Monitoring Paired Binary Surgical Outcomes Using Cumulative Sum Charts

Stefan H. Steiner and Richard J. Cook
Department of Statistics and Actuarial Sciences
University of Waterloo
Waterloo, Ontario
Canada N2L 3G1
tel. (519) 888-4567 x6506
e-mail: shsteine@setosa.uwaterloo.ca
e-mail: rjcook@navigator.math.uwaterloo.ca

Vern T. Farewell
Department of Statistical Sciences
University College
London, England
tel. 0171-419-3647
e-mail: vern@stats.ucl.ac.uk

Summary

Correlated binary data are encountered in many areas of medical research, system reliability, and quality control. For monitoring failures rates in such situations, simultaneous bivariate Cumulative Sum (CUSUM) charts with the addition of secondary control limits are proposed. Using an approach based on a Markov chain model, the run length properties of such a monitoring scheme can be determined for sudden, or gradual, changes in the failure rates. The proposed control charts are easy to implement, and are shown to be very effective at detecting small changes in the rate of undesirable outcomes, especially when the changes are gradual. This procedure is illustrated using bivariate outcome data arising from a series of pediatric surgeries. The methodology is sufficiently general that it may be adapted for multivariate normal, binomial, or Poisson responses.

1. Introduction

In many areas of medical research it is necessary to utilize more than one outcome variable to adequately characterize a response to a medical intervention. In such settings, it is often of interest to monitor the outcomes with a view to quickly detecting emerging changes in the response pattern. In the context of randomized clinical trials, Cook and Farewell[1], and Cook[2,3] proposed monitoring schemes based on multivariate group sequential designs which are attractive in that they maintain the individuality of the component test statistics. In this article we focus on situations in which a process is being monitored and two correlated outcomes variables are used to characterize the response. The process in the motivating example is a surgical procedure, but could be other medical or industrial processes. The detection of substantial adverse changes in the process should result in some investigation of the cause, and possibly process changes in order to mitigate the negative effect.

De Leval et al.[4] discussed the motivating problem for the developments that follow. The goal is to monitor failure rates of a surgical procedure for a bivariate outcome. Over a period of six years an arterial switch operation was performed on 104 newborn babies. Since the death rate from this surgery was relatively low (around 2-5%), de Leval et al.[4] introduced the idea of a surgical *near miss,* a term borrowed from the airline industry. In this context, *near miss* was defined as the need to reinstitute cardiopulmonary bypass after a trial period of weaning. By recording both *near misses* and deaths from the surgery we obtain more information regarding the success or "quality" of the surgery. Thus, after surgery one response indicates whether there was a *near miss* and the other indicates survival status. The original data is reproduced in chronological sequence in Appendix A. Over the monitoring period a cluster of unexplained deaths occurred between the 53rd and 68th patients. The surgeon desired to determine whether the cluster could be attributed to chance, or if there was some underlying cause for the flurry of adverse surgical outcomes. In addition, the surgeon wished to adopt a monitoring procedure that could quickly detect increases in the death rate so that, if necessary, some remedial measure, such as retraining, or adopting a procedural change, could be rapidly introduced. Note that in this example we assume that the results of the surgical procedure are quickly determined.

In analyzing this problem the possibility of covariates such as patient characteristics, procedural characteristics, and surgical team fatigue, were explored. However, none of the proposed covariates was found to have a significant effect on the failure rates. In the absence of covariate information de Leval et al.[4] propose two sequential probability ratio test (SPRT) charts to monitor the bivariate surgical outcomes. The purpose of one SPRT was to monitored the death rate alone, while the other SPRT was designed to monitor the occurrence of either a death or a *near miss*. If either of these two SPRTs signal, this was taken as evidence that either the death rate or the combined death and *near miss* rate had increased, and some remedial measure was necessary.

This methodology, which is subsequently referred to as the de Leval approach, works fairly well in this application, but has a number of shortcomings in general. First, the procedure is not easily generalized to cases where both the monitored outcomes are of equal importance. Using the de Leval approach, the *near miss* rate is not monitored separately, and thus increases in the *near miss* rate may not be easily detected. In addition, the proposed SPRTs may signal both an increase or a decrease in the failure rates, however, in this application, no action is suggested if failure rates appear to have decreased. Allowing the monitoring procedure to accumulate evidence that the rates have decreased makes it potentially less sensitive to rate increases. This happens if, by chance, the SPRT is below the nominal value when the failure rate shifts upwards. Finally, the charts are designed using error rates that specify the largest probability that the wrong decision will be made. However, in this context, we are more concerned with run lengths for monitoring schemes of an ongoing process, or in other words, how long it will take before we detect a specific change in the failure rates.

This remainder of this article is organized as follows. In Section 2 we briefly review established monitoring methodology and discuss alternative approaches to the problem. Details regarding the design of the proposed simultaneous CUSUMs with secondary limits (SCUSUMs) are provided in Section 3. In Section 4 we provide more information on the surgical failures example, and show how SCUSUMs may be applied. Finally, Section 5 summarizes the results and discusses some interesting extensions. In the Appendix the run length properties of SCUSUMs are derived.

2. Background

2.1 Univariate Cumulative Sum (CUSUM) Procedure

Since the subsequent material requires an understanding of univariate CUSUM charts a brief review is presented. The original formulation of a CUSUM is due to Page[5]. Two sided implementations suggested by Barnard[6] involved the use of a graphical device, called a V-mask, to determine if the monitored process had changed. In situations where concern is focused on only detecting increases (or decreases) in the failure rate, the tabular form of the CUSUM is more appropriate. Due to the ease of implementation, this article focuses on the tabular CUSUM. A standard univariate tabular CUSUM involves monitoring:

$$X_t = \max\left(0,\, X_{t-1} + W_t\right), \qquad t = 1, 2, 3, \ldots, \tag{1}$$

where $X_0 = 0$, and W_t is the sample weight[7]. The CUSUM repeatedly tests the hypotheses H_0: $\theta = \theta_0$ versus H_1: $\theta = \theta_1$, where θ is the parameter of interest. The process is assumed to be in state H_0 as long as $X_t < h$, and is deemed to have shifted to state H_1 if $X_t \geq h$ at any time *t*. For example, a CUSUM may be used to monitor the surgical death rate. In that case θ_0 represents an acceptable probability of death, say $\theta_0 = .02$, while θ_1 represents the undesirable death rate, say $\theta_1 = .05$.

The design of the CUSUM is given by the choices for W_t and h. Let y_t represent the current sample outcome, for example the outcome for patient *t*, and denote the probability distribution of the possible sample outcomes as $f\left(y_t;\theta\right)$. Moustakides[8] showed that the optimal choice for the tabular CUSUM weights W_t for all *t* is the log-likelihood ratio $\ln\left(f\left(y_t;\theta_1\right)/f\left(y_t;\theta_0\right)\right)$. This choice is optimal in the sense that, among all schemes with the same operating characteristics in-control it provides the best possible sensitivity. In the death rate monitoring example, $f(z;\theta) = \theta^z(1-\theta)^{1-z}$, where z equals unity if a death occurs and zero otherwise. Then, the log-likelihood ratio is $z\ln\left(\theta_1/\theta_0\right) + (1-z)\ln\left(\left(1-\theta_1\right)/\left(1-\theta_0\right)\right)$ or $0.916z - 0.031(1-z)$, substituting the above values for θ_0 and θ_1. Thus, the weights are

$$W_t = \begin{cases} -.031 \text{ if patient } t \text{ lives} \\ .916 \text{ if patient } t \text{ dies} \end{cases}.$$

Good choices for *h* are based on the expected or average run length (ARL) of the CUSUM under H_0 and H_1. The run length of the CUSUM is defined as the time (or number of observations) required before the CUSUM first exceeds the control limit. In quality control terminology, a CUSUM that exceeds the control limit is said to have "signaled." Note that CUSUMs are designed to monitor the responses sequentially until sufficient evidence of process deterioration is detected. Ideally, while the process is in state H_0 the run lengths should be long, since signals represent false alarms under H_0. On the other hand, if the process has shifted to H_1, or any other undesirable process settings, we would like short run lengths. Issues surrounding the design of univariate CUSUM procedures are discussed for normally distributed outcomes[9,10] and in the binomial data case[11].

Tabular CUSUM charts may be thought of as a geometric series of Wald Sequential Probability Ration Tests (SPRTs), where the SPRTs start on the lower boundary, and we observe a number of SPRTs that end in acceptance of H_0, followed by one SPRT that ends in rejection of H_0[5]. Through this equivalence, the run length properties of the CUSUM are related to the probability the underlying SPRT rejects H_0. Since we have a geometric series of SPRTs ending with rejection of H_0, the ARL of the CUSUM can be expressed as $1/p$, where p is the probability the underlying SPRT ends in rejection.

2.2 Multivariate Monitoring Schemes

In the field of quality control, there is a fairly rich literature concerning procedures for monitoring bivariate or multivariate processes, but these methods have received comparatively little attention in the medical area. Most of the established monitoring methods incorporate a control chart as a graphical representation to ease implementation and understanding. Here we review these and demonstrate that none are entirely satisfactory in situations such as the surgical failures example. For a general review see Montgomery[7].

The first and simplest approach uses multiple univariate charts, where each outcome variable is monitored using a separate control chart. Examples of univariate control charts include $\overline{X}$ charts, *p* charts, and CUSUM charts. Multiple univariate charts are used in many situations since they are easy to understand and implement. However, by monitoring based on separate control charts in the usual way, we ignore the joint distribution of the outcomes. As a result, the ability of multiple univariate charts to detect process changes without producing frequent false alarms may be impaired, particularly if the correlation is substantial.

Another option is to utilize a global test statistic derived from all outcomes and to monitor based on this summary statistic. Examples of so called multivariate charts include multivariate *p* charts[12] and multivariate CUSUM charts[13]. In some contexts, multivariate charts may be desirable since they are based on only one chart and the correlation between the variables can be taken into account in the chart design. However, somehow the different outcomes must be combined into one global test statistic, which may not be easy to justify, especially if the outcomes are not of comparable importance. Also, multivariate charts are often considered unattractive since the underlying cause of an out-of-control signal may be more difficult to determine.

A final general strategy is to assume an underlying parametric model that explains the relative probabilities of observing the different outcomes. In this way, a bivariate (or multivariate) problem where the outcomes are mutually exclusive and can be ranked in terms of importance may be recast into a univariate grouped data problem. Steiner, Geyer and Wesolowsky[14,15] have discussed monitoring methods that handle grouped data. Grouped data methods are attractive since they are simple to implement and interpret. However, to be applicable we must assume an underlying continuous distribution exists that explains the relative probabilities of the various outcomes, and this may not be reasonable.

3. Simultaneous CUSUM Charts with Secondary Control Limits

3.1 Formulation of the Simultaneous Secondary Limit CUSUM (SCUSUM)

For monitoring the success of a medical intervention in terms of two outcome variables we define the simultaneous CUSUM statistics based on two univariate tabular CUSUMs, i.e.

$$S_{Yt} = \max\left(0, S_{Y,t-1} + W_{Y_t}\right), \quad t = 1, 2, 3, \ldots$$
$$S_{Zt} = \max\left(0, S_{Z,t-1} + W_{Z_t}\right), \quad t = 1, 2, 3, \ldots , \qquad (2)$$

where $S_{Y0} = S_{Z0} = 0$, and W_{Y_t} and W_{Z_t} are respectively, the sample weights for the Y and Z charts. Using SCUSUM charts, the two CUSUMs defined by (2) are monitored simultaneously, and the procedure signals if either the individual CUSUM values S_{Yt} or S_{Zt} are greater than predetermined primary control limits h_y and h_z. We also introduce so called secondary control limits h_{yy} ($h_{yy} \leq h_y$) and h_{zz} ($h_{zz} \leq h_z$) for S_{Yt} and S_{Zt} respectively. The SCUSUM also signals when both CUSUM values are simultaneously greater than their respective secondary control limits. The secondary limits serve to aid in the rapid detection of small increases in the event rates of both outcomes, and their position will be influenced by the degree of correlation between the variables. To summarize, the SCUSUM chart signals at time t if any of the following three conditions are satisfied:

(i) $S_{Yt} \geq h_y$ (and $S_{Zt} < h_{zz}$), or
(ii) $S_{Zt} \geq h_z$ (and $S_{Yt} < h_{yy}$), or (3)
(iii) $S_{Yt} \geq h_{yy}$ and $S_{Zt} \geq h_{zz}$ simultaneously.

Individually each of the two CUSUM charts will be designed to test: $H_0: \theta = \theta_0$ versus $H_1: \theta = \theta_1$, where θ represents a vector of parameters of interest, θ_0 is a vector of in-control parameter values, and θ_1 is a vector of undesirable parameter values. In the context of bivariate binary data, θ may represent the vector (π_y, π_z), for example, where π_y and π_z equal the probabilities of failure for outcomes Y and Z respectively.

To determine explicit expressions for the weights based on the log-likelihood ratio (Moustakides, 1986) we must model the distribution of the outcome variables. Although SCUSUMs are generally applicable to any type of paired data, we focus in this article on paired binary data which arise in our motivating problem. Let Y and Z represent two correlated outcome variables as given by (4).

$$Y = \begin{cases} 1 \text{ if a failure of type Y occurs} \\ 0 \text{ otherwise} \end{cases}, \quad Z = \begin{cases} 1 \text{ if a failure of type Z occurs} \\ 0 \text{ otherwise} \end{cases} \tag{4}$$

The possible sample (patient) outcomes are represented by the ordered pair (y, z).

There are four possible outcomes for this ordered pair, namely (0,0), (0,1), (1,0) and (1,1). Let $q_{yz} = \Pr(Y = y, Z = z)$. Then the joint distribution of Y and Z is determined by q_{00}, q_{01} and q_{10} with $q_{11} = 1 - q_{00} - q_{01} - q_{10}$. Also define $\pi_y = \Pr(Y = 1) = q_{10} + q_{11}$ and $\pi_z = \Pr(Z = 1) = q_{01} + q_{11}$. Using this notation the conditional distributions are given by $\Pr(Z = z \mid Y = y) = q_{yz} / (q_{y0} + q_{y1})$, $\Pr(Y = y \mid Z = z) = q_{yz} / (q_{0z} + q_{1z})$.

In practice one has a choice in selecting which aspect of the bivariate process to monitor. Here we consider the rate of near misses, and the component of the death rate not explained by changes in the correlation and the near miss rate. This approach of constructing the joint distribution by the product of a marginal distribution for one outcome and a conditional distribution of the other (conditional on the realization of the first event) was examined by Klotz[16] and is developed further in the context of longitudinal data by Darlington and Farewell[17].

To formalize in this setting we let

$$\begin{aligned} \pi_y &= \Pr(Y = 1) = \frac{e^{\alpha_y}}{1 + e^{a_y}} \\ \pi_{z|y} &= \Pr(Z = 1 \mid Y = y) = \frac{e^{\alpha_z + \beta y}}{1 + e^{\alpha_z + \beta y}} \end{aligned} \tag{6}$$

Here β represents the log odds ratio of z = 1 for y = 1 versus y = 0 and serves as a natural measure of the dependence between y and z. Conditioning in the reverse order is also possible but not appropriate for our monitoring example due to the natural ordering of the outcomes. Indeed, In principle it is possible to run the same sort of procedure without the conditioning. To do this one would replace the probabilities in (6) with the two marginal probabilities and derive the appropriate CUSUM weights as before.

We assume β is fixed and we devise a monitoring scheme for detecting changes in the $\Pr(Y = 1)$ and $\Pr(Z = 1 \mid Y)$. Since the outcomes are paired Bernoulli random variables, the likelihood of $\theta = (\alpha_y, \alpha_z, \beta)$ is written:

$$\pi_y^y(1-\pi_y)^{1-y}\,\pi_{z|y}^z(1-\pi_{z|y})^{1-z} \;=\; \left(\frac{e^{\alpha_y y}}{1+e^{\alpha_y}}\right)\left(\frac{e^{\alpha_z z+\beta yz}}{1+e^{\alpha_z+\beta y}}\right), \qquad (7)$$

Assuming that under H_0 the parameters are given by α_{y0}, α_{z0}, β and under H_1 the model parameters are given by α_{y1}, α_{z1}, β the log-likelihood ratio weights for outcome (y,z) are given in Table 1.

Table 1: CUSUM Log-Likelihood Ratio Weights

(y, z)	Weight
(0, 0)	$\log\left(1+e^{\alpha_{y0}}\right)-\log\left(1+e^{\alpha_{y1}}\right)+\log\left(1+e^{\alpha_{z0}}\right)-\log\left(1+e^{\alpha_{z1}}\right)$
(0, 1)	$\log\left(1+e^{\alpha_{y0}}\right)-\log\left(1+e^{\alpha_{y1}}\right)+\left(\alpha_{z1}-\alpha_{z0}\right)+\log\left(1+e^{\alpha_{z0}}\right)-\log\left(1+e^{\alpha_{z1}}\right)$
(1, 0)	$\left(\alpha_{y1}-\alpha_{y0}\right)+\log\left(1+e^{\alpha_{y0}}\right)-\log\left(1+e^{\alpha_{y1}}\right)+\log\left(1+e^{\beta+\alpha_{z0}}\right)-\log\left(1+e^{\beta+\alpha_{z1}}\right)$
(1, 1)	$\left(\alpha_{y1}-\alpha_{y0}\right)+\log\left(1+e^{\alpha_{y0}}\right)-\log\left(1+e^{\alpha_{y1}}\right)+\left(\alpha_{z1}-\alpha_{z0}\right)+\log\left(1+e^{\beta+\alpha_{z0}}\right)-\log\left(1+e^{\beta+\alpha_{z1}}\right)$

Note that with this setup a signal due to (3i) represents evidence that π_y has increased, a signal due to (3ii) represents evidence that π_z has increased, and any signal due to (3iii) suggests that both π_y and π_z have increased. Thus, for given values of α_{y0}, α_{z0}, β, α_{y1} and α_{z1} the appropriate weights to use in (2) for each of the CUSUM can be determined from Table 1. Note that the alternative values of the parameters, i.e. α_{y1} and α_{z1}, will be different for the two charts.

3.2 Determination of the Control Limits

To complete the specification of the SCUSUM we need to determine the four control limits h_y, h_z, h_{yy} and h_{zz}. As discussed in Section 2, a CUSUM is a sequential monitoring procedure that has close links to sequential probability ratio tests. However, since CUSUMs are designed to continue until a signal occurs, their design does not involve the specification of a type I error rate and power for detecting the effect of interest. Instead interest lies in the expected run lengths under specific parameter configurations. Specifically CUSUM control limits are chosen to yield desirable run length properties[9], where the run length is defined as the number of observations before a signal. That is to say, for example, one asks the question, how frequently should the out-of-

control signal be triggered when the process is actual operating under acceptable conditions? We want the CUSUM procedure to have a long average run length under acceptable conditions and a short average run length when the process has shifted. Appendix B shows how the run length properties of SCUSUMs may be derived when the observations are correlated binary random variables. For the proposed SCUSUMs the probability of signaling due to (3i), (3ii) or (3iii), denoted p_1, p_2 and p_3 respectively, can also be used to aid the design. In particular, under the null hypothesis, it is natural to set the control limits such that all three possible ways of signaling are approximately equal. In this way, the chart will be sensitive to increases in the failure rates of both outcomes. Appendix B also shows how p_1, p_2 and p_3 can be determined.

In many medical applications, such as the surgical failures example, the monitored processes are extremely critical. In such situations, decisions regarding the acceptability of the failure rates are best made after each patient in order to most quickly detect adverse effects. Using such an approach, CUSUM charts can be designed to have certain desired run length properties under the null hypothesis, but out-of-control run length properties follow directly. We propose the following three step iterative approach to determining appropriate values for h_y, h_z, h_{yy} and h_{zz}:

1) Set the initial values for h_y and h_z, based on the desired in-control average run length (ARL) of the two univariate charts.
2) Set the initial values for h_{yy} and h_{zz} at some fixed proportion of the primary limits, say 60%. The proportion should be based on the correlation between the two outcomes. For large positive correlations the secondary limits will be quite close to the primary limits.
3) Given the current choices for h_y, h_z, h_{yy} and h_{zz} determine the ARL and probabilities of signaling under the H_0 using Appendix B. Iteratively adjust the control limits to obtain close to the desired ARL and p_1, p_2 and p_3 approximately equal.

4. Surgical Failures Example

4.1 Application

Define indicator variables as in (4) where $Y = 1$ corresponds to a *near miss* and $Z = 1$ corresponds to a death. In Table 2 we summarize the surgical failures data from Appendix A using

the four possible surgical outcomes for each patient. A patient can be classified as both a death and a *near miss* since any death within 24 hours of the surgery is considered a surgical death.

Table 2: Summary of Surgical Failure Data

		Death: no $(z=0)$	Death: yes $(z=1)$
Near Miss	no $(y=0)$	85	4
	yes $(y=1)$	10	5

The odds ratio for this 2x2 table is 10.6, with an approximate 95% confidence interval given by (2.44, 46.1). This suggests the outcomes are correlated, and that the odds of a death is approximately 10 times higher for a patient who is classified as a "near miss."

The SCUSUM is designed to directly monitor both the death and *near miss* rates. To specify the chart we need to determine appropriate sample weights as given in (2), and choose values for the primary and secondary control limits for each chart.

We start by modelling the failure rates. In this situation, since the outcome death is typically established after the determination of *near miss*, it is reasonable to model the probability of death conditionally on the *near miss* status. Defining π_y as the probability of a *near miss*, and $\pi_{z|y}$ as the conditional probability of death, we adopted model (6). Using this model the likelihood of $\theta = (\alpha_y, \alpha_z, \beta)$ is given by (7).

Based on previous surgical experience[4], a death rate and a "death or near miss" rate of around 2 and 10 percent, respectively, are the failures rates we would expect when in-control. Based on these event rates and the available data we adopted the parameter values $\alpha_{y0} = -2.3$, $\alpha_{z0} = -4.5$ and $\beta = 2.5$. Since these parameter values yield acceptable surgical performance they define H_0. We will later comment on sensitivity analysis regarding the initial choice for β.

Appropriate alternative hypotheses for the CUSUM charts are determined by considering increases in either the death or *near miss* rates that are unacceptable and should be detected quickly.

For the CUSUM chart designed to detect increases in *near miss* rate and the death rate alternate hypotheses H_{1y} and H_{1z} respectively are proposed, where

H_{1y}: $\alpha_{y1} = -1.7$ [Pr(near miss) = $\pi_y = .15$], $\alpha_{z0} = -4.5$, $\beta = 2.5$, and
H_{1z}: $\alpha_{y0} = -2.3$, $\alpha_{z1} = -2.9$ [Pr(death | not near miss) = $\pi_{z|y=0} = .05$], $\beta = 2.5$.

Note that H_{1y} and H_{1z} reflect increases in only the marginal *near miss* or death rates respectively, and that we assume the log odds ratio (correlation) parameter β is constant. The parameter values chosen for the alternative hypothesis correspond to increases in the failure rate that are deemed undesirable, and are of a sufficient magnitude that quick detection of such a change is desired.

The log-likelihood ratio is defined as $\log(LR) = \log\left[\left(L \mid H_1\right)/\left(L \mid H_0\right)\right]$. For the *near miss* and death charts respectively this yields:

$$\begin{aligned} \log\left(LR \mid H_{1y}\right) &= \left(\alpha_{y1} - \alpha_{y0}\right)y + \log\left(1 + e^{\alpha_{y0}}\right) - \log\left(1 + e^{\alpha_{y1}}\right) \\ \log\left(LR \mid H_{1z}\right) &= \left(\alpha_{z1} - \alpha_{z0}\right)z + \log\left(1 + e^{\beta y + \alpha_{z0}}\right) - \log\left(1 + e^{\beta y + \alpha_{z1}}\right) \end{aligned} \qquad (8)$$

Expressions (8) are functions of only the outcomes *y* and *z*. Substituting into (8) all possible combinations (*y*, *z*) leads to the optimal log-likelihood ratio weights. The weights are given in Table 3. For ease of implementation the actual weights used in the procedure, W_{Y_t} and W_{Z_t}, are integers that closely match the relative sizes of the true log-likelihood ratio weights. Note that in Table 3 there are only two distinct weights for the *near miss* chart, since the probability of a *near miss* is not effected by the occurrence of a death.

Table 3: Log-Likelihood Ratio Weights

Outcome (y, z)	(0, 0)	(0, 1)	(1, 0)	(1, 1)
$\log\left(LR \mid H_{1y}\right)$	–.07	–.07	.53	.53
W_{Y_t}	–1	–1	7	7
$\log\left(LR \mid H_{1z}\right)$	–.04	1.6	–.39	1.2
W_{Z_t}	–1	37	–9	29

To complete the design of the simultaneous CUSUMs we need to specify values for the control limits. This may be accomplished using the general iterative design procedure outlined in

the previous section. To begin we must specify an appropriate level for the average in-control run length of the procedure. Based on the current rate of surgeries an in-control average run length of around 280 was suggested, this represents approximately one false alarm every 16 years. There is a tradeoff associated with the choice of the desired in-control average run length. Choosing larger values makes false alarms less frequent, but also results in a procedure that takes longer to signal increases in the failure rates. As a result, it is useful to check that the out-of-control average run length of the procedure for given control limits is fairly small when the out-of-control situation corresponds to increases in the failure rates that would be important to detect quickly. Determining the appropriate corresponding control limits using the numerical routines from the Appendix gives $h_y = 32$, $h_z = 70$, $h_{yy} = 17$ and $h_{zz} = 38$. These control limits choices yield an in-control average run length of 284 patients where false alarms due to each of the three possible modes of signaling are approximately equally likely.

Using these control limits, the data from the original 104 patients, and the weights from Table 3 the resulting SCUSUM is given in Figure 1. At patient number 55 the SCUSUM signals (denoted by the two circles in the figure), due to the secondary limits, that the *near miss* rate and the death rate have both increased. Assuming this signal was ignored, and the monitoring procedure was continued, the chart subsequently signal an increase in the death rate and *near miss* rate at patient numbers 59 and 68 respectively. These primary control limit signals are identified by the crosses in Figure 1. Clearly, there is evidence that the surgical failure rates are substantially larger than the acceptable values given by H_0.

The SCUSUM chart in Figure 1 is relatively insensitive to the initial estimate for the log odds ratio parameter β. As part of a small sensitivity study we considered all values of β between 0.9 and 3.8. This range of β corresponds to the 95% confidence interval for the odds ratio of the surgical failures data given in Table 2. For each choice of β we determined the corresponding α_{y0} and α_{z0} values that correspond to death and "death or near miss" rates of 2 and 10 percent, and also determine the appropriate values for α_{y1} and α_{z1}, weights, and appropriate control limits to give the same expected in-control properties, the control chart for the actual surgical failures data signals at exactly the same patient numbers as in Figure 1.

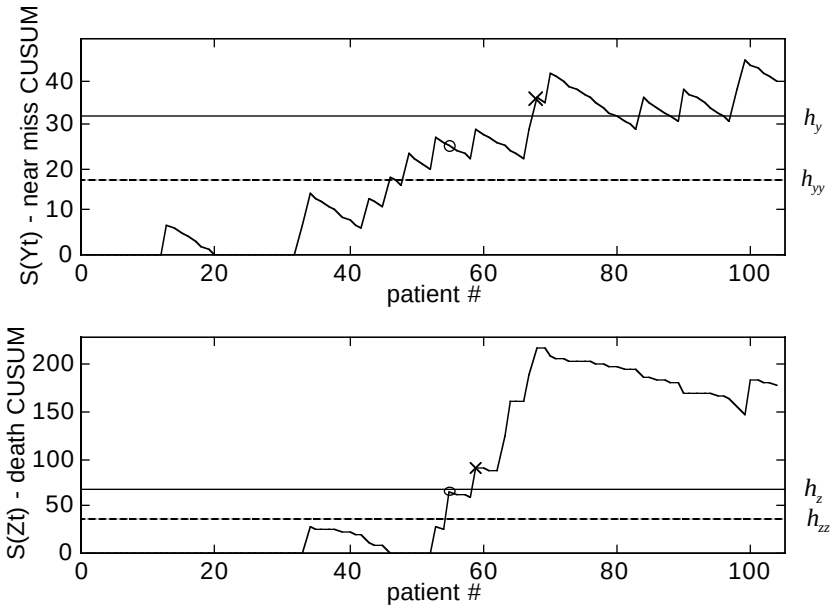

Figure 1: SCUSUM with Surgical Failures Data

4.2 Characteristics of the SCUSUM Procedure

It is of interest to explore the run length properties and effectiveness of the different control limits for this proposed procedure. Denote the probability the SCUSUM signals due to signaling rules (3i), (3ii) (3iii) as p_1 = Pr(signal – near miss), p_2 = Pr(signal – death) and p_3 = Pr(signal – joint) respectively, where $p_1 + p_2 + p_3 = 1$. Figures 2 to 5 are derived using the Markov chain methodology discussed in Appendix B.

Figure 2 shows ARL contours of the proposed SCUSUMs at various combinations of *near miss* and death rates. As one would expect, the ARL is very large for small failure rates, and decreases rapidly as the failure rates increase. In addition, Figure 2 shows contours of p_1, p_2 and p_3. As expected, the probability of signaling an increase in the death rate when the death rate, but not the *near miss* rate, has increased is very large. Similarly, the probability of signaling an increase in the *near miss* rate is tied to the actual *near miss* rate. Also, the secondary limits are most effective when both the error rates have increased. For example, when π_y = .20 and $\pi_{z|y=0}$ = .05 the probability the signal was due to the secondary limits (p_3) is around .43.

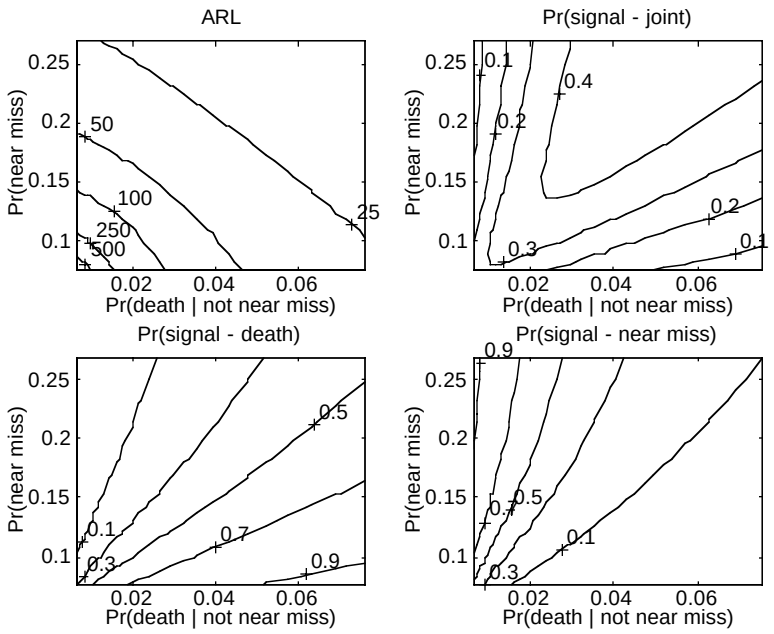

Figure 2: SCUSUM Properties (h_y = 32, h_z = 70, h_{yy} = 17, and h_{zz} = 38)

Figure 3 shows the equivalent plots when the log odds ratio β equals zero. This corresponds to the case where the outcomes are uncorrelated. Note that the sample weights used in the procedure were derived using a log odds ratio of β = 2.5, thus the results from Figure 3 provide some sensitivity analysis. From the contour plots it is evident that the benefits of the secondary limits are reduced. When β = 0, out-of-control situations yield longer run lengths, and there is less chance that the SCUSUM signals due to the secondary limits.

Figure 4 shows contours of ARL, p_1, p_2 and p_3 when the secondary limits are not used. Comparing the contour plots in Figures 2 and 4 shows the loss in efficiency that occurs when the secondary limits are abandoned. Specifically, the plots suggest that the secondary limits decrease the out-of-control ARL, and that the secondary limits are especially effective when both the failure rates increase by a small amount.

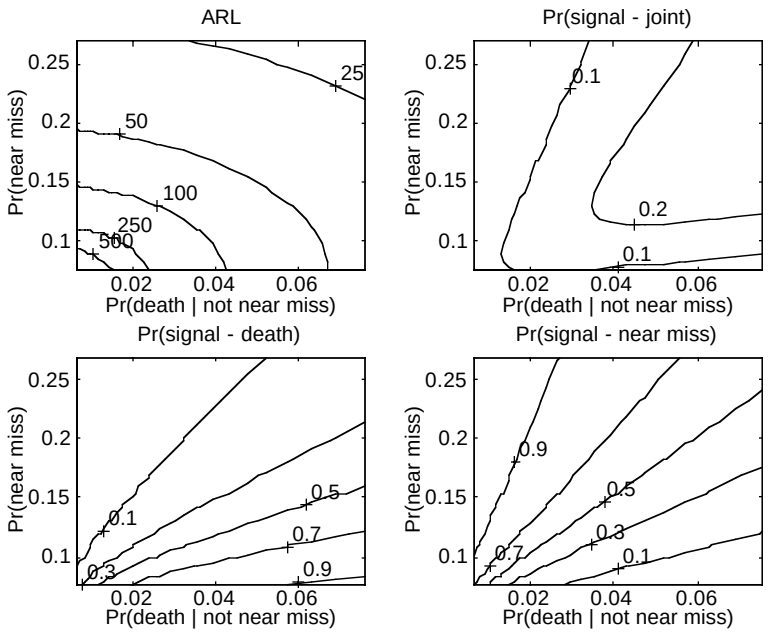

Figure 3: SCUSUM Properties with $\beta = 0$ (no correlation)

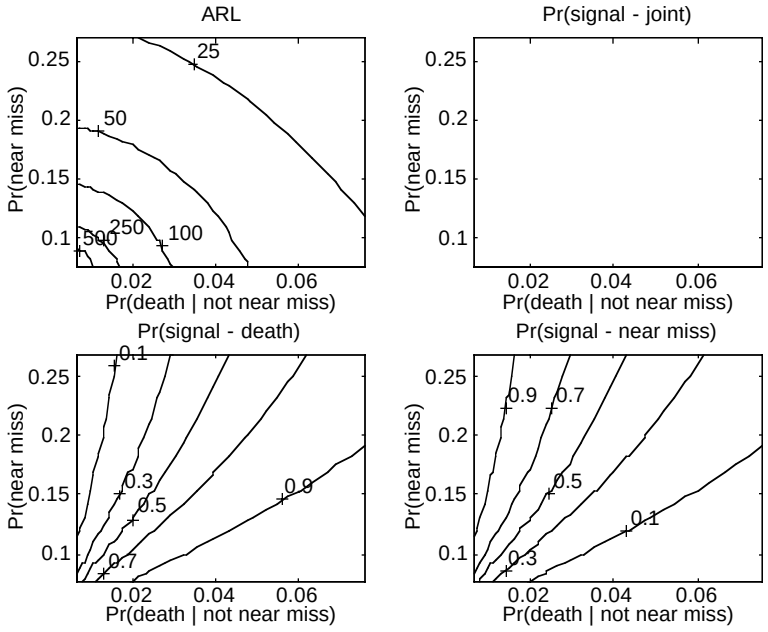

Figure 4: SCUSUM Properties without Secondary Limits, i.e. $h_y = h_{yy} = 32$, $h_z = h_{zz} = 70$

Figures 2-4 show the ARL and signal probabilities when the values of π_y and $\pi_{z|y}$ are fixed. This is the classical run length analysis and is useful when quality problems manifest themselves as sudden changes. However, another model for the onset of quality problems is based on the failure rates slowly deteriorating. Such a drift in quality would be of interest, for

example, if the surgeon's skills are expected to deteriorate with time (perhaps due to aging), or when a piece of equipment slowly wears out. Using results from Appendix B, ARL, p_1, p_2 and p_3 values can also be found for arbitrary changes in the failure rates over time.

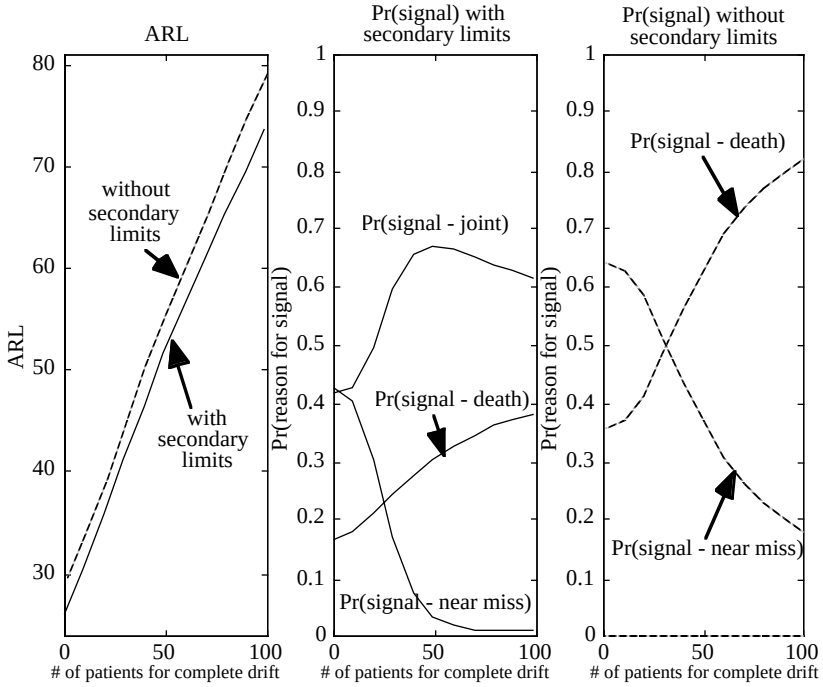

Figure 5: Solution Properties under Gradual Drift

Figure 5 plots ARL, p_1, p_2 and p_3 results when both the failure rates gradually drift from H_0 to triple the original rate. In Figure 5, the horizontal axis of all the plots gives the speed of the drift in terms of the number of patients operated on before the drift is completed. Note that in the two right-most frames of Figure 5 the signal probabilities must sum to one, since the CUSUM eventually signals. As expected, when the deterioration in the quality is slow, it is more difficult to detect, and thus requires longer ARLs. Figure 5 simultaneously shows results for the SCUSUM, and the multiple univariate approach where the secondary limits are not used. The figure suggests that SCUSUMs are also effective when the failure rate increases over time. The ARLs are lower when using the SCUSUM thus showing the benefits of the secondary limits.

4.3 Comparison with Procedure Proposed by de Level et al.

In Section 4.1 we observed that the proposed scheme would signal a problem after patient number 55. This is an improvement over the performance of the de Leval procedure where

problems were detected (at the 95% confidence level) first at patient number 59. A more comprehensive comparison of SCUSUM and the de Leval procedure is provided by the ARL contour plots of Figure 6. The contour plot on the left reproduces the ARL contour plot given as part of Figure 2 for the SCUSUM with control limits $h_y = 32$, $h_z = 70$, $h_{yy} = 17$, and $h_{zz} = 38$. Simulated ARL results for the de Leval procedure, where each contour point represents the average of 500 trials, is shown on the right. To make the procedures more easily comparable, the middle contour plot represents the ARL of an SCUSUM procedure with weights given by Table 3, and control limits given by $h_y = 25$, $h_z = 70$, $h_{yy} = 15$, and $h_{zz} = 38$. Note that these control limits are slightly smaller for the *near miss* chart than those proposed in Section 4.1. With this slightly adjusted SCUSUM, the ARL under H_0 very closely matches the ARL under H_0 of the de Leval scheme. In addition, the ARLs of the adjusted SCUSUM are very close to those of the de Leval procedure when increases occur only in the *near miss* rate. With the adjusted SCUSUM chart, as the death rate increases, the ARL drops much more rapidly than with the de Leval scheme. Since quickly detecting increases in the death rate is the primary propose of the monitoring the shorter ARL values for large death rates are a substantial advantage.

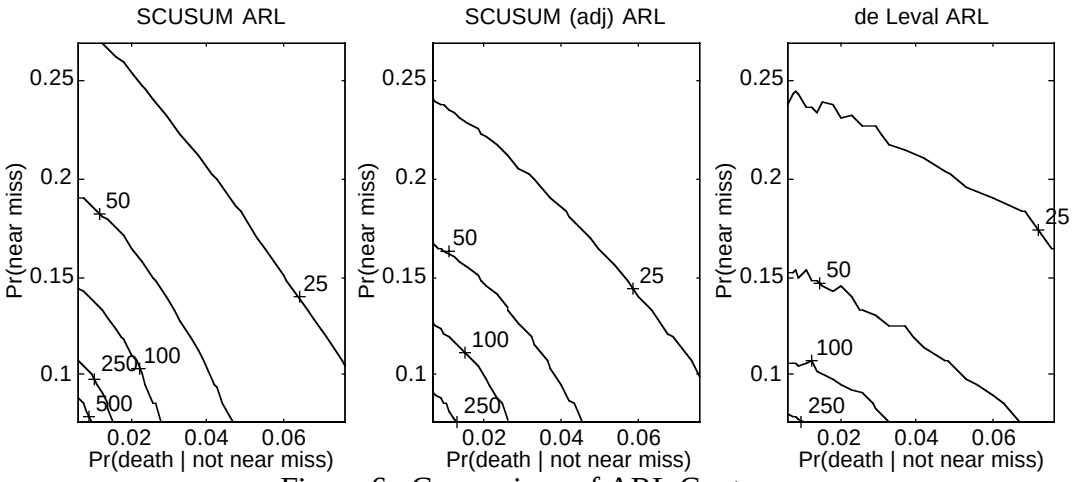

Figure 6: Comparison of ARL Contours

5. Summary and Discussion

Simultaneous CUSUM charts with secondary limits are developed and shown to be useful for monitoring processes with bivariate outcomes. The proposed CUSUM charts are preferable to multiple univariate charts since the correlation between the outcomes is taken into account in the

design of the chart. In addition, simultaneous secondary limit CUSUMs are also preferable to creating global statistic multivariate control charts, since it is often not clear how best to weight the different outcomes.

The run length properties of SCUSUM are found using Markov chain analysis. SCUSUMs are shown to be especially good at detecting simultaneous moderate increases (or decreases) in both quality process characteristics. The methodology is applicable in all situations where we have sequential paired (correlated) data, and are also especially useful for detecting gradual deterioration in both failure rates. The resulting control chart is very easy to implement and understand.

There are many interesting extensions possible for this work. In some situations it may be reasonable to assume that some underlying factor influences the occurrence of both types of failure. In that case a random effects model like (5) may be appropriate. Using a random effects model, the CUSUM weights could be set equal to the marginal log-likelihood ratio. Also, of great practical interest would be the addition of covariates. In the surgical outcome example, there were many covariates available, such as the occurrence of a related heart defect Ventricular Septal Defect (VSD), or the type of anesthetic used, that potentially could have a significant effect on the failure rates. Naturally, if important covariates are identified they should also be accounted for in the monitoring procedure. For covariates that have well known effects on the failure rates, perhaps as established through logistic regression, the CUSUM weights could be adjusted based on the value of the regression coefficients and the covariate itself. For example, the addition of the binary covariate VSD to model (6) results in a doubling of the number of CUSUM weights in Table 3.

In applications where the process outcome is not as critical as in the surgical failures example, it may be feasible take samples of size larger than one. In that case, for each of the two outcomes we would observe a binomial rather than Bernoulli random variable. Allowing samples of size n introduces another design parameter. By changing the sample size we would gain more flexibility in finding CUSUM designs that yield desirable ARL properties under both the null and alternative hypotheses.

In our motivating example the data were available immediately (within 24 hours) after surgery and therefore a bivariate binary response model was adopted. In other monitoring scenarios, such as when death is defined as a death within a month of the surgery, only censored observations may initially be available. In such a situation the monitoring procedure must be adapted since many failures are noted immediately, while successes are delayed. It may be possible to derive sliding scale weights that could be assigned to surgical survivor who can not yet be declared successes.

The principle behind the use of secondary limits is also appropriate if both the observed variables are count or continuous data, or some combination of data types. The secondary limits allow the rapid detection of quality decreases in both outcome variables. Unfortunately, however, for data other than paired binary, the procedure given in Appendix B for determining the SCUSUM's run length properties may no longer be computationally feasible. Finally, SCUSUM charts may also be applicable for three or more outcome variables of interest. With three or more secondary limit CUSUM charts, the combine chart may be defined to signal if any of the charts plotted above their primary limit, or if any two of the CUSUM statistics were above their corresponding secondary limit. Determination of the appropriate control limits in the case with three or more variables is somewhat more difficult and is worthy of further study.

Appendix A

The table below reproduces the surgical failure data from de Leval et al.[4].

Table A1: Surgical Failures Data

Pat #	Death	Near Miss	Pat #	Death	Near Miss	Pat #	Death	Near Miss	Pat #	Death	Near Miss
1	0	0	27	0	0	53	1	1	79	0	0
2	0	0	28	0	0	54	0	0	80	0	0
3	0	0	29	0	0	55	1	0	81	0	0
4	0	0	30	0	0	56	0	0	82	0	0
5	0	0	31	0	0	57	0	0	83	0	0
6	0	0	32	0	0	58	0	0	84	0	1
7	0	0	33	0	1	59	1	1	85	0	0
8	0	0	34	1	1	60	0	0	86	0	0
9	0	0	35	0	0	61	0	0	87	0	0
10	0	0	36	0	0	62	0	0	88	0	0
11	0	0	37	0	0	63	1	0	89	0	0
12	0	0	38	0	0	64	1	0	90	0	1
13	0	1	39	0	0	65	0	0	91	0	0
14	0	0	40	0	0	66	0	0	92	0	0
15	0	0	41	0	0	67	1	1	93	0	0
16	0	0	42	0	0	68	1	1	94	0	0
17	0	0	43	0	1	69	0	0	95	0	0
18	0	0	44	0	0	70	0	1	96	0	0
19	0	0	45	0	0	71	0	0	97	0	0
20	0	0	46	0	1	72	0	0	98	0	1
21	0	0	47	0	0	73	0	0	99	0	1
22	0	0	48	0	0	74	0	0	100	1	0
23	0	0	49	0	1	75	0	0	101	0	0
24	0	0	50	0	0	76	0	0	102	0	0
25	0	0	51	0	0	77	0	0	103	0	0
26	0	0	52	0	0	78	0	0	104	0	0

Appendix B

In this Appendix, Markov chain methodology is used to derive the run length properties of two simultaneous CUSUM charts with secondary control limits. Define the state space of the model as (Y, Z), where Y and Z represent the current values for the first and second CUSUM respectively. Define the SCUSUM as given by (2) and (3). We assume that $h_y > h_{yy}$ and $h_z > h_{zz}$, and that the control limits and all possible sample weights are integer. Discretizing the state space so that all values are integer we define the states as order pairs where

$$s_1 = (0,0),\ s_2 = (1,0),\ ....,\ s_{h_y} = (h_y - 1, 0),$$

$$s_{h_y+1} = (0,1),\ ...,\ s_{2h_y} = (h_y - 1, 1),\ ...,$$

$$s_{h_{zz}h_y+1} = (0, h_{zz}),\ ...,\ s_{h_{zz}h_y+h_{yy}} = (h_{yy} - 1, h_{zz}),$$

$$s_{h_{zz}h_y+h_{yy}+1} = (0, h_{zz}+1),\ ...,\ s_{h_{zz}h_y+(h_z-h_{zz})h_{yy}} = (h_{yy} - 1, h_z - 1).$$

The total number of states g equals $h_y h_{zz} + (h_z - h_{zz})h_{yy} + 3$, where the last three state correspond to the three possible out-of-control conditions given by (3). Depending on the values of h_y, h_z, h_{yy} and h_{zz}, g can be very large.

The transition probability matrix is given by

$$P = \begin{bmatrix} p_{11}, p_{12}, \cdots, p_{1g} \\ p_{21}, \cdots, p_{2g} \\ \vdots \qquad\qquad \vdots \\ p_{g1}, \cdots, p_{gg} \end{bmatrix} = \begin{bmatrix} R & , (I-R)\mathbf{1} \\ 0, \ldots, 0, & 1 \end{bmatrix},$$

where I is the g by g identity matrix, $\mathbf{1}$ is a g by 1 column vector of ones, and p_{ij} equals the transition probability from state s_i to state s_j. The last three rows and columns of the matrix P correspond to the three possible absorbing states that represents an out-of-control signal. The R matrix equals the transition probability matrix with the rows and columns that correspond to the absorbing (out-of-control) states removed. The R matrix is g by g and generally very large. In fact, given h_y = 32, h_z =70, h_{yy} = 17 and h_{zz} = 38 it is 1760x1760 elements. However, assuming paired binary outcomes there are only four possible outcomes, and thus there are at most four non-zero elements in each row of the matrix. As a result, the matrix R is very sparse.

Letting γ denote the run length of the CUSUM, and assuming the matrix R is constant, we have

$$\Pr(\gamma \le t) = \left(I - R^t\right)\mathbf{1}, \text{ and thus}$$
$$\Pr(\gamma = t) = \left(R^{t-1} - R^t\right)\mathbf{1} \quad \text{for } t \ge 1.$$

Therefore, the expected, or average, run length is

$$E(\gamma) = \sum_{t=1}^{\infty} t \Pr(\gamma = t) = \sum_{t=1}^{\infty} \left(R^t \mathbf{1}\right) = (I-R)^{-1}\mathbf{1}. \qquad \text{(A1)}$$

Higher moments of the run length can be found in a similar manner.

In general, solving (A1) is done without explicitly finding the inverse of $(I-R)$. A much more efficient approach is to solve the system of linear equations implied by (A1) via LU

decomposition and partial column pivoting to take advantage of the matrix sparseness. Routines given in MATLAB® can accomplish this task.

CUSUM charts are designed to eventually signal that the process has shifted to some undesirable process level. Let f_{ik} = probability of ending in absorbing state k, $k = g-2, g-1, \text{or } g$, when starting in state i. Let $f_{(0,0)\rightarrow\text{near miss signal}}$, $f_{(0,0)\rightarrow\text{death signal}}$, and $f_{(0,0)\rightarrow\text{joint signal}}$ denote the probability of signaling due to each of the three possible reasons for a signal. Then,

$$f_{ik} = p_{ik} f_{kk} + \sum_{j=1}^{g-3} p_{ij} f_{jk} \qquad \text{for } i = 1, 2, \ldots, g-2, g-3, \ldots, k. \tag{A2}$$

where $f_{g-2,g-2} = f_{g-1,g-1} = f_{gg} = 1$.

This system of linear equations may be solved for $f_{(0,0)\rightarrow\text{near miss signal}}$, $f_{(0,0)\rightarrow\text{death signal}}$, and $f_{(0,0)\rightarrow\text{joint signal}}$ using procedures similar to those used for (A1). Using MATLAB® on a Power Macintosh 7600 $E(\gamma)$ and f_{ik} values for any given probabilities of failure were determined in less than 10 seconds.

This numerical procedure can be used in the following manner to obtain appropriate control limits for an SCUSUM chart. The numerical routine require as inputs: the weights W_{Y_t}, W_{Z_t}, the parameters α_{y0}, α_{z0}, and β, and trial control limits h_y, h_z, h_{yy} and h_{zz}. Note that the weights and the parameter values remain fixed while we iteratively try different control limits. The numerical routine outputs the average run length, and $f_{(0,0)\rightarrow\text{near miss signal}}$, $f_{(0,0)\rightarrow\text{death signal}}$, and $f_{(0,0)\rightarrow\text{joint signal}}$ (denoted p_1, p_2 and p_3 in Section 4.2) that correspond to the given trial control limits. We are attempting to find the levels of the control limits that give some desired in-control ARL, ARL_0 and where $p_1 = p_2 = p_3$. We adjust the trial control limits based on the output from the routine as according to the suggestion in Table B1.

Table B1: Iterative Adjustments for Control Limit Determination

Outcome	Adjustment Needed
ARL smaller (larger) than ARL_0	increase (decrease) the control limits
p_1 too small (too large)	increase (decrease) h_y and increase (decrease) h_{yy} proportionally
p_2 too small (too large)	increase (decrease) h_z and increase (decrease) h_{zz} proportionally
p_3 too small (too large)	increase (decrease) h_{yy} and h_{zz}

In a more general case, the matrix R is a function of time, say R_t, and we have a non-homogenous Markov chain problem. This occurs under tool wear, or if we wish to model the deteriorating skills of an aging surgeon. We assume that t_{final} represents the time at which the probabilities of failure stop changing, and let R_{final} denote the transition probability matrix at time t_{final} and later. Then, using the same algebraic manipulations as used to derive (A1) we get

$$E(RL) = \sum_{t=1}^{t_{\text{final}}-1} \left(\prod_{s=1}^{t} R_s \, \mathbf{1} \right) + \left(\prod_{s=1}^{t} R_s \right) \left(I - R_{\text{final}} \right)^{-1} \mathbf{1}. \qquad \text{(A3)}$$

Define a_{tk} as the probability of being absorbed in absorbing state k, $k = 1,2,3$, at time t when starting in the first state s_1 = (0,0). The element of (a_{i1}, a_{i2}, a_{i3}) are found as the last three elements of the first row of the matrix $R_1 R_2 \ldots R_t$ for $t < t_{\text{final}}$. Then, assuming a changing transition probability matrix $\Pr(\text{absorb in state } k) = a_{1k} + a_{2k}\left(1 - \Pr(\gamma \le 2)\right) + a_{3k}\left(1 - \Pr(\gamma \le 3)\right) + \ldots + a_{t_{\text{final}}k}\left(1 - \Pr(\gamma \le t_{\text{final}})\right) + f_{ik(\text{final})} \Pr(\gamma > t_{\text{final}})$, where the last term is the probability of being absorbed into state k after the transition matrix stops changing. $f_{ik(\text{final})}$ can be found by solving the system of equations defined by (A2) with the transition probabilities p_{ij} given in R_{final}. The MATLAB® routines necessary to perform the calculations are available from the authors.

Acknowledgments

This research was supported, in part, by the Natural Sciences and Engineering Research Council of Canada and the Medical Research Council of Canada. The author would like to thank Dr. M. De Leval for providing access to the data, and his interest in this research. MATLAB® is a registered trademark of the MathWorks.

References

1. Cook, R.J. and Farewell, V. T. “Guidelines for Monitoring Efficacy and Toxicity in Clinical Trials,” *Biometrics*, **50**, 1146-1152 (1995).
2. Cook, R.J. “Interim Monitoring of Bivariate Responses Using Repeated Confidence Intervals, *Controlled Clinical Trials*, **15**, 187-200 (1994).

3. Cook, R.J. (1996), "Coupled Error Spending Functions for Bivariate Sequential Tests," *Biometrics*, **52**, 442-450.
4. De Leval, Marc R., François, K., Bull, C., Brawn, W.B., Spiegelhalter, D. "Analysis of a cluster of surgical failures," *The Journal of Thoracic and Cardiovascular Surgery*, March, 914-924 (1994).
5. Page, E. S. "Continuous Inspection Schemes," *Biometrika*, **41**, 100-114 (1954).
6. Barnard, G.A., "Control Charts and Stochastic Processes," *Journal of the Royal Statistical Society B*, **21**, 239-271 (1959).
7. Montgomery, D.C. *Introduction to Statistical Quality Control*, Second Edition, John Wiley and Sons, New York (1991).
8. Moustakides, G.V. "Optimal stopping times for detecting changes in distributions," *Annals of Statistics*, **14**, 1379-1387 (1986).
9. Woodall, W. H. "The design of CUSUM quality control charts," *Journal of Quality Technology*, **18**, 99- 102 (1986).
10. Gan, F.F. "An optimal design of CUSUM quality control charts," *Journal of Quality Technology*, **23**, 279-286 (1991).
11. Gan, F.F. "An optimal design of CUSUM control charts for binomial counts," *Journal of Applied Statistics*, **20**, 445-460 (1993).
12. Patel, H.I. "Quality Control Methods for Multivariate Binomial and Poisson Distributions," *Technometrics*, **15**, 103-112 (1973).
13. Crosier, R.B. "Multivariate Generalizations of Cumulative Sum Quality Control Schemes," *Technometrics*, **30**, 291-303 (1988).
14. Steiner, S.H., Geyer, P.L., Wesolowsky, G.O. "Shewhart Control Charts to Detect Mean Shifts Based on Grouped Data," *Quality and Reliability Engineering International*, **12**, 345-353 (1996).
15. Steiner, S.H., Geyer, P.L., Wesolowsky, G.O. "Grouped Data Sequential Probability Ratio Tests and Cumulative Sum Control Charts," *Technometrics*, **38**, 230-237 (1996).

16. Klotz, J. Statistical inference in Bernoulli trials with dependence. Annals of Statistics, **1**, 373-379 (1973).
17. Darlington, G.A. and Farewell, V.T. Binary longitudinal data analysis with correlation a function of explanatory variables. Biometrical Journal, **34**, 899-910 (1992).