The Canadian Journal of Statistics
Vol. 43, No. 3, 2015, Pages 403–419
La revue canadienne de statistique

Flexible risk-adjusted surveillance procedures for autocorrelated binary series

Edit GOMBAY[1]*, Abdulkadir A. HUSSEIN[2] and Stefan H. STEINER[3]

[1]*Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Alberta, Canada T6G 2G1*
[2]*Department of Mathematics and Statistics, University of Windsor, Windsor, Ontario, Canada N9B 3P4*
[3]*Department of Statistics and Actuarial Science, 200 University Avenue West, Waterloo, Ontario, Canada N2L 3G1*

Key words and phrases: Binary time series; health care performance; maximum partial likelihood estimator; sequential test.

MSC 2010: Primary 62L10; secondary 62M10

Abstract: Risk-adjusted cumulative sum (RACUSUM) charts are popular for the surveillance of binary health care outcomes such as 30-day mortality rates following cardiac surgery. RACUSUM charts are built on the assumptions that the binary outcomes are independent and the baseline rates are known constants. However, these two assumptions are often violated, thus undermining the validity of the surveillance procedure. In this paper, the authors propose risk-adjusted surveillance procedures using a binary logistic regression model which allows $AR(p)$-type autocorrelations among the binary outcomes. Two versions are presented: one with known, the other with estimated baseline parameters. The authors use Monte Carlo experiments to evaluate the power and the probability of false alarm (Type I error) of the surveillance procedures. Data on 30-day mortality rates following cardiac surgery are used for illustration. *The Canadian Journal of Statistics* 43: 403–419; 2015 © 2015 Statistical Society of Canada

Résumé: Les cartes à somme cumulée ajustées au risque (RACUSUM) sont des outils populaires pour la surveillance d'événements binaires tels que le décès dans les 30 jours suivant une chirurgie cardiaque. La construction des cartes RACUSUM est faite sous l'hypothèse que les événements binaires surveillés sont indépendants et que le risque de base est connu et constant. Ces deux hypothèses sont souvent violées, ce qui mine la validité de la procédure de surveillance. Les auteurs proposent une procédure de surveillance ajustée au risque en utilisant un modèle de régression logistique qui permet une autocorrélation de type $AR(p)$ entre les événements binaires modélisés. Ils présentent deux versions de leur modèle: dans la première, le risque de base est connu, mais dans la deuxième, il est estimé. À l'aide d'expériences de Monte Carlo, ils évaluent la puissance et la probabilité d'une fausse alarme (erreur de type I) de la procédure de surveillance. Ils illustrent leur méthode à l'aide de données sur la mortalité dans les 30 jours suivant une chirurgie cardiaque. *La revue canadienne de statistique* 43: 403–419; 2015 © 2015 Société statistique du Canada

1. INTRODUCTION

Statistical methods for the surveillance of health care outcomes have recently gained momentum. In particular, risk-adjusted cumulative sum (RACUSUM) charts have been developed for the surveillance of binary outcomes such as the 30-day mortality rates of patients undergoing heart surgery (see, for instance, Steiner et al., 2000; Frisen, 2003; Grigg & Farewell, 2004; Grigg & Spiegelhalter, 2010; Gombay, Hussein, & Steiner, 2011).

* *Author to whom correspondence may be addressed.*
E-mail: egombay@ualberta.ca

In general, apart from assumptions concerning the functional form of the risk model, there are two major assumptions underlying the RACUSUM procedure. The first is the assumption that the binary observations are independent over time, the second that the historically estimated baseline parameters can be treated as known numbers. Concerning the independence assumption it has been shown in a simulation study that, if the monitored sequence of binary observations has some sort of serial dependence, then the theoretical average run lengths (ARL) of RACUSUM charts are greatly affected (Hussein et al., 2015). Similarly, errors incurred in estimating the baseline rates can affect negatively the performance of the RACUSUM chart (Jones & Steiner, 2012). As far as the authors are aware the current literature on sequential testing does not address both of these issues simultaneously. Furthermore, the new methods of this paper have the desirable feature of simultaneous surveillance of several coefficients that is not available elsewhere. Other methods, such as cumulative sum (CUSUM) charts, monitor for the presence of change without providing information about its cause.

Höhle (2010) proposed a CUSUM procedure based on the generalized likelihood ratio statistic for surveillance of categorical time series. Although such methodology is quite general and applicable to a large class of categorical time series regression models, it carries several major drawbacks that are common to CUSUM procedures. These are: (1) the parameters under the two simple hypotheses are assumed to be known, (2) inflated false alarm rates, and (3) numerical complexity in computing average run lengths requiring knowledge of the distributions of the covariates. The literature on change-point analysis in the context of continuous responses is vast. The reader is referred to Gombay (2008), Gombay & Serban (2009), and citations therein.

Our objective in this manuscript is to propose new sequential risk-adjusted surveillance procedures for the coefficients of a logistic regression model with the following features that are not available in the current literature: (1) the binary responses are allowed to have an $AR(p)$-type serial dependence, (2) the error due to the estimation of the baseline parameters from a historical sample is accounted for, (3) the probability of false alarms is controlled, and (4) several regression coefficients can be monitored simultaneously. While in Fokianos et al. (2014) retrospective (offline) methods were considered, the current paper proposes sequential (online) algorithms. Although both of these proposals (Fokianos et al. and the current method) are based on the same likelihood functions, the sequential, prospective nature of the current proposal requires further theoretical considerations to verify the validity of the algorithms.

The new procedures will be proposed in Section 2. In Section 3, we use Monte Carlo simulations as well as real data on 30-day mortality rates following cardiac surgery to demonstrate the performance of the procedures. The technical proofs are contained in the Appendix.

2. MODEL AND SURVEILLANCE PROCEDURES

2.1. The Model and Hypotheses of Change

Monitoring 30-day mortality rates through risk-adjusted surveillance methods is a special instance of monitoring the parameters of a general logistic regression model. In particular, monitoring whether or not the odds of an adverse event for a particular surgeon are different from those of the baseline, after controlling for patient case mix, is tantamount to testing the hypothesis of change in the parameters of a logistic regression model. In this manuscript we will therefore present general surveillance procedures for monitoring changes in the coefficients of a logistic regression model, and then illustrate how this can be adapted to the case of monitoring 30-day mortality rates.

Consider a binary time series $\{Y_t\}$ with probability of success $\pi_t(\beta)$ depending on an unknown vector $\beta \in \Re^p$ of parameters along with a corresponding p-dimensional vector of covariates $\{Z_t\}$. Following Kedem & Fokianos (2002), let us denote the history of the binary process and past covariate information at time t by $\{\mathcal{F}_{t-1}\}$: a filtration generated by Z_{t-1} and, possibly, by some

variables X_t that may be known. Covariates Z_{t-1} are allowed to include lagged values of the binary response Y_t, thus permitting an $AR(p)$-type serial dependence over time. The conditional probability mass function of the series $\{Y_t\}$ is given by

$$f(y_t;\beta|\mathcal{F}_{t-1}) = \exp\left\{ y_t \log\left(\frac{\pi_t(\beta)}{1-\pi_t(\beta)}\right) + \log(1-\pi_t(\beta))\right\}, \tag{1}$$

while the dependence on the covariate vector $\{Z_t\}$ is modelled through the logit link function as

$$g(\pi_t(\beta)) = \eta_t = \log\left(\frac{\pi_t(\beta)}{1-\pi_t(\beta)}\right) = \beta^{'} Z_{t-1}, \tag{2}$$

where $\beta \in \Re^p$.

In order to formulate the surveillance procedure in the change-point setup, suppose that a sequence of observations $y_1, y_2, \ldots$ generated by the logistic model (1) is available for testing the following hypotheses of change in the β parameter:

$$H_0 : \beta = \beta_0, \ \text{ for all } \pi_t(\beta),\ t = 1, 2, \ldots, n,$$
$$H_A : \beta \neq \beta_0, \ \text{ for all } \pi_t(\beta),\ t = 1, 2, \ldots, n,$$

where β_0 is the baseline vector of coefficients. In some cases β_0 is a known parameter, in others it is estimated as $\hat{\beta}_m = (\hat{\beta}_{1m}, \hat{\beta}_{2m}, ..., \hat{\beta}_{pm})$ using a historical sample of size m collected prior to the initiation of the surveillance process.

In this manuscript we propose procedures based on testing the above hypotheses of change via the partial score statistics. Usually, it is not desirable to have an open-ended monitoring procedure for many practical reasons. We therefore set a horizon (a maximal sample n) by which, if H_0 has not been rejected, the surveillance procedure will be re-started. This makes our surveillance procedure a truncated one, thus providing better control over the probability of false alarms as in Gombay, Hussein, & Steiner (2011). The choice of n is beyond the scope of the current study, but it is related to the same choice in group sequential tests where it has been extensively studied. In fact, our Test 1 of Section 2.2 is the continuous version of Pocock's (1977) group sequential test.

In general, inferences concerning the binary time series model (1) are based on the binomial log-partial likelihood function

$$L(\beta) = \sum_{t=1}^{k} l_t(\beta) = \sum_{t=1}^{k} \left[y_t \log \frac{\pi_t(\beta)}{1-\pi_t(\beta)} + \log(1-\pi_t(\beta))\right],$$

with p-dimensional score vector

$$S_k(\beta) = \sum_t \nabla_\beta l(\beta) = \sum_{t=1}^{k} Z_{t-1}(Y_t - \pi_t(\beta)) = \sum_{t=1}^{k} Z_{t-1}\left(Y_t - \frac{\exp(\beta' Z_{t-1})}{1+\exp(\beta' Z_{t-1})}\right), \tag{3}$$

where k is the number of observations available at the time of analysis. The maximum partial likelihood estimator (MPLE) based on the m historical observations, denoted by $\hat{\beta}_m$, is a solution (provided that it exists) of the score equations using only the historical data. The asymptotic properties of the score vector and the MPLE $\hat{\beta}_m$ have been studied by Fokianos, Gombay, & Hussein (2014) who showed that the score vector can be approximated by a p-dimensional Brownian motion with optimally small rate of error. Those results are necessary for the validity of the surveillance procedure described in the next subsections.

2.2. Surveillance When the Baseline Parameters are Known

In what follows the superscript (i) indicates the ith component of a vector. The surveillance procedure proposed in this section is based on the standardized score vector where the standardizing

matrix is the inverse of the observed Fisher information matrix. To construct the surveillance procedure we first introduce the observed Fisher information matrix of the score vector (3) computed at the known baseline parameters, β_0, as

$$T_k(\beta_0) = \frac{1}{k}\sum_{t=1}^{k} Z_{t-1}Z'_{t-1}\pi_t(\beta_0)(1-\pi_t(\beta_0)). \tag{4}$$

It has been shown in Fokianos, Gombay, & Hussein (2014) that, under conditions (A–C) of Appendix A.1, the observed information matrix, $T_k(\beta_0)$, is a consistent estimator of the true Fisher information

$$T = E(Z_{t-1}Z'_{t-1}\pi_t(\beta)(1-\pi_t(\beta))). \tag{5}$$

In Appendix A.2 we will prove that, under these same conditions, the standardized score vector is well approximated by an Ornstein–Uhlenbeck process. Consequently, we obtain an approximate distribution for the maximum of the standardized score vector via results in Vostrikova (1981) as follows:

$$P\{\sup_{1<k\le n} k^{-1/2}|[T_k^{-1/2}(\beta_0)S_k(\beta_0)]^{(i)}|>u\} \cong \frac{\exp(-u^2/2)y}{\sqrt{2\pi}}\left\{N\left(1-\frac{1}{u^2}\right)+\frac{4}{u^2}+O\left(\frac{1}{u^4}\right)\right\}, \tag{6}$$

where $N = \log n$ and $\cong$ means that the ratio of the two sides converges to one as $n \to \infty$. This allows us to define the following surveillance procedure:

Test 1: *The null hypothesis of in-control is rejected at the k*th *observation of the binary sequence, $1 < k \le n$, if for some i, $i = 1, 2, \ldots, p$, the absolute value of the standardized score component corresponding to the i*th *coefficient crosses the horizontal boundary $C_1(\alpha^\star, n)$. That is, as soon as for some i, $i = 1, 2, \ldots, p$,*

$$k^{-1/2}\left|\left(T_k^{-1/2}(\beta_0)S_k(\beta_0)\right)^{(i)}\right| \ge C_1(\alpha^\star, n).$$

In this testing procedure, $\alpha^\star = 1-(1-\alpha)^{1/p}$ is the probability of false alarm in monitoring the ith component of the logistic regression coefficient, α is the overall probability of false alarm in testing for a change in any coefficient, and n is the surveillance horizon after which the monitoring process will be reset. The critical level (threshold for surveillance), $C_1(\alpha^*, n)$, can be obtained from Equation (6); the statistic is calculated with the known β_0 values and the incoming observations $\{y_t, Z_t\}$. At each stage k the standardizing matrix T_k is recalculated. To avoid initial small sample estimation problems, testing should start at some $k = n_0$, and should obviously continue until the horizon n is reached or an alarm is raised by crossing the threshold. In our studies an initial sample size $n_0 = 30$ worked well.

2.3. Surveillance When the Baseline Parameters are Estimated

Now suppose that we have a historical sample of size m and the baseline parameters are estimated by $\hat{\beta}_m$. It has been shown in Fokianos, Gombay, & Hussein (2014) under conditions (A–D) of Appendix A.1 that the observed information matrix, $\hat{T}_m = \hat{T}_m(\hat{\beta}_m)$, is a consistent estimator of the true Fisher information (5) (where beta denotes the true model parameter vector). The key theoretical result needed for our proposed surveillance procedure is that the maximum of the standardized score vector, computed at the historically estimated baseline parameters, can be approximated component-wise by the supremum of the standard Brownian motion process. This

result, which will be proven in Appendix A.3, is formulated as

$$m^{-1/2} \max_{1<k\le n} \left(\hat{T}_m^{-1/2} \left(1+\frac{k}{m}\right)^{-1} S_k(\hat{\beta}_m) \right)^{(i)} \xrightarrow{D} \sup_{0\le s\le j/(j+1)} W(s), \tag{7}$$

where $W(.)$ denotes a one-dimensional standard Brownian motion, $\hat{T}_m = \hat{T}_m(\hat{\beta}_m)$ is the historically observed information matrix, and $j = n/m$. The constant j is always well defined since the historical sample size (m) and the monitoring horizon (n) are both known at the planning stage. Now we are in a position to formalize the proposed surveillance procedure.

Test 2: *The null hypothesis of in-control is rejected at the k*th *observation of the binary sequence, $1 < k \le n$, if for some i, $i = 1, 2, \ldots, p$, the absolute value of the standardized score component corresponding to the i*th *coefficient crosses the horizontal boundary $C_2(\alpha^\star, j)$. That is, as soon as for some i, $i = 1, 2, \ldots, p$,*

$$m^{-1/2} \left| \left(1+\frac{k}{m}\right)^{-1} \left(\hat{T}_m^{-1/2} S_k(\hat{\beta}_m) \right)^{(i)} \right| \ge C_2(\alpha^\star, j). \tag{8}$$

The threshold $C_2(\alpha^*, j)$ is computed by taking the $1-\alpha^\star$ quantile of the distribution of $\sup_{0<s<1} \sqrt{j/(j+1)}|W(s)|$, where $W(s)$ is a one-dimensional standard Brownian motion. Again, component-wise level $\alpha^\star = 1-(1-\alpha)^{1/p}$ is the probability of false alarm in monitoring the ith regression coefficient of the model (1), while α is the overall probability of false alarm for a change in any coefficient.

It is worth emphasizing that our surveillance procedures have the ability to monitor each regression coefficient separately while controlling the overall false alarm rate, due to the asymptotic independence of the standardized score vector used in building the surveillance procedure. This gives the flexibility of choosing only some, or all, of the coefficients for monitoring purposes.

3. EMPIRICAL STUDIES

3.1. Changes in Mortality Rate After a Cardiac Surgery

We illustrate the proposed methodology by using data collected at a UK center for cardiac surgery. The data consist of patients' pre-operative covariate information such as age, gender, history of hypertension, etc., which were summarized as patient Parsonnet scores (see Steiner et al., 2000). These data have been used in the past for the purpose of monitoring surgeons' relative performance as compared to population baseline via RACUSUM and other methods. It is commonly assumed that 30-day mortality can be adequately described by a logistic regression with the Parsonnet score as the only covariate in the model. However, our analysis suggests that the data pertinent to some of the surgeons have an $AR(p)$-type dependence structure. We consider the data pertinent to Surgeon #6 who operated on approximately 1,655 patients over the period 1992–1998. We take the data of 1992–1993 as our historical sample of size $m = 450$ and we prospectively monitor the regression coefficients by using data over the period 1994–1998. Let Y_t be the indicator of mortality within 30 days of the surgery, that is, $Y_t = 1$ if the patient died within 30 days of surgery, and zero otherwise. Let π_t be the probability of such an adverse event for the tth and X_t their Parsonnet score. For model identification, we used a variable selection approach based on backward elimination and the AIC, as suggested in Kedem & Fokianos (2002), and obtained $\text{logit}(\pi_t) = \beta_1 + \beta_2 X_t + \beta_3 Y_{t-2}$ as the best fitting model with an AIC of 168.15. Our purpose is to demonstrate the ease with which our algorithms can be applied, so we assume that this is the correct model, even though the use of this selection criterion may be questioned. Surgeon

TABLE 1: Estimated coefficients of the model $\text{logit}(\pi_t) = \beta_1 + \beta_2 X_t + \beta_3 Y_{t-2}$ for Surgeons #6 and #7, based on historical samples of size $m = 450$ and $m = 243$, respectively.

Coefficient	Estimate	Std. error	P-value
	Estimated coefficients for Surgeon 6		
β_1	−3.781	0.325	2×10^{-16}
β_2	0.093	0.020	4×10^{-6}
β_3	1.597	0.614	0.009
	Estimated coefficients for Surgeon 7		
β_1	−4.726	0.617	2×10^{-14}
β_2	0.120	0.024	7×10^{-7}
β_3	2.177	0.809	0.007

#7 had data with similar characteristics. The fitted model's estimated parameters, their standard errors and corresponding P-values are reported in Table 1. We do not include more covariate components than necessary, such as a Y_{t-1} term, as increasing the number of parameters leads to tests with smaller power. This can be seen from the formula for α^*, which decreases as the dimension increases.

We applied our two prospective surveillance procedures to the same segment of the data (collected in the period 1994–1998). For Test 1, the known baseline values were obtained from the collective performances of a large number of surgeons, considered as acceptable population parameters. As shown in Figure 1, Test 1 is significant and it detected a change in the β_2 component at observation $k = 693$. Having the full data sequence we can show that the other two components' monitoring would not indicate instability. On the other hand, when monitoring the surgeon's performance against its own history by Test 2 with baseline estimated as $\hat{\beta}_m$, $m = 450$, the surveillance procedure, as seen in Figure 2, does not indicate any change in the coefficients.

3.2. Monte Carlo Simulations

We carried out Monte Carlo simulations to assess the empirical false alarm rates (Type I errors) and the power of the surveillance procedures. The data generating model was fitted to the data for surgeon #7 with coefficients reported in Table 1. Parsonnet scores X_t were randomly sampled (with replacement) from the 628 available Parsonnet scores for the patients operated on by Surgeon #7. We used $n = 300$ and $n = 400$ as monitoring horizons while the historical sample size was varied over $m = 300$ to $m = 800$ with steps of 100 and $m = 1{,}000$, 2,000. A second set of simulations used a $n = 9{,}000$ monitoring horizon with varying historical sample sizes closely matching situations when a lot of data are available. Each scenario was repeated 2,000 times to generate the probabilities of false alarm and the power of the surveillance procedure. All the simulations and tests were implemented in FORTRAN using the IMSL libraries and the results are reported in Tables 2–6.

In general, the false alarm rate (Type I error) varies somewhat as the historical sample size m changes, but it is fairly close to the target of 0.05 and improves as m increases. Also, for large n, m, the asymptotic independence of the surveillance procedures for the various coefficients can be seen from the results of the simulations. For example, in Table 2, when $n = 400$ and the historical sample size is $m = 2{,}000$, the overall probability of rejecting the true H_0 is $\hat{\alpha} =$

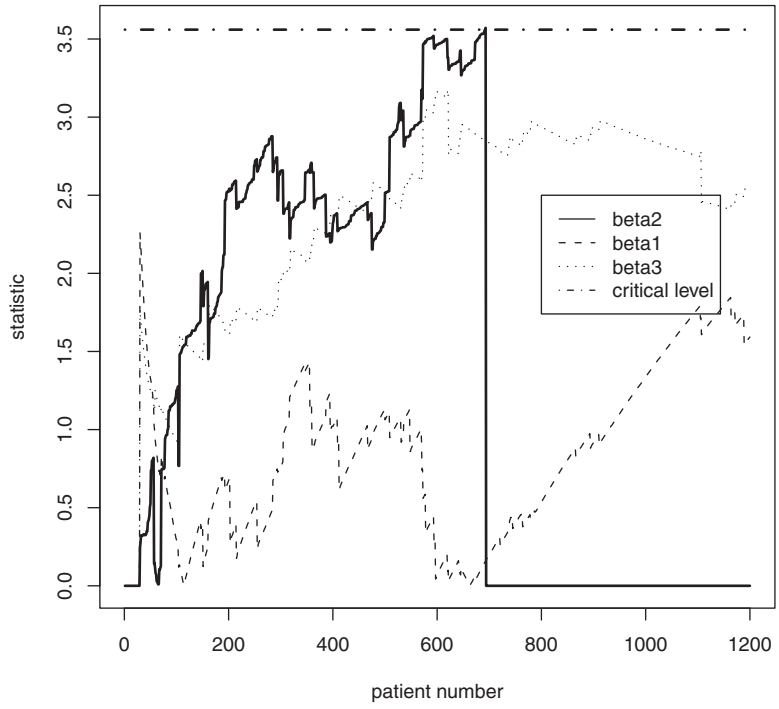

FIGURE 1: Performance of Surgeon #6 monitored with Test 1 over a finite horizon $n = 1{,}200$, using baseline parameters obtained form a large number of surgeons' performances. The critical level 3.56 is reached at $k = 693$ where testing would stop with change in β_2 detected. (We show the full history of the other two components for information only.)

0.0485 which is very close to the nominal $\alpha = 0.05$, and the probability of not rejecting H_0 is $(1 - 0.019)(1 - 0.014)(1 - 0.0195) = 0.9484$, which is close to $1 - \hat{\alpha} = 0.9515$ and to the theoretical $1 - \alpha = 0.95$, thus supporting our theoretical asymptotic independence. The values of m at which the asymptotic independence works and the theoretical critical level is good enough may vary from model to model, so we recommend that Monte Carlo experiments be used in order to evaluate the appropriate value of m.

Following the suggestion of a referee, we have also examined the effect of model selection on the Type I error of the procedures. For simplicity we only report results for Test 2 in Table 6. In these simulations, we generated a historical sample of $m = 2{,}000$ as well as a sample of size $n = 400$ for surveillance from each one of five true models given by $\text{logit}(\pi_t) = -4.70 + 0.12X_t + \beta_3 Y_{t-1}$ with β_3 varying over the values $\{0.2, 0.4, 0.6, 0.8, 1\}$. These values of β_3 are designed to result in a spectrum of significant and non-significant values. In each Monte Carlo run we fitted the true model and a model with $\beta_3 = 0$ to the historical data, choosing the one with the smaller AIC. We then prospectively monitored the coefficients of the chosen model by using Test 2. Over the 2,000 Monte Carlo runs, we recorded the average marginal P-values of the coefficient β_3 for the true model, the average proportion of times that the AIC selected the true model, and the overall as well as component-wise Type I errors of the surveillance procedure for the model coefficients. See Table 6 for the outcomes. When β_3 is marginally insignificant, the AIC chooses the correct model only 23–39% of the time, increasing to 95% when $\beta_3 = 1$, which is a highly significant case. Clearly, the overall Type I errors of Test 2 are not much affected, even when the wrong model is selected most of the time. The marginal Type I errors of the stable coefficients, (β_1, β_2) are slightly inflated while those of β_3 are slightly conservative.

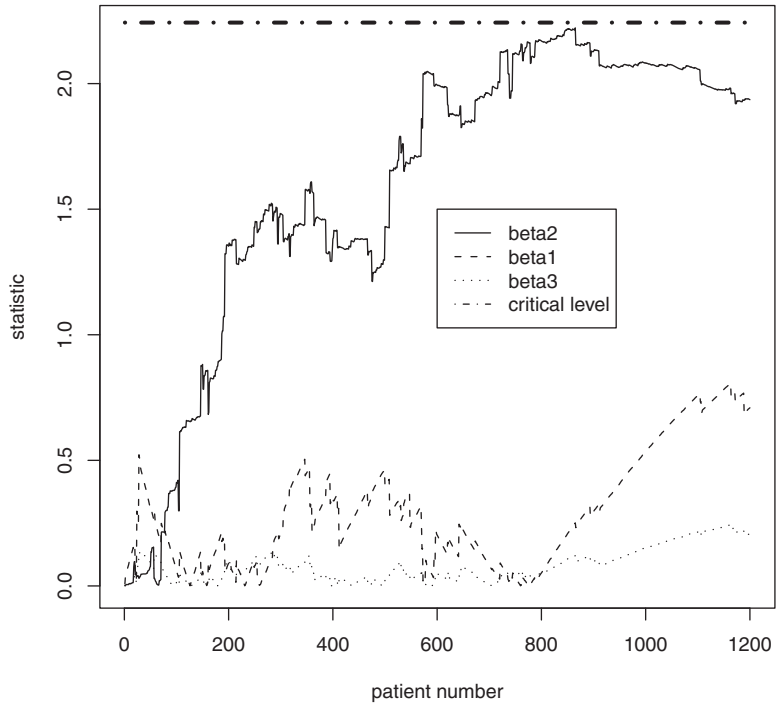

FIGURE 2: Performance of Surgeon #6 monitored with Test 2 over a finite horizon $n = 1{,}200$, using historical sample of size $m = 450$ of the same surgeon for estimating the baseline parameters. The critical level 2.24 is not reached for any component. (Maximum is 2.2219 at $k = 865$ for β_2.)

Figure 3 shows the simulated power properties of the surveillance procedure Test 2. Here we matched the in-control parameters to the data on surgeon #7 as $\beta_0 = (-4.70, 0.12, 2.2)'$, set $m = 600$ for the size of the historical data and $n = 1{,}000$ for the monitoring horizon. The Parsonnet scores were randomly generated from this surgeon's data. In these simulations we changed β_{02}, the coefficient of the Parsonnet score, from the in-control value $\beta_{02} = 0.12$ to an alternative value $\beta_{A2} = 0.17$ in steps of 0.01. The results, summarized in Figure 3, clearly show that the coefficient

TABLE 2: Simulated Type I error of Test 2 for $n = 400$ (bottom), $n = 300$ (top), overall nominal level $\alpha = 0.05$ with $\alpha^* = 0.01695$ for monitoring each of the three coefficients and baseline parameters estimated from historical sample of size m.

Coefficient	$m = 300$	$m = 400$	$m = 500$	$m = 600$	$m = 700$	$m = 800$	$m = 1{,}000$	$m = 2{,}000$
β_1	0.0490	0.0340	0.0300	0.0280	0.0240	0.0205	0.0230	0.0190
β_2	0.0290	0.0200	0.0235	0.0210	0.0235	0.0175	0.0185	0.0165
β_3	0.0405	0.0305	0.0315	0.0210	0.0235	0.0280	0.0260	0.0170
Combined	0.0980	0.0740	0.0750	0.0640	0.0665	0.0580	0.0610	0.0510
β_1	0.0405	0.0415	0.0290	0.0265	0.0230	0.0265	0.0210	0.0190
β_2	0.0255	0.0265	0.0240	0.0190	0.0150	0.0150	0.0210	0.0140
β_3	0.0415	0.0280	0.0225	0.0215	0.0260	0.0295	0.0250	0.0195
Combined	0.0915	0.0850	0.0695	0.0645	0.0585	0.0665	0.0585	0.0485

TABLE 3: Simulated Type I error of Test 2 for $n = 9{,}000$, overall nominal level $\alpha = 0.05$ with $\alpha^* = 0.01695$ for monitoring each of the three coefficients and baseline parameters estimated from historical sample of size m.

Coefficient	$m = 3{,}000$	$m = 5{,}000$	$m = 7{,}000$	$m = 9{,}000$	$m = 11{,}000$
β_1	0.0225	0.0190	0.0205	0.0215	0.0185
β_2	0.0130	0.0160	0.0230	0.0180	0.0150
β_3	0.0160	0.0210	0.0180	0.0140	0.0180
Combined	0.0495	0.0550	0.0605	0.0520	0.0510

TABLE 4: Power simulation results for Test 2 with $m = 600$ and $n = 1,000$.

Test component	$\beta_0 = 0.12$	$\beta_{A2} = 0.13$	$\beta_{A2} = 0.14$	$\beta_{A2} = 0.15$	$\beta_{A2} = 0.16$	$\beta_{A2} = 0.17$
β_1 Test	0.0285	0.0355	0.0530	0.0905	0.1500	0.2520
β_2 Test	0.0215	0.0900	0.3810	0.7765	0.9680	0.9970
β_3 Test	0.0280	0.0460	0.0845	0.1265	0.1885	0.245
Overall	0.0715	0.1460	0.4345	0.7935	0.9700	0.9970

Overall nominal level $\alpha = 0.05$ with $\alpha^* = 0.01695$ for each of the three monitoring. In control parameters are estimated with historical data of size m. Only the parameter in the column heading changed.

TABLE 5: Power simulation results for Test 1 with $n = 1{,}000$.

Test component	β_0	$\beta_{A2} = 0.13$	$\beta_{A2} = 0.14$	$\beta_{A2} = 0.15$	$\beta_{A2} = 0.16$
β_1 Test	0.0375	0.0360	0.0520	0.0795	0.1445
β_2 Test	0.0050	0.0945	0.6410	0.0980	0.9990
β_3 Test	0.0310	0.0370	0.0360	0.0310	0.0315
Overall	0.0735	0.1625	0.6675	0.9830	1.000
	$\beta_{A1} = -4.0$	$\beta_{A1} = -3.7$	$\beta_{A1} = -5.5$	$\beta_{A3} = 3.0$	$\beta_{A3} = 3.5$
β_1 Test	0.8615	0.9985	0.01750	0.1320	0.4460
β_2 Test	0.9580	0.9995	0.8220	0.0460	0.1720
β_3 Test	0.0935	0.2080	0.0190	0.6670	0.9910
Overall	0.9935	1.000	0.8460	0.7115	0.9950

Overall nominal level $\alpha = 0.05$ with $\alpha^* = 0.01695$ for each of the three monitoring. Parameter β_0 is given. Only the parameter in the column heading changed.

affected by the change is the one triggering the alarm the fastest in most of the cases and hence contributes the most to the power of the surveillance procedure. For instance, when β_{02} is changed from 0.12 to $\beta_{A2} = 0.15$, the power corresponding to the monitoring of that component is 0.7765 while the overall power is 0.7935. The 0.017 difference accounts for the proportion of cases when alarm was raised by a component other than the one corresponding to β_2. As the difference between the parameter value after change and its in-control value increases, the difference in power

TABLE 6: Type I errors of Test 2 when the true model, $\text{logit}(\pi_t) = -4.70 + 0.12X_t + \beta_3 Y_{t-1}$, is selected by AIC criteria and distinguished from a model with $\beta_3 = 0$.

	β_3				
Coefficient	0.2	0.4	0.6	0.8	1
β_1	0.019	0.028	0.025	0.014	0.008
β_2	0.018	0.029	0.024	0.022	0.019
β_3	0.007	0.004	0.006	0.018	0.019
Combined	0.042	0.061	0.051	0.053	0.046
AIC correct choice	0.231	0.388	0.617	0.841	0.952
P-val of β_3	0.302827	0.084603	0.004895	0.00003	0.00000

The coefficient β_3 is varied from insignificant to significant values and simulated P-values as well as the proportion of times that the AIC catches the correct model are reported (rows headed by P-val and AIC). Here $n = 400$, $m = 2{,}000$.

between the overall monitoring and the monitoring of the specific component in which the change occurred decreases. Also, we show the power exhibited by coefficients that did not change, and they are quite small (as desired). This clearly indicates that, although we are monitoring multiple coefficients at the same time, the investigator can visually see which coefficients have changed.

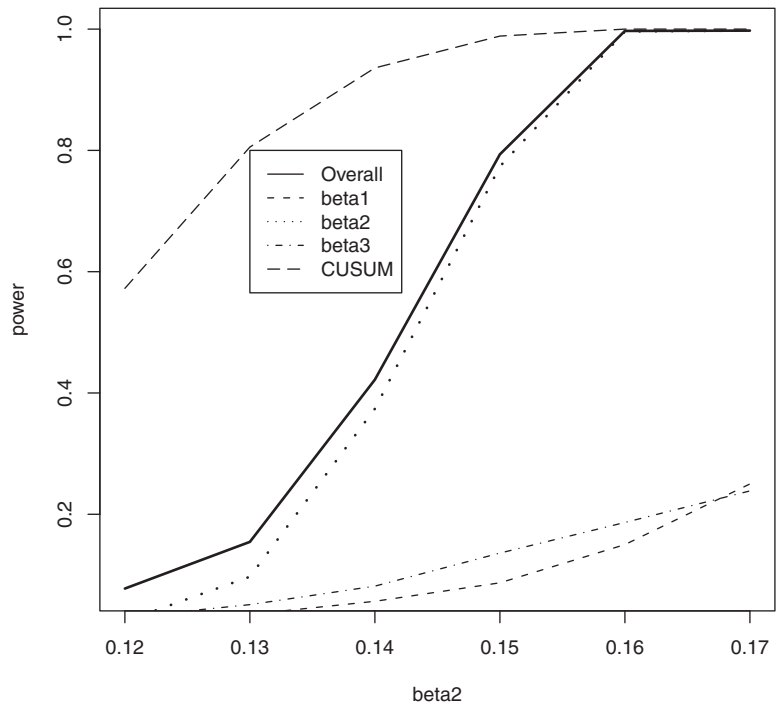

FIGURE 3: Simulated power of the surveillance procedure Test 2 with $n = 1{,}000$, $m = 600$, overall level $\alpha = 0.05$ and $\alpha^* = 0.01695$ for monitoring each of the three coefficients.

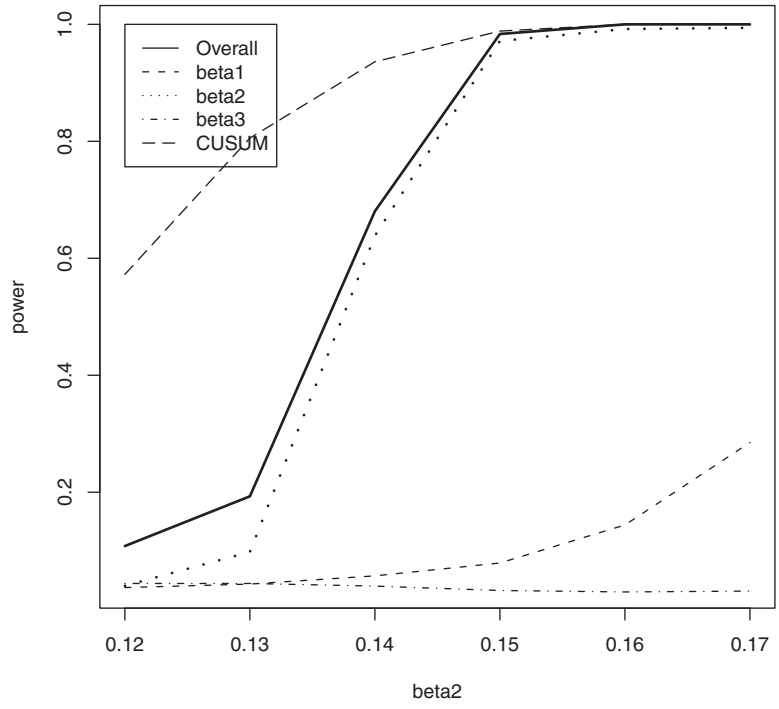

FIGURE 4: Simulated power of the surveillance procedure Test 1 with $n = 1{,}000$, overall level $\alpha = 0.05$ and $\alpha^* = 0.01695$ for monitoring each of the three coefficients.

In Figure 3 we included the performance of RACUSUM for comparison. The very same data were monitored with critical level h designed for average run length under H_0, ARL = 1,000, identical to our truncation point. The alternative required in the process was set as $R = 2$ for the increase in odds ratio, which is $R = 1$ under the null hypothesis. Although RACUSUM is open ended, we monitored the data only up to n=1,000, but even so, the figure shows that the Type I error exceeds 0.5, making the process unreliable. In other words, although the graph shows greater power of RACUSUM initially, much of it is attributable to the inflated empirical level making the power comparison unrealistic. Table 4 gives a detailed record of Monte Carlo studies concerning the power properties of the two-sided Test 2. Again, to match the parameters of the surgeon data study, we set $m = 600$ for the size of the historical data and $n = 1{,}000$ for the truncation point.

Figure 4 presents results of simulations of monitoring data by Test 1. The same data were generated as for Figure 3, except the in-control parameter vector is not estimated, but given as $\beta_0 = (-4.72, 0.12, 2.18)'$. Testing started at $n_0 = 30$ and from that point on the standardizing matrix T was estimated. More details about Test 1 are provided in Table 5, and they show general good performance.

The tests are designed using large sample approximations, so it is not surprising that they are somewhat, but not seriously, anti-conservative, and have good powers. Often the component of the test statistics process that is corresponding to the changing parameter is the earliest to trigger stopping. However, there may be values of β where this is not observable. For this reason, we recommend Monte Carlo studies to assess the performance of the test, which is possible as the algorithm is fast and easy to implement.

4. CONCLUSION AND COMPARISONS

In this paper we proposed two risk-adjusted surveillance procedures for monitoring the coefficients of a logistic regression model. The proposed methods are simple to implement and the output of both is an easily interpretable graph. These new methods give sequential surveillance procedures that enrich the family of available algorithms in several important ways. First, Test 2 does not assume that the baseline parameters of the logistic regression model are known, instead such parameters are estimated from historical data and so it accounts for the variability therein. Second, the proposed surveillance procedures allow for dependence on the past outcomes of the binary series, thus accommodating possible autocorrelations. Third, these tests can be used to monitor all the logistic regression model coefficients simultaneously, but can also monitor a selected subset only. This allows investigators to not only keep an eye on the mortality rate changes due to surgeons but also changes in the relationship between rates and other covariates such as the Parsonnet scores.

Currently, the most popular monitoring statistics are based on Page's CUSUM strategy, where independence of observations is assumed and specification of the alternative parameter value under the alternative H_A is required in the likelihood ratio statistic. Gombay, Hussein, & Steiner (2011) compared the performance of the efficient-score-based procedure to the RACUSUM-based surveillance in case of independent observations via Monte Carlo simulations. In our current simulation study we included RACUSUM to show what would happen if users chose that method instead of our tests. Note that the dependence structure is not accommodated in RACUSUM or in Gombay, Hussein, & Steiner (2011), so this is a new feature of this study. Furthermore, Kulldorff et al. (2011) demonstrated that the performance of the CUSUM test is very sensitive to the choice of the alternative parameter value. In contrast, the user does not have to specify such an alternative parameter value in the score vector, hence its performance does not depend on an often arbitrary selection. There are various versions of the CUSUM process in use, but their studies seldom focus on Type I errors as we do. An exception is, for example, Nishina & Nishiyuki (2003) who compare the performance of two one-sided CUSUM statistics designed to detect change in independent Normal observations with known initial mean and variance.

We provided Monte Carlo simulations to examine the performance of the procedures in terms of false alarm rates (Type I errors), that is, the error of signalling change in a parameter which is in fact stable. We also applied the procedure to the monitoring of 30-day mortality rates after cardiac surgery. Because of the simple structure of the test, it is easy to perform simulation studies to explore its performance for various models and various parameter combinations, and we recommended it before starting any surveillance.

APPENDIX

A.1. Regularity conditions

For ease of notation let the ith component of a p-dimensional vector x be denoted as x^i, $i = 1, \ldots, p$. Under the null hypothesis of no change, we need the following conditions on the covariate process and parameter β.

(A) $\{Z_k\}$ is ergodic and stationary in the sense that for all $k \geq 0$ $(Z_{k+1}, Z_{k+2}, \ldots)$ has the same distribution as $(Z_0, Z_1, \ldots)$.

(B) $E[Z_{k-1}^i]^4 < \infty$, $i = 1, \ldots, p$, where Z_{t-1}^i, $1 \leq i \leq p$, are the components of vector Z_{t-1}.

(C) For all components $i, j = 1, \ldots, d$

$$E\left|\frac{1}{n}E\left(\sum_{t=1}^{n}(z^i z^j)^2|\mathcal{F}_0\right) - E\left(z^i z^j\right)^2\right| \to 0,\ n \to \infty.$$

(D) The true value of β is in an open subset of the parameter space Ω, $\Omega \subset \Re^p$.

Conditions (A), (B), and (D) are standard. Condition (C) is a technical requirement in the proofs; its heuristic meaning is that the distant past is forgotten in this sense. See Serfling (1968) on more explanations, and possible alternative conditions. Note that from condition (B) by the ergodic theorem we have

$$\frac{1}{n}\sum_{t=1}^{n} Z_t^i Z_t^j \to^{a.s} E(Z_t^i Z_t^j), \quad n \to \infty,$$

$$\frac{1}{n}\sum_{t=1}^{n} Z_t^i Z_t^j \pi_t(\beta)(1-\pi_t(\beta)) \to^{a.s} E(Z_t^i Z_t^j \pi_t(\beta)(1-\pi_t(\beta))), \quad n \to \infty,$$

$$\frac{1}{n}\sum_{t=1}^{n} Z_t^i Z_t^j Z_t^l \to^{a.s} E(Z_t^i Z_t^j Z_t^l), \quad n \to \infty,$$

for all $i, j, l \in \{1, 2, \ldots, p\}$.

A.2. Theoretical justification for the validity of Test 1

In Fokianos, Gombay, & Hussein (2014), it was shown that under conditions (A–C) there exists a vector of mean-zero Brownian motions with covariance matrix T such that, if β is the true vector of coefficients in the regression model (1), the score vector in (3) admits the following approximation

$$S_n(\beta) - W(n) \ll n^{1/2-\kappa} \quad a.s. \tag{9}$$

for some $\kappa > 0$.

Let $W_1(n)$ be a standard one-dimensional Brownian motion. Covariance calculations for $k^{-1/2}W_1(k)$, $n^{-1/2}W_1(n)$ show that with transformation $t = \log n$, $s = \log k$, process $n^{-1/2}W_1(n)$ is stationary and Gaussian with covariance function $\exp(-1/2|t-s|)$, also known as an Ornstein–Uhlenbeck process. By (9)

$$k^{-1/2}S_k(\beta) - k^{-1/2}W(k) \ll k^{-\kappa}, \quad a.s. \tag{10}$$

Note that the components of $k^{-1/2}T^{-1/2}W(k)$ are independent, so we can use the results of Darling & Erdős (1956) to approximate the maximum functional of the components of $k^{-1/2}T^{-1/2}S_k(\beta)$ over the interval $[1, n]$ using the above transformation and $N = \log n$.

Theorem. *Let $U(t) = t^{-1/2}W_1(t)$, then with $a(N) = (2\log N)^{1/2}$ and $b(N) = 2\log N + (1/2)\log\log N - (1/2)\log\pi$ we have for all $-\infty < y < +\infty$ that*

$$\lim_{N\to\infty} P\left\{ a(N) \sup_{0\le t\le N} |U(t)| - b(N) \le y \right\} = \exp(-2e^{-y}). \tag{11}$$

From this, we get for any component i, $1 \le i \le p$, that

$$\lim_{N\to\infty} P\left\{ a(N) \sup_{1\le k\le n} k^{-1/2}|[T^{-1/2}(\beta)S_k(\beta)]^i| - b(N) \le y \right\} = \exp(-2e^{-y}). \tag{12}$$

For ease of notation, let $\Sigma = T^{-1/2}(\beta)$ and $T_k^{-1/2}(\beta) = \hat{\Sigma}_k$. We have to show that replacing matrix Σ by its estimate $\hat{\Sigma}_k$ does not change the asymptotic distribution.

In Fokianos, Gombay, & Hussein (2014), it was shown that for $\hat{\Sigma}_k$ of (4) as $k \to \infty$

$$T_k(\beta) = \frac{1}{k}\sum_{t=1}^{k} Z_{t-1}Z'_{t-1}\pi_t(\beta)(1-\pi_t(\beta)) \to^{a.s} E(Z_{t-1}Z'_{t-1}\pi_t(\beta)(1-\pi_t(\beta))) = T. \tag{13}$$

From this, we get that

$$\hat{\Sigma}_k \to^{a.s} \Sigma, \tag{14}$$

and for each component $i, j, 1 \le i, j, \le p$, of the matrices $\hat{\Sigma}$ and Σ

$$\max_{1\le k\le n} |\hat{\Sigma}_{k(i,j)} - \Sigma_{i,j}| = O_P(1), \tag{15}$$

an almost surely finite valued random variable.

Applying (11) on the interval $(1, \log n)$ we get that $a(N)\sup_{0\le t\le \log(N)} |U(t)| - b(N)$ converges to $-\infty$ in the sense that for any $y \in \Re$

$$\lim_{N\to\infty} P\left\{a(N) \sup_{0\le t\le \log(N)} |U(t)| - b(N) \le y\right\} = 1, \tag{16}$$

and adding an error of size $O_P(1)$ does not alter the limit as for any $y \in \Re$

$$\lim_{N\to\infty} P\left\{a(N)[\sup_{0\le t\le \log(N)} |U(t)| + O_P(1)] - b(N) \le y\right\} = 1. \tag{17}$$

On the interval $[\log n, n]$, as $\log n \to \infty$ from (14)

$$\max_{\log n<k\le n} |\hat{\Sigma}_{k(i,j)} - \Sigma_{i,j}| = o_P(1), \tag{18}$$

and for any $y \in \Re$

$$\lim_{N\to\infty} P\left\{a(N)[\sup_{\log N\le t\le N} |U(t)| + o_P(1)] - b(N) \le y\right\} = \exp(-2e^{-y}), \tag{19}$$

as

$$\lim_{N\to\infty} P\{a(N)o_P(1) - b(N) \le y\} = 1. \tag{20}$$

This validates the asymptotics for Test 1. Using arguments as above, we could also prove that replacing β by $\hat{\beta}_m$ the asymptotic distribution does not change as $m \to \infty$. However, as our simulations show, the level is distorted at the sample sizes of our study, so we do not recommend Test 1 in such situations.

A.3. Theoretical justification for Test 2

Again, for our starting point we use (9), and use the expansion

$$m^{-1/2}S_{m+k}(\hat{\beta}_m) = m^{-1/2}\left\{S_{m+k}(\beta_0) + (\hat{\beta}_m - \beta_0)\sum_{t=1}^{m+k} Z_{t-1}Z'_{t-1}\pi_t(\beta_0)(1-\pi_t(\beta_0))\right\} + R_{m+k}, \tag{21}$$

where R_{m+k} is the error of approximation. As $\sum_{1\le t\le m} Z_{t-1}(Y_t - \pi_t(\hat{\beta}_m)) = 0$, by using notation $\hat{q}_t = Z_{t-1}(Y_t - \pi_t(\hat{\beta}_m))$, $q_t = Z_{t-1}(Y_t - \pi_t(\beta))$ and $Q_t = Z_{t-1}Z'_{t-1}\pi_t(\beta)(1-\pi(\beta))$ we have for the monitoring statistic

$$m^{-1/2}\sum_{t=m+1}^{m+k}\hat{q}_t = m^{-1/2}\left\{\sum_{t=1}^{m+k}\hat{q}_t - \left(\frac{k}{m}+1\right)\sum_{1\le i\le m}\hat{q}_t\right\}, \tag{22}$$

and by (21),

$$\begin{aligned} m^{-1/2}\sum_{t=m+1}^{m+k}\hat{q}_t = {} & m^{-1/2}\left\{\sum_{t=1}^{m+k} q_t - \left(\frac{k}{m}+1\right)\sum_{1\le i\le m} q_t\right\} \\ & + m^{-1/2}(\hat{\beta}_m - \beta)\left\{\sum_{t=1}^{m+k} Q_t - \left(\frac{k}{m}+1\right)\sum_{t=1}^{m} Q_t\right\} \\ & + R_{m+k} - \left(\frac{k}{m}+1\right)R_m. \end{aligned} \tag{23}$$

Noting that $t = k/m$, by (9) as $m \to \infty$,

$$m^{-1/2}\left\{\sum_{t=1}^{m+k} q_t - \left(\frac{k}{m}+1\right)\sum_{1\le i\le m} q_t\right\} \xrightarrow{D} W(1+t) - (1+t)W(1),\ t \le j, \tag{24}$$

where $W(t)$ is a p-dimensional Brownian motion with covariance matrix $T(\beta)$, and for a fixed $t \le j$

$$\begin{aligned} & m^{1/2}(\hat{\beta}_m - \beta)\left\{\frac{t+1}{m(t+1)}\sum_{t=1}^{m+mt} Q_t - (t+1)\frac{1}{m}\sum_{t=1}^{m} Q_t\right\} \\ & = m^{1/2}(\hat{\beta}_m - \beta)(t+1)\left\{A(mt) - A(m)\right\}. \end{aligned}$$

By Assumption (B), we have that as $m \to \infty$, $A(mt)$ and $A(m)$ converge almost surely to the same finite constant C, so their difference converges to zero, almost surely. As $m^{1/2}(\hat{\beta}_m - \beta) = O_P(1)$, we get

$$m^{1/2}(\hat{\beta}_m - \beta)(t+1)\left\{A(mt) - A(m)\right\} = o_P(1). \tag{25}$$

The error term R_{m+k} consists of components like

$$(m+k)^{1/2}(\hat{\beta}_m-\beta)^{i_1}(\hat{\beta}_m-\beta)^{i_2}\frac{1}{m+k}\sum_{t=1}^{m+k}[\sum Z_{t-1}^{i_1}Z_{t-1}^{i_2}Z_{t-1}^{i_3}]c_{i_1,i_2,i_3}\left(\frac{k}{m}\right),$$

$k \le mj$, with the uniformly bounded coefficients c_{i_1,i_2,i_3}. Therefore, by condition (B), $(1/(m+k))\sum_{t=1}^{m+k} Z_{t-1}^{i_1}Z_{t-1}^{i_2}Z_{t-1}^{i_3}$ converges almost surely to a constant, hence $R_{m+k} = o_P(1)$.

For the dominant term's limit, we can conclude that

$$\left\{\frac{W(1+t)-(1+t)W(1)}{1+t}, 0<t<j\right\} =^D \left\{W(s), 0<s<\frac{j}{(j+1)}\right\}.$$

This justifies the approximation of the test statistic in (7). By Slutsky's theorem, $T(\beta)^{-1/2}$ can be replaced by $\hat{T}(\hat{\beta}_m)^{-1/2}$ of (4) without altering the limit distribution.

ACKNOWLEDGEMENT

This research was in part supported by NSERC Canada Discovery Grants.

BIBLIOGRAPHY

Darling, D. A. & Erdős, P. (1956). A limit theorem for the maximum of normalized sums of independent random variables. *Duke Mathematical Journal*, 23, 143–155.

Fokianos, K., Gombay, E., & Hussein, A. (2014). Retrospective change detection for binary time series. *Journal of Statistical Planing and Inferences*, 145, 102–112.

Frisen, M. (2003). Statistical surveillance. Optimality and methods. *International Statistical Review*, 71(2), 403–434.

Gombay, E. (2008). Change detection in autoregressive time series. *Journal of Multivariate Analysis*, 99(3), 451–464.

Gombay, E., Hussein, A., & Steiner, S. H. (2011). Monitoring binary outcomes using risk-adjusted charts: A comparative study. *Statistics in Medicine*, 30(23), 2815–1826.

Gombay, E. & Serban, D. (2009). Monitoring parameter change in AR(p) time series models. *Journal of Multivariate Analysis*, 100(4), 715–725.

Grigg, O. A. & Farewell, V. (2004). A risk-adjusted sets method for monitoring adverse medical outcomes. *Statistics in Medicine*, 23(10), 1593–1602.

Grigg, O. A. & Spiegelhalter, D. J. (2010). Clinical surveillance and patient safety. *Performance Measurement for Health System Improvement: Experiences, Challenges and Prospects*. Cambridge University Press.

Höhle, M. (2000). Online change-point detection in categorical time series in statistical modelling and regression structures. In *Festschrift in Honour of Ludwig Fahrmeir*, Kneib, T., & Tutz, G., editors. Springer, Basel.

Hussein, A., Abdullah, K., Severien, N., & Campostrini, S. (2015). Performance of risk-adjusted cumulative sum charts when some assumptions are not met. *Communications in Statistics – Simulation and Computation* (in press).

Jones, M. A. & Steiner, S. H. (2012). Assessing the effect of estimation error on risk-adjusted CUSUM chart performance. *International Journal for Quality in Health Care*, 24(1), 176–81.

Kedem, B. & Fokianos, K. (2002). *Regression Models for Time Series Analysis*, Wiley, New York.

Kulldorff, M., Davis, R. L., Kolczak, M., Lewis, E., Lieu, T., & Platt, R. (2011). A maximized sequential probability ratio test for drug and vaccine safety surveillance. *Sequential Analysis*, 30, 58–78.

Nishina, K. & Nishiyuki, Sh. (2003). False alarm probalility fundtion of Cusum charts. *Economic Quality Control*, 18, 101–112.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64, 191–199.

Serfling, R. J. (1968). Contributions to central limit theory for dependent variables. *The Annals of Mathematical Statistics*, 39(4), 1158–1175.

Steiner, S. H., Cook, R. J., Farewell, V. T., & Treasure, T. (2000). Monitoring surgical performance using risk-adjusted cumulative sum charts. *Biostatistics*, 1(4), 441–452.

Vostrikova, L. J. (1981). Detection of a "Disorder" in a Wiener process, *Theory of Probability and Its Applications*, 26, 356–362.

Received 5 July 2014
Accepted 11 March 2015