Statistical methods for the prospective detection of infectious disease outbreaks: A review

Steffen Unkel[†1], C. Paddy Farrington[1],
Paul H. Garthwaite[1], Chris Robertson[2] and Nick Andrews[3]

[1]Department of Mathematics and Statistics
The Open University, Milton Keynes, UK

[2]Department of Mathematics and Statistics
University of Strathclyde, Glasgow, UK

[3]Statistics Unit, Centre for Infections
Health Protection Agency, London, UK

January 10, 2011

Abstract

Unusual clusters of disease must be detected rapidly for effective public health interventions to be introduced. Over the past decade there has been a surge in interest in statistical methods for the early detection of infectious disease outbreaks. This growth in interest has given rise to much new methodological work, ranging across the spectrum of statistical methods. This paper presents a comprehensive review of the statistical approaches that have been proposed. Applications to both laboratory and syndromic surveillance data are provided to illustrate the various methods.

Keywords: Biosurveillance; Clusters; Control chart; Epidemics; Infectious diseases; Outbreak; Prospective detection; Surveillance.

[†] Correspondence should be addressed to: Steffen Unkel, Department of Mathematics and Statistics, Faculty of Mathematics, Computing & Technology, The Open University, Walton Hall, Milton Keynes, MK7 6AA, United Kingdom (e-mail:S.Unkel@open.ac.uk).
This study was supported by grants from the National Institute for Health Research and Medical Research Council.

1 Setting the scene

The past decade has witnessed a large increase in research activity on the statistical issues related to prospective detection of outbreaks of infectious diseases. The major challenges in this expanding field derive from its focus on *prospective* detection, namely, detection of outbreaks as they arise, in a sufficiently timely fashion to enable effective control measures to be taken. The growth in this area, sometimes now referred to as *biosurveillance* (e.g., Shmueli and Burkom 2010), has been so rapid as to spawn conferences and a learned society, the International Society for Disease Surveillance (`http://www.syndromic.org`), founded in 2005.

Outbreak investigations go back at least to John Snow's iconic removal of the handle of London's Broad Street pump during the 1854 cholera epidemic. In the modern era, following a trend apparent in all areas of epidemiology, statistical methods have come to the fore in outbreak detection and control. For several decades (see Tillett and Spencer (1982) for an early example), statistical techniques have been used to provide early warnings of outbreaks, supplementing more traditional surveillance based on a network of alert public health physicians. Since the early 1990s, the increasingly widespread availability of computerised databases which can be interrogated for evidence of emerging outbreaks has greatly facilitated the use of statistical outbreak detection, and has witnessed the creation of automated detection systems to process data on very large numbers of infections at frequent time intervals. Since the turn of the twenty-first century, two factors have combined to give further impetus to developments in this area: new concerns about the possible threat of large-scale bioterrorism, and heightened public and media awareness about emerging or re-emerging infections, including hospital-acquired infections such as methicillin-resistant *Staphylococcus aureus* (MRSA) and *Clostridium difficile*, and global epidemics such as severe acute respiratory syndrome (SARS) in 2002-2003 and the 2009 H1N1 swine influenza pandemic. Similar statistical surveillance methods have also been used for the early detection of new antimicrobial resistant strains of infectious pathogens.

Under these influences, a new focus has emerged, namely the surveillance of syndromes, which complements the previous emphasis on surveillance of infections. Thus, much of the new literature in the field relates to syndromic surveillance, which exploits more diverse

sources of data, such as calls to telephone or internet helplines, medical consultations, and pharmacy sales, that are believed to reflect in more timely fashion changes in behaviour that may stem from a large-scale bioterrorist incident. So far, no such incident has occurred. However, the need to detect outbreaks from more mundane sources - such as contaminated foodstuffs, breakdowns in water treatment plants, low vaccine efficacy, or imported infections - remains, and has led to further developments, as witnessed by the inclusion of routine methods for statistical outbreak detection in laboratory-based surveillance systems in several European countries (Hulth et al. 2010).

Both syndromic and laboratory-based prospective surveillance methods for outbreak detection pose diverse statistical challenges, relating to data sources, evaluation, multiplicity control, and follow-up, as well as the statistical techniques used to detect anomalies in data series. Periodic reviews of these methods, with varying emphases, have appeared in the statistical literature, notably Sonesson and Bock (2003), Farrington and Andrews (2004), Buckeridge et al. (2005) and Shmueli and Burkom (2010). There have also been numerous developments in related fields such as pharmacovigilance and institutional performance monitoring.

The aim of the present review is to provide an account of the statistical methodologies that have been proposed for detecting anomalies in data series, specifically in the context of prospective outbreak detection. These methods are used to identify unusual patterns in data, which may result from infectious disease outbreaks. Prospective detection involves identifying anomalies as they arise, so as to enable control measures to be implemented, if deemed appropriate.

Our aim is not to cover the entire range of statistical issues or data sources relevant to outbreak detection (Fienberg and Shmueli 2005) - comprehensive coverage is probably no longer possible in a single review paper. Nor do we seek to document all the variants of each methodological approach, as developed in response to particular circumstances. Rather, we focus entirely on broad classes of statistical methods for detecting aberrations. Our motivation for doing so is to inform a detailed study of some of the outbreak detection systems used in the UK. Surveillance for health care-associated infections, such as surgical-site infections or ventilator-associated pneumonia, is outside the remit of this review, although it is mentioned in places, as some aspects of health-care surveillance overlap disease outbreak

detection. Much health care surveillance, however, addresses different aims, such as the monitoring of hospital performance and the evaluation of quality indicators. The paper is structured in the following Sections: regression techniques (Section 2), time series methodology (Section 3), methods inspired by statistical process control (Section 4), methods incorporating spatial information (Section 5), and multivariate outbreak detection (Section 6). We stress that this classification is chosen only to help our presentation of material, and is not based on any rigorous taxonomy. Several methods could be classified under more than one heading. We include a brief review of evaluation methodologies in Section 7. Concluding comments are given in Section 8.

2 Regression techniques

Regression methods of outbreak detection have been widely used, both for detecting outbreaks in surveillance systems based on laboratory reports and notified infections, and for syndromic surveillance. Their application differs from other areas of biostatistics, in that they are used primarily to obtain standardized residuals. The distribution of these residuals in the absence of an outbreak is then used to determine a threshold value.

Regression methods can be regarded as extending the Shewhart chart (Shewhart 1931), in which a process variable y_t which is normally distributed $\mathcal{N}(\mu, \sigma^2)$ when in control is monitored by tracking the values of $y_t - \mu$, an alert being declared when $|y_t - \mu| > k\sigma$ for some pre-specified value of k.[1]

When applied to outbreak detection, only the upper control limit $\mu + k\sigma$ is usually of interest. Regression methods generalize the Shewhart chart in three respects: the in-control mean μ and possibly the in-control standard deviation σ vary with time; both these quantities must be estimated from historical data; and the distribution of y_t may not be normal.

The performance of regression outbreak detection methods may be expected to reflect the performance of Shewhart charts: applied to an observation at time t, they are effective at detecting large outbreaks starting at time t, but rather less effective at detecting more grad-

[1] Throughout the paper, random variables and their realizations are not distinguished by using uppercase and lowercase letters. This is because in the multivariate case the reader will be more concerned about whether a vector or matrix of data is involved than with the distinction between random variables and their observed values.

ual outbreaks starting at some time earlier than t.

In Subsection 2.1 and Subsection 2.2 parametric and semiparametric regression methods are described. Most regression methods are based on a threshold value, above which reports are declared aberrant. How these thresholds may be obtained is explained in Subsection 2.3, whereas in Subsection 2.4 non-thresholding methods are considered.

2.1 Parametric models

Perhaps the simplest regression model for outbreak detection is that described by Stroup et al. (1989), in which the expected disease count at month t, $\mathrm{E}(y_t)$, is calculated as the mean of observed counts at months $t-1, t, t+1$ over some pre-specified number of years. This ensures that seasonal effects are automatically adjusted for by design rather than by explicit modelling, thus providing some element of robustness. However, this model does not incorporate time trends. Stroup et al. (1989) apply this model, using normal errors, to data on notifiable infections. The results are summarized in a simple graphic published routinely in *Morbidity and Mortality Weekly Report* of the Centers for Disease Control and Prevention (CDC) (Stroup et al. 1993).

A commonly used fully parametric outbreak detection regression model is based on that of Serfling (1963), who modelled historical baselines using a trigonometric function with linear trend of the form:

$$\mathrm{E}(y_t) = \mu + \alpha t + \sum_{i=1}^{r} \{\beta_i \sin(\omega_i t) + \gamma_i \cos(\omega_i t)\} \tag{1}$$

and normal errors with constant variance. Serfling (1963) used the regression equation (1) to estimate excess mortality due to influenza based on weekly data on pneumonia-influenza deaths. This model has subsequently been used to detect the onset of epidemics of influenza (Costagliola et al. 1991, 1994). An automated version of the Serfling model with cubic trend and three trigonometric terms (i.e. $r = 3$), with model selection based on the Akaike information criterion (AIC), has been developed for prospective and retrospective surveillance, and is available as a web-based application (Pelat et al. 2007) at `http://www.u707.jussieu.fr/periodic_regression`. Figure 1 shows a sample output screen from the system of Pelat et al. (2007), displaying the result of a prospective analysis of weekly counts of *Salmonella enteritidis* Phage type (PT) 4 in England, Wales and

Northern Ireland from 2000 to 2009 and model-based extrapolation for the year 2010 with an epidemic threshold. The user has first to select a subset of the whole data series (called

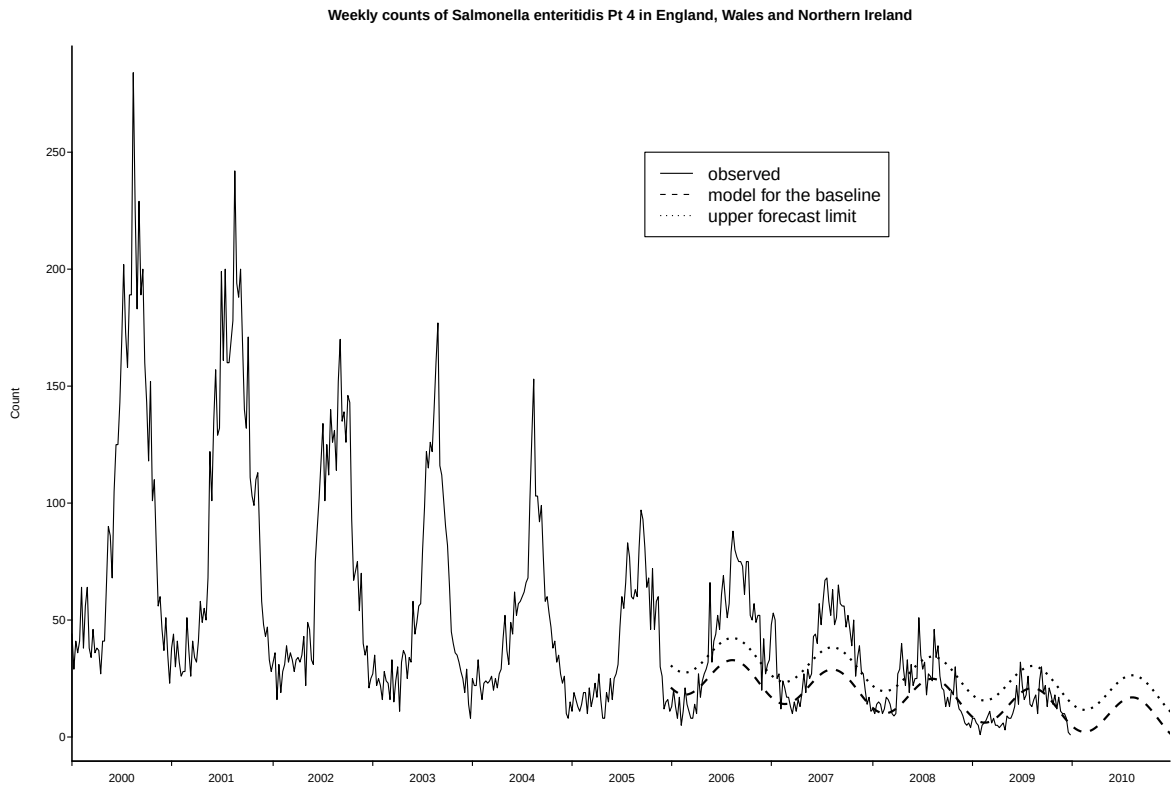

Figure 1: Sample output screen from the system of Pelat et al. (2007) (see text for explanation).

'training period') which is used to estimate the baseline level. In Figure 1 the training period consists of the years 2006 to 2009. Then the 20% highest values in the training data are excluded to account for past outbreaks (default value: 15%). A Serfling-type regression equation is used to model the baseline. A threshold is obtained by taking an upper percentile for the prediction distribution, here the upper 90th percentile is chosen. The system declares an aberration as soon as an observation exceeds this threshold.

When data are sparse, the normal errors regression model is inappropriate. Parker (1989) instead used Poisson regression with logarithmic link to monitor the mortality associated with abortions. Such a model can be elaborated at will; Jackson et al. (2007) describe a Poisson log-linear model for syndromic surveillance with terms for day of the week, month, linear time trend, and holidays. If denominator population data are available, binomial lo-

gistic models can be fitted in much the same way, the flexibility of the generalised linear model (GLM) approach allowing further extensions to include random effects, for example to represent spatial variation (cf. Section 5).

The log-linear regression model of Farrington et al. (1996), like that of Stroup et al. (1989), adjusts for seasonal effects by design, and explicitly allows for linear trends. Since much surveillance data exhibiting considerable overdispersion, the model is quasi-Poisson with

$$\mathrm{E}(y_t) = \mu_t \ , \qquad \log(\mu_t) = \alpha + \beta t \ , \qquad \mathrm{var}(y_t) = \phi\mu_t \ , \tag{2}$$

where ϕ is the dispersion parameter. To try to reduce the influence of past outbreaks, baseline values with high residuals are given lower weights in the regression. Estimates of model parameters in (2) are obtained by a quasi-likelihood method.

The method of Farrington et al. (1996) is used routinely by the Health Protection Agency (HPA) to detect outbreaks in laboratory-based surveillance data in England and Wales, and is referred to as the Lab-Base exceedance system hereafter. Elaborations of the Lab-Base exceedance system have been described for use with laboratory data in Scotland (McCabe et al. 2003) and The Netherlands (Widdowson et al. 2003).

2.2 Semiparametric models

All the regression models so far described use a parametric model to represent the historical data. A contrasting strategy is to use a non-parametric model for the historical baseline, as widely used in monitoring mortality and other effects in environmental time series (Dominici et al. 2003). The SPOT (Salmonella Potential Outbreak Targeting) system described by Stern and Lightfoot (1999) uses a smoothing method to obtain baselines and standard deviation. Five years' historical data are first smoothed, and the baseline value for each time point in the yearly cycle is taken to be the median of the five smoothed values. The standard deviation is obtained by smoothing the residuals (raw values minus smooth values). This system implicitly assumes Gaussian errors by using an alarm threshold of two standard deviations above the baseline (with a filter for low counts).

Wieland et al. (2007) proposed to model both the mean μ_t and the variance σ_t^2 at each time point t, using separate generalized additive models for both quantities. First, a generalized additive model (GAM) is fitted to historical data to obtain $\hat{\mu}_t$. Then a second GAM is

fitted to the residuals from the first GAM in order to obtain $\hat{\sigma}_t^2$. The threshold is taken as $\hat{\mu}_t + k\hat{\sigma}_t$ for some choice of k. The smooth terms in the GAMs were based on Gaussian kernel smoothers, with bandwidth chosen so as to minimize the mean predictive squared error on the historical data.

Mean-regression methods lack robustness to the presence of outbreaks in the baseline values, which bias $\hat{\mu}_t$ upwards and hence reduce the sensitivity of the system. Sensitivity is the probability that a true outbreak is detected: a high sensitivity is desirable for outbreaks of public health importance. Serfling (1963) identified past outbreaks by visual inspection and omitted them from the model; this will tend to bias $\hat{\mu}_t$ downwards and thus reduce the specificity. Specificity is the probability that a time period without an outbreak is correctly identified as such. The complement of this, 1 - specificity, must be kept low to avoid user fatigue and loss of credibility. The Lab-Base exceedance system (Farrington et al. 1996) downweights outliers in the baselines, but this reduces rather than eliminates the bias, as do non-parametric smoothing techniques.

An alternative is to use a wavelet transform of the baseline values to account for low frequency variation due to trends and seasonality, while remaining robust to high frequency variation resulting from past outbreaks or other artefacts, such as holiday dips. This approach was proposed by Zhang et al. (2003), whose simple and hence easily automated Wavelet-based Anomaly Detector (WAD) subtracts a baseline value obtained using the wavelet transform and bases thresholds on the distribution of the residuals (see also Wieland et al. 2007). More complex wavelet-based methods, which may also be applied in a time series framework, were used by Goldenberg et al. (2002) and discussed in Shmueli (2005).

2.3 Obtaining the thresholds

Most regression-based methods specify a model for the mean at time t, and declare an alarm at time t if the observed value lies above some threshold determined by the sample statistics and the quantiles of a suitable distribution, for example the normal, Poisson or negative binomial. A commonly used procedure for large counts is to estimate the baseline value $\hat{\mu}_t$ and the process variance at time t, $\hat{\sigma}_t^2$, and, assuming normal errors, to define the upper

threshold by

$$u_t = \hat{\mu}_t + k\hat{\sigma}_t \ ,$$

where in many applications it is further assumed that the process variance is constant, $\sigma_t^2 = \sigma^2$. A more accurate approach is to base the threshold on an upper $100(1-\alpha)\%$ prediction limit which takes account both of the process variance and the uncertainty in the estimation of the baseline value. Thus, for example, the Poisson model of Parker (1989) combines the estimated process variance $\hat{\mu}_t$ with the regression error variance $\text{var}(\hat{\mu}) = \hat{\mu}^2\text{var}(\log(\hat{\mu}_t))$, obtained using the delta method, to obtain the total variance

$$\text{var}_T(y_t) = \hat{\mu}_t\{1 + \hat{\mu}_t\text{var}(\log(\hat{\mu}_t))\} \ .$$

The quasi-Poisson method of Farrington et al. (1996) uses a similar approach. Clearly, regression methods do not account for serial correlation in the baselines. Kafadar and Stroup (1992) investigated bootstrap and jacknife procedures to estimate $\text{var}(\hat{\mu}_t)$ but suggested that such adjustments are inadvisable unless the autocorrelation structure is known or can be well estimated.

A further problem with thresholds of the form $\mu_t + k\sigma$ is that the normal theory upon which they are based is usually inappropriate, especially when the background means μ_t are small. The accuracy of the threshold - namely the extent to which $\text{P}(y_t > u_t)$ matches the nominal $100\alpha\%$ level upon it is based - will usually vary with μ_t, hence the sensitivity and specificity will vary with μ_t. This is undesirable for systems designed to monitor several different data series with a wide range of expected values.

Consequently, some detection algorithms apply a transformation to approximate normality, or approximate symmetry, derive the threshold on the transformed scale, and transform the threshold back to the original scale. The quasi-Poisson method of Farrington et al. (1996) uses the 2/3 power transformation, which yields approximate symmetry for Poisson data. The upper threshold is defined as

$$u_t = \hat{\mu}_t \left\{ 1 + \frac{2}{3} z_\alpha \left(\frac{\hat{\phi}_t \hat{\mu}_t + \text{var}(\hat{\mu}_t)}{\hat{\mu}_t^2} \right)^{1/2} \right\}^{3/2} \ ,$$

where z_α is the $100(1-\alpha)$-percentile of the standard normal distribution. The rationale is that, for rare organisms, $\phi \simeq 1$ and counts are distributed approximately Poisson, while

for frequent organisms normal approximations are valid. There are many possibilities for obtaining normal approximation to the error distribution: for example Cooper et al. (2004) use a sinh-based transformation applied to a binomial proportion to achieve a similar result.

2.4 Non-thresholding methods

Most regression methods for outbreak detection use a threshold u_t at time t to determine whether the current observation y_t is aberrant. An alternative is to test the null hypothesis that y_t belongs to the same distribution as the baseline values; Parker (1989) discusses various tests, including the likelihood ratio test, for Poisson data.

Other criteria for detecting outbreaks may be specified, based on the qualitative features of an outbreak, such as the start of an outbreak, at which point a previously stationary time series begins to increase, and the peak of an outbreak, at which the counts stop increasing and start decreasing (Andersson et al. 2008). Bock et al. (2008) describe methods for identifying the peak of an epidemic, and Frisén and Andersson (2009) and Frisén et al. (2010a, 2009) describe parametric and semi-parametric regression methods for detecting the onset of an epidemic.

The semiparametric model to detect epidemic onset assumes that the disease counts belong to a specified distribution within the regular exponential family. Under the semiparametric version of the model, it is assumed that the disease counts y_s up to time t ($s \leq t$) either have constant mean (the null hypothesis), or increase monotonically from time $t = d$ where d is known (the alternative hypothesis). For Poisson counts that increase from the start of the series ($d = 1$), the likelihood ratio statistic is

$$LR = \prod_{s=1}^{t} \left(\frac{\hat{\mu}_s^1}{\hat{\mu}_s^0} \right)^{y_s} \quad , \tag{3}$$

where $\hat{\mu}_s^0$ is the maximum likelihood estimator (MLE) of the process mean at time s under the null hypothesis, and $\hat{\mu}_s^1$ is the MLE under the alternative hypothesis. An alert is declared if $LR > k$ for some pre-specified value k. Frisén et al. (2010a, 2009) describe applications to outbreaks of influenza and tularaemia in Sweden. Computer software for this surveillance method is available (`http://www.statistics.gu.se/surveillance`), called 'Outbreak Detection P', as both a SAS program and as a VBA (Visual Basic for Applications) macro in

Microsoft Excel for Windows. Figure 2 shows the 'alarm graph' obtained from an analysis of the weekly counts of *Salmonella enteritidis* PT 4 in England, Wales and Northern Ireland in the year 2009 by means of the 'Outbreak Detection P' program. The likelihood ratio (alarm) statistic (3) exceeds the warning limit (100) for the first time in week 24 and the alarm limit (5000) in week 25. The semiparametric method by Frisén and Andersson (2009) is

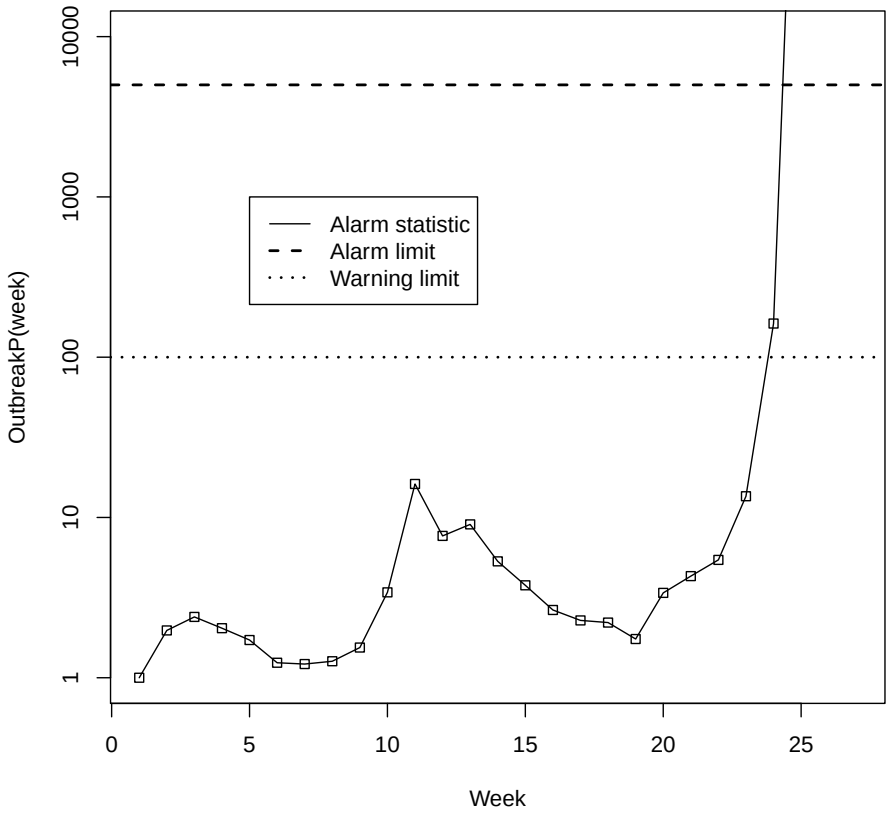

Figure 2: Sample alarm graph obtained from the 'Outbreak Detection P' program (see text for explanation).

also implemented in the R software package `surveillance` (Höhle 2007; Höhle and Mazick 2010). The `surveillance` package provides interfaces of several well-known procedures for prospective outbreak detection, among others the ones of Stroup et al. (1989), Farrington et al. (1996) and the system used by the Robert Koch Institute, Germany (Hulth et al. 2010). For a complete overview of its content, see `http://surveillance.r-forge.r-project.org`.

3 Time series methodology

Unlike most regression techniques, time series methods acknowledge the correlation structure of the data. For syndromic and laboratory data which are generally collected daily or weekly the principal correlations are autocorrelations at a lag of 1 time period (serial correlation) and correlations associated with the seasonal pattern in the data, which can be a combination of weekly or yearly seasonality. When time series data are available over a relatively long period of time, it is important to estimate the trend and seasonal components as the autocorrelation structure can only be identified using a stationary time series. Methods for estimating the trend and seasonal effects are described in Subsection 3.1. Failure to properly account for the autocorrelation will result in a misspecified model which may have bias in the estimated effects and prediction intervals which are too narrow, leading to a larger number of potential exceptions. The Box-Jenkins type methodology was designed to take the autocorrelation structure into account. These models are considered in Subsection 3.2. Much statistical innovation in the field of outbreak detection has tended towards time series models of increasing complexity, including Bayesian and hidden Markov models. The latter are discussed in Subsection 3.3.

3.1 Trend and seasonal estimation

With outbreak surveillance, the estimation of trend is best accomplished through a relatively simple procedure that is flexible and does not make any great demands on intervention by the operator. For time series, where there is a considerable amount of historical data and where the seasonal pattern is regular, a Serfling model (Serfling 1963) based upon sines and cosines may be used to estimate the trend and seasonal component.

Two common time series methods used in surveillance are simple exponential smoothing (e.g., Healy 1983; Ngo et al. 1996) and the Holt-Winters procedure (Holt 1957; Winters 1960). Simple exponential smoothing assures the data have no trend or seasonality. It forms predictions by taking a weighted average of past observations, where the weights decrease exponentially the further they are into the past (cf. Section 4).

The Holt-Winters technique is a generalization of simple exponential smoothing that allows for local trend and seasonal factors. Thus, for a series with multiplicative seasonal variation

of period p (e.g. for monthly observation $p = 12$), the h-step prediction function for the Holt-Winters model is given by (Chatfield and Yar 1988):

$$\hat{y}_{t+h} = (a_t + hb_t)s_{t-p+h} \quad (h = 1, \ldots, p), \tag{4}$$

where a_t is the overall average level of the series, b_t represents the trend and s_t represents the seasonal component, given by

$$a_t = \alpha\frac{y_t}{s_{t-p}} + (1-\alpha)(a_{t-1} + b_{t-1}) \ , \tag{5}$$

$$b_t = \beta(a_t - a_{t-1}) + (1-\beta)b_{t-1} \ , \tag{6}$$

$$s_t = \gamma\frac{y_t}{a_t} + (1-\gamma)s_{t-p} \ . \tag{7}$$

There are analogous formulae for additive seasonal variation. The smoothing constants α, β and γ $(0 < \alpha, \beta, \gamma < 1)$ in (5)–(7) can be estimated by minimizing the sum of squares of the one-step ahead prediction errors.

Holt-Winters is an extremely flexible method for estimating trend and has been used in many surveillance systems. It has performed well in many forecasting competitions in comparison to other more complex methods (Chatfield and Yar 1988). In syndromic surveillance systems it has been used to model outpatient attendance, in comparison to adaptive (using a moving baseline window) and non-adaptive regression models (Burkom et al. 2007), and found to provide a better fit to the data. It has also been used to account for the effects of temporal autocorrelation and spatial correlation when modelling calls to NHS24 and laboratory reports (Wagner 2010).

3.2 ARIMA and INAR models

Autoregressive integrated moving average (ARIMA) models (Box and Jenkins 1970) have been used for detecting outbreaks of infectious disease (e.g., Choi 1981; Helfenstein 1986; Reis and Mandl 2003; Watier et al. 1991). Fitting an ARIMA model requires the time series to be stationary. As much syndromic and laboratory data are likely to have seasonal and trend components it is necessary to remove these components from the original data before estimating the autocorrelation. Furthermore, the statistical testing procedure is based upon the normal distribution and this will really only be valid for infections or syndromes which occur frequently. When using outbreak surveillance for relatively rare events, or for more

common events in smaller areas, then integer-valued methods may be more appropriate.
The integer-valued autoregressive (INAR) model (e.g., Weiß 2009) is based upon the convolution operator $\circ$ (Steutel and van Harn 1979), where

$$\alpha \circ x = \sum_{k=1}^{x} y_k \ , \tag{8}$$

and the y_k are independent and identically distributed Bernoulli random variables with probability $\mathrm{P}(y_k = 1) = \alpha$ and x is a non-negative discrete random variable. Using (8), an INAR model of order p, denoted by INAR(p), can be defined as

$$x_t = \sum_{i=1}^{p} \alpha_i \circ x_{t-i} + \epsilon_t \ , \tag{9}$$

where ϵ_t is a (non-negative) random shock. It follows from (9) that an INAR(1) model is:

$$\begin{aligned} x_t &= \alpha_1 \circ x_{t-1} + \epsilon_t \\ &= y_1 + y_2 + \ldots + y_{x_{t-1}} + \epsilon_t \ . \end{aligned} \tag{10}$$

The model (10) states that the number of new cases in the interval $(t-1, t]$ is made up of two components - the x_{t-1} cases transmit the infection independently with probability α_1 and, as a consequence, $\sum y_k$ new cases arise, and a random number ϵ_t of new cases are generated via independent sources. Parameter estimates may be obtained using, for example, the Yule-Walker estimation technique.
When fitting ARIMA and INAR models to meningococcal incidence in the Montreal region of Canada (Cardinal et al. 1999), INAR(5) and AR(5) models were required for data aggregated in thirteen 4 week periods each year, from 1986 to 1993. There was no evidence of a trend or seasonality. In Scotland, using daily data on calls to NHS24 about vomiting there was evidence of a weekly pattern and a trend (Wagner 2010). The data were best fitted by an AR(6) model on a seasonally differenced series. ARIMA models were used to describe the daily visits to emergency departments at a hospital in Boston over a period of 10 years from 1992 (Reis and Mandl 2003) and an ARMA(2,1) model was required for total visits, while an ARMA(1,1) model was required for respiratory visits.
An extensive investigation of the use of ARIMA modelling, in comparison to statistical process control methods (cf. Section 4) has been carried out (Williamson and Weatherby Hudson 1999). They found that ARIMA modelling was unable to model 8 out of 17 syndrome time

series because of non-stationarity, partly arising from sparse data. Furthermore each series had to be investigated separately. For the series which were successfully modelled, one step ahead forecasts were satisfactory for forecasts up to 3 years in the future though better forecasts were obtained using continuously updated models.

In a syndromic-based detection approach, Reis and Mandl (2003) analyzed health care utilization patterns by a two-step time series modelling approach. A trimmed-mean seasonal model was used to capture both the yearly and weekly trends in daily utilization rates. The residuals from the trimmed-mean seasonal model were then fitted by an ARIMA model. An AR(6) model combined with a Serfling model with indicators variables for weekends and holidays was used to model attendance at ambulatory care centres for influenza like symptoms (Miller et al. 2003).

In these illustrations, the traditional ARIMA models require a relatively large number of parameters for the autocorrelation. Furthermore, a model the results for one syndrome is not easily adapted for use with another syndrome, so the whole process of model identification must be carried out each time, making the process difficult to automate.

With each new observation the parameters of the ARIMA model should be re-estimated. It is not clear how this can be achieved automatically, since model identification often relies on looking at residual plots. A practical approach may be to keep using the same model for a period of time, say one month for daily data and only refit the model each month. This might be practical for any one series but not for many. Thus it is likely that ARIMA methods might be better suited to the retrospective analysis of time series data, rather than prospective use within an outbreak surveillance system.

Heisterkamp et al. (2006) proposed the use of a so-called hierarchical time series model, which does not require a long time series of data for parameter estimation. The observed counts are assumed to be Poisson distributed, whereas an unobserved process is assumed for the time trend in the expected number of cases. The models for the unobserved process range from a completely stationary process over time, to an autoregressive model of order 3. This provides a flexible model for the trend in a times series. Likelihood ratio tests are used to discriminate between the different models in the hierarchy. It is claimed that the hierarchical time series model is able to detect outbreaks faster than the Lab-Base exceedance system (Heisterkamp et al. 2006).

3.3 Bayesian and hidden Markov models

Le Strat and Carrat (1999) proposed the use of a hidden Markov model (HMM) (e.g., Cappé et al. 2005) for monitoring epidemiological data. The basic idea is to segment the time series of disease counts into an epidemic and non-epidemic phase. In the HMM of Le Strat and Carrat (1999) each observed y_t $(t = 1, \ldots, n)$ is associated with a latent variable $z_t \in \{0, 1\}$ that determines the conditional distribution of y_t. That is, $y_t|z_t \sim f_k(y_t; \theta_k)$, where $k \in \{0, 1\}$, f_k is a pre-specified density (e.g. Gaussian or Poisson) and θ_k are parameters to be estimated. The unobserved state space, z_t $(t = 1, \ldots, n)$ is modelled by a two-state homogeneous Markov chain of order 1 with stationary transition probabilities

$$p_{kl} = \mathrm{P}(z_{t+1} = l|z_t = k),$$

where $k, l \in \{0, 1\}$ denote the two states of z_t (1: epidemic; 0: non-epidemic). For example, p_{01} is the probability of switching from the non-epidemic to the epidemic state. Note that in this Markov-dependent mixture model, y_t is conditionally independent of all the remaining variables, given z_t. Le Strat and Carrat (1999) also considered HMMs with more than two hidden states and performed model selection using the Bayesian information criterion (BIC). Parameter estimates were obtained by means of a modified version of the EM algorithm (Dempster et al. 1977). Model extensions to account for time trends and seasonality were proposed using the cyclic regression function of Serfling (1963). An on-line version of the retrospective approach of Le Strat and Carrat (1999) is implemented in the R package `surveillance` (Höhle 2007; Höhle and Mazick 2010).

Rath et al. (2003) and Madigan (2005) presented further exploration of HMM for surveillance, the latter incorporating the Bayesian perspective, which requires prior distributions to be specified for model parameters. Markov Chain Monte Carlo (MCMC) methods were used for parameter estimation.

In another Bayesian approach, Martínez-Beneito et al. (2008) proposed a Markov switching model (e.g., Cappé et al. 2005) for prospective surveillance of weekly influenza incidence rates. In a Markov switching model the observed variables depend not only on the hidden state variables but also on the lagged observable variables. This setting makes the Markov switching model more suitable for time-series analysis than HMM. In Martínez-Beneito et al. (2008) the conditional distribution of the first-order difference series, formed by the differ-

ences between rates in consecutive weeks, is modelled either as a first-order autoregressive process or as a Gaussian white noise process, depending on whether the system is in an epidemic or non-epidemic phase. Let $y^1_{t,j} = y_{t+1,j} - y_{t,j}$ denote the first-order difference that corresponds to the difference between the rates in weeks $t+1$ and t in season j. Conditional on the value of z_t,

$$\begin{aligned}
y^1_{1,j}|z_1 = 0 &\sim \mathcal{N}(0, \sigma^2_{0,j}) \\
y^1_{1,j}|z_1 = 1 &\sim \mathcal{N}(0, \sigma^2_{1,j}) \\
y^1_{t,j}|z_t = 0 &\sim \mathcal{N}(0, \sigma^2_{0,j}) \ , \quad t \geq 2 \\
y^1_{t,j}|z_t = 1 &\sim \mathcal{N}(\rho y^1_{t-1,j}, \sigma^2_{1,j}) \ , \quad t \geq 2 \ .
\end{aligned}$$

The advantage of using Markov switching models is that the differenced series is detrended , enabling autoregressive modelling to be used to analyse the data. The methodology described in Martínez-Beneito et al. (2008) is implemented in a web-based application called `FluDetWeb` (Conesa et al. 2009), which is used for the early detection of the onset of influenza epidemics. To illustrate this approach, we performed a prospective analysis using influenza-like illness (ILI) data from the Valencian Sentinel Network (VSN). The data set consists of eleven time series formed by the weekly ILI incidence rates (per 100,000 inhabitants in the Communitat Valenciana, Spain) during the seasons from 1996-1997 and 2006-2007. The data set can be downloaded via `http://www.geeitema.org/doc/meviepi/influenza.html`. An influenza season lasts 30 weeks (from the 42nd week of one year to the 19th week of the following year). In Figure 3, the weekly ILI incidence rates for the seasons 2005-2006 and 2006-2007 are compared. The black dots in Figure 3 (ii) indicate that the posterior probability of being in an epidemic phase exceeded 0.5 in week 15, 16 and 17 of the 2006-2007 influenza season. The posterior probabilities were found by means of the application `FluDetWeb`.

Conesa et al. (2010) introduced a framework of models with the idea of using them on any kind of surveillance data. In particular, the process of the observed cases is modelled via a Bayesian hierarchical Poisson model in which the intensity parameter is a function of the incidence rate. Different options for modelling the mean of the rates are described, including the option of modelling the mean at each phase as autoregressive processes of order 0, 1 and 2 (David Conesa 2010, personal communication).

Lu et al. (2010) developed a Markov switching model with jumps to handle the effect caused

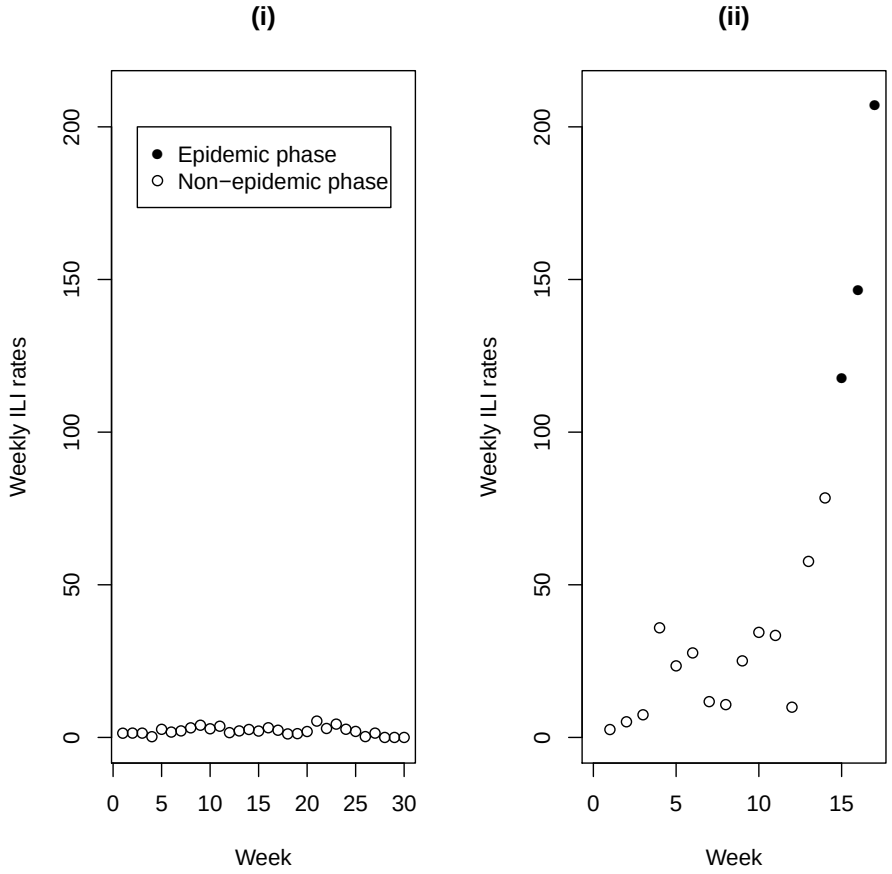

Figure 3: Comparison of weekly ILI incidence rates in the Communitat Valenciana, Spain, between the 2005-2006 (i) and 2006-2007 (ii) influenza season.

by past outbreaks. This model utilizes two additional hidden state variables in each period. The first hidden state variable models the disease outbreak state, and the second hidden state variable models the presence of extreme values. If an extreme value exists, the third hidden state variable represents the size of the extreme value. This is done to absorb any effect caused by sporadic extreme values in the training data.

In Cowling et al. (2006), dynamic linear models (West and Harrison 1997) were used as an approach to the early detection of the onset of the influenza epidemic period that requires only few weeks of baseline data.

For retrospective detection, Held et al. (2006b) proposed a two-component model where the counts are viewed as the sum of a 'parameter-driven' (endemic) and an 'observation-driven' (epidemic) component. The model for the parameter-driven component is a Poisson or negative binomial regression model. The observation-driven component is modeled with

an autoregressive parameter. Model estimates are obtained through Bayesian inference via MCMC techniques. The retrospective approach of Held et al. (2006b) can be used in a prospective setting.

The above models have clear similarities in that they are designed to monitor a system that can exist in two states, where the time of the change from one state to another is unknown. The models can be generalized to three or more states but, in terms of their utility for outbreak surveillance, the need for more than two states is questionable, as the main aim is to detect a change. The main differences among the models is in their methods of modelling trend, seasonality and autocorrelation. For most of these models the simplest method of estimating the parameters is through Bayesian MCMC. While this is feasible for one infection or syndrome, computation time may present difficulties for surveillance with many endpoints where the models require daily updating.

4 Methods inspired by statistical process control

The methods of statistical process control (e.g., Montgomery 2009; Oakland 2008) have a long history of application to problems in public health surveillance (Woodall 2006). Several proposed approaches for the on-line detection of outbreaks of infectious diseases are directly inspired by, or related to, methods of statistical process control. This is not surprising because the problem of detecting unusual clusters of diseases in epidemiological data prospectively is similar to that of detecting aberrances in industrial production processes as they arise. The main tools for tracking the characteristics of a production process over time are control charts. These are discussed in Subsection 4.1. In Subsection 4.2 and Subsection 4.3, further methods are considered which share a flavor of the statistical process control methodology, namely temporal scan statistics and methods based on the time to failure.

4.1 Control charts

The first control chart was proposed by Shewhart (1931) (cf. Section 2). The Shewhart chart only utilizes information about the last time point. Later, Page (1954) and Roberts (1959) derived control charts with memory, the cumulative sum (CUSUM) and exponentially weighted moving average (EWMA) control chart, respectively. To start with the former,

let $\{y_t, t = 1, 2, \dots\}$ denote the time series of the counts being monitored. Assuming that $y_t \sim \mathcal{N}(\mu_t, \sigma_t^2)$, the one-sided (standardized) Gaussian CUSUM at time t is defined iteratively by

$$C_t = \max\left\{0, C_{t-1} + \left(\frac{y_t - \mu_t}{\sigma_t} - k\right)\right\} \quad , \tag{11}$$

where $C_0 = 0$ and $k > 0$ is a constant that depends on the aberration size of interest. It is often chosen to be $1/2$ (Rogerson and Yamada 2004b). The baselines μ_t can be calculated from counts in comparable periods in previous years. These counts are also used to estimate the standard deviation σ_t. In the absence of any systematic departure from the expected values μ_t, (11) tends to remain at or close to zero. If $C_t > h$, where h is a specified threshold value, the process is declared to be out-of-control. Usually, the CUSUM is then reset to zero and the process starts again. There are many variants of this basic procedure, though. For example, one system restarts the process with the CUSUM set to half the alerting threshold, to increase sensitivity to early signals (Lucas and Crosier 2000). Methods based on the CUSUM formula are implemented in the Early Aberration Reporting System (EARS) of the CDC, which is used throughout the USA as a syndromic surveillance system (Hutwagner et al. 2003).

Figure 4 (ii) shows the CUSUM for *Salmonella enteritidis* in England, Wales and Northern Ireland in the year 2009. An outbreak from the 28th week onwards is detected by the CUSUM (11). The values in Figure 4 (i) are the weekly counts from the previous years 2000 to 2008 which were used to calculate μ_t and σ_t. The threshold h was chosen on the basis of a predetermined acceptable value for the in-control average run length ARL_0, i.e., the average time between alerts when there is no outbreak. The reciprocal of the ARL_0 is the false positive (or false discovery) rate, i.e., the proportion of apparently aberrant reports not associated with outbreaks. Tables that can be used to find the value of h that is associated with chosen values of ARL_0 and k are available (see for instance Rogerson 2001).

In the case of rare events, the CUSUM approach (11) is not adequate, since the counts do not have a normal distribution. One remedy is to use the Poisson CUSUM (Lucas 1985). Other methods used in disease surveillance to detect an increase in the mean of a Poisson distribution include, for example, the short memory scheme of Shore and Quade (1989), which is based on the distribution of cases in the current and previous periods. Kenett and

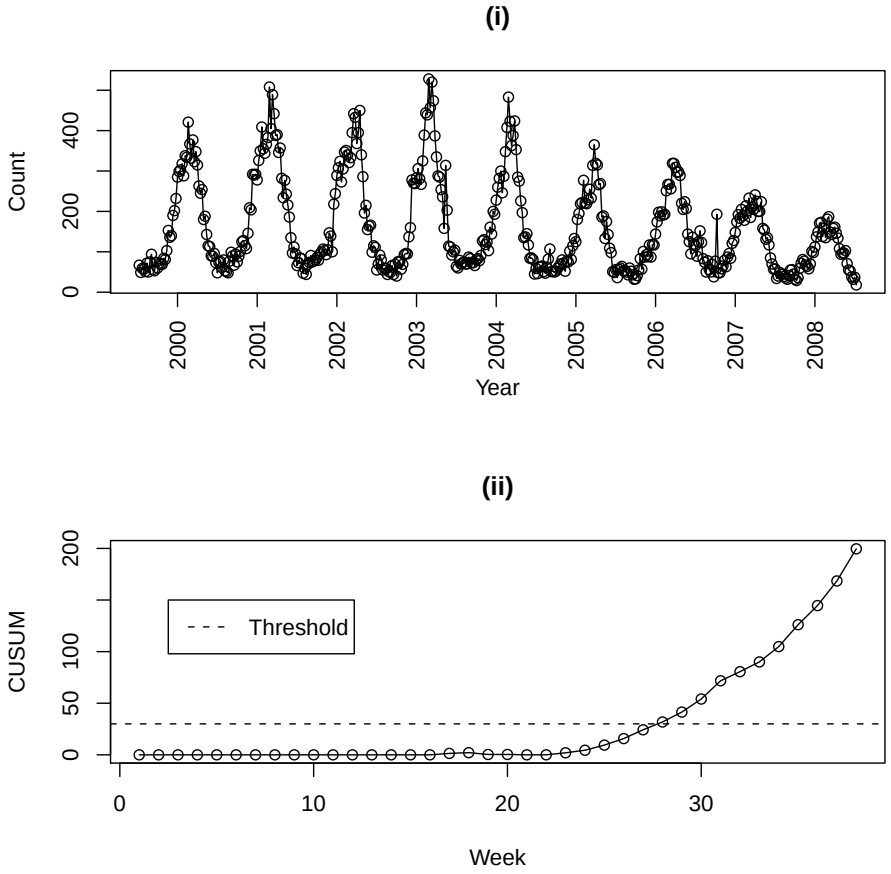

Figure 4: Weekly counts of *Salmonella enteritidis* PT 1, 4, 6, 6A, 8, 14B and 21 in England, Wales and Northern Ireland from 2000 to 2008 (i) and CUSUM of weekly counts in the year 2009 (ii) ($k = 0.5, h = 30$).

Pollak (1996) use the Shiryaev-Roberts statistic (Shiryaev 1963; Roberts 1966) and apply it to a non-homogeneous Poisson process. Whereas Gaussian or Poisson CUSUMs are designed to analyze counted data, binomial CUSUMs (e.g., Reynolds and Stoumbos 2000) can be used to monitor proportions.

Because CUSUMS are sensitive to small sustained changes in the mean numbers of reports, they are well suited to detecting relatively long-lasting epidemics, such as influenza. However, for the same reason, they are sensitive to small changes in reporting efficiency and other artefacts of the reporting process. Thus, they may lack robustness when used with surveillance data unless the baselines are frequently reset.

The EWMA control chart gives less weight to more historical data. The EWMA is defined

by the following recursive equation:

$$z_t = (1-\gamma)z_{t-1} + \gamma y_t \ , \tag{12}$$

where $z_0 = 0$ and the weight parameter $\gamma \in (0, 1]$. The weighting for each older data point decreases exponentially, giving much more importance to recent observations, while not discarding older observations entirely. For $\gamma = 1$, (12) is the same as the method by Shewhart (1931). The asymptotic (one-sided) variant of the EWMA chart will give an alarm at

$$t_A = \min\{t : z_t > L\sigma_z\} \ , \tag{13}$$

where $L > 0$ is a constant and σ_z is the asymptotic standard deviation of z_t (Sonesson 2003). Alternatively, one can use the exact standard deviation (which is increasing in time) instead of the asymptotic of the alarm limit (13). For the EWMA chart, Elbert and Burkom (2009) and Burkom et al. (2007) proposed the Holt-Winters technique for generalized exponential smoothing (cf. Subsection 3.1) to account for trends and seasonal features in syndromic data. Dong et al. (2008) constructed three types of EWMA methods that do not require an assumption of identical distributions of the counts to detect a positive shift in the incidence rate. Adaptions of the EWMA method for Poisson and binomial data are available (Borror et al. 1998; Gan 1991). Using the exponential smoothing technique and properties of numerical derivatives, Nobre and Stroup (1994) developed a method which bases monitoring on changes in the numerical gradient of the variable under surveillance with respect to time. Höhle and Paul (2008) present count data regression charts, which accommodate seasonal variation in the mean of the infectious disease counts. Assume that the observed counts originate from a negative binomial distribution parameterized by its mean μ and dispersion parameter θ. Note that for $\theta \to 0$ the Poisson distribution with mean μ is obtained. For the in-control situation $y_t \sim NegBin(\mu_{0,t}, \theta)$, where

$$\log(\mu_{0,t}) = \alpha + \beta t + c(t) \ , \tag{14}$$

and $c(t)$ is a cyclic function that may be modelled, for example, by trigonometric terms (Serfling 1963). That is, the in-control mean is assumed to be time-varying and linear on the log-scale. The out-of-control situation is characterized by a multiplicative shift $\mu_{1,t} = \mu_{0,t} \times$

$\exp(\kappa)$ with $\kappa \geq 0$, which corresponds to an additive increase of the mean of the log-scale. It is assumed that the in-control parameters are known, while κ is unknown and is estimated via maximum-likelihood. A generalized likelihood ratio statistic is computed to detect, on-line, whether a shift in the intercept occurred. Extensions of the basic seasonal count data regression chart are available that take account of autocorrelation between observations (Höhle and Paul 2008) or the population size of the age-strata (Höhle and Mazick 2010). Other modified CUSUM methods that allow for time-varying Poisson means were proposed by Rossi et al. (1999) and Rogerson and Yamada (2004a).

The use of control charts has also been widely advocated for the surveillance of health care-associated infections (Benneyan 1998a,b; Woodall 2006; Carey 2003; Limaye et al. 2008). In this context, CUSUM charts are more frequently useful than EWMA charts (Woodall 2006), but Shewhart charts appear to be the charts that have found greatest application, often being used to show the proportion of incidents in fixed periods of time. For example, they have been used in this way to monitor anesthesia-related adverse events (Fasting and Gisvold 2003) and risk-adjusted mortality rates of patients in hospital following admission for acute myocardial infarction (Coory et al. 2008). Morton et al. (2001) considered the application of Shewhart, CUSUM and EWMA charts for continuous real-time monitoring of various hospital acquired infections, such as vascular surgical site infection and *Klebsiella pneumoniae*. It was concluded that Shewhart and EWMA charts are together ideal for monitoring bacteraemia and multiresistant organism rates, while Shewhart and CUSUM charts together are suitable for surgical infection surveillance.

4.2 Temporal scan statistics

Scan statistics (e.g. Glaz et al. 2001) can be used to detect and evaluate clusters of disease cases in either a purely temporal, purely spatial or space-time setting (Woodall et al. 2008). In a temporal setting, this is usually done by gradually scanning a window across time, noting the number of observed and expected observations inside the interval. The scan statistic has long been used for retrospective detection of temporal clusters in epidemiology (Wallenstein 1980). Kulldorff (2001), Ismail et al. (2003) and Naus and Wallenstein (2006) adapted the scan statistic for the use in prospective temporal surveillance.

There are two general types of prospective temporal scan-based methods. One type involves counting the number of incidences in a single region in the most recent time period (or window) of a fixed length (Ismail et al. 2003; Naus and Wallenstein 2006). Let y_n denote the observation at the current time point n and let L be the fixed window size. The scan statistic can be viewed as an unweighted moving sum (Han et al. 2010; Joner et al. 2008):

$$S_n = \sum_{t=n-L+1}^{n} y_t \ . \tag{15}$$

An alert is flagged as soon as (15) exceeds a threshold h, that is, the first time that $S_n > h$, where h is typically chosen in conjunction with an acceptable value of ARL_0, although choosing h so that the type I error is a predetermined value α has also been suggested (Naus and Wallenstein 2006).

In the prospective temporal scan method of Kulldorff (2001), the length of the window is not a constant but varies over a range of values (see also Wallenstein and Naus 2004). Since the temporal scan statistic by Kulldorff (2001) can be viewed as a special case of his spatio-temporal procedure, a discussion of this method is deferred till Section 5. Public health surveillance data are often non-stationary with seasonal and other effects that are seldom found in industrial process control data. Wallenstein and Naus (2004) proposed a temporal scan method that can account for seasonal effects.

4.3 Methods based on inter-event times

To model rare events, instead of monitoring aggregate total cases in discrete time periods, Chen (1978) proposed the so-called *sets technique* which bases detection on the intervals between reports. Devised for monitoring congenital abnormalities, in that context the sets technique is based on the number of healthy newborns delivered in the interval between the birth of an infant with the specific malformation and the birth of the next infant with that malformation. The group of births in one interval is named a set, and its size is a geometrically distributed random variable.

The sets technique may be applied to time intervals as well (e.g., Chen et al. 1997). Assuming that events arise in a homogeneous Poisson process, inter-event times follow an exponential distribution with mean μ, say. The detection threshold is specified by parameters (n, τ),

and an aberrance is declared if the intervals between $n+1$ consecutive events are all less than τ. The false detection probability is $(1-\exp\{-\tau/\mu\})^n$, and the sensitivity of detection is $(1-\exp\{-\gamma\tau/\mu\})^n$ when the rate of events increases by a factor γ. The background rate, and hence μ, may be determined from historical data. Appropriate values of n and τ are selected to yield acceptable sensitivity and false detection probabilities. The analysis is performed whenever a new event occurs.

To reduce the expected frequency of false alarms obtained by the sets technique, Chen et al. (1993) suggested some confirmatory procedures. These declare an alarm as either confirmed or rejected according to data observed subsequent to the alarm. Reduction in the false detection probability is achieved at the expense of an increase in the delay in detecting a true alarm.

Note that methods which base detection on total reports will fail when events are very rare, because even a single report will then be unusual in a statistical sense. In such cases, one might either specify a minimum outbreak size that must be exceeded for the count to qualify as an aberration (Farrington et al. 1996), impose a lower bound on the standard error used to normalize residuals, or alternatively use the sets monitoring technique. Noteworthy modifications of the sets technique are available: an extension proposed by Sitter et al. (1990, 1991) and the cumulative score procedure (Munford 1980; Wolter 1987). Sego et al. (2008) proposed the Bernoulli CUSUM chart for the surveillance of rare health events instead.

Other techniques based on the time to failure have been proposed, such as time-between-event (exponential) CUSUM or EWMA schemes (e.g., Gan 1994, 1998). Exponential control charts arise naturally in the context of monitoring the occurrence rate of rare events, since inter-event times for a homogeneous Poisson process are exponentially distributed random variables. The exponential CUSUM will detect a significant increase in the rate of incidence (corresponding to a decrease in the mean time between events) earlier than the Poisson CUSUM because, unlike the latter, it can trigger an alarm before the end of the time period under consideration.

5 Methods incorporating spatial information

5.1 General

As well as giving a date, in almost any surveillance system the reported incidence of a disease will specify a location. Using the spatial information given by the location can potentially enable localised outbreaks of a disease to be detected, or variations in regional patterns to be identified. To use spatial data, surveillance methods must have some notion of the distance between observations or some spatial structure. Several methods require only a cut-off value that categorises pairs of observations as either being 'close' or 'not close' (e.g., Rogerson 2001; Kulldorff 2001). In others, an appropriate adjacency matrix or distance metric is defined. Suppose that there are m geographical units and let $\mathbf{S}$ be a $m \times m$ symmetric matrix of values s_{kl} representing the closeness of geographical units k and l, with $0 \leq s_{kl} \leq 1$. For example, Rogerson (1997) used the metric

$$s_{kl} = \exp(-d_{kl}/\tau) \ , \tag{16}$$

where d_{kl} is the geographic distance between reporting units k and l and τ is a specified constant. Tango (1995) suggested setting τ in (16) equal to five but results are fairly insensitive to its precise value (Bithell 1992). When the locations of observations are known individually, spatial structure has been imposed by using kernel estimation to fit a smooth surface that represents the intensity of reported cases (Diggle et al. 2005). Both fixed and variable kernel bandwidths have been used, but it seems sensible to have bandwidths that are narrower in large towns than in sparsely populated country areas. When a region is divided into sub-areas, spatial correlation can also be induced by defining a relationship between neighbouring areas. For example, if u_k denotes the "location effect" of area k, then a common choice for modelling spatial correlation is to assume that u_k follows the conditional autoregressive (CAR) model proposed by Besag et al. (1991). This model states that each u_k is normally distributed around the mean value of u amongst its immediate neighbours. That is, $u_k \,|\, \{u_l,\, l \neq k\} \sim \mathcal{N}(\overline{u}_k, \sigma^2)$, where

$$\overline{u}_k = \sum_{l=1}^{m} u_k w_{kl} / \sum_{l=1}^{m} w_{kl}$$

and

$$w_{kl} = \begin{cases} 1 & \text{if areas } k \text{ and } l \text{ are adjacent} \ , \\ 0 & \text{otherwise} \ . \end{cases}$$

In principle, it is suboptimal to ignore spatial information in disease surveillance. Indeed, it may seem suboptimal to adopt any surveillance system that does not model the space-time patterns of diseases as completely as possible when such information is available and reliable. However, the use of spatial information is computationally demanding, and it may not be practical for surveillance systems that monitor several hundred diseases and disease organisms – computer limitations will restrict the complexity of calculations that can be performed for each disease. Early approaches to the detection of localised disease outbreaks focused on only detecting outbreaks, enabling the methods to be comparatively simple. They include methods based on CUSUMs, and the scan statistic. These methods are described in Subsections 5.2 and 5.3, respectively. We then describe methods based on regression models in Subsection 5.4. These methods have the more ambitious aim of fitting space-time models to disease occurrences, typically using the Bayesian paradigm so that MCMC can be used as a tool to estimate model parameters. The resulting methods are computationally very intensive. However, there are examples where methods that require MCMC have been applied in the routine surveillance of some specific diseases (e.g., Diggle et al. 2005), though strategies to reduce computation were necessary.

5.2 Spatial CUSUM charts

CUSUM charts are a standard approach for detecting outbreaks and changes in spatial pattern. CUSUMs cumulate information from each case. In the spatial context, the information cumulated typically involves spatio-temporal proximity to other points. In order to isolate the specific contribution of the latest observation, conditioning on previous information is required. Thus the formula for the CUSUM given in equation (11) must be modified. The modified formula is

$$C_t = \max\left\{0,\, C_{t-1} + \left(\frac{y_t - \mathrm{E}(y_t|\nu_{t-1})}{\mathrm{var}(y_t|\nu_{t-1})^{1/2}} - k\right)\right\} \ , \tag{17}$$

where C_t is the CUSUM at time t, k is a specified constant, and $\mathrm{E}(y_t|\nu_{t-1})$ and $\mathrm{var}(y_t|\nu_{t-1})$ are the expectation and variance of y_t conditional on relevant information ν_{t-1} from the first

$t-1$ reporting periods, respectively. An alarm is flagged if the CUSUM exceeds a threshold. Statistics developed for the retrospective analysis of spatial disease data have been adapted to give the y_t. The Knox test (Knox 1964) aims to identify space-time interactions by categorising any two cases as close in space, close in time, close in both, or close in neither. The number of observations that are close in both space and time is referred to as the Knox statistic and a space-time interaction is indicated if it is unusually large. Rogerson (2001) adapted the Knox statistic for use in surveillance by equating its value after t cases to y_t in the CUSUM in equation (17). He showed that ν_{t-1} may be replaced by y_{t-1} (i.e. y_{t-1} contains all the relevant information in the first $t-1$ observations). He also gave formulas for the conditional means and variances needed to form the CUSUM. However, empirical testing by Marshall et al. (2007) suggests that these estimates of means and variances can be poor and that computer simulation should be used to estimate both them and the threshold at which an alarm is triggered.

Rogerson (1997) adapted Tango's statistic (Tango 1995) in a way similar to the method used with the Knox statistic. Tango's statistic for spatial clustering has the form

$$y = (\mathbf{r}-\mathbf{p})^{\top}\mathbf{S}(\mathbf{r}-\mathbf{p}) \ , \tag{18}$$

where $\mathbf{r}$ and $\mathbf{p}$ are $m \times 1$ vectors containing the observed and expected proportions of cases at each location, respectively, and $\mathbf{S}$ has the meaning assigned in Subsection 5.1. The statistic (18) was designed as a retrospective test of spatial clustering and is completely insensitive to a global change in the rate of disease occurrence. In part this is a disadvantage, but it also means that the statistic remains valid when there is seasonality and/or annual trend.

Raubertas (1989) gave a method that forms neighbourhoods in a way that allows a reporting unit to belong to more than one neighbourhood. The data within a neighbourhood are pooled in weighted averages, using a measure of closeness as weights. For each neighbourhood a CUSUM is formed that monitors incidence rate and an alarm is triggered if any CUSUM exceeds a threshold. This early method continues to influence surveillance methods in varied contexts (e.g., Sparks 2010).

ClusterSeer (Jacquez et al. 2002) and GeoSurveillance (Yamada et al. 2009) are software packages that implement surveillance methods based on spatial CUSUMs. The GeoSurveillance package was used to perform spatio-temporal surveillance of the weekly counts of all

Salmonella enteritidis cases in England, Wales and Northern Ireland in 2009. The geographical units are twelve regions defined by the HPA, namely "North East", "Yorkshire & Humberside", "East Midlands", "East", "London", "South East", "South West", "West Midlands", "North West", "Channel Islands & Isle of Man", "Wales" and "Northern Ireland". Figure 5 displays the results for four of these regions and gives evidence that spatial heterogeneity is inherent in the data (see also Figure 4 in which data aggregated over the twelve geographical areas are presented).

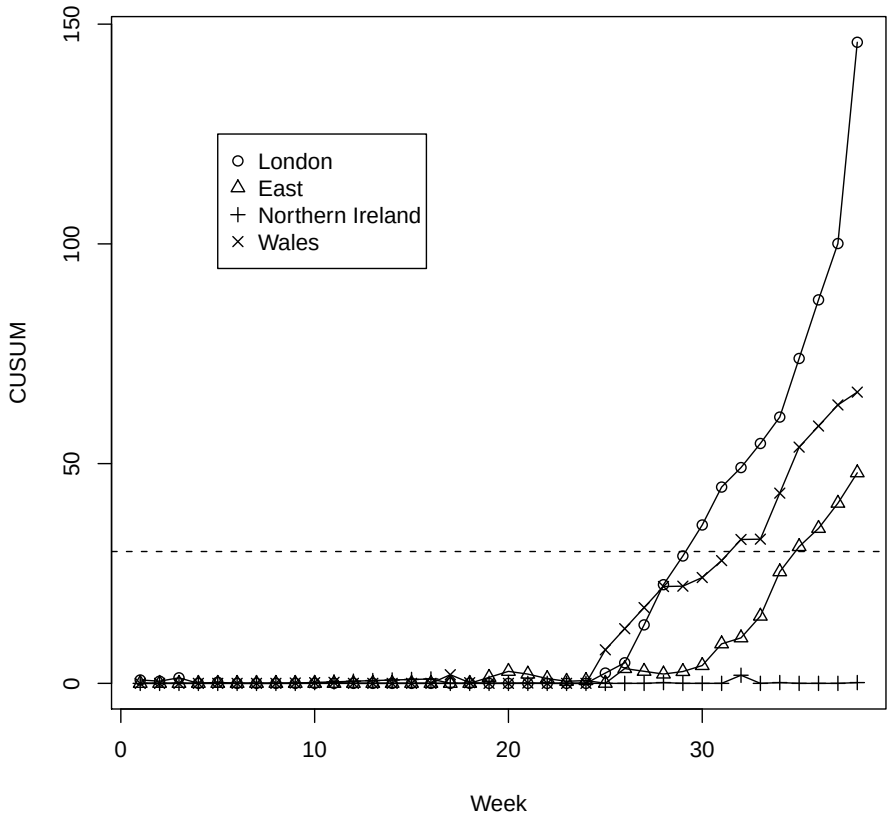

Figure 5: CUSUMs of weekly counts of *Salmonella enteritidis* PT 1, 4, 6, 6A, 8, 14B and 21 for four regions in England, Wales and Northern Ireland in the year 2009.

5.3 Space-time scan statistics

For a temporal scan, a search window is gradually moved across time, looking for a window in which the number of observed cases is unexpectedly high (cf. Subsection 4.2). Similarly,

for a spatial scan, a circular window is moved over a map of the study area, looking for a position where the circle contains an unexpectedly high number of cases. For a space-time scan the two are combined to form a cylindrical 'window' whose height is the time-dimension. This form of search window has been used for the retrospective detection of local disease outbreaks since at least Wallenstein et al. (1989). Kulldorff (2001) adapted the method for prospective surveillance.

The temporal component of the scan windows used by Kulldorff (2001) has a varying start time but reaches to the end of the current monitoring period, as the aim is to detect local disease outbreaks that are currently active. The spatial components of the windows have varying centres and radii. A search finds the scan window that has the most unexpectedly high number of cases, judged by the likelihood ratio

$$LR = \frac{L(z)}{L_0} \ , \tag{19}$$

where $L(z)$ is the maximum likelihood for cylinder z if the scan window has its own occurrence rate and L_0 is the maximum likelihood if the scan window has the underlying base rate as its occurrence rate. The likelihoods in (19) are generally based on the assumption that the number of cases follows a Poisson distribution. The maximum likelihood ratio is defined to be the space-time scan statistic and a p-value for its significance is obtained through Monte Carlo hypothesis testing.

Various methods aimed at improving the scan statistic have been suggested. Kleinman et al. (2005) focused on the baseline rates of disease incidence in the population-at-risk, which are used in calculating the scan statistic. They proposed one estimate of these rates that incorporates an adjustment for day-of-week, month and holidays, and a second estimate that includes an additional adjustment for local history of illness. Kulldorff et al. (2005) suggested a scan statistic that does not require the size of the population-at-risk to be estimated and only needs data on cases. The scan statistic is a log likelihood ratio, very commonly based on the Poisson assumption, and the p-value is an empirical probability of a statistic as large as that observed based on Monte Carlo repetitions. Its assumptions are similar to those made with the Knox statistic. Assunção and Correa (2009) adapt the Shiryaev-Roberts statistic (Shiryaev 1963; Roberts 1966) as a scan statistic, modelling a disease outbreak as a change point in a cylindrical scan window. They improve the computation speed of their method

through formula for updating estimates when a case occurs, rather than having to calculate estimates from scratch.

Sonesson (2007) fits the space-time scan statistics of Kulldorff (2001) and Kulldorff et al. (2005) into a CUSUM framework and examines properties of the resultant methods. Spatial scan statistics (without a temporal component) have also been used for disease surveillance by considering case incidence in a short fixed time period, such as the last seven days (Mostashari et al. 2003). In a different direction, using ellipses (rather than circles) for the spatial component of a space-time scan window has been suggested (Kulldorff et al. 2006). Also, non-parametric spatial scan windows to detect irregularly shaped clusters have also been explored (Assunção et al. 2006; Duczmal and Assunção 2004; Takahashi et al. 2008; Tango and Takahashi 2005). However using such windows significantly increases computer time.

The space-time scan statistic is currently the most widely used method for detecting the emergence of localised disease clusters (Shmueli and Burkom 2010). In one application it was used with a dataset on lower respiratory disease in Boston. The work showed the importance in some circumstances of modifying base rates to allow for seasonal effects and day-of-the-week effects – otherwise there appeared to be significant clusters on most weekdays between October and February (Kleinman et al. 2005). In another application the space-time scan statistic proved better at detecting emerging clusters of male thyroid incidence in New Mexico than a scan statistic that was purely spatial (Kulldorff 2001). Kulldorff et al. (2005) demonstrated the effectiveness of their method with data on diarrhea and influenza in New York City: four of the five clearest alarms given by their method were likely local precursors of city-wide outbreaks and the number of false alarms was modest. Space-time scan statistics have also been used with data on Burkitt's lymphoma (Assunção and Correa 2009), brain cancer (Kulldorff et al. 1998), and tularemia (Sonesson 2007).

SaTScan (Kulldorff 2010) and ClusterSeer (Jacquez et al. 2002) are software packages that implement scan statistic methods. SaTScan is used by various American federal, state and city agencies for retrospective and prospective cluster detection (Zhang and Lin 2009), including the New York City Department of Health, who use it for syndromic surveillance. SaTScan is versatile in its definition of a spatial location. To monitor infections disease outbreaks in hospitals, for example, a "spatial" location has consisted of individual wards and

services (such as medicine, oncology) or groups of wards or services sharing in patient care (such as cardiology and cardiac surgery services), regardless of physical proximity (Huang et al. 2010).

5.4 Spatio-temporal regression methods

A variety of spatio-temporal regression methods have been proposed. An important distinction between them is whether they analyse aggregated data or data at an individual level. Often the region of interest is broken into small areas and the aggregated number of cases in each area in each time period forms the response. This form of area-level data has usually been modelled as a discrete spatial model in which neighbourhood relationships between the areas are defined. The second form of data can arise with sparse data, when the location of each case is recorded at an individual level. These individual-level data have been modelled as a Cox point process (Cox 1955).

For an area-level model, let y_{kt} denote the number of cases in area k in time period t. Sometimes the model specifies that y_{kt} follows a binomial distribution (e.g., Diggle et al. 2004; Kleinman et al. 2004) but, more commonly, it is assumed that y_{kt} follows a Poisson distribution. Interest then centres on estimating the Poisson mean (Lawson et al. 2003; Vidal Rodeiro and Lawson 2006; Watkins et al. 2009; Zhou and Lawson 2008), which varies with k and t. Lawson and his co-workers generally assume that $\mathrm{E}(y_{kt}) = e_{kt}\theta_{kt}$, where θ_{kt} is the unknown true relative risk in area k and time period t and e_{kt} is the expected number of cases in the kth area in that period. The values of e_{kt} must be specified before the model is fitted or else there is an identifiability problem. The e_{kt} could be based on the "at risk" population demographics in each area, perhaps with an adjustment for seasonality or trend. Information about the e_{kt} might also be gleaned from monitoring a different disease that has a similar at-risk population structure to the disease of interest (Lawson 2005). The logarithm of the relative risk is decomposed into spatio-temporal components:

$$\log \theta_{kt} = u_k + v_k + \tau_t + \gamma_{kt} \ ,$$

where u_k represents spatially correlated extra variation, v_k represents uncorrelated extra variation, τ_t describes temporal variation, and γ_{kt} is the space-time interaction. It is generally assumed that τ_t and γ_{kt} follow random walks to allow a smooth variation in time

(Knorr-Held 2000): $\tau_t \sim \mathcal{N}(\tau_{t-1}, \sigma_\tau^2)$ and $\gamma_{kt} \sim \mathcal{N}(\gamma_{k,t-1}, \sigma\gamma^2)$. In general, the CAR model of Besag et al. (1991) is used to model the spatial correlation between the u_k.

The complexity of this model makes it virtually essential to use Bayesian methods for model fitting. Vague prior distributions are given to the model parameters and MCMC methods are used to sample from the posterior distribution and estimate parameters. To test for changes in spatial pattern, the estimates of parameters could be monitored (Lawson et al. 2003) but, much more commonly, observations are compared with one-step-ahead predictions (Kleinman et al. 2004). To detect gradual signals of interest, buffers between modeling intervals and test intervals can be used.

Other modelling stategies have also been proposed. Zhou and Lawson (2008) give a computationally cheap approach in which separate spatial models are fitted to the data from each collection period. These models are combined by forming exponentially weighted moving averages and a sharp change in the weighted average in any neighbourhood suggests an outbreak. Another computationally quick method is given by Kleinman et al. (2004), who fit generalised linear mixed models that include time-of-day and seasonality components, but which only allow uncorrelated heterogeneity between areas. Several researchers have used a model in which the disease process can switch between two (unobserved) states: endemic (non-outbreak) and epidemic (outbreak). To model the process HMMs are used and an alarm is flagged on the basis of a Bayes factor that reflects the relative likelihoods of each state (Lawson et al. 2003; Madigan 2005; Watkins et al. 2009).

Modelling individual-level data has attracted far less attention. Clark and Lawson (2006) proposed a novel approach using non-parametric regression via kernel smoothers. However the major work with individual-level data was in the *AEGISS* (Ascertainment and Enhancement of Gastrointestinal Infection Surveillance and Statistics) project, reported in Diggle et al. (2004) and Diggle et al. (2005). The aim of this project was to develop a monitoring tool that could identify anomalies in the space-time distribution of non-specific gastrointestinal infections. The data it collected was the location, x, and date, t, of each individual case. As the point process model for these data, Diggle et al. (2005) used a non-stationary log-Gaussian Cox process in which the spatio-temporal intensity, $\lambda(x,t)$ was decomposed as

$$\lambda(x,t) = \lambda_0(x)\mu_0(t)R(x,t) \ , \tag{20}$$

where $\lambda_0(x)$ is a smoothly varying surface describing the normal disease pattern that was estimated using kernel smoothing, $\mu_0(t)$ is temporal variation, modelled parametrically to reflect day-of-week and season effects, and $R(x,t)$ is the residual space-time variation.
Both $\lambda_0(x)$ and $\mu_0(t)$ in (20) could be estimated from historical data but up-to-date predictions of $R(x,t)$ were required. These predictions were based on the most recent five days data. Naturally they had far more uncertainty attached to them than the estimates of $\lambda_0(x)$ and $\mu_0(t)$, so these latter estimates were treated as deterministic quantities. To detect outbreaks, the region was divided into neighbourhoods and MCMC was used to determine the probability in each neighbourhood that $R(x,t)$ exceeded a prespecified threshold. The results were reported daily on maps with colours denoting probability levels.
The complexity of models places a burden on computer resources and the problem is exacerbated if MCMC is used to implement Bayesian methods. As noted by Lawson et al. (2004, p. 952), *. . . for any Bayesian model, computational speed-ups must be sought to make implementation realistic in a surveillance context.* In line with this, it seems sensible to follow Diggle et al. (2005) and treat any quantity that is estimated from historical data as deterministic. Also, historical data can generally be restricted to a moving window of the last 3, 5 or 8 years (Lawson et al. 2003) unless the data are very sparse. Similarly, a window of just the most recent past (perhaps only a few days) has generally been used when much more weight should be given to recent data (e.g., Diggle et al. 2005). Numerical approximations have also been used to good effect. For example, Kleinman et al. (2004) estimated parameters through quasi-likelihood, which is computationally less demanding but which introduces some bias. In many situations these pragmatic measures would simply be an option, but in the context of real-time surveillance they are a necessity.

6 Multivariate outbreak detection

6.1 Scope of multivariate detection

Most outbreak detection systems track more than one data series. For example, in the UK, systems of laboratory surveillance (Farrington et al. 1996; McCabe et al. 2003), syndromic surveillance (Baker et al. 2003; Robertson 2006), and systems for institutional surveillance

(Marshall et al. 2004) may typically monitor dozens or even hundreds of different data series. When the different data series (and, most importantly, outbreaks within them) are likely to be unrelated, it usually makes most sense to consider them as separate univariate series. However, in some cases, several series will relate to the same underlying process, and hence process changes are likely to be strongly correlated. This applies, for example, to indicators of influenza (Stroup et al. 1988; Griffin et al. 2009; Mann 2009), reports of gastrointestinal illness from different sources (Kulldorff et al. 2007), or time series of counts of the same infection in different age groups (Held et al. 2005). In such circumstances multivariate methods of outbreak detection are likely to be fruitful in exploiting dependencies, both between the underlying processes and between the timing of outbreaks.

An overview of methods for multivariate surveillance is given in Sonesson and Frisén (2005) and Frisén (2010), who classified genuinely multivariate approaches into categories that include reduction of dimensionality, joint modelling, and vector accumulation methods. These are discussed in turn in Subsection 6.1, Subsection 6.2 and Subsection 6.3, respectively. Sonesson and Frisén (2005) and Frisén (2010) also mention so-called 'parallel surveillance' methods. A parallel approach monitors each variable separately by means of a univariate surveillance method. An alarm for the multivariate process is declared if some condition is fulfilled, e.g. the first time that any of the univariate processes gives an alarm. These methods will not be considered further here.

6.2 Dimension reduction

Dimension reduction methods for multivariate surveillance data could in principle include standard tools such as principal component analysis (Jolliffe 2002). These have been used for detecting aberrations in other fields (Ku et al. 1995), though they may lead to problems of interpretation. A more popular approach is to reduce the multivariate data at each time point to a scalar, which is then monitored by univariate surveillance methods. Let $\mathbf{y} = \{\mathbf{y}_t, t = 1, 2, \dots\}$ be the multivariate process under surveillance, where $\mathbf{y}_t = (y_{1t}, y_{2t}, \dots, y_{pt})^\top$ is observed with in-control mean $p \times 1$ vector $\boldsymbol{\mu}$ and $p \times p$ covariance matrix $\boldsymbol{\Sigma}$. An early multivariate surveillance scheme is that based on Hotelling's T^2 (Hotelling 1947;

Jackson 1959). The process parameter at time t for multivariate data $\mathbf{y}_t$ is

$$T^2(t) = (\mathbf{y}_t - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{y}_t - \boldsymbol{\mu}) \ . \tag{21}$$

The statistic (21) is a multivariate extension of the Shewhart chart, an alarm being declared if $t_A = \min\{t : T^2(t) > h\}$, where h is a specified threshold. For a bivariate process (x_t, y_t) with changepoint (t_x, t_x), Andersson (2009b) shows that the conditional expected delay, defined as $t_A - t_{(1)}$ given that $t_A > t_{(1)}$, where $t_{(1)} = \min\{t_x, t_y\}$, depends only on $|t_x - t_y|$. A further possibility is to undertake univariate analyses on each data set, and combine the p-values for the marginal tests into a single 'consensus' value. One such method uses Fisher's rule to obtain the summary statistic F from n individual p-value p_i:

$$F = 2 \sum_i^n \ln(p_i) \ .$$

If the n tests are independent then the null distribution of F is χ^2 with $2n$ degrees of freedom. This and other methods for combining p-values, along with Hotelling's T^2, are discussed in an outbreak detection setting by Burkom et al. (2005).

6.3 Joint modelling methods

Kulldorff et al. (2007) developed a multivariate space-time scan statistic based on the sum of the log likelihood ratio statistics for the univariate processes. This generalizes an earlier univariate version (Kulldorff 1997). Thus, suppose that the total number of reports for series j is N_j $(j = 1, \ldots, p)$. For a space-time cylinder z with $n_{j,z}$ cases from series j and expected cases $\mu_{j,z}$ obtained under a Poisson model, the likelihood ratio for a 'high' cluster is

$$LR_j(z) = \left(\frac{n_{j,z}}{\mu_{j,z}}\right)^{n_{j,z}} \left(\frac{N_j - n_{j,z}}{N_j - \mu_{j,z}}\right)^{N_j - n_{j,z}} \tag{22}$$

if $n_{j,z} > \mu_{j,z}$ and 1 otherwise. The multivariate scan statistic for detecting 'highs' is then

$$T = \max_z \sum_j \log LR_j(z) \ , \tag{23}$$

which adjusts automatically for multiple testing inherent in considering multiple series as well as multiple cylinders. Note that equation (22) differs very slightly from Kulldorff's formulation: his $LR_j(z)$ is multiplied by an indicator function, but this will cause problems

when taking logarithms.

To illustrate this approach, the multivariate space-time scan statistic (23) was applied to syndromic surveillance data in Scotland. The data give the number of calls to NHS24 by postcode district and day from 19 February to 1 April 2007 originated within the Glasgow postcode area (see Figure 6). This area consists of 50 postcode districts. Data are

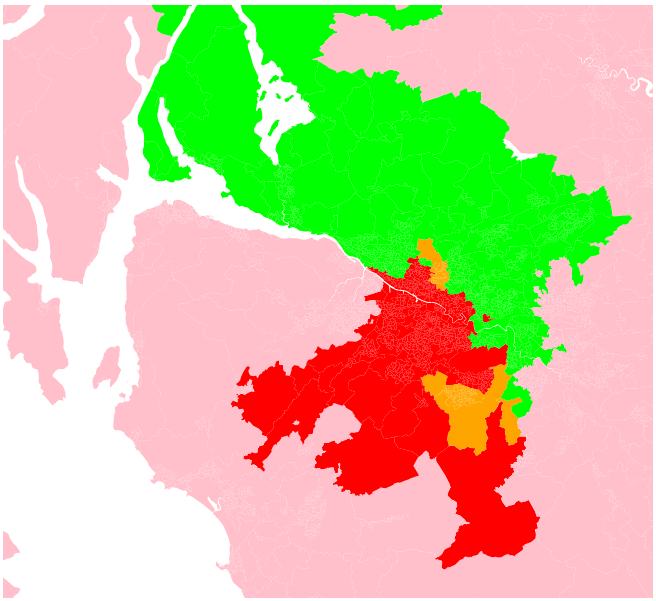

Figure 6: Map of the 50 districts of the Glasgow postcode area (green, red and orange) in the West of Scotland (pink). The two clusters are indicated in red (1 April) and red+orange (26 March). Green are the Glasgow postcode districts that are not in the clusters.

recorded for the two symptoms diarrhoea and vomiting which are syndromic indicators for norovirus infection. A space-time scan statistic analysis was performed for each of the 7 days from 26 March to 1 April, using the data from 19 February 2007. We used 3 days as the maximum temporal window size. Calculations were performed using the SaTScan software (Kulldorff 2010). The results of the analyses are presented in Table 1. The strongest signal was on 1 April, with the cluster consisting of 2 days and 25 postcode districts. With a recurrence interval of 2 years and 269 days, this cluster is unlikely to be a chance occurrence. A second cluster with an identical recurrence interval was detected on March 26, containing 27 postcode districts during 3 days.

Table 1: The two most likely clusters for norovirus in the Glasgow postcode area, during 26 March to 1 April 2007, as generated by the multivariate space-time scan statistic (Obs. (observed), Exp. (expected), RR (relative risk), RI (recurrence interval)).

Cluster characteristics			Diarrhoea			Vomiting			
Date	#postc.	#days	Obs.	Exp.	RR	Obs.	Exp.	RR	RI
26 March	27	3	49	49.29	0.99	192	123.84	1.59	2 years 269 days
1 April	25	2	47	28.28	1.69	164	76.02	2.22	2 years 269 days

Other joint modelling approaches have been suggested, based on a joint model for the entire multivariate data series. In one such method, the alarm function is based on the likelihood ratio derived from the joint distribution of the multivariate process (Andersson 2009a). This is a multivariate version of the method of Frisén and de Maré (1991). Schiöler and Frisén (2010) present a multivariate extension of the semiparametric univariate method of Frisén and Andersson (2009), based on the sufficient reduction approach of Frisén et al. (2010b) for step changes.

Several other joint modelling methods have been used for infectious disease data, but have not been applied to outbreak detection *per se*, though extensions in this direction should be possible. Held et al. (2005) extended to the multivariate case a model incorporating a branching process presented in Held et al. (2006b). This model was further extended in Paul et al. (2008) to analyse data from several pathogens. A multivariate spatial model for different gastrointestinal infections was presented in Held et al. (2006a).

Sebastiani et al. (2006) used dynamic Bayesian networks to study the interplay of four different data sources that are monitored for influenza surveillance. Mann (2009) developed a multivariate HMM with a shared hidden process to model several markers of influenza.

6.4 Vector accumulation methods

Vector accumulation methods include multivariate extensions of the EWMA and CUSUM charts, referred to as MEWMA and MCUSUM, respectively. Generalizing the EWMA is relatively straightforward (Lowry et al. 1992), via the recursive scheme

$$\mathbf{z}_t = \mathbf{\Lambda}(\mathbf{z}_t - \boldsymbol{\mu}) + (\mathbf{I}_p - \mathbf{\Lambda})\mathbf{z}_{t-1} \ , \tag{24}$$

where $\mathbf{\Lambda}$ is a diagonal $p \times p$ matrix of values in $(0, 1]$ and $\mathbf{I}_p$ is an identity matrix of order p. The chart from (24) goes out of control when

$$T_t^2 = \mathbf{z}_t^\top \mathbf{\Sigma}_{\mathbf{Z}_t}^{-1} \mathbf{z}_t > h$$

for some control limit h, where $\mathbf{\Sigma}_{\mathbf{Z}_t}$ is the covariance matrix of $\mathbf{z}_t$. In contrast, generalising the standard univariate two-sided CUSUM is complicated by the fact that there are two cumulative sums for each variable. Crosier (1988) and Pignatiello and Runger (1990) developed multivariate CUSUMS as generalisations of new univariate CUSUMS requiring a single cumulative sum per variable. Golosnoy et al. (2009) investigated the properties of the MCUSUM chart of Pignatiello and Runger (1990) and suggested further enhancements. Ngai and Zhang (2001) generalized the standard univariate CUSUM in the sense that the p-dimensional version reduces to it when $p = 1$. These various MCUSUMS all apply only to independent multivariate observations. Bodnar and Schmid (2007) further extended these methods to multivariate time series, taking into account dependencies in the underlying process. Moving beyond Gaussian processes, an example of a rank-based MCUSUM can be found in Qiu and Hawkins (2001), with further development of this non-parametric scheme provided by Qiu and Hawkins (2003).

7 Comparison and evaluation of prospective outbreak detection systems

As documented in earlier Sections, a rich array of methodologies for outbreak detection is available. This raises the question: which to use? It is not feasible to make detailed recommendations as to which method is 'best', because this will depend critically on the specific details of the application and implementation, as well as its purpose and context. In particular, the key issues that are likely to affect any assessment of the relative merits of different methods are (a) the scope of the system, in particular how many parallel data series are to be monitored, which can range from one to several thousand; (b) the quality of the data available, including the method of data collection, and the delay between event occurrence and reporting; (c) the spatio-temporal features of the data, such as count frequency, trend structure, seasonality, epidemicity, time step and spatial resolution; (d) the features of the outbreaks that may occur, for example explosive or gradual onset, brief or long duration, and

level of severity, or a mix of these; (e) the use to which the system is to be put, including the post-signal processing protocols; (f) the availability of processing power and human resources to support the system; (g) the choice of metric to evaluate results.

Accordingly, we venture no recommendations as to the strengths and weaknesses of the different methods. Instead, and in the spirit of this review, we focus on some of the statistical issues involved in evaluating and comparing methods. This remains to some extent an undeveloped area, in spite of much discussion and some progress; see Buckeridge et al. (2005) and Fraker et al. (2008) for overviews. It is influenced by two distinct perspectives: technical approaches derived from statistical process control, and the more empirical perspective of epidemiology.

7.1 Optimality properties

In the statistical process control literature, performance of a detection system is usually assessed by criteria based on time to event, such as the average run length (ARL). The ARL to detect an ongoing outbreak that started at the same time as the surveillance is ARL_1. The criterion of minimal ARL_1, given a fixed ARL_0, is sometimes used in public health settings (Grigg et al. 2003; Musonda et al. 2008), but the usefulness of this criterion for detecting outbreaks has been questioned by Frisén (1992, 2003), owing in part to the restriction that the measure only relates to outbreaks starting concurrently with the surveillance.

Frisén (1992) considers other criteria, such as the probability of a false alarm before some specified time t after the start of surveillance, the probability of detection of an outbreak starting at time t before time $t+d$ for some d, and the positive predictive value of an alarm, namely the probability that an outbreak is occurring, given that the system has signalled one. These and related measures are discussed in Sonesson and Bock (2003). Recently, measures of performance based on other temporal criteria have been proposed (Fraker et al. 2008). One such is the recurrence interval, namely the time interval over which the expected number of false alarms is 1, when the process is in control.

It is sometimes possible to determine which statistical detection method is optimal under a given criterion. The main theoretical results on optimality properties are those of Frisén and de Maré (1991) and Frisén (1992, 2003). Consider a series of independent random variables x_s at times s with mean μ_0 at $s < T$ and mean $\mu_1 > \mu_0$ at $s \geq T$, where T is the

(unobserved) start of the outbreak. An alarm function p is defined based on suitable null and alternative hypotheses, defining a test with rejection region of the form $p(x_s) > k$. This test maximizes the probability that an alarm is detected, among all tests with a specified false alarm probability. Optimality in this sense derives from a version of the Neyman-Pearson Lemma, and indeed the test may be expressed as a likelihood ratio statistic. Thus

$$p(x_s) = \sum_{t=1}^{s} w(s,t)L(s,t) > k_s$$

where

$$w(s,t) = \frac{\mathrm{P}(T=t)}{\mathrm{P}(T \leq s)} \qquad \text{and} \qquad L(s,t) = \frac{f_{x_s}(x_s|T=t)}{f_{x_s}(x_s|T>t)}$$

where f_x is the density of x and k_s is the critical value of the test at time s.

Within this framework, Frisén (2003, 2007) shows that the Shewhart chart method, for which $p(x_s) = L(s,s)$, maximises the probability of detecting the outbreak at time $T = s$, and, asymptotically (as μ_1 increases), minimizes time to detection, for a fixed false detection probability. In the Shiryaev-Roberts method (Shiryaev 1963; Roberts 1966), dependence on T through $w(s,t)$ is eliminated by setting $w(s,t) = 1$, which may be justified asymptotically as corresponding to vanishingly low probability of an outbreak. More generally, however, the performance of a method will depend on the time T at which the outbreak occurs after surveillance has begun. Minimax optimality criteria minimize the expected delay in detection, for the worst possible choice of T. Some CUSUM methods, for which $p(x_s) = \max_t L(s,t)$, are minimax-optimal in this sense (Frisén 2003).

However, the usefulness of such theoretical optimality properties is limited by the restrictive assumptions under which they are derived. In practice, surveillance data are noisy, often nonstationary or autocorrelated, outbreaks are of varying durations and intensities but seldom indefinite, the distributions of the observed counts are not generally Poisson. Arguably, the notion of the process being 'in control' or in a 'steady state' bears little relation to reality in the context of surveillance. While Frisén (2003) discusses optimality properties in the presence of such complexities, a sufficiently robust general theory remains elusive. In addition, it is often advisable to consider several performance measures concurrently. For these reasons, more empirical criteria are usually employed to compare methods.

7.2 The epidemiological perspective

The epidemiological perspective derives from the evaluation of diagnostic tests in clinical medicine, and is based on concepts such as sensitivity, specificity, and predictive value, applied to surveillance time units rather than individual patients. An extensive discussion of these types of measures, relating largely to the era before computerized surveillance, but of lasting relevance, may be found in Thacker et al. (1988), who identify the following criteria: usefulness, cost, sensitivity, specificity, representativeness, timeliness, simplicity, flexibility and acceptability. These measures relate to the performance of a surveillance system as a whole, including the process of data collection and the wider public health impact of the system, rather than the evaluation of different statistical algorithms.

To evaluate statistical algorithms, it is customary to use a combination of numerical indices, most commonly sensitivity, specificity (or positive predictive value), and timeliness, namely the delay between the start of the outbreak and its detection. However, none of these quantities is straightforward to standardize or operationalize in the context of outbreak detection, owing to the contextual factors set out at the beginning of this section; their big advantage is that they relate directly to the preoccupations of system users.

Several approaches have been taken to combine these indices so as to obtain a single performance measure which can be compared meaningfully across detection systems. The standard method is to calculate a ROC (receiver operating characteristic) curve, and use the AUC (area under the curve) as a summary measure. ROC curves are obtained by plotting sensitivity against 1 - specificity, and thus ignore the key timeliness variable. Buckeridge et al. (2005) discuss a variant, the AMOC curve, obtained by plotting a measure of timeliness (such as the time to outbreak detection on an inverted scale) against 1 - specificity. Zhang et al. (2003) used both ROC and AMOC curves to compare performance of several outbreak detection systems.

Kleinman and Abrams (2006) have proposed a weighted ROC curve, or WROC curve, in which each point of the curve is weighted by a timeliness measure, and a weighted AUC is derived which thus takes timeliness into account. They suggest two ways of doing this. In the first, it is assumed that there is a reference time after the outbreak starts by which it must be detected, and measure the proportion P_{sav} of time saved relative to this reference

time. Thus, suppose that an outbreak starts at time t, and that the pre-determined reference time is $t + D$. If the outbreak is detected at time $s \geq t$, or not detected (in which case $s = \infty$), then

$$P_{sav} = 1 - s/D \qquad \text{if} \qquad s \leq D, \qquad P_{sav} = 0 \qquad \text{otherwise.}$$

Martínez-Beneito et al. (2008) used this approach to evaluate different systems for detecting influenza epidemics. In the second method of Kleinman and Abrams (2006), timeliness is a 0-1 variable according to whether the outbreak was detected within a specified time or not, and is therefore similar to the successful detection probability of Frisén (1992). Under both schemes, the weighted AUC equals 1 if the system is fully accurate and timely. Kleinman and Abrams (2006) also propose several 3-dimensional generalizations of the ROC curve, the third dimension being a timeliness measure. The AUC is replaced by the volume under the curve, VUC. Cowling et al. (2006) used the VUC approach in comparing different outbreak detection systems for influenza.

Further complications arise when the system is intended to work with more than one data series. In order to control the false discovery rate, namely the proportion of signals that do not correspond to outbreaks, either a Bonferroni correction may be applied, or the methods of Benjamini and Hochberg (1995) may be used. Marshall et al. (2004) discuss these issues in the context of hospital performance monitoring. In contrast, Farrington et al. (1996) use an empirical approach in which, each week, the number of flagged organisms (out of many hundred reported) is limited to some manageable ranked number with the smallest p-values (Farrington 2004), since the true timeliness of outbreak detection depends also on post-signal confirmatory procedures. Grigg and Spiegelhalter (2008) show how the methods of Benjamini and Hochberg (1995) can be applied to p-values derived from CUSUMs. For spatial surveillance, which may also involves monitoring several data series, a further elaboration of the ROC curve has been proposed, in which the fraction of outbreak locations detected is plotted against 1 - specificity Buckeridge et al. (2005).

Evaluation methods for large surveillance systems is an area requiring further work; it is unlikely that a single plot or index will suffice. For example, evaluations of systems applied to multiple organisms should ideally allow for differences in public health importance of different outbreaks (as determined by their size, severity, and cost). Frisén (2003) discusses

utilities, but in the context of delays in detection rather than outbreak severity. This is a topic requiring further development, in particular to incorporate prior knowledge about the likely severity of an outbreak for a particular organism and age group combination, and derive suitable importance weights.

7.3 Study designs

Most of the empirical investigations of outbreak detection algorithms have been done using real data, wholly simulated data, or real data with superimposed simulated outbreaks (Buckeridge et al. 2005). Evaluations based on real data suffer from the obvious problem that the true outbreaks are generally not known. Kleinman and Abrams (2006) proposed to evaluate a new surveillance system by comparing the outbreak signals it generates with outbreaks in real data identified by an existing system of recognized reliability. They suggest three tests, including a permutation test in which outbreak locations and dates are randomly reassigned, to assess whether the new system works in the sense that it detects outbreaks better than chance, or that the signals are generated earlier than by the existing system.
An appealing approach is to inject simulated outbreaks into real data series. This combines the realism of idiosyncratically noisy surveillance data, with knowledge of the outbreaks, the epidemic curves and durations of which are under the investigator's control. This is the approach taken by Neill (2009). Others have used fully simulated data, with superimposed outbreaks (Hutwagner et al. 2005a,b).
Relatively few evaluations of automated statistical surveillance methods against traditional approaches have been undertaken: Leal and Laupland (2008) have attempted a systematic review, while Huang et al. (2010) carried out a retrospective cohort study. Kleinman and Abrams (2006) have commented that, in view of the multiplicity of evaluation measures now available, an evaluation metric is now required for evaluating evaluation metrics.

8 Final remarks

In this paper we have sought to review the methods that have been proposed for identifying outbreaks of infectious diseases as they arise, in sufficiently timely fashion to allow interventions to take place. In limiting ourselves to these statistical methods, we chose not to deal

with several issues of critical importance, which require further research and involvement of statisticians. One key issue is how to design effective and flexible user interfaces - ideally including a choice of statistical algorithms. Another is how to create effective protocols to handle the signals that emerge from a statistical surveillance system. A third is how to deal with the data imperfections, such as reporting delays, which inevitably affect prospective detection systems.

This review shows that a very broad range of statistical techniques have been proposed for prospective outbreak detection. They range from the very simple to the very complex, involve both testing and modelling approaches, and increasingly make use of modern simulation-based inference, often in a Bayesian framework. The choice of which statistical technique to use will depend critically on the nature of the intended application. In particular, systems designed for the surveillance of a single infection or syndrome should arguably be tuned to the specific features of that infection and may need frequent user intervention. In contrast, systems designed for routine application to hundreds or thousands of possible infections, with diverse frequencies and temporal patterns, will require robustness and automation. Systems will also vary according to what features they are designed to detect, for example trends and seasonal variation may or may not be of interest according to context. No single statistical technique is likely to be ideal in every setting.

Statistical outbreak detection is a multidisciplinary science, involving epidemiologists and computer programmers, as well as statisticians. Its rapid development over recent years presents an opportunity for statisticians, through the numerous techniques now available, and the increased acceptance of statistical methodology in this field. But it also presents a challenge to the statisticians: to demonstrate that these new statistical methodologies provide added value in public health terms, by effectively supplementing other surveillance methods for detecting outbreaks.

Acknowledgements

The editor, two anonymous reviewers and Bill Woodall made valuable comments and suggestions on the first draft of this paper. We are grateful to David Conesa, Marianne Frisén, Michael Höhle and Ron Kenett for making us aware of some references that we would oth-

erwise have missed.

References

Andersson, E. (2009a), "Effect of dependency in systems for multivariate surveillance," *Communications in Statistics - Simulation and Computation*, 38, 454–472.

— (2009b), "Hotelling's T^2 method in multivariate on-line surveillance: on the delay of an alarm," *Communications in Statistics - Theory and Methods*, 38, 2625–2633.

Andersson, E., Bock, D., and Frisén, M. (2008), "Modeling influenza incidence for the purpose of on-line monitoring," *Statistical Methods in Medical Research*, 17, 421–438.

Assunção, R. and Correa, T. (2009), "Surveillance to detect emerging space-time clusters," *Computational Statistics & Data Analysis*, 53, 2817–2830.

Assunção, R., Costa, M., Tavares, A., and Ferreira, S. (2006), "Fast detection of arbitrarily shaped disease clusters," *Statistics in Medicine*, 25, 723–742.

Baker, M., Smith, G. E., Cooper, D., Verlander, N. Q., Chinemana, F., Cotterill, S., Hollyoak, V., and Griffiths, R. (2003), "Early warning and NHS Direct: a role in community surveillance?" *Journal of Public Health Medicine*, 25, 362–368.

Benjamini, Y. and Hochberg, Y. (1995), "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society, Series B*, 57, 289–300.

Benneyan, J. C. (1998a), "Statistical quality control methods in infection control and hospital epidemiology, Part I: Introduction and basic theory," *Infection Control and Hospital Epidemiology*, 19, 194–214.

— (1998b), "Statistical quality control methods in infection control and hospital epidemiology, Part II: Chart use, statistical properties, and research issues," *Infection Control and Hospital Epidemiology*, 19, 265–283.

Besag, J., York, J., and Mollie, A. (1991), "A Bayesian image restoration, with two applications in spatial statistics (with discussion)," *Annals of the Institute of Statistical Mathematics*, 43, 1–59.

Bithell, J. F. (1992), "Statistical methods for analysing point-source exposures," in *Geographic and Environmental Epidemiology: Methods for Small Area Studies*, eds. Elliot, P., Cuzick, D., English, D., and Stern, R., Oxford University Press: Oxford, pp. 1–35.

Bock, D., Andersson, E., and Frisén, M. (2008), "Statistical surveillance of epidemics: peak detection of influenza in Sweden," *Biometrical Journal*, 50, 71–85.

Bodnar, O. and Schmid, W. (2007), "Surveillance of the mean behavior of multivariate time series," *Statistica Neerlandica*, 61, 383–406.

Borror, C. M., Champ, C. W., and Rigdon, S. E. (1998), "Poisson EWMA control charts," *Journal of Quality Technology*, 30, 352–361.

Box, G. E. P. and Jenkins, G. M. (1970), *Time Series Analysis: Forecasting and Control*, Holden-Day: San Francisco.

Buckeridge, D. L., Burkom, H. S., Campell, M., Hogan, W. R., and Moore, A. (2005), "Algorithms for rapid outbreak detection: a research synthesis," *Journal of Biomedical Informatics*, 38, 99–113.

Burkom, H. S., Murphy, S., Coberly, J., and Hurt-Mullen, K. (2005), "Public health monitoring tools for multiple data streams," *Morbidity and Mortality Weekly Report (Supplement)*, 54, 55–62.

Burkom, H. S., Murphy, S. P., and Shmueli, G. (2007), "Automated time series forecasting for biosurveillance," *Statistics in Medicine*, 26, 4202–4218.

Cappé, O., Moulines, E., and Rydén, T. (2005), *Inference in Hidden Markov Models*, Springer: New York.

Cardinal, M., Roy, R., and Lambert, J. (1999), "On the application of integer-valued time series models for the analysis of disease incidence," *Statistics in Medicine*, 18, 2025–2039.

Carey, R. G. (2003), *Improving Healthcare with Control Charts: Basic and Advances SWPC Methods and Case Studies*, ASQ Quality Press: Milwaukee, WI.

Chatfield, C. and Yar, M. (1988), "Holt-Winters forecasting: some practical issues," *The Statistician*, 37, 129–140.

Chen, R. (1978), "A surveillance system for congenital malformations," *Journal of the American Statistical Association*, 73, 323–327.

Chen, R., Connelly, R. R., and Mantel, N. (1993), "Analysing post-alarm data in a monitoring system in order to accept or reject the alert," *Statistics in Medicine*, 12, 1807–1812.

Chen, R., Iscovich, J., and Goldbourt, U. (1997), "Clustering of leukaemia cases in a city in Israel," *Statistics in Medicine*, 16, 1873–1887.

Choi, K. (1981), "An evaluation of influenza mortality surveillance, 1962-1979. 1: Time series forecasts of expected pneumonia and influenza deaths," *American Journal of Epidemiology*, 113, 215–226.

Clark, A. B. and Lawson, A. B. (2006), "Surveillance of individual level disease maps," *Statistical Methods in Medical Research*, 15, 353–362.

Conesa, D., López-Quílez, A., Martínez-Beneito, M. A., Miralles, M. T., and Verdejo, F. (2009), "FluDetWeb: an interactive web-based system for the early detection of the onset of influenza epidemics," *BMC Medical Informatics and Decision Making*, 9:36.

Conesa, D., Martínez-Beneito, M. A., Amorós, R., and López-Quílez, A. (2010), "Bayesian hierarchical Poisson model with a hidden Markov structure for the detection of influenza epidemic outbreaks," Working paper, Departament d'Estadística i Investigació Operativa, Universitat de València.

Cooper, D. L., Smith, G., Baker, M., Chinemana, F., Verlander, N. Q., Gerard, E., Hollyoak, V., and Griffiths, R. (2004), "National symptom surveillance using calls to a telephone health advice service - United Kingdom, December 2001-February 2003," *Morbidity and Mortality Weekly Report (Supplement)*, 53, 179–183.

Coory, M., Duckett, S., and Sketcher-Baker, K. (2008), "Using control charts to monitor quality of hospital care with administrative data," *International Journal for Quality in Health Care*, 20, 31–39.

Costagliola, D., Flahault, A., Galinec, D., Garnerin, P., Menares, J., and Valleron, A.-J. (1991), "A routine tool for detection and assessment of epidemics of influenza-like syndromes in France," *American Journal of Public Health*, 81, 97–99.

— (1994), "When is the epidemic warning cut-off point exceeded?" *European Journal of Epidemiology*, 10, 475–476.

Cowling, B. J., Wong, I. O. L., Riley, S., and Leung, B. M. (2006), "Methods for monitoring influenza data," *International Journal of Epidemiology*, 35, 1314–1321.

Cox, D. R. (1955), "Some statistical methods related with series of events (with discussion)," *Journal of the Royal Statistical Society Series B*, 17, 129–157.

Crosier, R. B. (1988), "A new two-sided cumulative sum quality control scheme," *Technometrics*, 28, 187–194.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum-likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society Series B*, 39, 1–38.

Diggle, P. J., Knorr-Held, L., Rowlingson, B., Su, T., Hawtin, P., and Bryant, T. (2004), "Spatio-temporal point processes: methods and applications," in *Monitoring the Health of Populations: Statistical Principles & Methods for Public Health Surveillance*, eds. Brookmeyer, R. and Stroup, D. F., Oxford University Press: Oxford, pp. 233–266.

Diggle, P. J., Rowlingson, B., and Su, T.-L. (2005), "Point process methodology for on-line spatio-temporal disease surveillance," *Environmetrics*, 16, 423–434.

Dominici, F., Sheppard, L., and Clyde, M. (2003), "Health effects of air pollution: a statistical review," *International Statistical Review*, 71, 243–276.

Dong, Y., Hedayat, A. S., and Sinha, B. K. (2008), "Surveillance stratgeies for detecting changepoint in incidence rate based on exponentially weighted moving average methods," *Journal of the American Statistical Association*, 103, 843–853.

Duczmal, L. and Assunção, R. M. (2004), "A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters," *Computational Statistics & Data Analysis*, 45, 269–286.

Elbert, Y. and Burkom, H. S. (2009), "Development and evaluation of a data-adaptive alerting algorithm for univariate temporal biosurveillance data," *Statistics in Medicine*, 28, 3226–3248.

Farrington, C. P. (2004), "Monitoring clinical performance: invited comments on the papers by Grigg and Farewell and Marshall et al." *Journal of the Royal Statistical Society Series A*, 167, 562–563.

Farrington, C. P. and Andrews, N. (2004), "Outbreak detection: Application to infectious disease surveillance," in *Monitoring the Health of Populations: Statistical Principles & Methods for Public Health Surveillance*, eds. Brookmeyer, R. and Stroup, D. F., Oxford University Press: Oxford, pp. 203–231.

Farrington, C. P., Andrews, N. J., Beal, A. D., and Catchpole, M. A. (1996), "A statistical algorithm for the early detection of outbreaks of infectious disease," *Journal of the Royal Statistical Society Series A*, 159, 547–562.

Fasting, S. and Gisvold, S. E. (2003), "Statistical process control methods allow the analysis and improvement of anesthesia care," *Canadian Journal of Anesthesia*, 50, 767–774.

Fienberg, S. E. and Shmueli, G. (2005), "Statistical issues and challenges associated with rapid detection of bio-terrorist attacks," *Statistics in Medicine*, 24, 513–529.

Fraker, S. E., Woodall, W. H., and Mousavi, S. (2008), "Performance metrics for surveillance schemes," *Quality Engineering*, 20, 451–464.

Frisén, M. (1992), "Evaluations of methods for statistical surveillance," *Statistics in Medicine*, 11, 1489–1502.

— (2003), "Statistical surveillance. Optimality and Methods," *International Statistical Review*, 71, 403–434.

— (2007), "Properties and use of the Shewhart method and its followers," *Sequential Analysis*, 26, 171–193.

— (2010), "Principles for multivariate surveillance," in *Frontiers in Statistical Quality Control 9*, eds. Lenz, H.-J., Wilrich, P.-T., and W., S., Springer: Berlin, Heidelberg, pp. 133–144.

Frisén, M. and Andersson, E. (2009), "Semiparametric surveillance of monotonic changes," *Sequential Analysis*, 28, 434–454.

Frisén, M., Andersson, E., and Pettersson, K. (2010a), "Semiparametric estimation of outbreak regression," *Statistics*, 44, 107–117.

Frisén, M., Andersson, E., and Schiöler, L. (2010b), "Sufficient reduction in multivariate surveillance," *Communications in Statistics - Theory and Methods*, to appear.

Frisén, M., Andersson, E., and Schiöler (2009), "Robust outbreak surveillance of epidemics in Sweden," *Statistics in Medicine*, 28, 476–493.

Frisén, M. and de Maré, J. (1991), "Optimal surveillance," *Biometrika*, 78, 271–280.

Gan, F. F. (1991), "Monitoring observations generated from a binomial distribution using modified exponentially weighted moving average control charts," *Journal of Statistical Computation and Simulation*, 37, 45–60.

— (1994), "Design of optimal exponential CUSUM control charts," *Journal of Quality Technology*, 26, 109–124.

— (1998), "Designs of one- and two-sided exponential EWMA Charts," *Journal of Quality Technology*, 30, 44–69.

Glaz, J., Naus, J., and Wallenstein, W. (2001), *Scan Statistics*, Springer: New York.

Goldenberg, A., Shmueli, G., Caruana, R., and Fienberg, S. (2002), "Early statistical detection of anthrax outbreaks by tracking over-the-counter medication sales," *Proceedings of the National Academy of Sciences of the United States*, 99, 5237–5240.

Golosnoy, V., Ragulin, S., and Schmid, W. (2009), "Multivariate CUSUM chart: properties and enhancements," *Advances in Statistical Analysis*, 93, 263–279.

Griffin, B. A., Jain, A. K., Davies-Cole, J., Glymph, C., Lum, G., Washington, S. C., and Stoto, M. A. (2009), "Early detection of influenza outbreaks using the DC Department of Health's syndromic surveillance system," *BMC Public Health*, 9:483.

Grigg, O. A., Farewell, V. T., and Spiegelhalter, D. J. (2003), "Use of risk-adjusted CUSUM and RSPRT charts for monitoring in medical contexts," *Statistical methods in Medical Research*, 12, 147 – 170.

Grigg, O. A. and Spiegelhalter, D. J. (2008), "An empirical approximation to the null unbounded steady state distribution of the cumulative sum statistic," *Technometrics*, 50, 501–511.

Han, S. W., Tsui, K.-L., Ariyajunya, B., and Kim, S. B. (2010), "A comparison of CUSUM, EWMA, and temporal scan statistics for detection of increases in Poisson rates," *Quality and Reliability Engineering International*, 26, 279–289.

Healy, M. J. R. (1983), "A simple method for monitoring routine statistics," *The Statistician*, 32, 347–349.

Heisterkamp, S. H., Dekkers, A. L. M., and Heijne, J. C. M. (2006), "Automated detection of infectious disease outbreaks: hierarchical time series models," *Statistics in Medicine*, 25, 4179–4196.

Held, L., G., G., Frank, C., and Rue, H. (2006a), "Joint spatial analysis of gastrointestinal infectious diseases," *Statistical Methods in Medical Research*, 15, 465–480.

Held, L., Hofmann, M., Höhle, M., and Schmid, V. (2006b), "A two-component model for counts of infectious diseases," *Biostatistics*, 7, 422–437.

Held, L., Höhle, M., and Hofmann, M. (2005), "A statistical framework for the analysis of multivariate infectious disease surveillance counts," *Statistical Modelling*, 5, 187–199.

Helfenstein, U. (1986), "Box-Jenkins modelling of some viral infectious diseases," *Statistics in Medicine*, 5, 37–47.

Höhle, M. (2007), "surveillance: An R package for the monitoring of infectious diseases," *Computational Statistics*, 22, 571–582.

Höhle, M. and Mazick, A. (2010), "Aberration detection in R illustrated by Danish mortality monitoring," in *Biosurveillance: Methods and Case Studies*, eds. Kass-Hout, T. and Zhang, X., CRC Press: London, pp. 215–237.

Höhle, M. and Paul, M. (2008), "Count data regression charts for the monitoring of surveillance time series," *Computational Statistics & Data Analysis*, 52, 4357–4368.

Holt, C. C. (1957), "Forecasting trends and seasonals by exponentially weighted moving averages," ONR Memorandum 52/1957: Carnegie Institute of Technology.

Hotelling, H. (1947), "Multivariate quality control," in *Techniques of Statistical Analysis*, eds. Eisenhart, C., Hastay, M. W., and Wallis, W. A., New York: McGraw-Hill, pp. 111–184.

Huang, S. S., Yokoe, D. S., Stelling, J., Placzek, H., Kulldorff, M., Kleinman, K., O'Brien, T. F., Calderwood, M. S., Vostok, J., Dunn, J., and Platt, R. (2010), "Automated detection of infectious disease outbreaks in hospitals: a retrospective cohort study," *PLOS Medicine*, 7: e1000238.

Hulth, A., Andrews, N., Ethelberg, S., Dreesman, J., Faensen, D., van Pelt, W., and Schnitzler, J. (2010), "Practical usage of computer-supported outbreak detection in five European countries," *Eurosuveillance*, 15 (36).

Hutwagner, L., Browne, T., Seeman, G. M., and Fleischhauer, A. T. (2005a), "Comparing aberration detection methods with simulated data," *Emerging Infectious Diseases*, 11, 314–316.

Hutwagner, L., Thompson, W. W., Seeman, G. M., and Treadwell, T. (2003), "The Bioterrorism Preparedness and Response Early Aberration Reporting System (EARS)," *Journal of Urban Health*, 80, i89–96.

— (2005b), "A simulation model for assessing aberration detection methods in public health surveillance for systems with limited baselines," *Statistics in Medicine*, 24, 543–550.

Ismail, N. A., Pettitt, A. N., and Webster, R. A. (2003), "'Online' monitoring and retrospective analysis of hospital outcomes based on a scan statistic," *Statistics in Medicine*, 22, 2861–2876.

Jackson, J. E. (1959), "Quality control methods for several related variables," *Technometrics*, 1, 359–377.

Jackson, M. J., Baer, A., Painter, I., and Duchin, J. (2007), "A simulation study comparing aberration detection algorithms for syndromic surveillance," *BMC Medical Informatics and Decision Making*, 7:6.

Jacquez, G. M., Greilling, D. A., Durbeck, H., Estberg, L., Do, E., Long, A., and Rommel, B. (2002), "*ClusterSeer User Guide 2: Software for Identifying Disease Clusters*," TerraSeer Press: Ann Arbor, Michigan.

Jolliffe, I. T. (2002), *Principal Component Analysis*, Springer: New York, 2nd ed.

Joner, M. D., Woodall, W. H., and Reynolds, M. R. (2008), "Detecting a rate increase using a Bernoulli scan statistic," *Statistics in Medicine*, 27, 2555–2575.

Kafadar, K. and Stroup, D. (1992), "Analysis of aberrations in public health surveillance data: estimating variances on correlated samples," *Statistics in Medicine*, 11, 1551–1568.

Kenett, R. S. and Pollak, M. (1996), "Data-analytic aspects of the Shiryaev-Roberts control chart: surveillance of a non-homogeneous Poisson process," *Journal of Applied Statistics*, 23, 125–137.

Kleinman, K. P. and Abrams, A. (2006), "Assessing surveillance using sensitivity, specificity and timeliness," *Statistical Methods in Medical Research*, 15, 445–464.

Kleinman, K. P., Abrams, A., Kulldorff, M., and Platt, R. (2005), "A model-adjusted space-time scan statistic with an application to syndromic surveillance," *Epidemiology and Infection*, 133, 409–419.

Kleinman, K. P., Lazarus, R., and Platt, R. (2004), "A generalized linear mixed models approach for detecting incident clusters of disease in small areas, with an application to biological terrorism," *American Journal of Epidemiology*, 159, 217–224.

Knorr-Held, L. (2000), "Bayesian modelling of inseparable space-time variation in disease risk," *Statistics in Medicine*, 19, 2555–2567.

Knox, E. G. (1964), "The detection of space-time interactions," *Applied Statistics*, 13, 25–29.

Ku, W., Storer, R., and Georgakis, C. (1995), "Disturbance detection and isolation by dynamic principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, 30, 179–196.

Kulldorff, M. (1997), "A spatial scan statistic," *Communications in Statistics: Theory and Methods*, 26, 1481–1496.

— (2001), "Prospective time periodic geographical disease surveillance using a scan statistic," *Journal of the Royal Statistical Society Series A*, 164, 61–72.

— (2010), "SaTScan version 8.2.1: software for the spatial, temporal, and space-time scan statistics," http://www.satscan.org.

Kulldorff, M., Athos, W., Feuer, E., Miller, B., and Key, C. (1998), "Evaluating cluster alarms: a space-time scan statistic and brain cancer in Los Alamos," *American Journal of Public Health*, 188, 1377–1380.

Kulldorff, M., Heffernan, R., Hartman, J., Assunção, R., and Mostashari, F. (2005), "A space-time permutation scan statistic for disease outbreak detection," *PLoS Medicine*, 2, 216–224.

Kulldorff, M., Huang, L., Pickle, L., and Duczmal, L. (2006), "An elliptic spatial scan statistic," *Statistics in Medicine*, 25, 3929–3943.

Kulldorff, M., Mostashari, F., Duczmal, L., Yih, W. K., Kleinman, K., and Platt, R. (2007), "Multivariate scan statistics for disease surveillance," *Statistics in Medicine*, 26, 1824–1833.

Lawson, A. B. (2005), "Spatial and spatio-temporal disease analysis," in *Spatial and Syndromic Surveillance for Public Health*, eds. Lawson, A. B. and Kleinman, K., John Wiley & Sons: Chichester, pp. 55–76.

Lawson, A. B., Clark, A., and Rodeiro, C. L. V. (2003), "Developments in general and syndromic surveillance for small area health data," *Journal of Applied Statistics*, 31, 951–966.

Le Strat, Y. and Carrat, F. (1999), "Monitoring epidemiologic surveillance data using hidden Markov models," *Statistics in Medicine*, 18, 3463–3478.

Leal, J. and Laupland, K. B. (2008), "Validity of electronic surveillance systems: a systematic review," *Journal of Hospital Infection*, 69, 220–229.

Limaye, S. S., Mastrangelo, C. M., Zerr, D. M., and Jeffries, H. (2008), "A statistical approach to reduce hospital-associated infections," *Quality Engineering*, 20, 414–425.

Lowry, C. A., Woodall, W. H., Champ, C. W., and Rigdon, S. E. (1992), "A multivariate exponentially weighted moving average control chart," *Technometrics*, 34, 46–53.

Lu, H.-M., Zeng, D., and Chen, H. (2010), "Prospective infectious disease outbreak detection using Markov switching models," *IEEE Transactions on Knowledge & Data Engineering*, 22, 565–577.

Lucas, J. M. (1985), "Counted data CUSUM's," *Technometrics*, 27, 129–144.

Lucas, J. M. and Crosier, R. B. (2000), "Fast initial response for CUSUM quality-control schemes: give your CUSUM a head start," *Technometrics*, 42, 102–107.

Madigan, D. (2005), "Bayesian data mining for health surveillance," in *Spatial and Syndromic Surveillance for Public Health*, eds. Lawson, A. B. and Kleinman, K. P., John Wiley & Sons: Chicester, pp. 203–221.

Mann, A. E. (2009), "Estimating the impact of influenza vaccinations and antigenic drift on influenza-related morbidity and mortality in England and Wales using hidden Markov models," PhD Thesis, University of London.

Marshall, C., Best, N., Bottle, A., and Aylin, P. (2004), "Statistical issues in the prospective monitoring of health outcomes across multiple units," *Journal of the Royal Statistical Society Series A*, 167, 541–559.

Marshall, J. B., Spitzner, D. J., and Woodall, W. H. (2007), "Use of the local Knox statistic for the prospective monitoring of disease occurrences in space and time," *Statistics in Medicine*, 26, 1579–1593.

Martínez-Beneito, M. A., Conesa, D., López-Quílez, A., and López-Maside, A. (2008), "Bayesian Markov switching models for the early detection of influenza epidemics," *Statistics in Medicine*, 27, 4455–4468.

McCabe, G. J., Greenhalgh, D., Gettingby, G., Holmes, E., and Cowden, J. (2003), "Prediction of infectious diseases: an exception reporting system," *Journal of Medical Informatics and Technologies*, 5, 67–74.

Miller, B., Kassenborg, H., Dunsmuir, W., Griffith, J., Hadidi, M., Nordin, J., and Danila, R. (2003), "Syndromic surveillance for influenzalike illness in ambulatory care network," *Emerging Infectious Diseases*, 10, 1806–1811.

Montgomery, D. C. (2009), *Introduction to Statistical Process Control*, John Wiley & Sons: Hoboken, New Jersey, 6th ed.

Morton, A. P., Whitby, M., McLaws, M. L., Dobson, A., McElwain, S., Looke, D., Stackelroth, J., and Sartor, A. (2001), "The application of statistical process control charts to the detection and monitoring of hospital-acquired infections," *Journal of Quality in Clinical Practice*, 21, 112–117.

Mostashari, F., Kulldorff, M., Hartman, J. J., Miller, J. R., and Kulasekera, V. (2003), "Dead bird clusters as an early warning system for West Nile virus activity," *Emerging Infectious Diseases*, 9, 641–646.

Munford, A. G. (1980), "A control chart based on cumulative scores," *Applied Statistics*, 26, 252–258.

Musonda, P., Hocine, M. N., Andrews, N. J., Tubert-Bitter, P., and Farrington, C. P. (2008), "Monitoring vaccine safety using case series cumulative sum charts," *Vaccine*, 26, 5358–5367.

Naus, J. and Wallenstein, S. (2006), "Temporal surveillance using scan statistics," *Statistics in Medicine*, 25, 311–324.

Neill, D. B. (2009), "An empirical comparison of spatial scan statistics for outbreak detection," *International Journal of Health Geographics*, 8:20.

Ngai, H.-M. and Zhang, J. (2001), "Multivariate cumulative sum control charts based on projection pursuit," *Statistica Sinica*, 11, 747–766.

Ngo, L., Tager, I. B., and Hadley, D. (1996), "Application of exponential smoothing for nosocomial infection surveillance," *American Journal of Epidemiology*, 143, 637–647.

Nobre, F. F. and Stroup, D. F. (1994), "A monitoring system to detect changes in public health surveillance data," *Journal of Epidemiology*, 23, 408–418.

Oakland, J. S. (2008), *Statistical Process Control*, Butterworth-Heinemann: Oxford, 6th ed.

Page, E. S. (1954), "Continuous inspection schemes," *Biometrika*, 41, 100–115.

Parker, R. A. (1989), "Analysis of surveillance data with Poisson regression: a case study," *Statistics in Medicine*, 8, 285–294.

Paul, M., Held, L., and Toschke, A. M. (2008), "Multivariate modelling of infectious disease surveillance data," *Statistics in Medicine*, 27, 6250–6267.

Pelat, C., Boëlle, P.-Y., Cowling, B. J., Carrat, F., Flahault, A., Ansart, S., and Valleron, A.-J. (2007), "Online detection and quantification of epidemics," *BMC Medical Informatics and Decision Making*, 7:29.

Pignatiello, J. J. and Runger, G. C. (1990), "Comparison of multivariate CUSUM charts," *Journal of Quality Technology*, 22, 173–186.

Qiu, P. H. and Hawkins, D. (2001), "A rank-based multivariate CUSUM procedure," *Technometrics*, 43, 120–132.

— (2003), "A nonparametric multivariate cumulative sum procedure for detecting shifts in all directions," *The Statistician*, 52, 151–164.

Rath, T. M., Carreras, M., and Sebastiani, P. (2003), "Automated detection of influenza epidemics with hidden Markov models," in *Advances in Intelligent Data Analysis V, LNCS 2810*, eds. Berthold, M. R., Lenz, H.-J., Bradley, E., Kruse, R., and Borgelt, C., Springer: Berlin, pp. 521–532.

Raubertas, R. F. (1989), "An analysis of disease surveillance data that uses the geographical locations of the reporting units," *Statistics in Medicine*, 8, 267–271.

Reis, B. Y. and Mandl, K. D. (2003), "Time series modeling for syndromic surveillance," *BMC Medical Informatics and Decision Making*, 3, No. 2.

Reynolds, M. R. and Stoumbos, Z. G. (2000), "A general approach to modeling CUSUM charts for a proportion," *IIE Transactions*, 32, 515–535.

Roberts, S. W. (1959), "Control chart tests based on geometric moving averages," *Technometrics*, 1, 239–250.

— (1966), "A comparison of some control chart procedures," *Technometrics*, 8, 411–430.

Robertson, C. (2006), "Protecting the leaders: syndromic surveillance for the G8 summit in Scotland," *Significance*, 3, 69–72.

Rogerson, P. A. (1997), "Surveillance systems for monitoring the development of spatial patterns," *Statistics in Medicine*, 26, 2081–2093.

— (2001), "Monitoring point patterns for the development of spatial patterns," *Journal of the Royal Statistical Society Series A*, 164, 87–96.

Rogerson, P. A. and Yamada, I. (2004a), "Approaches to syndromic surveillance when data consist of small regional counts," *Morbidity and Mortality Weekly Report (Supplement)*, 53, 79–85.

— (2004b), "Monitoring change in spatial patterns of disease: comparing univariate and multivariate cumulative sum approaches," *Statistics in Medicine*, 23, 2195–2214.

Rossi, G., Lampugnani, L., and Marchi, M. (1999), "An approximate CUSUM procedure for surveillance of health events," *Statistics in Medicine*, 18, 2111–2122.

Schiöler, L. and Frisén, M. (2010), "Multivariate outbreak detection," Manuscript submitted for publication.

Sebastiani, P., Mandl, K. D., Szolovits, P., Kohane, I. S., and Ramoni, M. F. (2006), "A Bayesian dynamic model for influenza surveillance," *Statistics in Medicine*, 25, 1803–1816.

Sego, L. H., Woodall, W. H., and Reynolds, M. R. (2008), "A comparison of surveillance methods for small incidence rates," *Statistics in Medicine*, 27, 1225–1247.

Serfling, R. (1963), "Methods for current statistical analysis of excess pneumonia-influenza deaths," *Public Health Reports*, 78, 494–506.

Shewhart, W. A. (1931), *Economic Control of Quality of Manufactured Product*, Van Nostrand Reinhold: Princeton.

Shiryaev, A. N. (1963), "On the detection of disorder in a manufactoring process," *Theory of Probability and its Applications*, 8, 247–265.

Shmueli, G. (2005), "Wavelet-based monitoring for modern biosurveillance," Working paper, RHS-06-002: Robert H. Smith School of Business, University of Maryland. http://ssrn.com/abstract=902878.

Shmueli, G. and Burkom, H. (2010), "Statistical challenges facing early outbreak detection in biosurveillance," *Technometrics*, 52, 39–51.

Shore, D. L. and Quade, D. (1989), "A surveillance system based on a short memory scheme," *Statistics in Medicine*, 8, 311–322.

Sitter, R. R., Hanrahan, L. P., Anderson, H. A., and DeMets, D. (1991), "The authors' reply," *American Journal of Epidemiology*, 134, 917–919.

Sitter, R. R., Hanrahan, L. P., DeMets, D., and Anderson, H. A. (1990), "A monitoring system to detect increased rates of cancer incidence," *American Journal of Epidemiology*, 132, 123–130.

Sonesson, C. (2003), "Evaluations of some exponentially weighted moving average methods," *Journal of Applied Statistics*, 30, 1115–1133.

— (2007), "A CUSUM framework for detection of space-time disease clusters using scan statistics," *Statistics in Medicine*, 26, 4770–4789.

Sonesson, C. and Bock, D. (2003), "A review and discussion of prospective statistical surveillance in public health," *Journal of the Royal Statistical Society Series A*, 166, 5–21.

Sonesson, C. and Frisén, M. (2005), "Multivariate surveillance," in *Spatial Surveillance for Public Health*, eds. Lawson, A. B. and Kleinman, K., John Wiley & Sons: Chicester, pp. 169–186.

Sparks, R. (2010), "Enhancing road safety through early detection of outbreaks in the frequency of motor vehicle crashes," *Safety Science*, 48, 135–144.

Stern, L. and Lightfoot, D. (1999), "Automated outbreak detection: A quantitative retrospective analysis," *Epidemiology and Infection*, 122, 103–110.

Steutel, F. W. and van Harn, K. (1979), "Discrete analogues of self-decomposability and stability," *Annals of Probability*, 7, 893–899.

Stroup, D., Wharton, M., Kafadar, K., and Dean, A. (1993), "Evaluation of a method for detecting aberrations in public health surveillance data," *American Journal of Epidemiology*, 137, 373–380.

Stroup, D. F., Thacker, S. B., and 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.

Stroup, D. F., Williamson, G. D., Herndon, J. L., and Karon, J. M. (1989), "Detection of aberrations in the occurence of notifiable diseases surveillance data," *Statistics in Medicine*, 8, 323–329.

Takahashi, K., Kulldorff, M., Tango, T., and Yih, K. (2008), "A flexibly shaped space-time scan statistic for disease outbreak detection and monitoring," *International Journal of Health Geographics*, 7:14.

Tango, T. (1995), "A class of tests for detecting general and focussed clustering of rare diseases," *Statistics in Medicine*, 14, 2323–2334.

Tango, T. and Takahashi, K. (2005), "A flexibly shaped spatial scan statistic for detecting clusters," *International Journal of Health Geographics*, 4:11.

Thacker, S. B., G., P. R., and L., T. F. (1988), "A method for evaluating systems of epidemiological surveillance," *World Health Statistical Quarterly*, 41, 11–18.

Tillett, H. E. and Spencer, I.-L. (1982), "Influenza surveillance in England and Wales using routine statistics," *Journal of Hygiene, Cambridge*, 88, 83–94.

Vidal Rodeiro, C. L. and Lawson, A. B. (2006), "Monitoring changes in spatio-temporal maps of disease," *Biometrical Journal*, 3, 463–480.

Wagner, A. (2010), "Extending Scottish Exception Reporting Systems Spatially and Temporally," Ph.D. thesis, University of Strathclyde.

Wallenstein, S. (1980), "A test for detection of clustering in time," *American Journal of Epidemiology*, 111, 367–372.

Wallenstein, S., Gould, M., and Kleinman, M. (1989), "Use of the scan statistic to detect time-space clustering," *American Journal of Epidemiology*, 130, 1057–1064.

Wallenstein, S. and Naus, J. (2004), "Scan statistics for temporal surveillance for biologic terrorism," *Morbidity and Mortality Weekly Report (Supplement)*, 53, 74–78.

Watier, L., Richardson, S., and Hubert, B. (1991), "A time series construction of an alert threshold with application to *S. bovismorbificans* in France," *Statistics in Medicine*, 10, 1493–1509.

Watkins, R. E., Eagleson, S., Veenendaal, B., Wright, G., and Plant, A. J. (2009), "Disease surveillance using a hidden Markov model," *BMC Medical Informatics and Decision Making*, 9:39.

Weiß, C. H. (2009), *Categorical Time Series Analysis and Applications in Statistical Quality Control*, dissertation.de Verlag: Berlin.

West, M. and Harrison, J. (1997), *Bayesian Forecasting and Dynamic Models*, Springer: New York, 2nd ed.

Widdowson, M.-A., Bosman, A., van Straten, E., Tinga, M., Chaves, S., van Eerden, L., and van Pelt, W. (2003), "Automated, laboratory-based system using the internet for disease outbreak detection," *Emerging Infectious Diseases*, 9, 1046–1052.

Wieland, S. C., Brownstein, J. S., Berger, B., and Mandl, K. D. (2007), "Automated real-time constant specificity surveillance for disease outbreaks," *BMC Medical Informatics and Decision Making*, 7:15.

Williamson, G. and Weatherby Hudson, G. (1999), "A monitoring system for detecting aberrations in public health surveillance reports," *Statistics in Medicine*, 18, 3283–3298.

Winters, P. R. (1960), "Forecasting sales by exponentially weighted moving averages," *Management Science*, 6, 324–342.

Wolter, C. (1987), "Monitoring intervals between rare events: a cumulative score procedure compared with Rina Chen's sets technique," *Methods of Information in Medicine*, 26, 215–219.

Woodall, W. H. (2006), "Use of control charts in health care and public health surveillance (with discussion)," *Journal of Quality Technology*, 38, 88–103.

Woodall, W. H., Marshall, J. B., Joner Jr., M. D., Fraker, S. E., and 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 Series A*, 171, 223–237.

Yamada, I., Rogerson, P., and Lee, G. (2009), "GeoSurveillance: a GIS-based system for the detection and monitoring of spatial clusters," *Journal of Geographical Systems*, 11, 155–173.

Zhang, J., Tsui, F.-C., Wagner, M. M., and Hogan, W. (2003), "Detection of outbreaks from time series data using wavelet transform," *AMIA Annual Symposium*, Omnipress: Madison, Wisconsin, 748–552.

Zhang, T. and Lin, G. (2009), "Spatial scan statistics in loglinear models," *Computational Statistics & Data Analysis*, 53, 2851–2858.

Zhou, H. and Lawson, A. B. (2008), "EWMA smoothing and Bayesian spatial modeling for health surveillance," *Statistics in Medicine*, 27, 5907–5928.