets rapportserie; 2006. Report No.: 2:2006.

Research Report

2007:1	Andersson, E.:	Effect of dependency in systems for multivariate surveillance.
2007:2	Frisén, M.:	Optimal Sequential Surveillance for Finance, Public Health and other areas.
2007:3	Bock, D.:	Consequences of using the probability of a false alarm as the false alarm measure.
2007:4	Frisén, M.:	Principles for Multivariate Surveillance.