Biostatistics (2000), **1**, 4, *pp*. 441–452
Printed in Great Britain

Monitoring surgical performance using risk-adjusted cumulative sum charts

STEFAN H. STEINER*, RICHARD J. COOK

Department of Statistics and Actuarial Sciences, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1
shsteine@uwaterloo.ca

VERN T. FAREWELL

Department of Statistical Sciences, University College London, UK

TOM TREASURE

St. George's Hospital Medical School London, UK

SUMMARY

The cumulative sum (CUSUM) procedure is a graphical method that is widely used for quality monitoring in industrial settings. More recently it has been used to monitor surgical outcomes whereby it 'signals' if sufficient evidence has accumulated that there has been a change in the surgical failure rate. A limitation of the standard CUSUM procedure in this context is that since it is simply based on the observed surgical outcomes, it may signal as a result of changes in the referral pattern, such as an increased proportion of high-risk patients, rather than due to a change in the actual surgical performance. We describe a new CUSUM procedure that adjusts for each patient's pre-operative risk of surgical failure through the use of a likelihood-based scoring method. The procedure is therefore ideally suited for settings where there is a variable mix of patients over time.

Keywords: Cumulative sum; Monitoring performance; Patient mix; Risk factors; Surgical outcomes.

1. INTRODUCTION

The need to formally monitor surgical outcomes has been brought to the forefront in some recent well-publicized cases (Treasure *et al.*, 1997; Waldie, 1998) where undesirable high rates of surgical complications remained undetected for an undue length of time. In such cases, the rapid detection of deterioration in surgical performance is critical since it will result in prompt investigation of the cause and procedural changes.

The cumulative sum (CUSUM) chart methodology was initially developed by Page (1954) for industrial problems where monitoring of the production process is of interest. In the industrial setting, CUSUM charts have been shown to be ideally suited to detecting small persistent process changes (Montgomery, 1991). In the medical context, CUSUMs have been proposed to monitor procedures in clinical chemistry (Nix *et al.*, 1986) and to monitor rare congenital malformations (Gallus *et al.*, 1986). Application of the

*To whom correspondence should be addressed

CUSUM to monitoring surgical performance was first proposed by Williams *et al.* (1992). The first application of a CUSUM chart to monitoring surgical performance is documented in De Leval *et al.* (1994) and Steiner *et al.* (1999) who considered the problem of monitoring outcomes in paediatric cardiac surgery. In this application none of the possible covariate information collected, such as patient characteristics, procedural characteristics, and surgical team fatigue, was found to have a significant effect on the failure rate. In the absence of informative covariates all patients were assumed to have the same surgical risk, and a standard CUSUM chart was applicable. However, in many medical contexts there is considerable variation in the characteristics of the people under study (i.e. heterogeneity). Patients with different clinical presentations and physiology have different prior risks. Thus, even for surgeons (or surgical teams) with an acceptable complication rate, the probability of a successful outcome may vary considerably across patients. In most medical applications this heterogeneity of patients must be taken into account in any monitoring scheme. A number of methods for surgical monitoring that take into account different levels of prior risk have recently been described. Lovegrove *et al.* (1997, 1999) and Poloniecki *et al.* (1998) suggest simple monitoring schemes based on a plot of the difference between the cumulative predicted and observed deaths. In both approaches the predicted number of deaths is estimated using the Parsonnet score (Parsonnet *et al.*, 1989) of each patient. The Parsonnet scoring system is widely applied in cardiac surgery and can be used to adjust different case mixes for risk. These charts, showing the difference between the cumulative predicted and observed deaths, provide valuable visual aids that show how the current surgical performance compares to past performance. It is difficult, however, to interpret the charts proposed by Lovegrove *et al.* (1997, 1999) since they do not specify how much variation in the plot is expected under good surgical performance, and hence how large a deviation from the expected should be a cause for concern. Poloniecki *et al.* (1998) suggest an intuitively sensible procedure for determining control limits. Essentially the scheme involves, after each observation, testing of the hypothesis that the failure rate in the last series of patients where we would expect 16 deaths is different than that in all the previous operations in the series. No formal adjustment for multiple testing is done although testing at a significance level of 0.01 is suggested. The authors acknowledge that this does 'not amount to a formal test of significance' and therefore consideration of, say, a false alarm rate does not make sense. Since the scheme involves continually updating our estimate of the desirable surgical performance using larger and larger series of historical data, the control limits effectively change over time. Updating is an advantage if we start with a poor estimate of the current performance, but also allows the possibility that small gradual changes in the performance will not be detected. This updating makes it very difficult to determine the chart's theoretical performance with respect to the usual run length criteria. However, we may determine the performance of a CUSUM based on an intuitively appealing observed–expected statistic. In Section 4 we compare the performance of a CUSUM with the observed–expected statistic and the CUSUM procedure proposed in Section 2.2.

To alleviate the problem of interpretation we propose the use of a new risk-adjusted CUSUM chart to monitor surgical outcomes, where the CUSUM procedure is adapted to address the level of pre-operative risk of each patient. The risk adjustment is made though a likelihood score. The risk-adjusted CUSUM procedure is illustrated with data from a UK centre for cardiac surgery. The data set is based on 6994 operations, from a single surgical centre over the seven-year period, 1992–1998. The data consist of information on each patient including date, surgeon, type of procedure and the pre-operative variables which comprise the Parsonnet score. These include age, gender, hypertension, diabetic status, renal function and left ventricular mass. To illustrate the methodology we focus on the 30-day post-operative mortality rate. In the data, 461 deaths occurred within 30 days of surgery, giving an overall mortality rate of 6.6%.

This article is organized as follows. In Section 2, both the standard (unadjusted) CUSUM and risk-adjusted CUSUM are defined. To illustrate the advantage of using a risk-adjusted CUSUM we apply the risk-adjusted CUSUM to the cardiac surgery example in Section 3. In Section 4 we turn to a more detailed discussion of the properties of the new procedure and design issues related to the risk-adjusted CUSUM. We make concluding remarks in Section 5.

2. Cumulative sum (CUSUM) procedure

2.1. Standard CUSUM

The CUSUM procedure is a well-established sequential monitoring scheme designed to detect changes in a process parameter of interest, denoted by, say, θ. The original formulation of the CUSUM is due to Page (1954). Two-sided implementations suggested by Barnard (1959) involved the use of a graphical device, called a V-mask. Unfortunately the V-mask is awkward to use in practice. An easier to use tabular form of the CUSUM can detect increases (or decreases) in θ. Using two tabular CUSUMs in conjunction accomplishes the goal of detecting any process changes. A standard tabular CUSUM involves monitoring

$$X_t = \max(0, X_{t-1} + W_t), \quad t = 1, 2, 3, \ldots, \tag{2.1}$$

where $X_0 = 0$, and W_t is the sample weight or score assigned to the tth subgroup. Subgroups are a collection of units taken from the production process at roughly the same time. Through a judicious choice of W_t the CUSUM can be designed to detect increases or decreases in θ. The CUSUM given by (2.1) sequentially tests the hypothesis $H_0 : \theta = \theta_0$ versus $H_A : \theta = \theta_A$. The value of θ_0 is typically determined by the current process performance, while θ_A represents an alternate value of interest, corresponding typically to inferior performance. 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_A if $X_t \geq h$ at any time t. The constant h is called the control limit of the CUSUM. In quality-control terminology, a CUSUM that exceeds the control limit is said to have 'signalled'. A signal means that the chart has accumulated enough evidence to conclude that the process parameter has changed.

Notice that although individual scores (W_t) may be negative, the tabular CUSUM based on X_t is restricted to non-negative values to make the CUSUM sensitive to runs of poor performance. CUSUMs are designed to monitor the responses sequentially until sufficient evidence of process deterioration is detected. Thus, theoretically the CUSUM will eventually signal, although the signal may be a false alarm. The run length of the CUSUM is defined as the time (or number of observations) required before the CUSUM first exceeds the control limit (i.e. signals). Good choices for the control limit h are based on the expected or average run length (ARL) of the CUSUM under H_0 and H_A. Ideally, while the process is in state H_0 the run length should be long, since in this context signals represent false alarms. On the other hand, if the process has shifted to θ_A, or any other process setting substantially different than θ_0, we would like short run lengths.

The ARL under H_0 may be considered analogous to the type I error rate of a traditional statistical test. However, there is no generally accepted level (like 5% tests) since what is acceptable varies considerably from application to application. Similarly, the ARL of the CUSUM when θ has changed substantially is analogous to the power of a traditional statistical test. Determining the average run length of a CUSUM is computationally intensive since it is based on all possible outcomes for a long series of surgeries. The ARL however may be closely approximated (see the Appendix).

The design of the CUSUM is given by the choice of sample weight W_t and control limit h. Moustakides (1986) showed that the optimal choice for the tabular CUSUM weights W_t is based on the log-likelihood ratio. For example, if we let y represent the current outcome (which could be the average or some other summary statistic of a subgroup of units collected around the same time) and denote the probability distribution of the possible subgroup outcomes as $f(y; \theta)$, the log-likelihood ratio is given by $\ln(f(y; \theta_A)/f(y; \theta_0))$. This choice is optimal in the sense that, among all schemes with the same ARL under H_0, the log-likelihood ratio weights give the smallest ARL under H_A. The choice of control limits for CUSUM procedures has been discussed for normally distributed outcomes (Woodall, 1986) and in the binomial data case (Gan, 1991).

When applying the standard CUSUM methodology to monitoring surgical performance, as in Steiner *et al.* (1999), we can be more specific regarding the definition of a subgroup and the form of the scores

W_t. First, due to the critical nature of surgery, we update the CUSUM after each patient to be sure to detect changes as quickly as possible. Thus, the outcome y corresponds to one of two possible outcomes (success or failure) for each patient, and the CUSUM is a sum of scores taken over *all* patients operated on from the start of monitoring. Assuming y_t is the outcome for patient t, and that $y_t = 1$ if patient t dies and $y_t = 0$ otherwise, we have $f(y_t|\theta) = p(\theta)^{y_t}[1 - p(\theta)]^{1-y_t}$, where $p(\theta_0) = c_0$, the estimated current failure rate, and $p(\theta_A) = c_A$, an important change in the failure rate. Note that c_A may represent a deterioration or improvement in the surgical failure rate. As a result, the CUSUM sequentially tests the hypothesis H_0 $p = c_0$ versus H_A $p = c_A$ and the CUSUM scores are:

$$W_t = \begin{cases} \log([1 - c_A]/[1 - c_0]) & \text{if } y_t = 0 \\ \log(c_A/c_0) & \text{if } y_t = 1 \end{cases}. \tag{2.2}$$

When designing the chart to detect increases in the surgical failure rate, the scores associated with failures will be positive, while successes receive a negative score. In this way, the weight for each patient is based on three factors: the current acceptable level of surgical performance (c_0), a chosen level of surgical failure rate reflecting a change in performance deemed interesting (c_A), and the actual surgical outcome for the patient (y_t).

2.2. Risk-adjusted CUSUM

In most surgical contexts the risk of mortality estimated pre-operatively will vary considerably from patient to patient. An adjustment for prior risk is therefore appropriate to ensure that mortality rates that appear unusual and arise from differences in patient mix are not incorrectly attributed to the surgeon. We can adjust the CUSUM based on prior risk by adapting the magnitude of the scores using the patient's surgical risk, estimated pre-operatively. The surgical risk varies for each patient depending on risk factors present. We define $p_t(\theta) = g(\theta, \mathbf{x}_t)$, where $\mathbf{x}_t = (x_{t1}, x_{t2}, \ldots, x_{tp})^T$ is a px1 vector reflecting the risk factors for patient t. The function g may be determined pre-operatively using a rating method such as Parsonnet risk factors (Parsonnet *et al.*, 1989), or may be based on a logistic regression model fitted to sample data. Since each patient has a different baseline risk level we define the hypotheses H_0 and H_A based on an odds ratio. Let R_0 and R_A represent the odds ratios under null and alternate hypotheses, respectively. To detect increases we set $R_A > R_0$. The choice of R_A is similar to defining the minimal clinically important effect in a clinical trial. If the estimated risk p_t is based on the current conditions we may set $R_0 = 1$. Given an estimated risk of failure equal to p_t, the odds of failure equals $p_t/(1 - p_t)$. Thus, for patient t under H_0 the odds of failure equals $R_0 p_t/(1 - p_t)$, whereas under H_A the odds of failure is $R_A p_t/(1 - p_t)$, which corresponds to a probability of failure equal to $R_A p_t/(1 - p_t + R_A p_t)$ under H_A. In this way the CUSUM repeatedly tests

$$H_0 : \text{ odds ratio } = R_0 \text{ versus}$$
$$H_A : \text{ odds ratio } = R_A.$$

Then, the two possible log-likelihood ratio scores for patient t are:

$$W_t = \begin{cases} \log\left[\dfrac{(1 - p_t + R_0 p_t)R_A}{(1 - p_t + R_A p_t)R_0}\right] & \text{if } y_t = 1 \\ \log\left[\dfrac{1 - p_t + R_0 p_t}{1 - p_t + R_A p_t}\right] & \text{if } y_t = 0 \end{cases} \tag{2.3}$$

Note that other weights are possible. For example, we may wish to create a CUSUM using weights based on calculating observed deaths minus the expected deaths. This scheme is similar to that proposed

by Lovegrove *et al.* (1997, 1999) and Poloniecki *et al.* (1998), and is equivalent to monitoring cumulative patient weights, where each patient's weight is $1 - p_t$ if the patient dies or $-p_t$ if the patient survives. However, as we shall illustrate in Section 4, the weights given by (2.3) are optimal (Moustakides, 1986) to detect shifts in the odds ratio to R_A.

The CUSUM signals whenever $X_t \geq h$. There then arises the critical question of how to respond when the chart signals. First, we cannot assume that all possible risk-adjusting factors have been taken into account in the mathematical model. The deaths that occurred prior to a CUSUM signal and which are therefore responsible for the alarm must be scrutinized by a clinically informed but independent assessor. It may be that there were risk factors inherent in the patient which were not detected in the risk-prediction model. For instance, the Parsonnet scoring system omits factors such as the technical difficulty of the coronary arteries and the presence of lung disease. Both were omitted for good reason. Both are amenable to subjective loading by the surgeon who can thus unwittingly (or knowingly) load the prior risk to such an extent that he or she always appears to perform better than predicted. These and other factors might be recognized by another surgeon. Factors which are occasional, or dependent on judgement, may be better left as that—a matter of judgement. However, even if the surgeon's performance is exonerated at a technical level, one may still ask if the clinical risk was fully appreciated. Failure of good clinical decision making is as great a problem as failure of technical expertise. If surgical performance consistently appears to be different than predicted, but this difference can be explained by new risk factors, these factors should be incorporated into the model used to estimate prior risk.

Alternatively, the assessor may find that the fault lies not with the surgeon but is attributable to another member of the team. There may have been incomplete pre-operative or poor post-operative care. The CUSUM can only signal that there have been more deaths than were expected but it cannot seek out the precise cause or its solution. Once the cause has been established there must begin a process to resolve the problem, which may include retraining, mentoring and further monitoring. These are outside the scope of this paper.

3. Example

To illustrate the characteristics of the risk-adjusted CUSUM we use the cardiac surgery example described in the introduction. Since CUSUM procedures are designed to quickly detect changes in the surgical performance we must first estimate the current level of performance. In the cardiac surgery example, the data includes the patient characteristics and surgical outcomes for all patients seen at a single surgical centre between 1992 and 1998. During that time no formal monitoring was done. To illustrate the proposed monitoring procedure we suppose that monitoring was begun in 1994, and use the first two years of data (corresponding to 1992 and 1993) to identify the risk factors and estimate their effects through a logistic regression model. In the first two years a total of 2218 surgeries were performed and we observed 143 deaths for a mortality rate of 6.5%.

Using backward elimination we found the logistic model given by (3.4), with only a Parsonnet score as an explanatory variate, was appropriate:

$$\text{logit}(p_t) = -3.68 + 0.77X_t \tag{3.4}$$

where X_t denotes the Parsonnet score for patient t, $t = 1, 2, \ldots$. Since the Parsonnet score itself is based on a combination of many other explanatory variates thought to be important in cardiac surgery, this is not surprising. Based on this model, the lowest risk patients in the group (Parsonnet score $= 0$) were estimated to have a risk of death of just 2.5% following surgery, while the patients with the highest risk (Parsonnet score $= 71$) had an estimated mortality rate of 86%. This suggests that adjustment for the patient mix is critical.

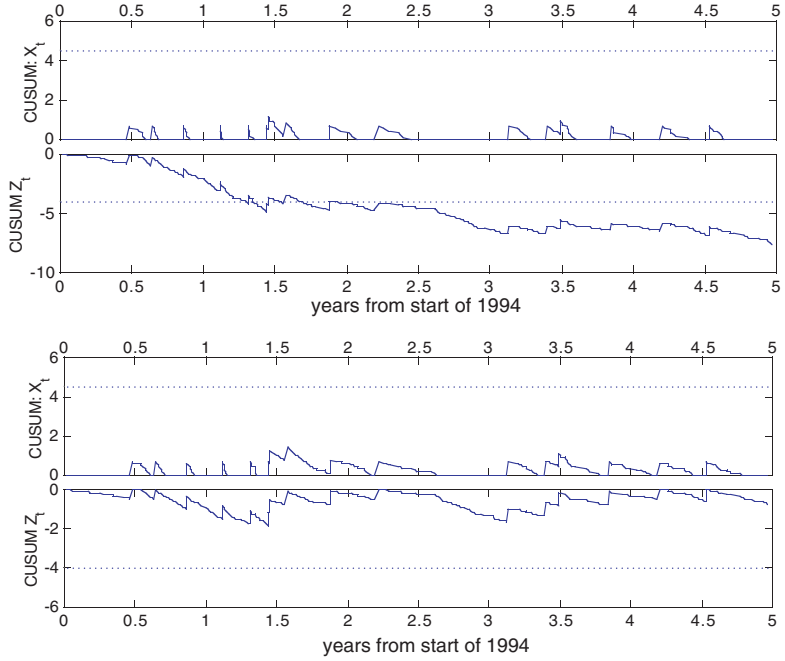

Fig. 1. Trainee surgeons CUSUM; unadjusted CUSUMs on top, risk-adjusted CUSUMs on the bottom.

In this application we use two CUSUM procedures. The first is designed to detect a doubling of the odds of death (i.e. we choose odds ratios $R_0 = 1$ and $R_A = 2$), and the second is designed to detect a halving of the odds of death (i.e. we choose odds ratios $R_0 = 1$ and $R_A = 0.5$). In this way, we should be quickly aware of any substantial changes in the failure rate. One could easily only monitor for increases in the failure rate if preferred. However, the CUSUM designed to detect decreases in the surgical failure rate is useful because if we have either over-estimated the failure rate, or the surgical performance has actually improved, the CUSUM designed to detect increases will be less sensitive. If the CUSUM designed to detect improvements in performance signals, we should re-estimate the failure rate to ensure protection for future changes in the rate.

The CUSUM scores are calculated using (3.4) in combination with (2.3), where the positive score is assigned in the case of a death, and the negative score is assigned in the case of survival. For example, the CUSUM chart designed to detect an increase in the failure rate gives the following possible patient scores: 0.67 and -0.024 for patients with a Parsonnet score of zero, and 0.26 and -0.43 for higher risk patients with a Parsonnet score of 50. Notice that the scores reflect the surgical risk assessed pre-operatively, since the 'penalty' for a death of a low-risk patient is more severe than for a death of a higher risk patient.

For ease of implementation and presentation, the CUSUM designed to detect decreases in the surgical failure rate will accumulate negative values, i.e. the updating formula will be given by

$$Z_t = \min(0, Z_{t-1} - W_t) \tag{3.5}$$

where $Z_0 = 0$ and W_t is still given by (2.3), rather than that given by (2.1). In addition, for the CUSUM designed to detect decreases in the surgical failure rate we set the control limit to a negative value. In this way, the CUSUM to detect decreases can be plotted underneath the CUSUM to detect increases. See Figure 1 for an example.

The choice of control limits for the CUSUM charts is discussed in more detail in the next section. In the cardiac surgery example, setting the control limits for the two CUSUM charts given by (2.1) and (3.5) at 4.5 and -4 respectively gives an average run length of around 9600 patients for each chart when the surgical performance is acceptable. Given the frequency of surgery in this example, this implies a signal from the monitoring procedure, on average, once every nine years even if no true changes in the death rate have occurred. The control limits can be altered to achieve other desired design objectives.

To illustrate the effectiveness and desirability of the risk adjustment we stratify the series based on surgeon. In this example, patient mix refers to the distribution of Parsonnet scores assigned to the surgeon under consideration. The patient mix for each surgeon differed, with more experienced surgeons typically receiving higher risk cases. Figure 1 shows the CUSUM where we only include the patients operated on by trainee surgeons throughout the period. The trainees had a mortality rate of just 2.5%, substantially lower than the overall rate, but they dealt with only the relatively straightforward cases. Furthermore, if during an operation, serious difficulties arose then a consultant (experienced) surgeon would take over and the case would then be attributed to the consultant. As a result, one might expect fewer deaths in a case series attributed to the trainees than expected, based on a risk measure like the Parsonnet score. Similarly, higher than expected rates might be expected for a consultant supervising trainee surgeons. Surgical training is discussed further by Anderson *et al.* (1996). Using a standard (unadjusted) CUSUM procedure we obtain the pair of CUSUMs given in the top plot of Figure 1. This plot suggests that the performance of trainee surgeons appears to be substantially better than their peers, since the CUSUM that monitors for improvement in performance signals around March of 1995. Due to the magnitude of the control limit we are fairly sure that this signal has not occurred simply due to good luck. Observing such a signal in the CUSUM, we would react by attempting to determine why trainee surgeons were doing so well, in the hope that the success could be replicated by other surgeons. For example, other surgeons might try to copy the surgical procedure of the 'good' surgeons. Based on changes in the surgical process we would reset the CUSUM values to zero for all surgeons to see if the changes were effective.

When we look at the risk-adjusted CUSUM series (bottom pair of CUSUMs in Figure 1) we see that, in fact, based on the patient mix, the trainee surgeons are doing as well as expected, but not better. The signal we observe in the standard CUSUM is due to the lower risk patient mix given to the trainees. As a result, the conclusion from the unadjusted CUSUM was in error, and all the time and effort devoted to searching for a cause of what appeared to be improved performance would have been for nothing.

We next examine the performance of an experienced surgeon. In Figure 2, the top pair of CUSUMs result from using a standard CUSUM with unadjusted scores, while the bottom set of CUSUMs come from using the risk-adjusted weights. In this case, based on the unadjusted scores, the experienced surgeon appears to be doing substantially worse than their peers. The unadjusted CUSUM signals an increase in the mortality rate just before the end of 1994. Observing such a signal we look to establish the cause of the 'problem'. However, when we adjust for the patient mix encountered by the experienced surgeon the signal disappears (see bottom of Figure 2). Thus, again the conclusion from the unadjusted CUSUM was a mistake and all the effort devoted to finding a cause of the problem would have been wasted.

4. Characteristics of the risk-adjusted CUSUM procedure

A key step in the design of the CUSUM procedure is setting the control limit. Using the Markov chain procedure presented in the Appendix we can closely approximate the ARL of a chart for any control limit, actual performance level and patient mix. Using this approximation, we choose the control limit so that the ARL, given the current (estimated) performance level and patient mix, is relatively long. Control limits further from zero will lead to longer ARLs, and what is considered long depends on the application. However, there is a tradeoff in the choice of the control limit. A control limit further from zero will also result in a longer ARL when the surgical performance has changed.

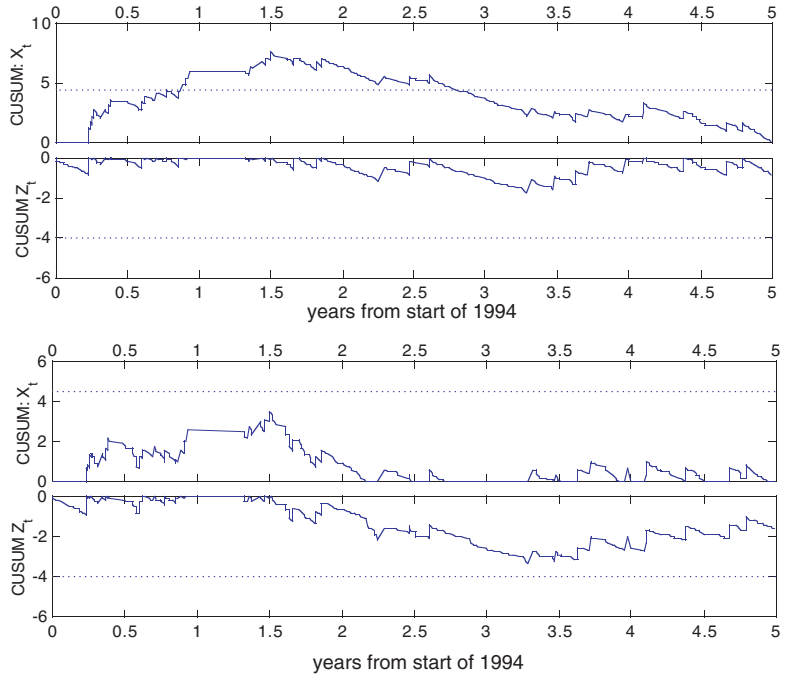

Fig. 2. Experienced surgeon CUSUM; unadjusted CUSUMs on the top, risk-adjusted CUSUMs on the bottom.

To explore this we can quantify the ability of the CUSUM procedure to detect increases (or decreases) in the odds of death. Figure 3 shows a plot of the average run length versus a measure of surgical performance (given in terms of the odds ratio) for different patient mixes. The ARLs are given for the CUSUM designed to detect increases in the failure rate. The acceptable level of surgical performance is given by an odds ratio equal to unity, while an increase in the odds ratio signifies a deterioration of performance. Note that the ARLs are given on a log scale. In Figure 3, the solid lines give the results for the current patient mix for all surgeons observed in the first two years of data. For comparison, we also show the ARL curve that results when using weights determined by cumulative predicted deaths minus the cumulative observed deaths as discussed at the end of Section 2. As suggested by theoretical considerations, our proposed weighting scheme based on log-likelihood ratios is superior, in that for substantial shifts in the odds ratio the ARL of our CUSUM is much shorter. For example, although both weighting methods are designed to have an in-control ARL of 9600, the CUSUM based on log-likelihood ratio weights has an ARL of 215 when the true odds ratio is 2, compared with an ARL of 485 for the predicted minus observed weights. To provide some sensitivity analysis for our proposed procedure, we also plot two dotted lines in Figure 3 which give the results for the patient mix of the two surgeons in our example with the most different patient mixes, based on the average Parsonnet scores, in the sample data.

To quantify the sensitivity of the proposed procedure to the initial estimate of the patient mix and regression parameters we use the bootstrap procedure (Efron and Tibshirani, 1993) as follows. We generate the bootstrap samples of 2218 observations from the data given in the first two years of the sample data by sampling with replacement. Based on the bootstrap sample, we estimate the parameters in the model given by (3.4). This defines the likelihood scores given by (2.3). Figure 4 shows the 5th and 95th percentiles for the generated CUSUM paths for the five years of data used to create Figures 1 and 2 based on 1000 such bootstrap samples.

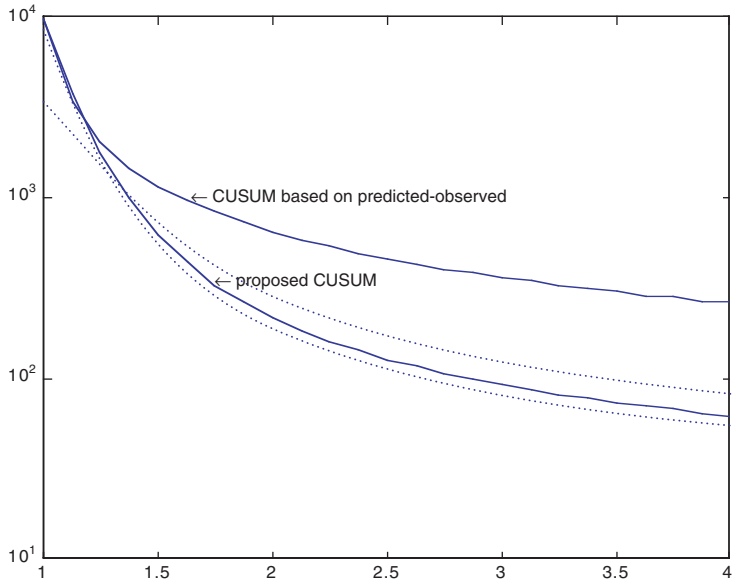

Fig. 3. Average run length (on a log scale) for different actual odds ratios. Solid line shows performance with current patient mix. Dotted lines—surgeon with the lowest and highest risk patients.

The results from Figure 4 suggest that our original conclusion, that both surgeons are performing as expected (once we adjust for patient mix), is relatively robust. In the case of the trainee surgeons, even in the worst case, the CUSUM path does not cross the control limits. For the experienced surgeon there is more uncertainty. Although the 95th percentile of the CUSUM path exceeds the upper control briefly near the middle of 1995, the CUSUM quickly declines again suggesting an anomaly. Contrast this plot with the unadjusted CUSUM shown in Figure 2 where the CUSUM path remained above the control limit for an extended period of time, and we would have concluded that the experienced surgeon was performing more poorly than their peers. In the bottom CUSUMs in Figure 4 near the end of the series, the extreme cases of the CUSUM path also cross the lower control limit. Here the CUSUM remains below the control limit for more time, suggesting some evidence that the experienced surgeon may be actually doing better than expected. We also generated the 5th and 95th percentile of the CUSUM paths based on bootstrap samples for the unadjusted CUSUM. These results are not shown here since the conclusions from the plots are unchanged.

5. Conclusions

We have proposed a new CUSUM chart to monitor surgical performance in which the scores are adjusted to reflect the pre-operative estimate of the surgical risk of each patient. This approach provides a logical way to accumulate evidence over many patients, while adjusting for a changing mix of patient characteristics that significantly affect the risk. This is particularly important when monitoring outcomes of surgery at referral centres where referral patterns may change over time. Through use of the CUSUM procedure the sensitivity of the chart can be set so that false alarms do not happen very frequently, but substantial changes in the failure rate are quickly detected. This approach is appealing since the ability of the chart to detect specific changes can be easily quantified.

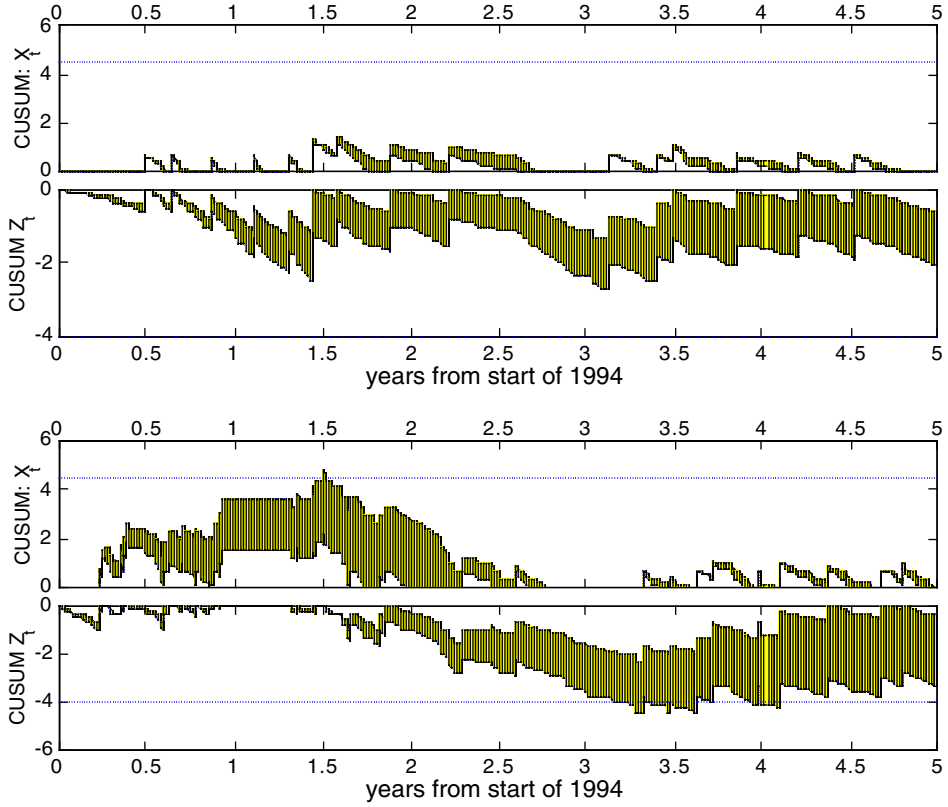

Fig. 4. Plot of the 5th and 95th percentile of the CUSUM paths based on 1000 bootstrap samples; trainee surgeons on the top, experienced surgeon on the bottom.

APPENDIX

In this Appendix, Markov chain methodology is used to derive approximate run length properties of a CUSUM chart as defined by (2.1). The approximation can, in theory, be made as precise as desired through scaling. To use the Markov chain methodology, we discretize the state space of all possible CUSUM values given by X_t into g + 1 states. The CUSUM is constrained to lie between 0 and h. To ensure all values are integers we define the states as 0 to g, where state 0 corresponds to the starting state, and state g corresponds to the absorbing state where $X_t > h$. This can be accomplished by scaling and rounding off. Using this new definition of the possible states of the CUSUM the transition probability matrix is given by:

$$Q = \begin{bmatrix} q_{00}, q_{01}, \ldots, q_{0g} \\ q_{10}, \quad \cdots \quad, q_{1g} \\ \vdots \qquad\qquad \vdots \\ q_{g0}, \quad \cdots \quad, q_{gg} \end{bmatrix} = \begin{bmatrix} R \quad, (I-R)\mathbf{1} \\ 0, \ldots, 0, \quad 1 \end{bmatrix},$$

where I is the g by g identity matrix, $\mathbf{1}$ is a g by 1 column vector of ones, and q_{ij} equals the transition probability from state i to state j. The q_{ij}s can be estimated using (2.2) and knowledge of the distribution

of p_t which is given by the patient mix. The last row and column of the matrix Q corresponds to the absorbing state that represents an out-of-control signal. As such, the R matrix equals the transition probability matrix with the row and column that correspond to the absorbing (out-of-control) state removed. To get good approximations we use g approximately equal to 250.

For example, we can illustrate the calculation of the transition probabilities. Consider patients whose Parsonnet score is zero. Using (3.4), such low-risk patients are estimated to have a 2.5% chance of mortality. So from (2.3), and assuming $OR_0 = 1$ and $OR_A = 2$, the possible scores for such a patient are -0.024 and 0.669. In the example, we choose the control limit of the CUSUM, h, equal to 4.5. As an example of the discretization we multiply by 60 and round off to the nearest integer. This results in a control limit of 270 ($g = 271$) and scores for the low-risk patient equal to either -1 if the patient survives, or 40 if the patient dies. To build the transition probability matrix from such information we must also know the distribution of Parsonnet scores for the patient mix. For example, 22% of the patients operated on by the trainee surgeons have a Parsonnet score equal to zero. Thus, for the trainees, assuming the surgical process is in-control and therefore that the actual risk of death for patients with Parsonnet score equal to zero is 2.5%, we obtain the -1 score 21.45% (22×0.975) of the time, and the score of 40, 0.55% (22×0.025) of the time. Looking at all the possible types of patients in this way we can determine the distribution of the discretized CUSUM scores, and thus determine the transition probabilities q_{ij}. Consider transitions from state zero and assuming that only patients with Parsonnet equal to zero can yield a score of 40, we set $q_{0,40} = 0.0055$.

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

$$\begin{aligned} \Pr(\gamma \leq t) &= (I - R^t)\mathbf{1}, \text{ and thus} \\ \Pr(\gamma = t) &= (R^{t-1} - R^t)\mathbf{1} \quad \text{for } t \geq 1. \end{aligned}$$

Therefore, the expected, or average run length is:

$$E(\gamma) = \sum_{t=1}^{\infty} t \Pr(\gamma = t) = \sum_{t=1}^{\infty} (R^t \mathbf{1}) = (I - R)^{-1}\mathbf{1}. \tag{A.1}$$

Higher moments of the run length can be found in a similar manner, but are not used here.

In general, solving (A.1) 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 (A.1) via LU decomposition.

ACKNOWLEDGEMENTS

This research was supported, in part, by the Natural Sciences and Engineering Research Council of Canada and the Medical Research Council of Canada. R. J. Cook is a Scholar of the Medical Research Council of Canada.

REFERENCES

ANDERSON, J. R., UNSWORTH-WHITE, M., VALENCIA, O., PARKER, D. J. AND TREASURE, T. (1996). Training and safeguarding patients. *Annals of the Royal College of Surgeons of England* **78** (Suppl), 116–118.

BARNARD, G. A. (1959). Control charts and stochastic processes. *Journal of the Royal Statistical Society*, Series B **21**, 239–271.

DE LEVAL, MARC R., FRANÇOIS, K., BULL, C., BRAWN, W. B. AND SPIEGELHALTER, D. (1994). Analysis of a cluster of surgical failures. *The Journal of Thoracic and Cardiovascular Surgery*, **104**, 914–924.

EFRON, B. AND TIBSHIRANI, R. J. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.

GALLUS, G., MANDELLI, C., MARCHI, M. AND RADAELLI, G. (1986). On surveillance methods for congenital malformations. *Statistics in Medicine* **5**, 565–571.

GAN, F. F. (1991). An optimal design of CUSUM quality control charts. *Journal of Quality Technology* **23**, 279–286.

LOVEGROVE, J., VALENCIA, O., TREASURE, T., SHERLAW-JOHNSON, C. AND GALLIVAN, S. (1997). Monitoring the results of cardiac surgery by variable life-adjusted display. *Lancet* **18**, 350, 1128–1130.

LOVEGROVE, J., SHERLAW-JOHNSON, C., VALENCIA, O., TREASURE, T. AND GALLIVAN, S. (1999). Monitoring the performance of cardiac surgeons. *Journal of the Operational Research Society* **50**, 684–689.

MONTGOMERY, D. C. (1991). *Introduction to Statistical Quality Control*, 2nd edn. New York: John Wiley and Sons.

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

NIX, A. B., ROWLANDS, R. J. AND KEMP, K. W. (1986). Internal quality control in clinical chemistry: a teaching review. *Statistics in Medicine* **6**, 425–440.

PAGE, E. S. (1954). Continuous inspection schemes. *Biometrika* **41**, 100–114.

PARSONNET, V., DEAN, D. AND BERNSTEIN, A. D. (1989). A method of uniform stratification of risks for evaluating the results of surgery in acquired adult heart disease. *Circulation* **779** (Suppl 1), 1–12.

POLONIECKI, J., VALENCIA, O. AND LITTLEJOHNS, P. (1998). Cumulative risk adjusted mortality chart for detecting changes in death rate: observational study of heart surgery. *British Medical Journal* **316**, 1697–1700.

STEINER, S., COOK, R. AND FAREWELL, V. (1999). Monitoring paired binary surgical outcomes using cumulative sum charts. *Statistics in Medicine* **18**, 69–86.

TREASURE, T., TAYLOR, K. AND BLACK, N. (1997). *Independent Review of Adult Cardiac Surgery—United Bristol*. Bristol: Health Care Trust, March.

WALDIE, P. (1998). Crisis in the Cardiac Unit. *The Globe and Mail*, Canada's National Newspaper, Oct. 27; Sect. A:3 (col. 1).

WILLIAMS, S. M., PARRY, B. J. AND SCHLUP, M. M. (1992). Quality control: an application of the CUSUM. *British Medical Journal* **304**, 1359–1361.

WOODALL, W. H. (1986). The design of CUSUM quality control charts. *Journal of Quality Technology* **18**, 99–102.

[*Received November 17, 1999*; *revised March 15, 2000*; *accepted for publication April 5, 2000*]