Research Report
Statistical Research Unit
Department of Economics
Göteborg University
Sweden

Semiparametric surveillance of outbreaks

Frisén, M. & Andersson, E.

Research Report 2007:11
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			

Semiparametric surveillance of outbreaks

MARIANNE FRISÉN*
Statistical Research Unit, Department of Economics, University of Gothenburg, Box 640, SE 405 30 Göteborg, Sweden
Marianne.Frisen@statistics.gu.se

EVA ANDERSSON
Statistical Research Unit, Department of Economics, University of Gothenburg, Box 640, SE 405 30 Göteborg, Sweden
and
Department of Occupational and Environmental Medicine, Sahlgrenska University Hospital, Box 414, SE 405 30 Göteborg, Sweden

SUMMARY

The detection of a change from a constant level to a monotonically increasing (or decreasing) regression is of special interest for the detection of outbreaks of, for example, epidemics. A maximum likelihood ratio statistic for the sequential surveillance of an "outbreak" situation is derived. The method is semiparametric in the sense that the regression model is nonparametric while the distribution belongs to the regular exponential family. The method is evaluated with respect to timeliness and predicted value in a simulation study that imitates the influenza outbreaks in Sweden. To illustrate its performance, the method is applied to Swedish influenza data for six years. The advantage of this semiparametric surveillance method, which does not rely on an estimated baseline, is illustrated by a Monte Carlo study. The proposed method is successively accumulating the information. Such accumulation is not made by the commonly used approach where the current observation is compared to a baseline. The advantage of information accumulation is illustrated.

Keywords: Monitoring, Change-points, Generalised likelihood, Ordered regression, Robust regression, Exponential family.

* To whom correspondence should be addressed.

1. INTRODUCTION

An outbreak of an epidemic disease should be detected as soon as possible after the onset. On-line monitoring of incidences can help detect the yearly outbreaks of influenza as well as new diseases, such as SARS and avian flu, and effects of bioterrorism. We will illustrate the methodology of outbreak surveillance by using influenza data from Sweden. An early detection of the onset of the outbreak is useful in order for health authorities to act timely.

In order to develop a system for quick and safe detection, the methodology of statistical surveillance can be used. A process is observed, and the cumulated information is continually evaluated in order to detect a change in an underlying process. For a review and discussion of prospective statistical surveillance in public health, see Sonesson and Bock, (2003). In industrial surveillance, control charts have been in use since 1930. However, the situation in public health surveillance requires other evaluations of how the aims are met. Woodall *and others*, (2008) have stressed the gains of cross-fertilisation. Surveillance systems for detecting outbreaks should have known properties concerning detection ability, risks of false alarms and predictive value, as described in Section 4.

Various approaches have been suggested for public health surveillance. Sometimes the spatial pattern is important and the surveillance is focused on detecting a spatial clustering of adverse health events, as discussed for example by Besag and Diggle, (1977), Kulldorff, (1997), Diggle *and others*, (2004), Lawson and Rodeiro, (2004), Sonesson, (2007) and Tibshirani and Wang, (2008). However, in some situations, such as the case of influenza in Sweden, the outbreak pattern is not characterised by simple clustering, Bock and Pettersson, (2006). Even though spatial patterns are very important, we will not deal with this issue in this paper. Instead, we concentrate on the detection of an increased incidence. A combination of spatial issues and an increased incidence was treated for example by Diggle *and others*, (1999). It might be useful to combine our surveillance statistic with a spatial approach. However, this is not done here.

Most methods suggested for detecting an increased incidence are based on some parametric model for the process. The most commonly used method to detect an increased incidence is to compare each observed incidence value with a baseline. A signal is given as soon as one observation exceeds a threshold, usually a 95% prediction interval (see for example Stroup *and others*, (1988)). The effect of misspecifying the baseline due to estimation errors will be examined in Section 6. There are also more advanced parametric models for the incidence during the outbreak. The cyclic regression function by Serfling, (1963) has frequently been used for seasonal diseases like influenza. It was used by Le Strat and Carrat, (1999), who applied a Hidden Markov Model (HMM) to model the switch between two different

states (non-epidemic and epidemic), where the switch occurs at an unknown time. The conditional mean for the process, given the state, was modelled by the Serfling model. The estimated periodicity of the disease was the same as that of the season (one year), so the effects cannot be separated. Sebastiani *and others*, (2006) base the surveillance on the comparison between an advanced model (based on data from previous years) for the "average" curve and the observations from the current season. By such comparisons, an alarm will be given if the disease starts unusually early in the current season, but not if the start is average or late. This will tell us whether the current season is extraordinary in comparison to an average season. However, here the aim is to detect the onset of an outbreak as early as possible.

In Andersson *and others*, (2007) and Bock *and others*, (2007) it is concluded that methods in which the expected non-epidemic value is modelled by a parametric function are not suitable for the surveillance of influenza incidence, since the parameters, describing the size, shape and onset time of the outbreak, vary much from year to year. Thus, we suggest a nonparametric model for the change in incidence at the onset of the outbreak.

A simple but reasonable model for the expected value of the incidence at the outbreak is that the expected value is constant at first and then, after the onset of the outbreak, monotonically increasing for some time. If there are seasonal effects or other disturbing covariates, the residuals from a model including these characteristics may be relevant for the surveillance. Most of the problems in on-line surveillance are the same for several diseases and also for applications other than medical ones, but here we have chosen the case of influenza outbreaks in order to be specific. In Section 2 the model and the semiparametric maximum likelihood estimation by Frisén *and others*, (2007a), which we will utilise for the surveillance method in Section 3.4, are described.

Baron, (2000) suggested a nonparametric method for the detection of a stochastically larger distribution. When working with detecting the onset of an influenza outbreak, Baron, (2002) stated that his nonparametric method would give a too long delay, and therefore he preferred a parametric method. The method we propose here is something in-between, since it is nonparametric but utilises monotonicity.

In many papers monotone gradual changes are discussed. For example, Fried and Imhoff, (2004) stated that the detection of a monotonic trend from a constant baseline is important in medical applications. They suggested a retrospective test for a flexible monotonic trend and applied this to moving windows. Chang and Fricker, (1999) treated the problem of detecting when the expected value exceeds a threshold, given that the trend is monotonic. Chang and Fricker derived a repeated likelihood ratio test for solving this problem, which is different from detecting when a monotonic trend starts..

Some surveillance methods are in fact using repeated hypothesis testing. Earlier surveillance methods are often variants of the Shewhart method, which is described in Section 3.5, in the sense that information is not aggregated. This is not always an efficient method. Serfling, (1963) and Quenel *and others*, (1994) suggested that there

should be an alarm as soon as there are two consecutive observations beyond the limit. A more fruitful approach could be the use of a surveillance method that gives optimal weights to the observations. Optimal aggregation of the sequentially obtained observations is essential. This will be demonstrated in Section 6.

In Section 3, we describe methods of surveillance with an emphasis on likelihood based methods, since the optimality properties of these are well known (Frisén, 2003). For our outbreak detection, we need a system for detecting a change from a constant level to an increasing function. Here we suggest and evaluate a surveillance method based on the maximum likelihood ratio for two states. We assume that the distribution belongs to the exponential family. Details are given for the normal and Poisson distributions. The state before the outbreak is characterised by a constant (but unknown) expected incidence. The state at the onset of the outbreak is characterised by a monotonically increasing expected incidence, but neither the shape nor the values of the increasing function are specified. Since the surveillance method is nonparametric with respect to the regression but parametric with respect to the distribution, it is semiparametric.

When an outbreak occurs quick and safe detection is essential. The onset should be detected with minimal delay, but at the same time, false alarms should be rare in order to ensure that the alarm has a high predictive value. Evaluation metrics are discussed in Section 4. The semiparametric method is applied to Swedish data in Section 5, where also a simulation study describes the properties of the method for this situation. In Section 6 we compare the commonly used Shewhart approach to our method, which accumulates information by combining likelihood expressions. Conclusions are made in Section 7.

2. MODELS AND SPECIFICATIONS

In Andersson *and others*, (2007), Swedish influenza data from six seasons (1999–2006) were analysed, and it was concluded that several of the characteristics of the yearly influenza cycle varied considerably from year to year. The baseline varies and is hard to estimate due to lack of data. This makes it difficult to describe the non-epidemic period by the same parametric function. The peak time varies, as do the outbreak time and the shape of the outbreak. All this makes it difficult to describe the epidemic period by a parametric function. Therefore, we suggest a nonparametric approach based on monotonicity restrictions (the outbreak regression).

We monitor the process X and at time t we observe x(t). The decision time is denoted by s. At each decision time s we use the available observations $x_s = \{x(1), ..., x(s)\}$ to discriminate between two states. The state D before the outbreak is characterised by a constant (but unknown) expected incidence. The state C at the onset of the outbreak is characterised by a monotonically increasing expected incidence. Let τ be the unknown time of the onset of the outbreak. Thus, at $\tau = j$ the

expected value μ changes from a constant (baseline) level to an increasing process. This case corresponds to state Cj, j = 1, 2, ..., s. We have $C = \{C1 \bigcup C2 \bigcup ...Cs\}$ or equivalently C = {τ≤s}. The state with the constant baseline is denoted D, where D={τ>s}. The states can be expressed by the expected value of the incidence, μ(t), as:

$$\begin{aligned} &\text{State D: } \mu(1) = \mu(2) = ... = \mu(s) \\ &\text{State C1: } \mu(1) \le \mu(2) \le ... \le \mu(s) \\ &\text{State C2: } \mu(1) < \mu(2) \le ... \le \mu(s) \\ &\text{State Cj: } \mu(1) = ... = \mu(j\text{-}1) < \mu(j) \le ... \le \mu(s) \text{ for } j > 2, \end{aligned} \qquad (2.1)$$

where j is the first time when we have an increased expected value. For both j = 1 and j = 2 the curve is increasing.

The situation where the regression is constant at first and then monotonically increasing will be called "outbreak regression". In many situations, the "normal", or "in-control", state can be described by a constant regression, and then, at a possibly unknown time, the process changes to an increasing regression. Apart from matters of public health, this can also be of interest when investigating whether data deviate from a specified econometric model by analysing whether the residuals are increasing after the change point. The opposite situation (the regression is first constant and then monotonically decreasing) can be treated in the same way but will not be discussed here.

There are several suggestions for nonparametric estimation. For example Gill and Baron, (2004) suggest a highly general nonparametric estimation method for monotonic functions. Most nonparametric methods are based on some kind of smoothing and least squares. Although these are excellent for graphics, which give good insights, maximum likelihood estimations have advantages for some purposes. For tests on monotonicity properties, there are methods based on kernel estimates (Bowman *and others, 1998*) and on likelihood ratios (Andersson and Frisén, 2002). In likelihood based surveillance, we need maximum likelihood estimators. (Frisén, 1986) gave the maximum likelihood estimator for a unimodal regression (monotonically increasing and then monotonically decreasing, or vice versa). Here we will use the maximum likelihood estimate for the outbreak regression situation. This estimator is presented in Frisén *and others*, (2007a) for the family of regular exponential distributions which includes both the normal and Poisson distributions.

3. LIKELIHOOD BASED SURVEILLANCE

Some characteristics separate a surveillance situation from a hypothesis testing situation. In hypothesis testing, we use the sample data to perform one test to judge whether we can reject the null hypothesis or not. In surveillance, we take repeated decisions to determine whether the process is in state D or if it has changed to state C.

The specifications of states D and C change with the decision time s, since D = {τ > s} and C = {τ ≤ s}. Generally, the aim of a surveillance system is to, at each decision time s, discriminate between two states; "the change has occurred" (state C) and "the change has not occurred" (state D). A surveillance system consists of an alarm statistic and an alarm limit.

3.1. *The full likelihood ratio method*

Shiryaev, (1963) showed that, for discriminating between the two events C={τ ≤ s} and D={τ > s}, the full likelihood ratio between C and D is optimal in the sense that the method gives a minimal expected delay for a fixed false alarm probability. The methods considered here are all based on the likelihood ratio.

The full likelihood ratio method gives an alarm for the first time s at which

$$\frac{f(x_s|C)}{f(x_s|D)} > k_s, \tag{3.1}$$

where f is the likelihood function, x_s = {x(1), x(2), ..., x(s)} and $k_s = k/(1-k)\cdot(P(D(s))/P(C(s))$. For the situation where $P(C) = 1-P(D)$, it was shown by Frisén and de Maré, (1991) that the likelihood ratio is equivalent to the posterior probability for surveillance

$$\{x_s : P(C|x_s) \geq k\} = \left\{x_s : \frac{f(x_s|C)}{f(x_s|D)} \geq \frac{P(D)\cdot k}{P(C)\cdot(1-k)}\right\}.$$

The definition of the event C is important. A very general situation is that we want to discriminate between "the change time is in the future", i.e. D = {τ>s}, and "the change has occurred", i.e. C={τ ≤ s}. The event C is composite, C = {{τ=1}, {τ=2},..., {τ=s}}. The partial likelihood ratio for one of these components is

$$L(s,t) = \frac{f(x_s|\tau = t)}{f(x_s|\tau > s)}.$$

The full likelihood ratio is based on all s partial likelihood ratios

$$w_1 \cdot L(s,1) + w_2 \cdot L(s,2) + ... + w_s \cdot L(s,s),$$

where wj = P(τ = j)/P(τ ≤ s).

The important change in the outbreak situation is a change in the expected value, μ, which depends on τ and is expressed as

$$\mu(t) = \begin{cases} \mu^{D}(t), & t < \tau \\ \mu^{Cj}(t), & t \geq \tau, \tau = j \end{cases}$$

If a parametric approach had been used, then μ might be specified as

$$\begin{aligned} \mu^{D}(t) &= \mu_0 \\ \mu^{Cj}(t) &= \exp(\beta_0 + \beta_1 \cdot (t - j + 1)) \end{aligned} \tag{3.2}$$

where μ_0, β_0 and β_1 are known constants.

If X follows a normal distribution, the full likelihood ratio method has the following alarm rule:

$$LRN(s) = \sum_{j=1}^{s} w_j \frac{\exp\left(-\frac{1}{2\sigma^2} \sum_{t=1}^{s} (x(t) - \mu^{Cj}(t))^2 \right)}{\exp\left(-\frac{1}{2\sigma^2} \sum_{t=1}^{s} (x(t) - \mu^{D}(t))^2 \right)} > k_s$$

and for a Poisson distribution, the alarm statistic can be written as:

$$LRP(s) = \sum_{j=1}^{s} w_j \frac{\prod_{t=1}^{s} \exp\left(-\mu^{Cj}(t)\right) \cdot \left(\mu^{Cj}(t)\right)^{x(t)}}{\prod_{t=1}^{s} \exp\left(-\mu^{D}(t)\right) \cdot \left(\mu^{D}(t)\right)^{x(t)}}.$$

3.2. *The Shiryaev Roberts approach of the likelihood ratio method*

When the intensity of the change, $P(\tau = j \mid \tau \geq j)$, tends to zero, the full likelihood ratio method (LR, see (3.1)) tends to the method suggested by Shiryaev, (1963) and Roberts, (1966). This method gives an alarm when

$$\frac{f(x_s \mid \mu = \mu^{C1})}{f(x_s \mid \mu = \mu^{D})} + \frac{f(x_s \mid \mu = \mu^{C2})}{f(x_s \mid \mu = \mu^{D})} + \ldots + \frac{f(x_s \mid \mu = \mu^{Cs})}{f(x_s \mid \mu = \mu^{D})}$$

exceeds a constant alarm limit. This approach can also be motivated by a non-informative density for τ.

3.3. *The maximum likelihood ratio approach*

The generalised likelihood ratio (GLR) surveillance method by Lai, (1995) uses the maximum likelihood estimator of the value after the change.

For a situation that is somewhat related to outbreak detection, namely turning point detection, Frisén, (2000) suggested a surveillance method based on nonparametric estimation without any parametric assumptions, only the natural order restrictions that are present at a turning point. The method is based on the maximum likelihood ratio

$$\frac{\max f(x_s|C)}{\max f(x_s|D)},$$

where the likelihood expressions are maximised by using the maximum likelihood estimators. This approach was found useful for example in Andersson, (2002, 2004), Andersson *and others*, (2005) and Bock *and others*, (2007).

3.4. *Semiparametric outbreak detection*

For the outbreak situation studied in this paper we use a maximum likelihood ratio method, and we base the method on detection of the violation of order restrictions, see Section 2. If no outbreak has occurred, we have that the observations (or residuals) belong to state D where $\mu(1) = \mu(2) = ... = \mu(s)$. At an onset of the outbreak at time j, we have state Cj: $\mu(1) = \mu(2) = ... = \mu(j-1) < \mu(j) \leq ... \leq \mu(s)$. The maximum likelihood estimates $\hat{\mu}^{D}$ and $\hat{\mu}^{Cj}$ are given in Frisén *and others*, (2007a) for the exponential family. If we were interested in the specific value $\tau = j$ we could use

$$\frac{\max f(x_s|Cj)}{\max f(x_s|D)} = \frac{f(x_s;\mu=\hat{\mu}^{Cj})}{f(x_s;\mu=\hat{\mu}^{D})}.$$

However, here we are interested in onsets at any time up to the decision time s, so that $C = \{\tau \leq s\}$. Since all other states Cj, $j \geq 2$ are on the border of C1, we have that

$$\max f(x_s|C) = \max f(x_s|C1) = f(x_s;\mu=\hat{\mu}^{C1})$$

and thus

$$\frac{\max f(x_s|C)}{\max f(x_s|D)} = \frac{\max f(x_s|C1)}{\max f(x_s|D)}, \qquad (3.3)$$

which is our suggested alarm statistic. We will subsequently use a constant alarm limit in the surveillance method. This corresponds to a non-informative density for the change point, as in the Shiryaev-Roberts approach.

In the present context, this approach has similarities with the CUSUM approach, which is expressed using likelihood ratios in Frisén, (2003). For the CUSUM approach, the alarm statistic is the maximum likelihood ratio with respect to τ,

$$\max_{j=\{1,2,\ldots,s\}}\left[\frac{f(x_s|Cj)}{f(x_s|D)}\right].$$

The expression above has similarities with the suggested statistic in (3.3), which can also be written as

$$\frac{\max f(x_s|C)}{\max f(x_s|D)} = \max_{j=\{1,2,\ldots,s\}}\left[\frac{\max f(x_s|Cj)}{\max f(x_s|D)}\right],$$

since in our case

$$\max_{j=\{1,2,\ldots,s\}}[\max f(x_s|Cj)] = \max f(x_s|C1) = f(x_s;\mu=\hat{\mu}^{C1}).$$

The full likelihood ratio method is optimal with respect to the expected delay (Shiryaev, 1963), and the CUSUM method is minimax optimal (Moustakides, 1986). However, when the models are not fully known and maximal likelihood expressions are used , as in (3.3), we cannot prove optimality. Instead, we have to examine whether the use of approaches, which are similar to the optimal ones, results in methods with good properties.

For the outbreak detection situation and the normal distribution, the method is denoted by OutbreakN, and the maximum likelihood alarm statistic (3.3) becomes

$$\frac{f(x_s;\mu=\hat{\mu}^{C1})}{f(x_s;\mu=\hat{\mu}^{D})} = \frac{\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{s}(x(i)-\hat{\mu}^{C1}(i))^2\right)}{\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{s}(x(i)-\hat{\mu}^{D}(i))^2\right)}. \qquad (3.4)$$

The normal distribution may be of interest (as an approximation) for diseases with a high baseline incidence. In most public health applications, however, the Poisson distribution is of special interest. Here the method is denoted by OutbreakP, and the alarm statistic is

$$\frac{f(x_s;\mu=\hat{\mu}^{C1})}{f(x_s;\mu=\hat{\mu}^{D})} = \exp\left\{\sum_{t=1}^{s}\left(\hat{\mu}^{D}(t)-\hat{\mu}^{C1}(t)\right)\right\}\cdot\prod_{t=1}^{s}\left(\frac{\hat{\mu}^{C1}(t)}{\hat{\mu}^{D}(t)}\right)^{x(t)} = \prod_{t=1}^{s}\left(\frac{\hat{\mu}^{C1}(t)}{\hat{\mu}^{D}(t)}\right)^{x(t)} \quad (3.5)$$

The time of alarm, t_A, is the first time when the Outbreak statistic exceeds a constant alarm limit.

It is not possible to base the Outbreak statistic on a single observation. Since the maximum likelihood method uses the ordering of the data, no alarm can be given when we have only one observation, x(1). Thus, the first decision is taken when we have two observations, x(1) and x(2).

The semiparametric Outbreak methods will be compared to the Shewhart method described in the next section.

3.5. *The Shewhart method*

In 1931, a method later known as "the Shewhart method" was presented (Shewhart, 1931). Originally, it was presented for the purpose of industrial quality control. The method is very simple and still the most commonly used in surveillance. Detailed descriptions are found in many textbooks, for example Wetherill and Brown, (1991) and Ryan, (2000).

The Shewhart method is often described in terms of a deviation from a known baseline μ^D. An alarm is called the first time s that

$$(x(s)-\mu^D) > k, \quad (3.6)$$

where k is the alarm limit which is often chosen as $3*\sigma$, where σ is the standard deviation.

The Shewhart method can also be seen as a special case of the full likelihood ratio method (Frisén, 2003, 2007). In a situation where we want to detect a change that has occurred at the current time point, we would specify C as $\{\tau = s\}$. In a situation where we have independent normal observations and a shift in the mean, the full likelihood ratio in (3.1) would be reduced to the Shewhart method as in (3.6), see Frisén and de Maré, (1991). A generalised Shewhart method could be expressed with the alarm criteria

$$L(s,s) > G,$$

where G is a constant. This means that the Shewhart method gives the minimum expected delay in the situation where we want immediate detection. We also have minimal error probabilities for each decision time s Frisén and de Maré, (1991). The Shewhart method is the limit of several advanced surveillance methods when these are optimised for a large shift, see Frisén and Wessman, (1999), Frisén, (2007). When

we expect a large change at the current time point, the Shewhart method is suitable and will have the best detection ability.

Even though very advanced modelling is sometimes used, the (generalised) Shewhart approach of not accumulating information over time is by far the most common also in public health surveillance. Examples of methods which are well developed and well recognised include those by Farrington and Andrews, (2004), Stern and Lightfoot, (1999) and most methods for spatial surveillance.

In Section 6 the semiparametric OutbreakN method is compared to the Shewhart method.

4. EVALUATION MEASURES

In hypothesis testing, we usually evaluate performance by power for a fixed size. In diagnostic tests, we often use specificity and sensitivity. In this outbreak detection situation, it may also seem safe to use such well-established metrics. Simple metrics are also required by medical authorities who have to handle the information in this new area, and there currently are many suggestions of simple metrics for surveillance. However, simple solutions to complex problems are not always useful. In surveillance we need measures that involve time, since timeliness is important and since the properties of a surveillance method often change with time (cf. Frisén, 1992 and Frisén, 2003).

Quick detection and few false alarms are desired properties of methods for surveillance. The time of the alarm, t_A, should come soon after the time of the change (τ) – but not before.

The false alarm frequency is here measured by the Average Run Length when no change has occurred. We have

$$ARL^0 = E[t_A|D],$$

which is the most commonly used false alarm measure in surveillance.

The delay of an alarm is most often measured by ARL^1, which is the average run length until the detection of a change at $\tau = 1$ (i.e. a change that occurs right at the start of the surveillance). Here we do not want to restrict the evaluation to $\tau = 1$, since we are interested in changes which can occur at any time. Thus, we use the more general measure of the conditional expected delay, CED.

$$CED(t) = E[t_A - \tau \,|\, t_A \geq \tau, \tau = t],$$

For most methods, the CED(t) will converge to a constant value when τ tends to infinity. This value is the Steady state Average Delay Time, SADT. It is, in a sense,

the opposite of ARL[1] since only very large values of τ are considered. SADT has been advocated for example by Srivastava and Wu, (1993).

When judging which method is best, it matters much if the evaluation is made for early changes or for late ones, as illustrated by the results in Sego *and others*, (2008). Compared to earlier authors, they came to the opposite conclusion about which method is the best. They used SADT, which evaluates the performance at late changes, while earlier papers have used ARL, which evaluates the performance at early changes.

Sometimes the time available for rescue actions is limited. The Probability of Successful Detection, suggested by Frisén, (1992), measures the probability of detection with a delay time no longer than a constant d

$$\mathrm{PSD}(d, t) = P(t_A - \tau \le d \,|\, t_A \ge \tau, \tau = t).$$

It may be useful to describe the ability to detect the change within a certain time limit, and PSD can be calculated for different time limits d. This has been done for example by Marshall *and others*, (2004) and Buckeridge *and others*, (2005).

The predictive value is a well-established measure in epidemiology. In surveillance, however, we need a variant that also incorporates time. If a method calls an alarm, it is important to know whether this alarm is a strong indication of a change or just a weak one. The difference in surveillance, as compared to situations involving only one decision, is that we can get an alarm at any time point, and therefore we need a measure of the predictive value at each of them. In order to judge the trust in an alarm at time t, it is necessary to consider the balance between the risk of false alarms, the detection ability and the probability of a change for that time point. If τ is regarded as a random variable, this can be done by the following predictive value of an alarm, which was suggested by Frisén, (1992):

$$\mathrm{PV}(t) = P(C|t_A = t) = \frac{\sum_{i=1}^{t} P(t_A = t \,|\, \tau = i) \cdot P(\tau = i)}{\sum_{i=1}^{t} P(t_A = t \,|\, \tau = i) \cdot P(\tau = i) + P(t_A = t \,|\, \tau > t) \cdot P(\tau > t)}.$$

In Section 5.1, the results from a simulation study on the properties of the OutbreakP method are presented. In addition, the OutbreakP method is applied to Swedish influenza data. In Section 6 we use the measures to compare different methods.

5. DETECTION OF THE INFLUENZA OUTBREAK

Epidemics, such as influenza, are for several reasons very costly to society and it is therefore of great value to monitor the epidemic period in order to allocate medical resources. Great emphasis should be put on the timeliness of a surveillance method. In this section, the properties of the OutbreakP method are presented, first by the results from a simulation study and then by the application of the method to observed Swedish influenza data.

5.1. *Simulation study to determine the properties of the semiparametric method*

In this study, the OutbreakP method (3.5) was applied to data generated from a model that mimics the Swedish LDI data. In all simulation studies in this paper there are at least 1,000,000 replicates. Observations on X(t) were generated from two different distributions, depending on whether $t<\tau$ (state D) or $t\geq\tau$ (state C), and we generated the data according to the structure described in (3.2). A Poisson distribution for X was suggested in Andersson *and others*, (2007) for the onset phase, and the model used was

$$X(t) \sim \begin{cases} \text{Poi}(\mu_0), & t < \tau \\ \text{Poi}(\mu(t)), & t \geq \tau \end{cases}$$

where Poi(*) refers to the Poisson distribution. The level at the constant phase, μ_0, was roughly estimated to $\mu_0 = 1$ from Swedish LDI data for eight years. The exponential curve $\mu(t) = \exp(\beta_0 + \beta_1(t-\tau+1)$ for the increasing phase was suggested in Andersson *and others*, (2007). The parameters, β_0 and β_1, were estimated to $\beta_0 = -0.26$ and $\beta_1 = 0.826$ from Swedish LDI data from the season 2003-2004, which was not extreme in any sense but "typical". The curve of the expected value is illustrated in Figure 1.

The properties of the method were determined in the simulation study and are illustrated in the figures below. The predicted value depends on whether the disease appears commonly or rarely (i.e. on the intensity of the outbreaks, the distribution of τ). Knowledge of the exact distribution of τ is seldom available, but since the predicted value contains very important information, we will nevertheless try to give a rough indicator. Here a constant intensity was used. This might not be the most probable density, but in order to detect outbreaks which occur at an unusual time we did not want to include information on which week is the most common for the onset. The level of the intensity was roughly estimated from all available historical data to be $\nu = 0.1$. In Figure 2, the PV curve is given both for $\nu = 0.1$ and for a lower intensity, $\nu = 0.05$, which weakens the PV. The alarm limit was chosen to 5,000 in order to give the method a high PV curve (higher than 0.99, so that alarms can be trusted. Since it is not possible for the OutbreakP method to signal an alarm at the first time point, no predicted value was calculated for $t_A = 1$.

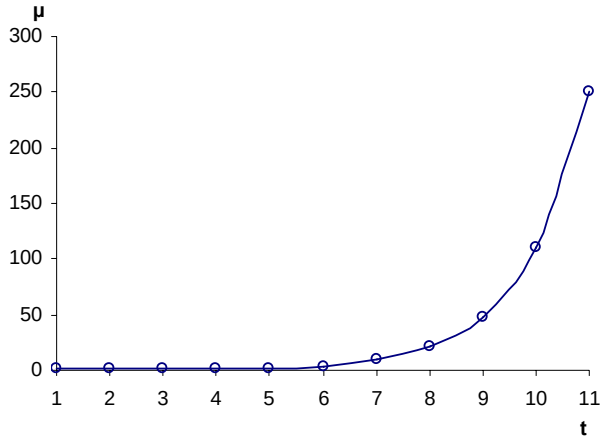

Figure 1. The expected value μ(t) of the incidence, using the model that mimics LDI. The model is here exemplified for the time τ=5 of the onset of the outbreak.

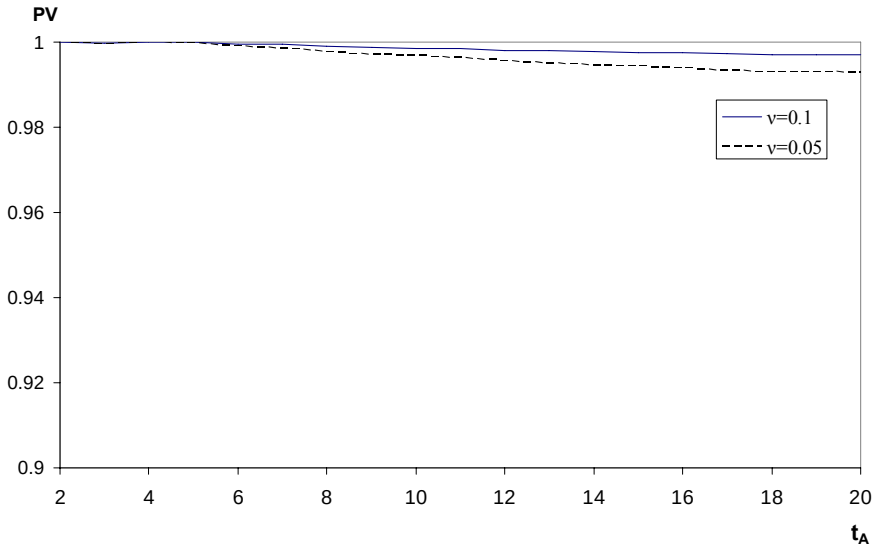

Figure 2. Predictive value (PV) as a function of the time of alarm, t_A, for the OutbreakP method.

A high alarm limit will result in few false alarms and a high predicted value. The drawback is a long delay before detection. The conditional expected delay, CED, and the probability of a successful detection, PSD, as discussed in Section 4, are given in Figures 3 and 4.

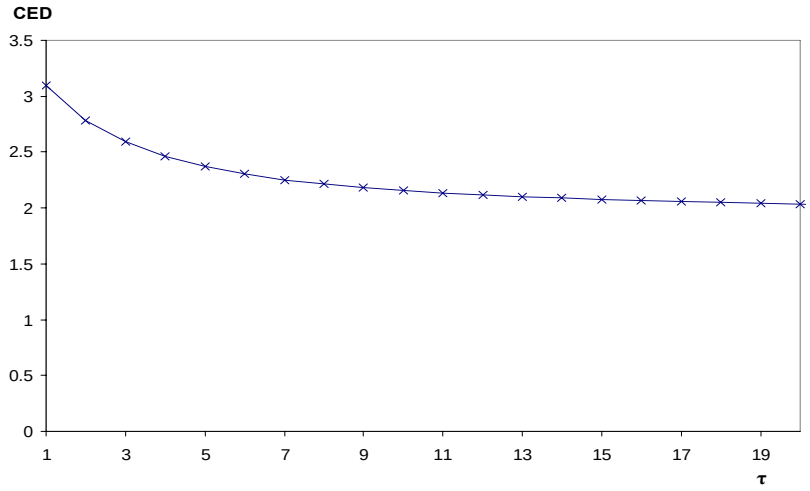

Figure 3. Conditional expected delay (CED) as a function of the outbreak time, τ, for the OutbreakP method.

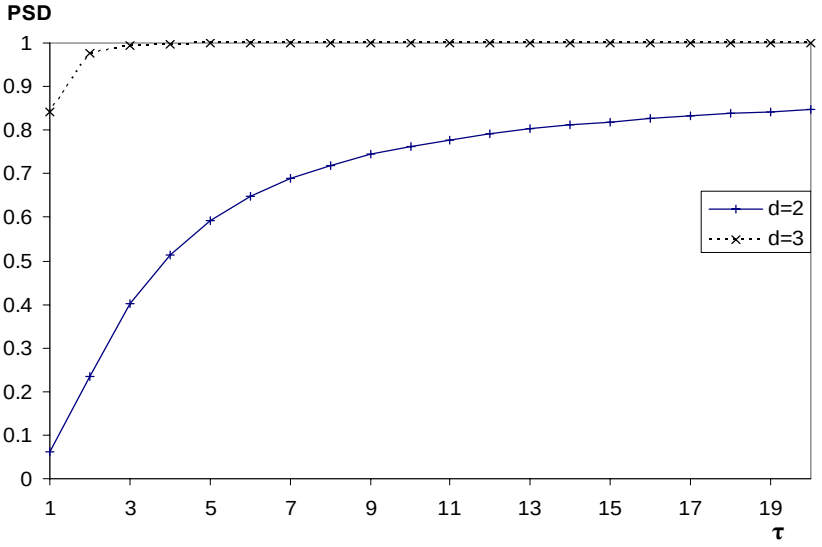

Figure 4. Probability of successful detection (PSD) within d time units from the onset, τ, as a function of τ for the OutbreakP method.

The method and alarm limit used in the simulation study were considered potentially useful for practical application since the predictive value was high.

5.2. *Application of the OutbreakP method to Swedish LDI data*

The OutbreakP method was applied to Swedish LDI data for six years. The alarm limit was the same as in the simulation study presented in Section 5.1, which means that, for a typical influenza season, the OutbreakP method has the properties (PV, CED, PSD) described in that section.

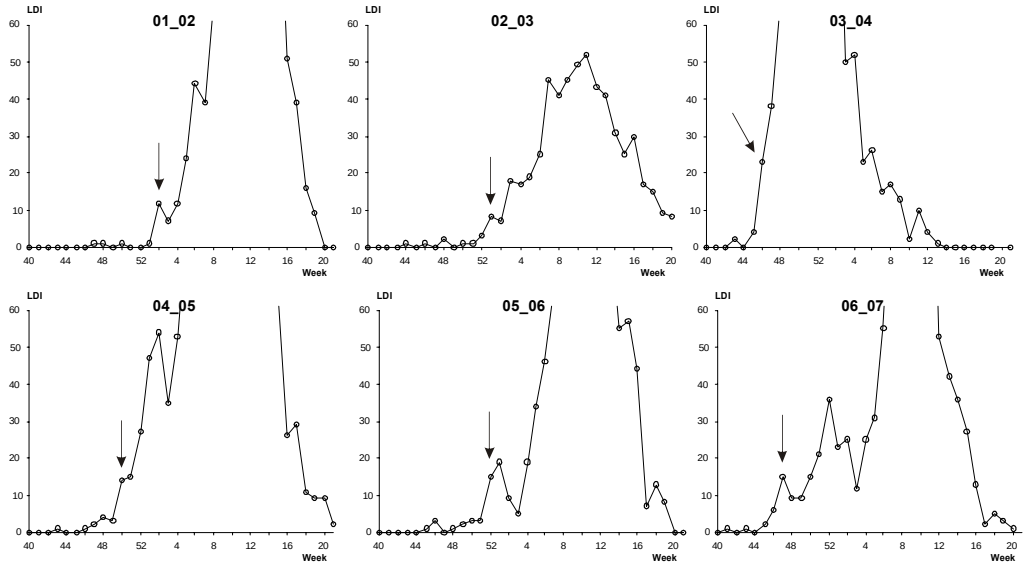

Figure 5. Swedish LDI data for six years. The scale is chosen in order to set focus on the low values at the onset. Thus, the peaks for most years cannot be seen. The arrows mark the time of alarm using the OutbreakP method.

Figure 5 demonstrates that the new method has potential. The alarms come at time points that seem natural. However, this does not mean that the statistical method is unnecessary and that a subjective judgement would work just as well. When studying the graphs above we make a retrospective judgement, whereas the OutbreakP method works prospectively. Making prospective judgements is much more difficult since less information is available at the decision time. In a real situation we would work prospectively, getting a new observation each week and aiming at an alarm as soon as we had enough evidence for an outbreak. In an experiment, reported in Frisén *and others*, (2007b), it was demonstrated that the statistical method worked better than subjective judgements.

Figure 6 also illustrates the OutbreakP method applied to the six influenza seasons. Here we present both the observed incidence and the alarm statistic used to produce the alarms in Figure 5. Figure 6 shows that the alarm statistic captures the pattern of an increasing incidence and thus gives early information about the onset of the outbreak.

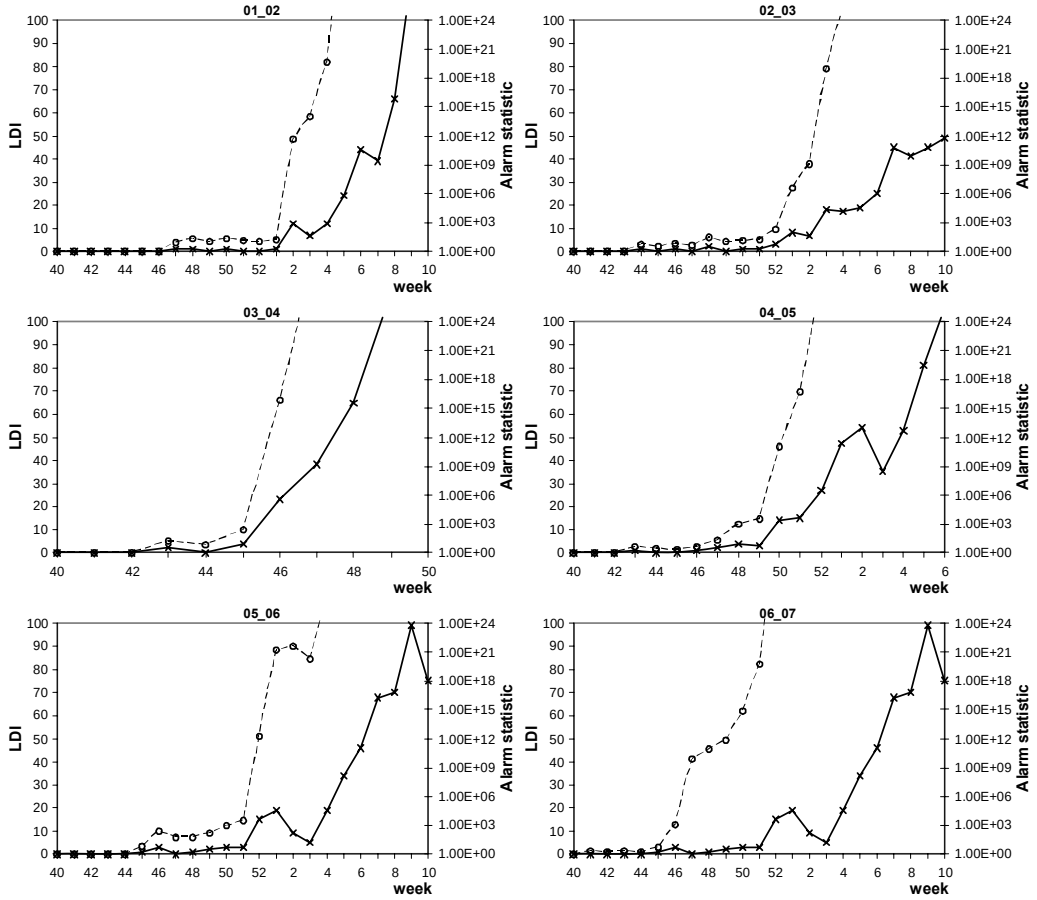

Figure 6. The OutbreakP method applied to Swedish LDI data for the latest six seasons (01-02 to 06-07). The left axis and the solid line correspond to the number of LDI cases. The right axis and the dotted curve correspond to the alarm statistic.

6. COMPARISONS BETWEEN METHODS

Above we illustrated the OutbreakP method by giving both the observed incidence and the alarm statistic. It should be remembered that the Shewhart method uses only the latest observation. Thus, the Shewhart alarm statistic has the same pattern as the observations themselves in Figure 6. As we can see, the alarm statistic of the OutbreakP method captures the pattern of an increasing incidence also when the incidence is low. This may serve as an illustration of the drawback of the Shewhart method, which only evaluates each time point without accumulating the information about the pattern.

The Shewhart approach of judging each time separately and not accumulating the information is frequently used. Thus, it is important to compare this approach to our method where the information is accumulated. Many methods are advanced in terms of seasonal adjustment or background variables. However, we will concentrate on the accumulation effect by comparing the approaches when applied to simple models. Usually, the residuals from a complex model are used in this simple way. The further comparison between the parametric Shewhart method and the semiparametric outbreak detection method is made by a simulation study of a simple situation, which agrees rather well with Swedish ILI data.

We will now compare the OutbreakN method, see (3.4), to the Shewhart method. The aim is to make the comparison more focused. In both OutbreakN and the Shewhart method, we need the variance σ^2, which is not necessary in OutbreakP. For the Shewhart method, (3.6), we furthermore need the knowledge of the baseline value, μ_0. The nonparametric OutbreakN method and the Shewhart method are compared with special concern about the effect of uncertainty of the baseline.

Observations are generated according to the following model

$$X(t) \sim \begin{cases} N(\mu_0;\sigma), & t < \tau \\ N(\mu(t);\sigma), & t \geq \tau \end{cases}$$

where $\mu(t) = \exp(\beta_0 + \beta_1(t-\tau+1)$ and $\mu_0 = 20$, $\beta_0 = 2.67$ and $\beta_1 = 0.68$ and $\sigma^2 = 100$. This curve was estimated in Frisén *and others*, (2007b) for the incidence of the number of influenza-like cases (ILI) during the winter 2003-2004. The normal distribution with a constant variance is chosen in order to illustrate important principal differences between methods rather than to give information about Swedish influenza. The sentinel system in Sweden still has the disadvantage of a low reporting tendency in the beginning and end of the influenza season as well as during holidays, see Andersson *and others*, (2007) and Andersson *and others*, (2008). However, progress is made in this area. With the data we have, the estimates of parameters resembling ILI data. are not as good as we would have whished.

For comparability, the alarm limits were chosen to give all methods the same value (27.4) of $E[t_A|D]$, where D is the nonepidemic state. Thus, the expected run length, given that there is no outbreak, is intended to be the same.

An important difference between the OutbreakN method and the Shewhart method is the requirement of a known baseline for the Shewhart method. We will study the effect of an estimation error of the baseline, but first the baseline value is assumed to be exactly known in the Shewhart method.

Exact knowledge of the baseline provides important information, and one could expect the Shewhart method to have much better properties than the nonparametric method, which does not utilize such knowledge. However, if the baseline in the model (μ^D in (3.6)) is estimated the situation is different. For Swedish data four or five weeks each year could be used for estimation, giving us at total of 25 observations. If the true model is the same as above, then the estimates (20+4 = 24 and 20–4 = 16) are both rather probable since they are both within 95% of the frequency distribution of the estimator.

We first discuss the properties of the OutbreakP method and the Shewhart method with a known baseline (μ^D=20), see Figures 7–9. The Shewhart method with a correct baseline has a better CED and PSD when the constant phase is short. However, the PV for Shewhart is worse except for very late alarms. This can be explained by the generally bad PV-property of the Shewhart method (Frisén, 2003) and by the fact that this method does not accumulate the information (see Fig 6).

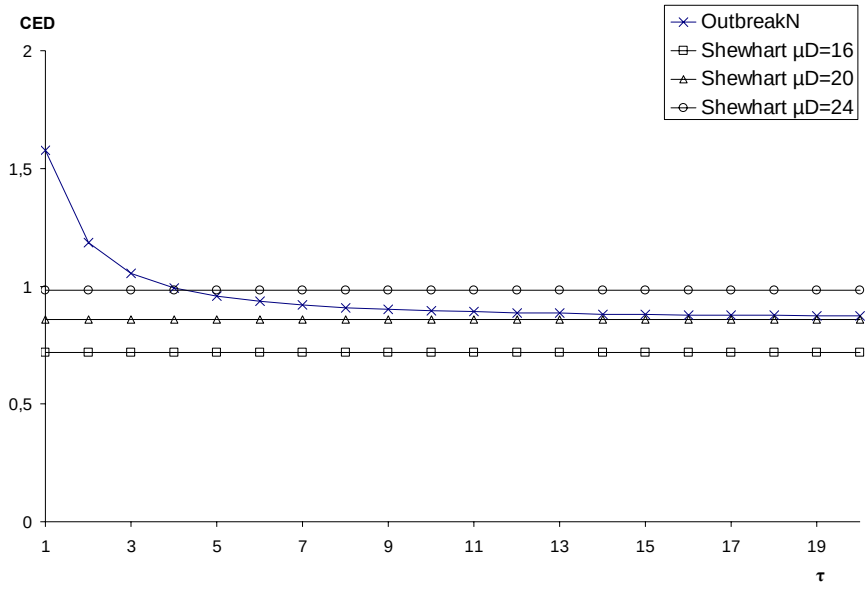

Figure 7. Conditional expected delay, CED(τ), for the methods OutbreakN and Shewhart, where the Shewhart method is compared for two different possible estimates of the baseline.

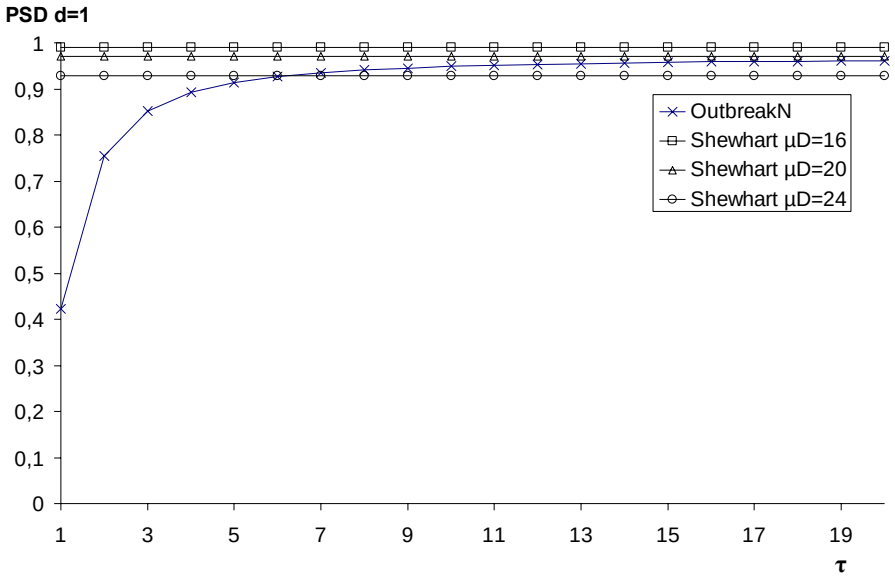

Figure 8. Probability of successful detection within 1 time unit (PSD for d=1) as a function of the time, τ, of the onset of the outbreak.

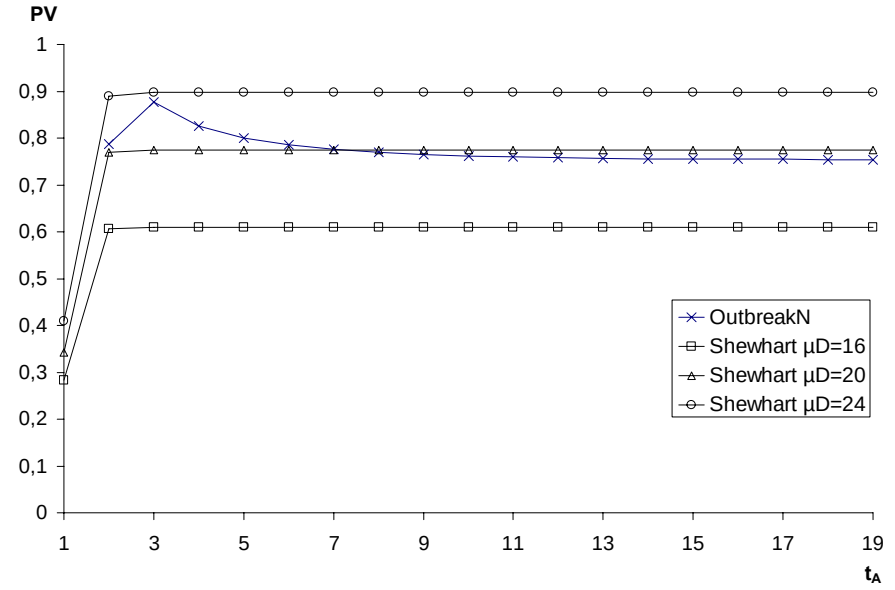

Figure 9. Predictive value, PV, as a function of the time of alarm, t_A.

We now turn to the wrongly specified Shewhart method. When the baseline is overestimated ($\mu^D = 24$ used in the Shewhart method instead of $\mu^D = 20$), the CED is

longer than for the correct baseline and hardly more satisfactory than that of the nonparametric method. When the baseline is underestimated ($\mu^D = 16$ used instead of $\mu^D = 20$), the PV is very low and considerably weaker than that of the nonparametric method. Thus, uncertainty about the baseline will mean that the properties of the method are highly uncertain. Our investigation confirms the results by Albers and Kallenberg, (2004) that very large sample sizes for the baseline are necessary in order to obtain reliable properties (in their case ARL).

Here we studied misspecifications, which have to be considered because of the stochastic variation. In many applications, however, also other errors than the stochastic ones have to be considered. In surveillance systems monitoring a large number of diseases baseline estimation will certainly prove problematic in some cases, and then the nonparametric method can be an alternative. The worst consequence of using a poorly estimated baseline might be that one does not know the properties but has to prepare for the least favourable results in each of the graphs here – that is, both an unsatisfactory predicted value and an unsatisfactory delay.

7. DISCUSSION

To detect the onset of an outbreak is important. Often, the information about the baseline is limited. Thus, it can be of value to have access to a method which does not require knowledge about the baseline but is focused on the increasing incidence at an outbreak. A semiparametric maximum likelihood ratio surveillance method was derived for the regular exponential family and described in detail with reference to the normal and Poisson distributions. Its properties, such as the delay and the predicted value, were determined by a simulation study where data of a similar pattern as the Swedish influenza data were generated. The method was also applied to influenza data from six seasons with satisfactory results.

Since many methods suggested for outbreak detection are based on the Shewhart approach where the residuals from some model are evaluated for each time point, we also made a special study of the effect of information accumulation, as in the suggested method. If the baseline is exactly known, then the Shewhart method (which uses this) performs better than the nonparametric method (which does not). The difference is large for the first time points, when little information is available from the data, but diminishes quickly. Even slight errors in the estimation of the baseline, used in the Shewhart method, have a large effect on the properties. For an overestimated baseline, the nonparametric method has better detection ability, and for an underestimated baseline, it has a higher predicted value. The worst consequence of using a poorly estimated baseline might be that one does not know the properties, which makes it difficult to interpret an alarm.

The comparison between the Shewhart and nonparametric methods was focused on the effect of an estimated baseline and the accumulation of information, while

other elements were kept as equal as possible. The common approach of giving an alarm when the incidence passes a fixed limit is the Shewhart method for a fixed variance. At the onset of an outbreak, however, a constant variance may not be a realistic assumption. The possibility to choose for example the Poisson distribution in the likelihood expression is an advantage, because of the small variance at the onset.

We derived a method which did not utilise any information about when it would be probable that the outbreak occurred. The likelihood principle makes it possible to include such knowledge. However, we chose a non-informative approach in this paper, since it may be valuable to detect outbreaks which occur at an unexpected time.

ACKNOWLEDGMENTS

Linus Schiöler has provided expert technical and computational help. Kjell Pettersson has given constructive comments. 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. *Conflict of interest:* None declared.

FOUNDING

Swedish Emergency Management Agency (0314/206).

REFERENCES

ALBERS, W. & KALLENBERG, W. C. M. (2004) Estimation in Shewhart control charts: effects and corrections. *Metrika,* **59,** 207-234.

ANDERSSON, E. (2002) Monitoring cyclical processes - a nonparametric approach. *Journal of Applied Statistics,* **29,** 973-990.

ANDERSSON, E. (2004) The impact of intensity in surveillance of cyclical processes. *Communications in Statistics-Simulation and Computation,* **33,** 889-913.

ANDERSSON, E., BOCK, D. & FRISÉN, M. (2005) Statistical surveillance of cyclical processes. Detection of turning points in business cycles. *Journal of Forecasting,* **24,** 465-490.

ANDERSSON, E., BOCK, D. & FRISÉN, M. (14 August 2007) Modeling influenza incidence for the purpose of on-line monitoring. *Statistical Methods in Medical Research***,** doi:10.1177/0962280206078986.

ANDERSSON, E., KUHLMANN-BERENZON, S., LINDE, A., SCHIÖLER, L., RUBINOVA, S. & FRISÉN, M. (2008) Predictions by early indicators of the time and height of yearly influenza outbreaks in Sweden. *Scandinavian Journal of Public Health***,** in press.

ANDERSSON, L. & FRISÉN, M. (2002) Verifications of Turning Points. *Journal of Nonparametric Statistics,* **14,** 623-645.

BARON, M. (2000) Nonparametric adaptive change-point estimation and on-line detection. *Sequential Analysis,* **19,** 1-23.

BARON, M. (2002) Bayes and asymptotically pointwise optimal stopping rules for the detection of influenza epidemics. IN C. GATSONIS, R. E. K., A. CARRIQUIRY, A. GELMAN, D. HIGDON, D. K. PAULER AND I. VERDINELLI (Ed.) *Case Studies in Bayesian Statistics.* New York, Springer-Verlag.

BESAG, J. & DIGGLE, P. (1977) Simple Monte Carlo Tests for Spatial Pattern *Applied Statistics,* **26,** 327-333.

BOCK, D., ANDERSSON, E. & FRISÉN, M. (11 September 2007) Statistical surveillance of epidemics: Peak detection of influenza in Sweden. *Biometrical Journal***,** doi:10.1002/bimj.200610362.

BOCK, D. & PETTERSSON, K. (2006) Exploratory analysis of spatial aspects on the Swedish influenza data. *Smittskyddsinstitutets rapportserie.* Stockholm, Report from the Swedish Institute for Infectious Disease Control.

BOWMAN, A. W., JONES, M. C. & GIJBELS, I. (1998) Testing monotonocity of regression. *J. Comp. Graph. Statist.,* **7,** 489-500.

BUCKERIDGE, D. L., BURKOM, H., CAMPBELL, M., HOGAN, W. R. & MOORE, A. W. (2005) Algorithms for rapid outbreak detection: a research synthesis. *Journal of Biomedical Informatics,* **38,** 99-113.

CHANG, J. T. & FRICKER, R. D. (1999) Detecting when a monotonically increasing mean has crossed a threshold. *Journal of Quality Technology,* **31,** 217-234.

DIGGLE, P., KNORR-HELD, L., ROWLINGSON, B., SU, T.-L., HAWTIN, P. & BRYANT, T. N. (2004) On-line Monitoring of Public Health Surveillance Data. IN BROOKMEYER, R. & STROUP, D. (Eds.) *Monitoring the Health of Populations: Statistical methods for Public Health Surveillance.* Oxford, Oxford University Press.

DIGGLE, P., MORRIS, S. & MORTON-JONES, T. (1999) Case-control isotonic regression for investigation of elevation in risk around a point source. *Statistics in Medicine,* **18,** 1605-1613.

FARRINGTON, C. P. & ANDREWS, N. J. (2004) Outbreak detection: application to infectious disease surveillance. IN BROOKMEYER, R. & STROUP, D. F. (Eds.) *Monitoring the Health of Populations.* Oxford, Oxford University Press.

FRIED, R. & IMHOFF, M. (2004) On the Online Detection of Monotonic Trends in Time Series. *Biometrical Journal,* **46,** 90-102.

FRISÉN, M. (1986) Unimodal regression. *The Statistician,* **35,** 479-485.

FRISÉN, M. (1992) Evaluations of Methods for Statistical Surveillance. *Statistics in Medicine,* **11,** 1489-1502.

FRISÉN, M. (2000) Statistical Surveillance of Business Cycles. Research Report, Department of Statistics, Göteborg University.

FRISÉN, M. (2003) Statistical surveillance. Optimality and methods. *International Statistical Review,* **71,** 403-434.

FRISÉN, M. (2007) Properties and Use of the Shewhart Method and Followers. *Sequential Analysis,* **26**.

FRISÉN, M., ANDERSSON, E. & PETTERSSON, K. (2007a) Estimation of outbreak regression. *Research Report.* Statistical Research Unit, Department of Economics, Göteborg University, Sweden : 2007:13.

FRISÉN, M., ANDERSSON, E. & SCHIÖLER, L. (2007b) Robust outbreak surveillance of epidemics in Sweden. *Research Report.* Statistical Research Unit, Department of Economics, Göteborg University, Sweden : 2007:12.

FRISÉN, M. & DE MARÉ, J. (1991) Optimal Surveillance. *Biometrika,* **78,** 271-80.

FRISÉN, M. & WESSMAN, P. (1999) Evaluations of likelihood ratio methods for surveillance. Differences and robustness. *Communications in Statistics. Simulations and Computations,* **28,** 597-622.

GILL, R. & BARON, M. (2004) Consistent estimation in generalized broken-line regression. *Journal of Statistical Planning and Inference,* **126,** 460.

KULLDORFF, M. (1997) A spatial scan statistic. *Communications in Statistics. Theory and Methods,* **26,** 1481-1496.

LAI, T. L. (1995) Sequential Changepoint Detection in Quality-Control and Dynamical Systems. *Journal of the Royal Statistical Society B,* **57,** 613-658.

LAWSON, A. & RODEIRO, C. (2004) Developements in general and syndromic surveillance for small area health data. *Journal of Applied Statistics,* **31,** 397-406.

LE STRAT, Y. & CARRAT, F. (1999) Monitoring epidemiologic surveillance data using hidden Markov models. *Statistics in Medicine,* **18,** 3463-3478.

MARSHALL, C., BEST, N., BOTTLE, A. & AYLIN, P. (2004) Statistical issues in the prospective monitoring of health outcomes across multiple units. *Journal of the Royal Statistical Society A,* **167,** 541-559.

MOUSTAKIDES, G. V. (1986) Optimal stopping times for detecting changes in distributions. *The Annals of Statistics,* **14,** 1379-1387.

QUENEL, P., DAB, W., HANNOUN, C. & COHEN, J. M. (1994) Sensitivity, Specificity and Predictive Values of Health- Service Based Indicators For the Surveillance of Influenza-a Epidemics. *International Journal of Epidemiology,* **23,** 849-855.

ROBERTS, S. W. (1966) A Comparison of some Control Chart Procedures. *Technometrics,* **8,** 411-430.

RYAN, T. P. (2000) *Statistical methods for quality improvement,* New York, Wiley.

SEBASTIANI, P., MANDL, K. D., SZOLOVITS, P., KOHANE, I. S. & RAMONI, M. F. (2006) A Bayesian dynamic model for influenza surveillance. *Statistics in Medicine,* **25,** 1803-1816.

SEGO, L., WOODALL, W. & REYNOLDS JR., M. (2008) A comparison of surveillance methods for small incidence rates. *Statistics in Medicine***,** in print.

SERFLING, R. (1963) Methods for current statistical analysis of excess pneumonia-influenza deaths. *Public Health Reports***,** 494-506.

SHEWHART, W. A. (1931) *Economic Control of Quality of Manufactured Product,* London, MacMillan and Co.

SHIRYAEV, A. N. (1963) On optimum methods in quickest detection problems. *Theory of Probability and its Applications.,* **8,** 22-46.

SONESSON, C. (2007) A CUSUM framework for detection of space-time diease clusters using scan statistics. *Statistics in Medicine,* **26,** 4770-4789.

SONESSON, C. & BOCK, D. (2003) A review and discussion of prospective statistical surveillance in public health. *Journal of the Royal Statistical Society A,* **166,** 5-21.

SRIVASTAVA, M. S. & WU, Y. (1993) Comparison of EWMA, CUSUM and Shiryaev-Roberts Procedures for Detecting a Shift in the Mean. *The Annals of Statistics,* **21,** 645-670.

STERN, L. & LIGHTFOOT, D. (1999) Automated outbreak detection: a quantitative retrospective analysis. *Epidemiology and Infection,* **122,** 103-110.

STROUP, D. F., THACKER, S. B. & HERNDON, J. L. (1988) Application of multiple time-series analysis to the estimation of pneumonia and influenza mortality by age 1962-1983. *Statistics in Medicine,* **7,** 1045-1059.

TIBSHIRANI, R. & WANG, P. (2008) Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics,* **9,** 18-29.

WETHERILL, G. B. & BROWN, D. W. (1991) *Statistical process control: Theory and practice*, Chapman and Hall.

WOODALL, W. H., MARSHALL, J. B., JONER, J. M. D., FRAKER, S. E. & ABDEL-SALAM, A.-S. G. (2008) On the use and evaluation of prospective scan methods for health-related surveillance *Journal of the Royal Statistical Society A,* **171,** 223–237.

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.
2007:5	Andersson, E., Bock, D. & Frisén, M.:	Modeling influenza incidence for the purpose of on-line monitoring.
2007:6	Bock, D., Andersson, E. & Frisén, M.:	Statistical Surveillance of Epidemics: Peak Detection of Influenza in Sweden.
2007:7	Andersson, E., Kühlmann-Berenzon, S., Linde, A., Schiöler, L., Rubinova, S. & Frisén, M.:	Predictions by early indicators of the time and height of yearly influenza outbreaks in Sweden.
2007:8	Bock, D., Andersson, E. & Frisén, M.:	Similarities and differences between statistical surveillance and certain decision rules in finance.
2007:9	Bock, D.:	Evaluations of likelihood based surveillance of volatility.
2007:10	Bock, D. & Pettersson, K.	Explorative analysis of spatial aspects on the Swedish influenza data.