Risk Adjusted Monitoring of Binary Surgical Outcomes

Stefan Steiner and Richard Cook
University of Waterloo
Vern Farewell
University College, London, England

March 2000

Risk Adjusted Monitoring of Binary Surgical Outcomes

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

Vern T. Farewell Ph.D.
Department of Statistical Sciences
University College, London, England

Abstract

A graphical procedure suitable for prospectively monitoring surgical performance is proposed. The approach is based on accumulating evidence from the outcomes of all previous surgical patients in a series using a new type of cumulative sum chart. Cumulative sum procedures are designed to "signal" if sufficient evidence has accumulated that the surgical failure rate has changed substantially. In this way, the chart rapidly detects deterioration (or improvement) in surgical performance, while not over reacting to the expected fluctuations due to chance. Through the use of a likelihood-based scoring method, the cumulative sum procedure is adapted so that it adjusts for the surgical risk of each patient estimated pre-operatively. The procedure is therefore applicable in situations where it is desirable to adjust for a mix of patients. Signals of the chart lead to investigations of the cause and to the timely introduction of remedial measures designed to avoid unnecessary future failures.

Key Words: 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[1,2] 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 should result in prompt investigation of the cause and procedural changes.

A number of methods for surgical monitoring have recently been described. Lovegrove et al.[3] and Poloniecki et al.[4] suggest simple monitoring schemes based on a plot of the difference between the cumulative expected number of deaths and cumulative observed deaths. These charts provide valuable visual aids that show how the current surgical performance compares to past performance. However, the charts 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.

De Leval et al.[5] and Steiner et al.[6] propose an alternative surgical monitoring procedure based on a cumulative sum (CUSUM) chart which uses a methodology borrowed from an industrial context where process monitoring has been extensively studied[7]. In the industrial setting, CUSUM charts have been shown to be ideally suited to detecting relatively small persistent changes in the event rate over time. Traditional CUSUM approaches, however, make no adjustment for different risk profiles because machine inputs are usually relatively homogenous and such adjustments are not required in industrial settings. In contrast, patients undergoing a particular surgical intervention are often very heterogeneous in their clinical presentation and physiology. This means that even for a surgeon with an acceptable complication rate, the probability of a successful outcome may vary considerably across patients.

We propose the use of a CUSUM chart to monitor surgical outcomes, where the CUSUM procedure is adapted to address the level of pre-operative risk. The procedure is illustrated with sample data, kindly supplied by Professor M. de Leval, from a UK study of neonatal arterial switch operations for transposition of the great arteries. In the example, patient survival status constitutes the response and gender and the arterial pattern or diagnosis are the two risk factors of

primary interest that characterize the patient mix. The data set is based on 230 operations, from a number of surgical centres over a three year period. To illustrate the methodology we use a random ordering of the observations, and monitor the post-operative mortality rate of this artificially ordered series of 230 surgeries as if they came from one centre over a three year period.

2. Standard CUSUM Procedure

A CUSUM procedure is a monitoring scheme that may be used to accumulate evidence regarding the recent level of surgical performance[6]. The idea is to monitor surgical performance prospectively to detect as quickly as possible if the level of performance has changed. The cumulative sum is a sum of scores where each patient contributes a score. The sum is taken over all patients operated on from the start of monitoring until the point of observation. Mathematically, a CUSUM chart involves plotting X_t versus t, where

$$X_t = \max(0, X_{t-1} + w_t) \qquad (1)$$

and $t = 1, 2, 3, \ldots$, $X_0 = 0$, and w_t is the score assigned to patient t. In the standard CUSUM a patient's score is based on their surgical outcome (success or failure), the acceptable overall death rate and a change in the death rate deemed to be important. The acceptable death rate could be estimated from previous data, or a desired rate could be obtained from other surgical centers. See De Leval et al.[5] and Steiner et al.[6] for an example. When designing the chart to detect increases in the surgical failure rate, we define scores associated with failures to be positive, while successes receive a negative score. We assume that at any point in time the surgical performance may change (improve or deteriorate). As such, although individual scores may be negative, the CUSUM is restricted to nonnegative values to make the CUSUM sensitive to recent runs of poor performance.

The CUSUM value (1) has accumulated the information from all previous surgeries, and will become large if the surgical performance level has deteriorated, but will fluctuate close to zero for a long time if no change has occurred. The surgical process is assumed to be acceptable as long as the CUSUM remains below a predetermined value, denoted h, called the control limit. When the

CUSUM exceeds the control limit we conclude enough evidence of a change in surgical failure rate has accumulated, and we say the CUSUM "signals." Signals from the CUSUM chart should trigger a review of surgical procedures, including possible retraining[5].

A CUSUM is designed to continually monitor the surgical performance until a signal occurs. The procedure will theoretically always eventually signal even if the surgical performance has not changed. This implies that the usual evaluation criteria for test procedures, like false alarm rates and power, are not applicable for CUSUMs. In a sense, for a CUSUM both the false alarm rate and power can be thought of as equal to one since if the procedure has not signaled yet we continue to monitor (i.e. take a larger sample size). The number of patients before the CUSUM first exceeds the control limit is called the run length of a CUSUM. We evaluate CUSUMs based on aspect of the run length distribution, such as for example, the average run length. Ideally, while the surgical failure rate has not changed (and is acceptable) the run length is long, since signals represent false alarms. On the other hand, if the failure rate has increased substantially, short run lengths are desirable to ensure remedial action is brought about in a timely fashion. When evaluating a CUSUM we consider the run length a random variable whose distribution represents all the possible values of the run length that may arise given a particular mortality rate and the effects of chance. Then, the average run length while the failure rate is acceptable may be considered analogous to the type I error rate of a traditional statistical test. Similarly, the average run length of the CUSUM when the surgical failure rate has increased substantially is analogous to the power of a traditional statistical test. Determining the average run length of a CUSUM at the design stage is computational intensive since it is based on all possible outcomes for a long series of surgeries, however, they may be closely approximated[6]. An appropriate value for the control limit, h, in any specific example is based on the desired average run length of the CUSUM while the failure rate is acceptable. Note that we should always react to a CUSUM signal even if that signal follows a long run length, since the signal is evidence of a *recent* change in the surgical performance.

3. Risk Adjusted CUSUM Procedure for Cardiac Surgery

Unlike a traditional CUSUM procedure, with our new procedure, the magnitude of the scores, given by w_t in (1), depends on each patient's surgical risk estimated pre-operatively. Thus, the score depends on four factors: the current acceptable level of surgical performance, a chosen level of surgical performance deemed undesirable, the patient's surgical risk estimated pre-operatively, and the actual surgical outcome for the patient. The scores (w_t) are derived based on the log likelihood ratio of the current risk compared with a specified change in risk, see (2) below. For example, we may decide we wish to optimize the chart to detect a doubling in the odds of failure. Assuming patient t has a surgical risk of death equal to p_t, the likelihood for patient t is given by $p_t^y(1-p_t)^{1-y}$, where y equals unity if a surgical failure occurs and zero otherwise. The surgical risk for each patient may be determined pre-operatively using a rating method such as Parsonnet risk factors[3,4], or may be based on a logistic regression model fit to some sample data. Given, an estimated risk of failure equal to p_t, the odds of failure equals $p_t/(1-p_t)$. The CUSUM is formal sequential procedure for assessing the null hypothesis H_0: odds ratio = OR_0 versus the alternative hypothesis H_A: odds ratio = OR_A. To detect increases we set $OR_A > OR_0$. The choice of OR_A effects the patient scores, but the ability of the procedure to quickly detect a changes in actual odds ratio is relatively insensitive to OR_A. If the estimated risk p_t is based on the current conditions we may set $OR_0 = 1$. For patient t, assuming an odds ratio of OR, the odds of failure equals $ORp_t/(1-p_t)$, which corresponds to a probability of failure equal to $ORp_t/(1-p_t+ORp_t)$. Then, the two possible log-likelihood ratio scores for patient t are

$$w_t = \begin{cases} \log\left[\dfrac{(1-p_t+OR_0p_t)OR_A}{(1-p_t+OR_Ap_t)OR_0}\right] & \text{if } y=1 \\ \log\left[\dfrac{1-p_t+OR_0p_t}{1-p_t+OR_Ap_t}\right] & \text{if } y=0 \end{cases} \qquad (2)$$

4. Characteristics of the Procedure

To illustrate the characteristics of the risk adjusted CUSUM we use the arterial switch example discussed in the introduction. In the data set a total 15 deaths occurred giving an overall death rate of 6.5%. The risk of death as a function of the explanatory variates was estimated through a logistic regression model. The estimated risk varied significantly with gender and the pre-operative arterial pattern or diagnosis, and effectively classified patients into 10 risk groups. The lowest risk patients in the group were estimated to have a risk of death of just 1.8% following surgery, while the patients with the highest risk had a mortality rate of 46%. This suggests that some adjustment for the patient mix is necessary. We designed the CUSUM chart to detect a doubling of the odds of death from the preoperative risk. In this example, a doubling of the odds results in the death rate of 3.5% and 63% for the lowest and highest risk groups respectively. Based on the likelihood ratio statistic this leads to the following possible patient scores: 0.68 and –0.02 for the lowest risk patients, and 0.31 and –0.38 for the highest risk patients, where the positive score is assigned in the case of a death, and the negative score is assigned in the case of survival. These scores are derived using (2) with $OR_0 = 1$, $OR_A = 2$, and p_t either .018 or .46. 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 high risk patient. Setting the control limit h at 2 gives an average run length of around 460 patients when the surgical performance is acceptable. Given the frequency of surgery in this artificial example, this implies a positive signal from the monitoring procedure, on average once every six years, even if no true changes in the death rate have occurred. If surgical procedures were more frequent, it may be desirable to select a longer average run length while the surgical mortality rate is acceptable. We add a similarly designed CUSUM chart to detect decreases in the odds of death ($OR_0 = 1$ and $OR_A = 0.5$). The CUSUM designed to detect improvements (decreases) in the surgical failure rate is useful because if it signals it suggests that the currently acceptable failure rate should be re-estimated. This may happen if either the actual failure rate has decreases or if our initial estimate of the acceptable failure rate was too high.

The CUSUM is designed to prospectively monitor the surgical performance, i.e. we would use the logistic equation for death rate estimated from the current data together with (2) and (1) to monitor our future performance. However, to illustrate the procedure, we create a CUSUM plot using the current data ignoring the fact that we used the series to design the CUSUM. This analysis corresponds to a check of whether the surgical performance was stable over the 230 patients. Figure 1 shows two examples of the resulting CUSUM charts designed to detect increases or decreases in the mortality rate. For ease of presentation, the CUSUM to detect decreases in odds of mortality, accumulates negative values when there are surgical successes. Thus, on each plot in Figure 1 we see two CUSUM charts. The top pair of CUSUM charts is the result from the randomly ordered set of 230 operations and shows no signals. The bottom plot shows the resulting CUSUM charts when all the nine deaths that previously occurred between patients 100 to 230 are concentrated (but randomly distributed) between patients 100-150. This corresponds to a surgeon having an odds ratio of approximately 3.5 for the series of patients numbered 100 to 150. The bottom pair of CUSUM charts signals an increase in the death rate at around patient number 115. This suggests unstable surgical performance over time since there is a run of poor performance that was not likely due only to chance.

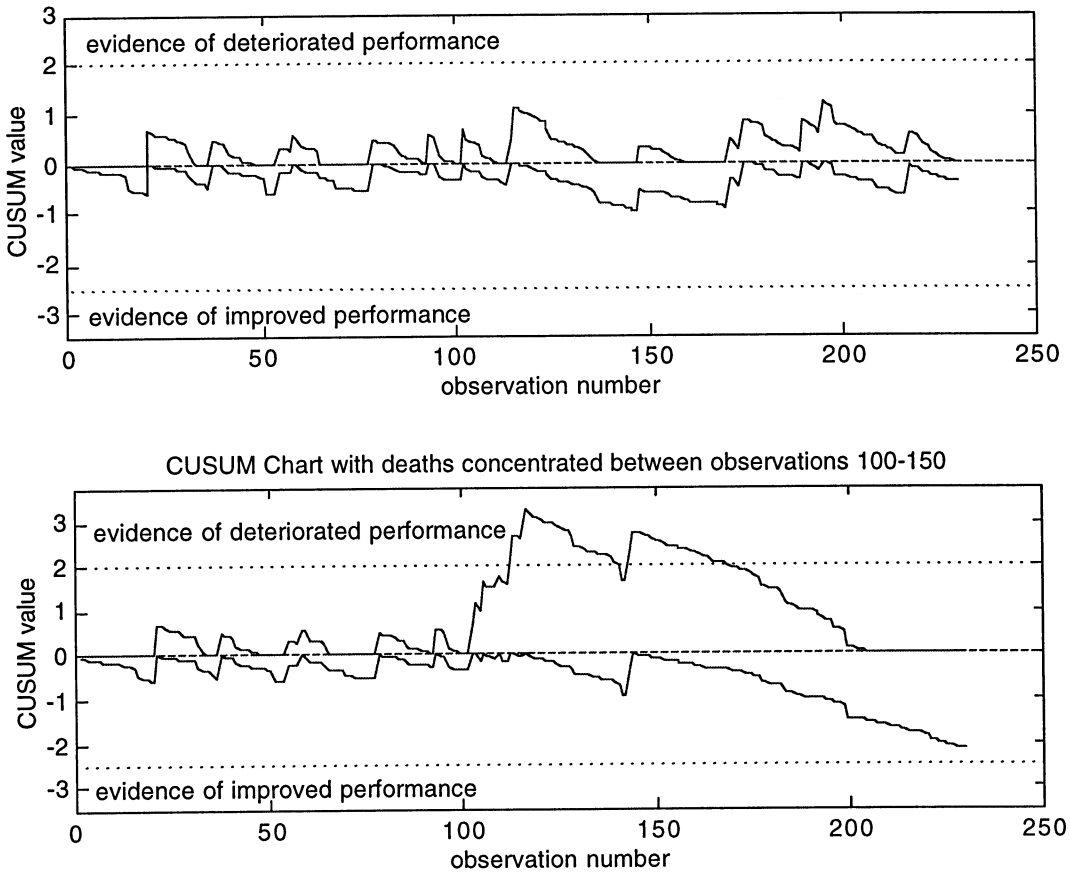

Figure 1: CUSUM charts designed to detect increases or decreases in performance
Top chart shows no evidence of a change in death rate
Bottom chart shows a string of deaths not likely caused by chance

To quantify how the proposed CUSUM adjusts for pre-operative risk we consider the extreme case where we observe a number of deaths in a row. Given the scores and the control limit for the example problem as defined previously, and assuming the CUSUM starts at zero, three low risk deaths in a row would trigger a signal, while it would take six high risk deaths in a row. Note that the procedure is very flexible and that by changing the control limit and/or the alternate hypothesis (OR_A) monitoring schemes with a wide variety of operating characteristics are possible.

We may also quantify the ability of the CUSUM procedure to quickly detect increases in the odds of death. More generally, Figure 2 shows plots of the average run length versus a measure of the actual surgical performance (given in terms of the odds ratio) for different patient mixes. The acceptable level of surgical performance is given by an odds ratio equal to unity, while increases in the odds ratio signifies a deterioration of performance. The solid line gives the results for the

current mix of high and low risk patients. For this particular example, extreme changes in patient mix substantially change the run length properties of the procedure when monitoring the death rate, as shown by the plot on the left hand side. This suggests that when monitoring the death rate, if patient mix changes dramatically the control limit of the monitoring procedure should be adjusted. This sensitivity is due to the large difference in risk of death between the lowest and highest risk patients. In other situations where the pre-operative risks are more similar, the run length curve is much less sensitive to the patient mix. As an example, the plot on the right hand side on Figure 2 shows the average run length curves when monitoring for either a death or the need to reinstitute cardiopulmonary bypass after a trial period of weaning, called a near miss in de Leval[5]. When using death or near miss as the response the estimated rates of failure for the lowest and highest risk categories are 19% and 52% respectively.

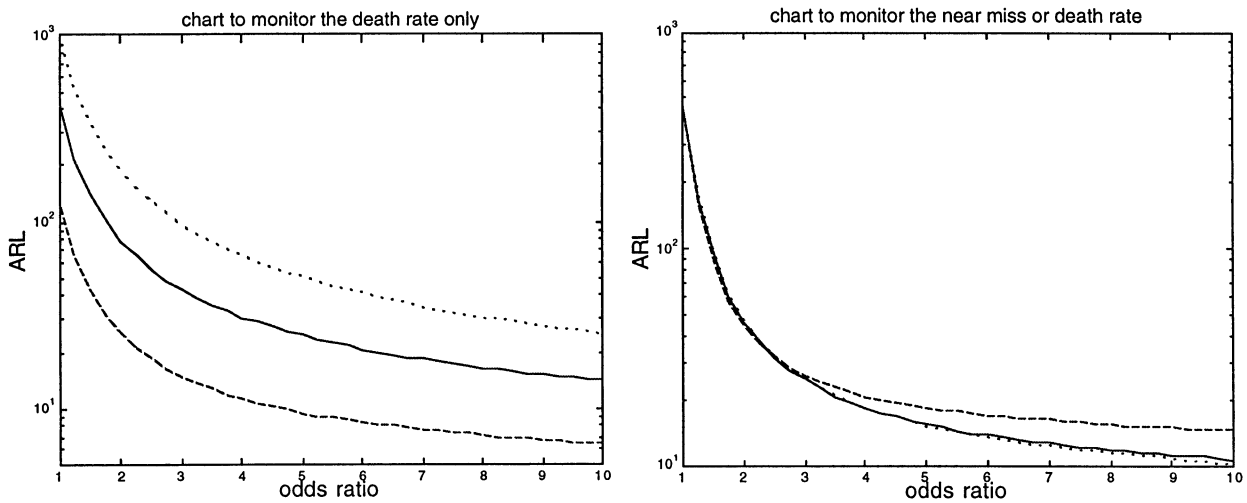

Figure 2: Average Run Length for Different Actual Odds Ratios
Solid line shows performance with current patient mix
Dotted line – all lowest risk patients, Dashed line – all highest risk patients

4. Conclusions

The use of a CUSUM chart with scores adjusted to reflect the estimated surgical risk of the patients is proposed to monitor surgical performance. This approach provides a logical way to accumulate evidence over many patients, while adjusting for 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. Note that the CUSUM methodology is also applicable when the covariates are continuous or a mix of continuous and categorical variables.

In summary, the proposed CUSUM chart is a valuable tool in the assessment and monitoring of surgical outcomes since it allows the early detection of problems, such as an increased failure rate. Evidence of any problems would lead to a review of surgical procedures and possibly some remedial measures, such as retraining, that could prevent unnecessary future failures.

Acknowledgments

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

References

1. Waldie, P. Crisis in the Cardiac Unit. The Globe and Mail, Canada's National Newspaper, 1998, Oct. 27; Sect. A:3 (col. 1).
2. Treasure T, Taylor K, Black N. Independent Review of Adult Cardiac Surgery – Unite Bristol. Bristol: Health Care Trust, March 1997.
3. Lovegrove J, Valencia O, Treasure T, Sherlaw-Johnson C, and Gallivan S. Monitoring the results of cardiac surgery by variable life-adjusted display. *Lancet* 1997; 18, 350(9085); 1128-1130.
4. Poloniecki, J, Valencia O, Littlejohns P. Cumulative risk adjusted mortality chart for detecting changes in death rate: observational study of heart surgery. *British Medical Journal*, 1998; 316: 1697-1700.

5. de Leval MR, François K, Bull C, Brawn WB, Spiegelhalter D. Analysis of a cluster of surgical failures. *The Journal of Thoracic and Cardiovascular Surgery* 1994 March: 914-924.
6. Steiner S, Cook R, Farewell V. Monitoring Paired Binary Surgical Outcomes Using Cumulative Sum Charts. *Statistics in Medicine* 1999; 18; 69-86
7. Montgomery DC *Introduction to Statistical Quality Control*, Second Edition. New York: John Wiley and Sons; 1991.