Risk-Adjusted Monitoring of Binary Surgical Outcomes

Stefan H. Steiner, Richard J. Cook and Vern T. Farewell

Med Decis Making 2001; 21; 163

Risk-Adjusted Monitoring of Binary Surgical Outcomes

STEFAN H. STEINER, PhD, RICHARD J. COOK, PhD, VERN T. FAREWELL, PhD

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 overreacting 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 preoperatively. 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. **(Med Decis Making 2001;21:163–169)**

The need to formally monitor surgical outcomes has been brought to the forefront in some recent well-publicized cases[1,2] where undesirably 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 and others[3] and Poloniecki and others[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 with 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 and others[5] and Steiner and others[6] propose an alternative surgical monitoring procedure based on a cumulative sum (CUSUM) chart that 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.[8] Traditional CUSUM approaches, however, make no adjustment for different risk profiles because machine inputs are usually relatively homogeneous, 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 overall 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 preoperative risk. The procedure is illustrated with sample data kindly supplied by Professor M. de Leval from a United Kingdom study of neonatal arterial switch operations for transposition of the great arteries. In

Received 10 August 1999 from the Department of Statistics and Actuarial Sciences, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1 (SHS, RJC); and the Department of Statistical Sciences, University College, London, United Kingdom (VTF). Revision accepted for publication 12 December 1999. 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.

Address correspondence and reprint requests to Dr. Steiner: Department of Statistics and Actuarial Sciences, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1; telephone: (519) 888-4567 x6506; e-mail: shsteine@uwaterloo.ca.

the example, patient survival status constitutes the response and gender and the arterial pattern or diagnosis are the 2 risk factors of primary interest that characterize the patient mix. The data set is based on 230 operations from a number of surgical centers over a 3-year period. To illustrate the methodology, we use a random ordering of the observations and monitor the postoperative mortality rate of this artificially ordered series of 230 surgeries as if they came from 1 center over a 3-year period.

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

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

$t = 1, 2, 3, \ldots, X_0 = 0$, and w_t is the score assigned to patient t. In the standard CUSUM, a patient's score is based on his or her 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 and others[5] and Steiner and others[6] for examples. When designing the chart to detect increases in the surgical failure rate, we define scores associated with failures to be positive whereas 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 (equation 1) has accumulated the information from all previous surgeries. It will become large if the surgical performance level has deteriorated, but it will fluctuate close to 0 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 due to chance. This implies that the usual criteria for the evaluation of test procedures, such as false positive error rates and power, are not appropriate for assessing the performance of CUSUMs. In a sense, for a CUSUM, both the false alarm rate and power can be thought of as equal to 1 because, if the procedure has not signaled yet, we continue to monitor (i.e., take a larger sample size) until a signal occurs. The number of patients seen before the CUSUM first exceeds the control limit is called the *run length* of a CUSUM. We evaluate CUSUMs based on aspects of the run length distribution such as the average run length. Ideally, if the surgical failure rate has not changed (and is acceptable), the run length is long because 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. Thus, when the failure rate is acceptable, the average run length is similar in some ways to the type I error rate of a traditional statistical test. Likewise, the average run length of the CUSUM when the surgical failure rate has increased substantially is somewhat 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 may be evidence of a recent change in the surgical performance.

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 equation 1, depends on each patient's surgical risk estimated preoperatively. Thus, the score depends on 4 factors: the current acceptable level of surgical performance, a chosen level of surgical performance deemed undesirable, the patient's surgical risk estimated preoperatively, 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 to a specified change in risk (see equation 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 = unity if a surgical failure occurs and 0 otherwise. The surgical risk for each patient may be determined preoperatively 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 equal $p_t/(1 - p_t)$. The CUSUM is a formal sequential procedure for assessing the null hypothesis H_0: odds ratio = 1, versus the alternative hypothesis H_A: odds ration = OR_A. To detect increases, we set $OR_A > 1$. The choice of OR_A affects the patient scores, but the ability of the procedure to quickly detect changes in actual odds ratios is relatively insensitive to OR_A. For patient t, assuming an odds ratio of OR_A, the odds of failure equal $OR_A p_t/(1 - p_t)$, which corresponds to a probability of failure under H_A equal to $OR_A p_t/(1 - p_t + OR_A p_t)$ Then, the 2 possible log-likelihood ratio scores for patient t are

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

Characteristics of the Procedure

To illustrate the characteristics of the risk-adjusted CUSUM, we use the arterial switch example discussed above. In the data set a total of 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 preoperative 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, whereas 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 a 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 patient, and 0.31 and –0.38 for the highest-risk patients, where the positive score is assigned in the case of death and the negative score is assigned in the case of survival. These scores are derived using equation 2 with OR_A = 2 and p_t either 0.018 or 0.46. Note that the scores reflect the surgical risk assessed preoperatively, since the "penalty" for death of a low-risk patient is more severe than for 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 6 years, even if no true changes in the death rate have occurred. If surgical procedures were more frequent, it might 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_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 reestimated. This may happen if either the actual failure rate has decreases or if our initial estimate of the acceptable failure rate was too high.

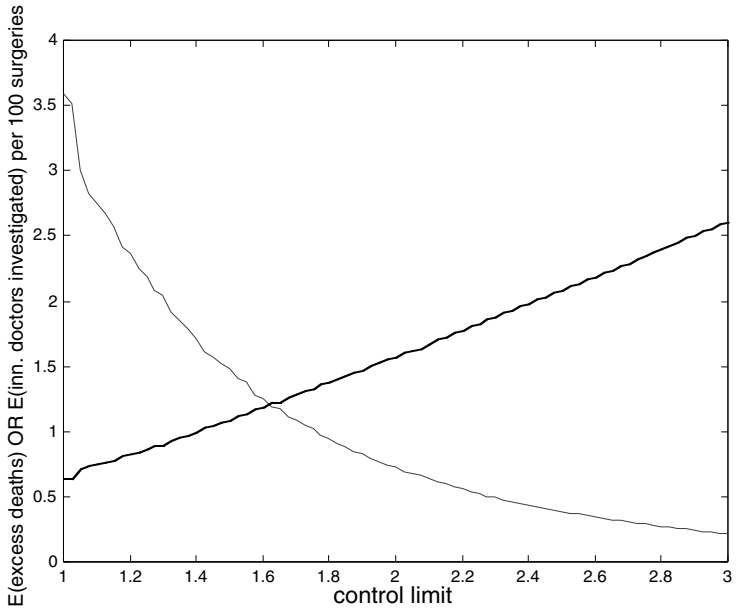

FIGURE 1. Trade-off inherent in the choice of control limit (h). The solid line indicates 100 times expected number of innocent doctors investigated. The dashed line gives the expected excess deaths.

The choice of the control limit (h) involves an inherent trade-off based on the in-control versus out-of-control average run length of the proposed procedure. Figure 1 illustrates the trade-off by showing the expected "excess deaths" that would result before a doubling of the odds of failure signals and the expected number of "innocent doctors investigated" due to false alarms as we change h. To create Figure 1, we assume the CUSUM scores, patient mix, and so on, of the example problem. The expected number of innocent doctors investigated is proportional to 1 over the average run length when the odds of failure have not changed. To put the quantities on a similar scale, we plot the expected number of innocent doctors investigated per 100 surgeries. The expected excess deaths that result if the failure rate changes to OR_A equals $(p_1 - p_0)ARL[OR_A]$, where p_1 is the overall failure rate when $OR = OR_A$, p_0 is the current overall failure rate, and $ARL[OR_A]$ is the average run length when $OR = OR_A$. Figure 1 allows us to quantify the effects of the choice of control limit in terms of the medical context. In the example, we chose a control limit equal to 2 for the CUSUM to detect increases in the odds of failure. From Figure 1, this results in an average of 0.73 innocent doctors investigated per 100 surgeries and an average of 1.57 excess deaths if the odds ratio of failure doubles.

The CUSUM is designed to prospectively monitor the surgical performance; that is, we would use the logistic equation for death rate estimated from the current data together with equations 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 2 shows 2 examples of the resulting pair of CUSUM charts designed to detect either 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 2 we see 2 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 9 deaths that previously occurred between patients 100 to 230 are concentrated (but randomly distributed) between patients 100 to 150. This corresponds to a surgeon's 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.

To quantify how the proposed CUSUM adjusts for preoperative 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

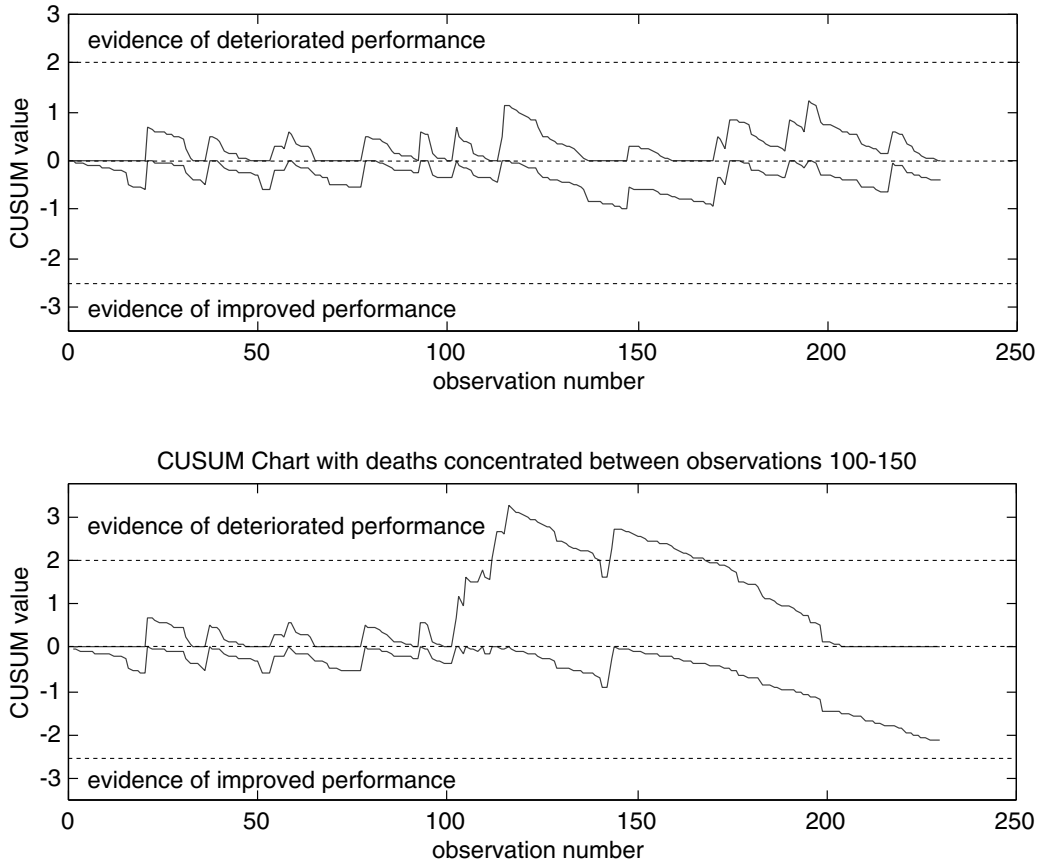

FIGURE 2. Cumulative sum (CUSUM) charts designed to detect increases or decreases in performance. *Top*, chart showing no evidence of a change in death rate; *bottom*, chart showing a string of deaths not likely caused by chance.

example problem as defined previously, and assuming the CUSUM starts at 0, 3 low-risk deaths in a row would trigger a signal, whereas it would take 6 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 3 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, whereas 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. 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 preoperative risks are more similar, the run length curve is much less sensitive to the patient mix. As an example, the plot on the right of Figure 3 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.

To focus attention on the performance of the monitoring procedure under H_0 and H_A, we may examine a plot of the approximate cumulative run length distribution. Figure 4 shows the cumulative probability of a signal for the in-control condition

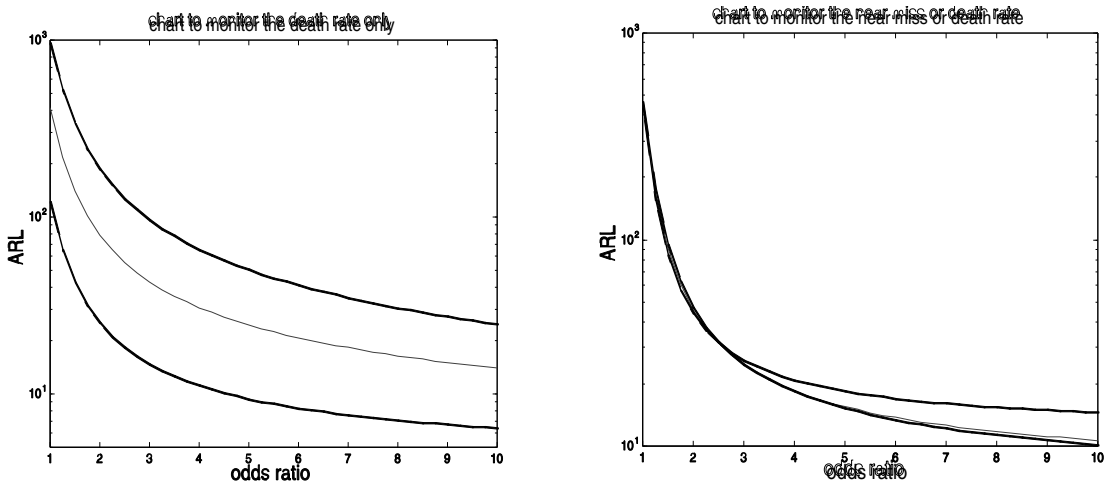

FIGURE 3. Average run length for different actual odds ratios. The solid lines indicate performance with current patient mix. The dotted lines indicate all lowest-risk patients. The dashed lines indicate all highest-risk patients.

(OR = 1) and when the odds ratio has doubled. The figure shows that even with no increased rate of failure, the CUSUM would signal around 51.4% of the time by 100 surgeries. Although this seems like a high rate of false positive signals, we must remember that in our example this represents around 15.5 months' worth of surgeries from a number of surgical centers. Also, through our choice of the control limit h, the in-control run length distribution can be changed to satisfy whatever CUSUM design characteristics are desirable. Similarly, we can see from Figure 4 that if the odds of a failure have doubled, the CUSUM will signal around 89% of the time within 50 patients. A caution in the interpretation of Figure 4 is necessary: Assume our CUSUM signals after a long run length of, say, 500 patients. We may be tempted to conclude that this must correspond to a false signal because, if the odds ratio had actually doubled, the CUSUM would likely have signaled much earlier based on Figure 4. But this rationale may well be incorrect. Recall that Figure 4 is based

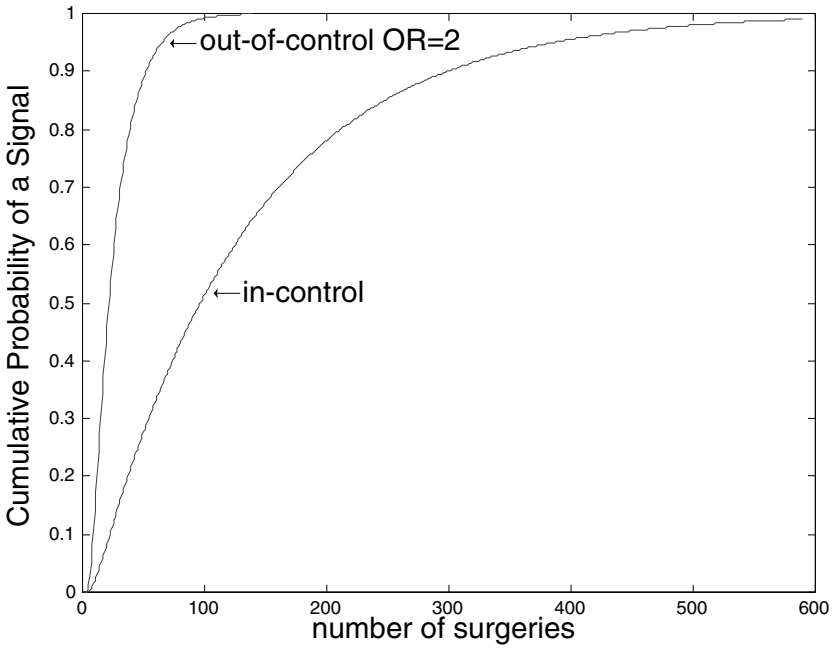

FIGURE 4. Cumulative run length distribution.

on the assumption that the rate of failure (through the odds ratio) stays constant. The observed CUSUM signal may be due to the rate of failure's changing somewhere in the series of 500 patients. For example, perhaps only for the last 50 patients has the odds ratio been 2. The purpose of a CUSUM chart is to quickly detect any changes in the failure rate no matter when they occur.

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 centers, 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 because the ability of the chart to detect specific changes can be easily quantified. Note that the CUSUM method- ology 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.

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, UK: Health Care Trust, 1997.
3. Lovegrove J, Valencia O, Treasure T, Sherlaw-Johnson C, Gallivan S. Monitoring the results of cardiac surgery by variable life-adjusted display. Lancet. 1997;18,350(9085): 1128–30.
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. J Thorac Cardiovasc Surg. 1994;March:914–24.
6. Steiner S, Cook R, Farewell V. Monitoring paired binary surgical outcomes using cumulative sum charts. Stat Med. 1999;18:69–86.
7. Montgomery DC. Introduction to Statistical Quality Control. 2nd ed. New York: John Wiley and Sons, 1991.
8. Hawkins DM, Olwell DH. Cumulative Sum Charts and Charting for Quality Improvement. New York: Springer, 1998.