Early Detection of a Change in Poisson Rate After Accounting For Population Size Effects

Sung Won Han, Yajun Mei, and Kwok-Leung Tsui

School of Industrial and Systems Engineering,
Georgia Institute of Technology,
765 Ferst Drive NW,
Atlanta, GA 30332-0205, USA
{shan,ymei,ktsui }@isye.gatech.edu

Abstract. Motivated by applications in bio and syndromic surveillance, this article is concerned with the problem of detecting a change in the rate of Poisson distributions after taking into account the effects of population size.

Keywords. CUSUM, on-line monitoring, non-homogeneous population.

1 Introduction and Motivation

This work was motivated by a set of data regarding male thyroid cancer cases (with malignant behavior) in New Mexico during 1973-2005, which has been studied before in the biosurveillance literature in other contexts; see, for example, Kulldorff (2001). Figure 1 plots three different curves related to this data set: (1) yearly total number of cancers with malignant behavior; (2) yearly population size (of males) in New Mexico; and (3) yearly (crude) incidence rate per 100,000 (male) population.

For this data set, one obvious starting point is to investigate whether the yearly number of cancer cases increases over time. Such a problem can be formulated as detecting a change in the mean (or intensity or rate) of Poisson distributions, and the classical results in the sequential change-point detection literature such as those in Lorden (1971) and Moustakides (1985) are applicable. Also refer to Peskir and Shiryaev (2006).

From the biosurveillance viewpoint, a more interesting analysis goal of this data set is to determine whether or not the *risk* for male thyroid cancer increases over time. The term *risk* in this context essentially means that the probability of developing thyroid cancer in a given year, which can be characterized by the number of incidence rate per 100,000 (male) population. This consideration inspires us to investigate the problem of detecting a change in the disease incidence rate after taking into account the effect of population size.

2 Mathematical Formulation

A very simplified probability model of the above problem is the Poisson model in which one observes independent two-dimensional random vectors (l_n, Y_n) over time n, where Y_n has a Poisson distribution of mean $\mu_n = l_n\lambda_n$. Here l_n, Y_n and λ_n can be thought of as the observed population size (in the units of 100,000 population), the number of observed disease cases, and the (unobservable true) incidence rate per 100,000 (male) population at the n-th year, respectively. Of course, in theory, it is better to model the observation Y_n's by binomial distributions with mean $l_n\lambda_n$, but it is well-known that the binomial distribution can be approximated by a Poisson distribution with the same mean, provided that the population size is large and the binomial probability parameter is small, so that the observed count is small relative to the population size. The data in our motivated example satisfy this requirement reasonably well, and thus the Poisson model is applicable.

In the context of sequential change-point detection problems, it is assumed that the λ_n's, e.g., the incidence rate per 100,000 (male) population, changes from one value λ_0 to another value λ_1 at some unknown time ν, and we want to detect such a change as soon as possible if it occurs. Note that under our setting, the l_n's (the population sizes) are observable, but their distributions are nuisance parameters that are left unspecified, as we are only interested in detecting a change in λ_n's.

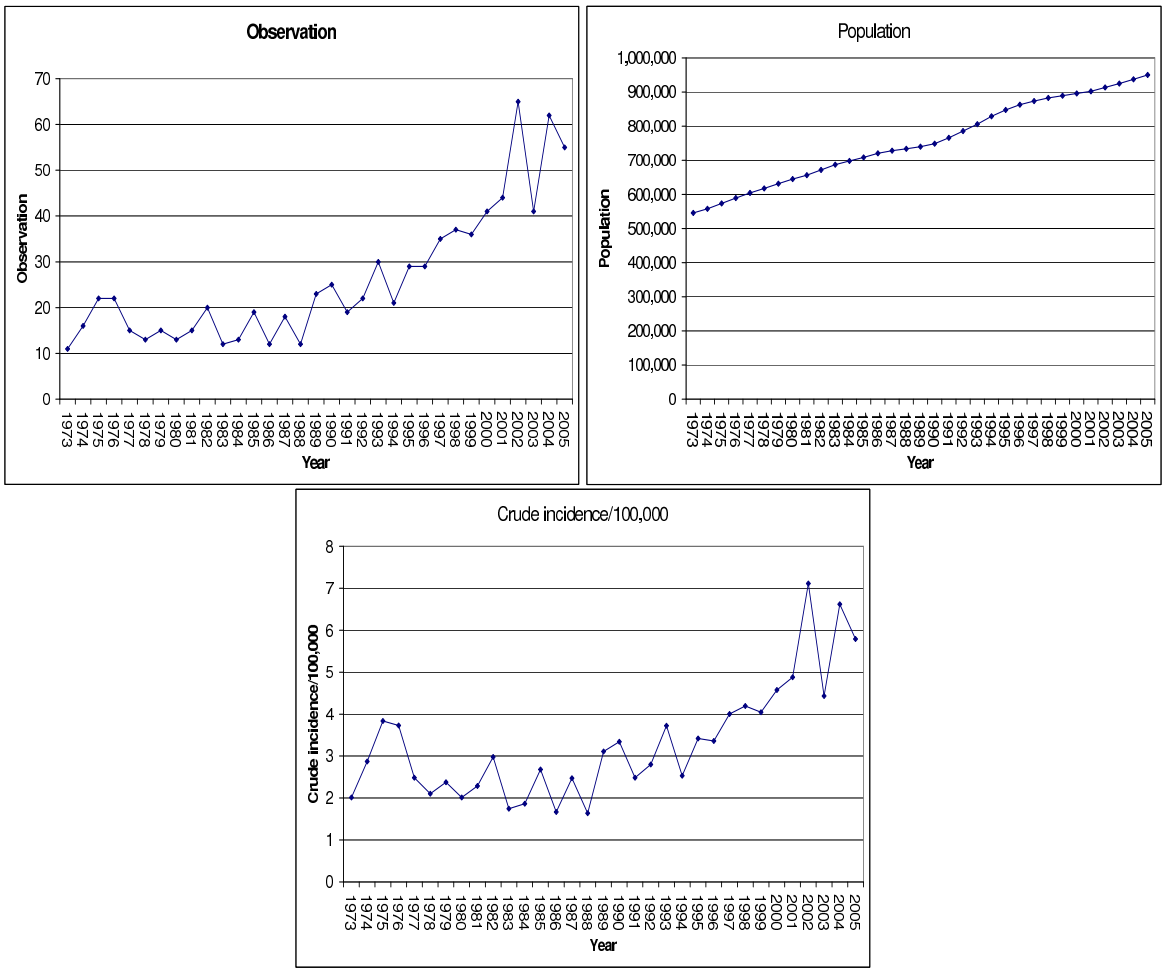

Fig. 1. A three set of time series data of male thyroid cancer in New Mexico during 1973-2005. Top: the left panel plots the total number of male thyroid cancers (Y axis) over years (X axis), and the right panel illustrates the trend of male population (Y axis). Bottom: the plot illustrates the crude incidence per 100,000 population (Y axis) over years (X axis).

Mathematically, a detection scheme is a stopping time T with respect to the the observed data $\{(l_n, Y_n)\}_{n\geq 1}$. That is, the decision $\{T = n\}$ only depends on the first n observations, or more rigourously, the event $\{T = n\}$ belongs to the sigma-algebra $\mathcal{F}_n$, which is assumed to be generated by the first n observations. Denote by $\mathbf{P}_\nu$ and $\mathbf{E}_\nu$ the probability measure and expectation when the change in the λ_n's occurs at time n. Also denote by $\mathbf{P}_\infty$ and $\mathbf{E}_\infty$ the probability measure and expectation when there are no changes in the λ_n's.

A standard minimax formulation is to find a detection scheme T that minimizes the detection delay in Lorden (1971)

$$\overline{\mathbf{E}}(T) = \sup_{\nu\geq 1} \operatorname{ess\,sup} \mathbf{E}_\nu\Big((T-\nu+1)^+|\mathcal{F}_n\Big)$$

subject to the constraint

$$\mathbf{E}_\infty(T) \geq \gamma, \tag{1}$$

where γ is a pre-specified (typically large) constant to control the false alarm rate of the scheme.

3 Methodology

In sequential change-point detection, or more generally in statistics, a basic tool to construct statistical tests or procedures is the generalized/maximum likelihood ratios (GLR) method. For the sequential change-point detection problem, it can be thought of testing the null hypothesis

$$H_0 : \nu = \infty \text{ (no change)}$$

versus the composite alternative hypothesis

$$H_1 : 1 \le \nu < \infty \text{ (a change occurs)},$$

and thus the logarithm of the GLR statistic of the first n observations, $\{(l_i, Y_i)\}_{i=1}^n$ is given by

$$\max_{1\le\nu<\infty} \frac{d\mathbf{P}_\nu}{d\mathbf{P}_\infty}\Big((l_1, Y_1), \cdots, (l_n, Y_n)\Big).$$

Note that when $\mu_j = l_j\lambda_j$ (with $j = 0$ or 1) is the true value (of the Poisson mean), the marginal density function of (l_i, Y_i) is $f(l_i, Y_i|\lambda_j) = h(l_i)\frac{e^{-l_i\lambda_j}(l_i\lambda_j)^{Y_i}}{(Y_i)!}$, where $h(\cdot)$ is the distribution of l_i's. Hence, under our setting, the logarithm of the GLR statistic becomes

$$W_n = \max_{1\le\tau\le n+1} \sum_{i=\tau}^{n} \log \frac{f(l_i, Y_i|\lambda_1)}{f(l_i, Y_i|\lambda_0)} = \max_{1\le\tau\le n+1} \sum_{i=\tau}^{n} \Big[Y_i \log \frac{\lambda_1}{\lambda_0} - l_i(\lambda_1 - \lambda_0)\Big], \tag{2}$$

where $\sum_{i=n+1}^{n} = 0$ as conventional, and the GLR detection scheme raises an alarm at time

$$T_{GLR}(a) = \text{ first } n \ge 1 \text{ such that } W_n \ge a, \tag{3}$$

for some constant a. It is easy to show that W_n in (2) also enjoys a recursive formula of the classical CUSUM statistics: $W_n = \max\Big\{0, W_{n-1} + \Big[Y_n \log \frac{\lambda_1}{\lambda_0} - l_n(\lambda_1 - \lambda_0)\Big]\Big\}$.

Besides the GLR detection scheme in (3), in this article we also consider two alternative detection schemes. The first one is based on Y_n/l_n, which is a natural estimator of the *risk* or the disease rate per 100,000 population. If we pretend that Y_n/l_n is Poisson distributed with mean λ_n (while this is not true under our setting, we can still use it to construct detection schemes), then the corresponding CUSUM statistic is

$$\hat{W}_n = \max_{1\le\tau\le n+1} \sum_{i=\tau}^{n} \Big[\frac{Y_i}{l_i} \log \frac{\lambda_1}{\lambda_0} - (\lambda_1 - \lambda_0)\Big]. \tag{4}$$

A comparison between W_n in (2) and $\hat{W}_n$ in (4) shows that $\hat{W}_n$ is a weighted version of W_n by normalizing the population sizes l_n's for each individual log-likelihood ratio statistic. Thus, we call the following procedure as weighted likelihood ratio (WLR) procedure:

$$T_{WLR}(b) = \text{ first } n \ge 1 \text{ such that } \hat{W}_n \ge b, \tag{5}$$

for some constant b.

The second alternative procedure we propose is to use the GLR-based statistic W_n in (2) but with adaptive thresholds that take into account of population sizes. Ideally, one would like to use the optimal thresholds or boundaries, say, by some Bayesian or non-Bayesian arguments, but such boundaries seem to be too complicated. Here we simply use the simple linear form of adaptive thresholds: $l_n c$. Specifically, the proposed adaptive threshold method (ATM) raises an alarm at time

$$T_{ATM}(c) = \text{ first } n \ge 1 \text{ such that } W_n \ge l_n c, \tag{6}$$

for some constant $c > 0$.

We should point out that when the population sizes l_n's are a constant $l > 0$, the above three detection schemes, $T_{GLR}(a), T_{WLR}(b)$ and $T_{ATM}(c)$, are equivalent (if $a = lb = lc$.) Moreover, it is well-known from Moustakides (1986) that they are (exactly) optimal in the problem of detecting a change in the mean (intensity) λ_i's of Poisson distribution from λ_0 to λ_1.

However, when the population sizes l_n's are not constant, then the above three detection schemes are not equivalent. Moreover, the theorem cannot be literally applied to establishing the optimality properties of Page's CUSUM procedure.

4 Theoretical Results: Step Function For Population Sizes

It is very challenging to investigate the change-point detection problem beyond the simplest i.i.d. models. In order to be tractable in theory as well as shed light on the practical situations, in this section (only), we consider the scenario when population sizes are modeled by a step function:

$$l_n = \begin{cases} l^{(0)}, \text{ if } n \le \omega \\ l^{(1)}, \text{ if } n \ge \omega \end{cases},$$

where ω is observable (since the l_n's are observable).

Intuitively, if $\omega >> \gamma$ in the false alarm constraint (1), e.g., $\omega = \infty$, then the problem is asymptotically equivalent to the scenario with the constant population size $l^{(0)}$. On the other hand, if $\omega << \log\gamma$, e.g., $\omega = 1$, then the problem is asymptotically equivalent to the scenario with the constant population size $l^{(1)}$. Also refer to Baron and Tartakovsky (2006).

In this article we are interested in the scenario when $\omega >> \log\gamma$ but $\omega << \gamma$. Such scenario is motivated from the growth curve (logistic) model for population size (see Section 5 below), which can be thought of as a smooth version of the step function.

When the population sizes are modeled by the step function, for our proposed detection schemes, two kind of changes are specially interesting: one is when $\nu = 1$ and the other is when $\nu = \omega$. Hence, motivated by Lorden's worst-case detection delay, we focus on the following detection delays:

$$D(T) = \max\left[\mathbf{E}_1(T),\ \text{ess sup}\ \mathbf{E}_\omega\Big((T-\omega+1)^+|\mathcal{F}_\omega\Big)\right].$$

The asymptotic properties of the proposed three detection schemes are summarized in the following theorem:

Theorem 1. *Assume that the population sizes l_n's follow the step function above, and assume that $\omega = \omega_\gamma$ satisfies $\log\gamma << \omega << \gamma$. Subject to the false alarm constraint in (1), we have*

$$D[T_{GLR}(a)] = (1+o(1))\frac{\log\gamma}{\min\{l^{(0)}, l^{(1)}\}I(\lambda_1,\lambda_0)}$$
$$D[T_{WLR}(b)] = (1+o(1))\frac{\log\gamma}{l^{(1)}I(\lambda_1,\lambda_0)}$$
$$D[T_{ATM}(c)] = (1+o(1))\frac{\log\gamma}{l^{(1)}I(\lambda_1,\lambda_0)},$$

as $\gamma \longrightarrow \infty$, where

$$I(\lambda_1,\lambda_0) = \lambda_1\log\frac{\lambda_1}{\lambda_0} - (\lambda_1-\lambda_0).$$

From the theorem, it is interesting to note that the detection delay of the GLR procedure $T_{GLR}(a)$ in (3) is asymptotically smaller than that of the $T_{WLR}(b)$ in (5) or $T_{ATM}(c)$ in (6) if and only if $l^{(0)} > l^{(1)}$. In particular, this indicates that the GLR procedure $T_{GLR}(a)$ may not be efficient if the population size increases, but it seems to be efficient if the population sizes decrease.

5 Example Revisited

We revisited the male thyroid data in New Mexico, and apply the proposed three detection schemes to the cancer data for the illustration purpose.

5.1 Model for Population Growth

In the literature, it is common (e.g., Pinheiro and Douglas, 2000) to model the growth curve by the following logistic model:

$$l_i = \psi(i) + \epsilon_i = \frac{\phi_1}{1+exp[-(i-\phi_2)/\phi_3]} + \epsilon_i, \tag{7}$$

where $E[\epsilon_i] = 0$ and $Var[\epsilon_i] = \sigma^2$. Here ϕ_1 indicates an asymptotic upper limit of population size, ϕ_2 the middle point of t in the S-shaped curve, and ϕ_3 the scale adjustment of time periods.

In our specific application of New Mexico, we fit this model to the population sizes in New Mexico by a nonlinear least-squares method (we treat year 1972 as time 0, and the population sizes are in the units of 100,000). Using the statistical software R version 2.8.0, the estimated parameters of the logistic models for the population sizes are

Table 1. Result of estimated parameters

Parameter	Estimate
ϕ_1	13.8065 ± 0.9552
ϕ_2	11.8532 ± 3.7438
ϕ_3	26.4037 ± 2.3127
σ	0.0907

After plotting the actual observed population sizes and the estimated growth curves in New Mexico during 1973-2005, we find that the two curves are very close to each other, implying that the logistic model is reasonable.

5.2 Parameters in Detection Schemes

To specify the pre-change rate λ_0 and the post-change rate λ_1, we consider the time period of 1973-1983 as training periods: the pre-change rate λ_0 and the post-change rate λ_1 are estimated by the median and the maximum of crude incidence per 100,000 during 1973-1983, respectively. For our data, we have $\lambda_0 = 2.4$ and $\lambda_1 = 3.8$.

We also assume that $\gamma = 300$ in the false alarm constraint (1), i.e., on average we want all detection schemes to raise a false alarm at least once every 300 years if the pre-change rate is $\lambda_0 = 2.4$. Of course, the choice of 300 is intended only for an illustration, and the idea can be easily extended to any other choices of false alarm constraints.

In order to satisfy the false alarm constraint (1) with $\gamma = 300$, for the three detection schemes, $T_{GLR}(a)$ in (3), $T_{WLR}(b)$ in (5) and $T_{ATM}(c)$ in (6), numerical simulations show that the corresponding threshold values are $a = 3.6870, b = 0.2975$, and $c = 0.2975$ (based on 100,000 replicates).

5.3 Detection Delays

If we control the false alarm constraint $\gamma = 300$, then the WLR and ATM procedures, $T_{WLR}(b)$ and $T_{ATM}(c)$ trigger an alarm in 1993, and the GLR procedure $T_{GLR}(a)$ raises an alarm until 1997. See Figure 2. Numerical simulations show that we reach the same conclusion if we control the false alarm constraint $\gamma = 100$ or 200.

In addition, for the purpose of comparison, for each of these three detection schemes, we also simulate $\text{ess sup}\Big[\mathbf{E}_\nu\big(T-\nu\Big|\mathcal{F}_\nu, T \geq \nu\big)\Big]$ at different change-point ν. The simulated detection delays are based on 50,000 replicates, and are illustrated in Figure 3.

From the plot, it is interesting to note that the GLR detection scheme $T_{GLR}(a)$ performs poorly if the change-point ν occurs at an earlier stage, but the properties of these three methods are similar if the change-point ν occurs at a very late stage. In fact, for the GLR detection scheme $T_{GLR}(a)$, its detection delays $\text{ess sup}\Big[\mathbf{E}_\nu\big(T-\nu\Big|\mathcal{F}_\nu, T \geq \nu\big)\Big]$ seem to be decreasing as a function of change-point ν. On the other hands, the detection schemes $T_{WLR}(b)$ and $T_{ATM}(c)$ seem to be "equalizer rule" in the sense that $\text{ess sup}\Big[\mathbf{E}_\nu\big(T-\nu\Big|\mathcal{F}_\nu, T \geq \nu\big)\Big]$ is same for different values of change-point ν. The property of "equalizer rule" is crucial to establish the exact optimality of the CUSUM procedure in the simplest i.i.d. models, and our results suggest that the GLR detection scheme $T_{GLR}(a)$ may not be efficient when the population sizes vary, especially when they increase over the years.

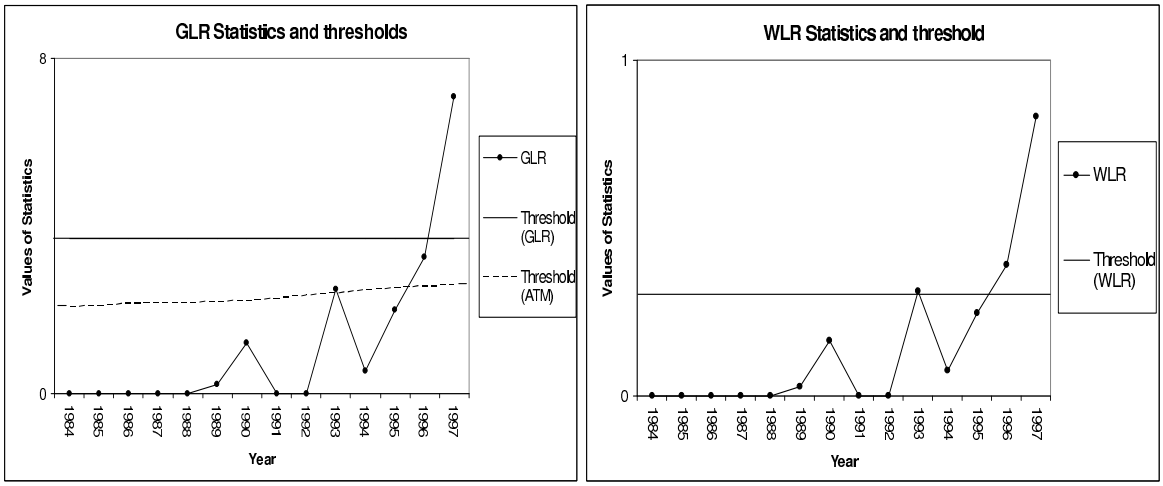

Fig. 2. Left Panel plots the GLR statistics W_n, where the solid line and the dotted line indicate the boundaries of $T_{GLR}(a)$ and $T_{ATM}(c)$, respectively. Right Panel is for the WLR statistic $\hat{W}_n$, and the solid line is the boundary of $T_{WLR}(b)$. Both $T_{WLR}(b)$ and $T_{ATM}(c)$ raise an alarm in year 1993, whereas $T_{GLR}(a)$ raises an alarm in year 1997.

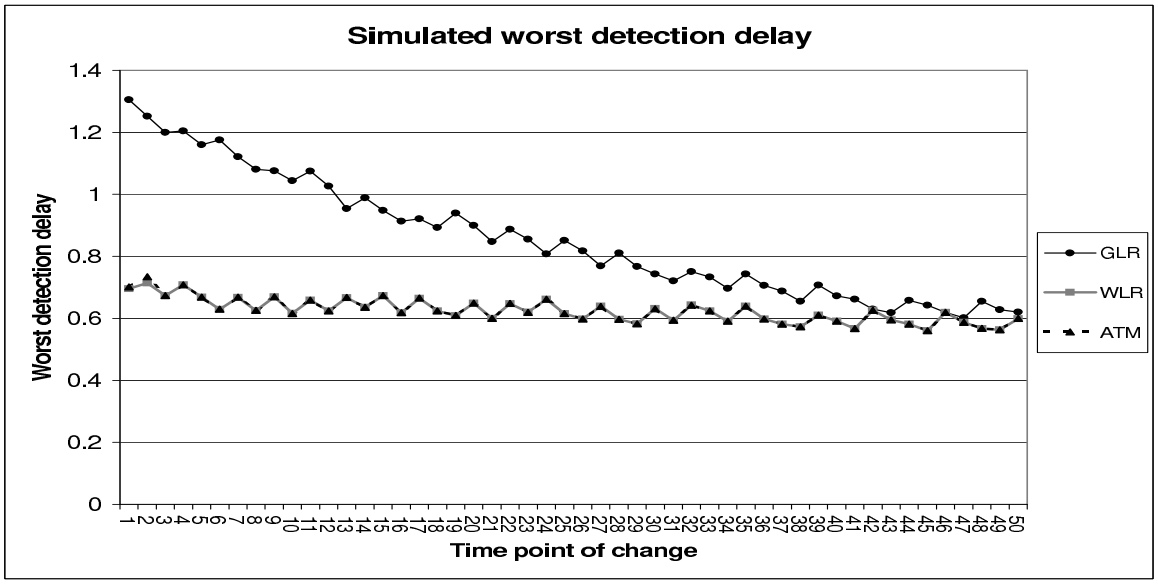

Fig. 3. The simulated detection delays of three detection schemes with respect to different values of true change-point ν. The simulations show that the GLR detection scheme $T_{GLR}(a)$ performs poorly if the change-point ν occurs at an earlier stage, but the properties of these three methods are similar if the change-point ν occurs at a very late stage.

Acknowledgements

The authors would like to thank Professor George Moustakides for simulating discussions. Y. Mei's research was supported in part by the AFOSR grant FA9550-08-1-0376 and the NSF Grant CCF-0830472.

References

Kulldorff, M. (2001). Prospective Time Periodic Geographical Disease Surveillance Using a Scan Statistic. *Journal of the Royal Statistical Society, Series A,* **164**: 61–72.

Pinheiro, J. C. and Douglas, M. B. (2000). *Mixed-Effects Models in S and S-PLUS*, New York: Springer.

Moustakides, G. V. (1986). Optimal Stopping Times for Detecting Changes in Distributions. *Annals of Statistics*, **14**, 1379–1387.

Lorden, G. (1971). Procedures for Reacting to a Change in Distribution. *Annals of Mathematical Statistics*, **42**, 1897–1908.

Baron, M. and Tartakovsky A. G. (2006). Asymptotic Optimality of Change-Point Detection Schemes Procedures in General Continuous-Time Models. *Sequential Analysis*, **25**, 257–296.

Peskir, G. and Shiryaev, A. (2006). *Optimal Stopping and Free-Boundary Problems.* Birkhuser Verlag, Basel, Switzerland.