Biometrics 69, 62–69
March 2013

DOI: 10.1111/j.1541-0420.2012.01822.x

A Risk-Adjusted O–E CUSUM with Monitoring Bands for Monitoring Medical Outcomes

Rena Jie Sun* **and John D. Kalbfleisch**

Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.
**email:* renajsun@umich.edu

Summary. In order to monitor a medical center's survival outcomes using simple plots, we introduce a risk-adjusted Observed–Expected (O–E) Cumulative SUM (CUSUM) along with monitoring bands as decision criterion.The proposed monitoring bands can be used in place of a more traditional but complicated V-shaped mask or the simultaneous use of two one-sided CUSUMs. The resulting plot is designed to simultaneously monitor for failure time outcomes that are "worse than expected" or "better than expected." The slopes of the O–E CUSUM provide direct estimates of the relative risk (as compared to a standard or expected failure rate) for the data being monitored. Appropriate rejection regions are obtained by controlling the false alarm rate (type I error) over a period of given length. Simulation studies are conducted to illustrate the performance of the proposed method. A case study is carried out for 58 liver transplant centers. The use of CUSUM methods for quality improvement is stressed.

Key words: Control limits; Cumulative SUM (CUSUM); Failure time data; Monitoring bands; Quality control; Quality improvement; Risk adjustments; Sequential tests; V-mask.

1. Introduction

Control charts are used to continuously monitor outcomes of a process, and hence to guide improvement in quality by providing timely feedback. CUmulative SUM (CUSUM) control charts were first introduced by Page (1954), in an industrial quality control setting. Over the last decade or so, CUSUMs have been suggested to monitor the performance of clinicians by, for example, measuring the occurrence of deaths or other outcomes after a surgical procedure. This approach enables early detection of an unacceptable number of deaths, and helps with the identification and correction of problems. Steiner et al. (2000) and Steiner, Cook, and Farewell (2001) developed a risk-adjusted one-sided CUSUM procedure based on the likelihood ratio in a logistic model. Axelrod et al. (2006) demonstrated the utility of the one-sided CUSUM method for analyzing 1-year binary mortality outcomes using a cohort of transplanted patients at multiple centers. However, a built-in 1-year lag is necessary in this approach. Biswas and Kalbfleisch (2008) developed a risk-adjusted one-sided CUSUM procedure that is based on a continuous timescale, incorporating a failure as soon as it occurs. In their method, a selected alternative hypothesis defines the one-sided CUSUM from a sequential probability ratio test (SPRT). They applied the procedure to detect "worse than expected" outcomes, but it can also be used to detect the alternative hypothesis "better than expected" in a separate one-sided chart. Gandy et al. (2010) discussed a timescale transformation under which some properties of the one-sided CUSUM can be obtained analytically.

The path of the one-sided CUSUM, however, does not clearly exhibit the true difference between observed and expected failures. For example, a horizontal path does not mean that the center is operating at the national average level, but rather that it has a risk approximately half way between the national average and the target risk used in constructing the chart. Collett et al. (2009) suggested supplementing the one-sided chart with an O–E CUSUM for which the slopes of the plot provide a simple estimate of the relative risk of death associated with the outcomes for the center under investigation. If $O(t)$ is the observed number of failures in (0,t] and $E(t)$ represents the expected number of failures; a plot of $O(t) - E(t)$ versus t or $E(t)$ is called an O–E CUSUM plot (Collett et al., 2009).

In this article, we consider such a risk-adjusted O–E CUSUM, and propose monitoring bands along the CUSUM path; when the CUSUM crosses either band, a signal occurs. This approach has the advantage of providing a true reading as to whether the rate of deaths at a center is above or below a chosen standard, while being a simple monitoring tool that is easy for clinicians to operate and interpret. The reader is referred to Figures 2.I and 3.I for example charts. The single plot suffices for summarizing the past data and trends, and provides signals in the same way as the two one-sided CUSUMs.

The monitoring bands are obtained from the V-mask approach which was proposed in the context of normally distributed outcomes by Barnard (1959). He suggested a CUSUM as a "reversed" SPRT and showed that a

predetermined shift of the process mean can be detected through the use of a cursor, called a V-mask, superimposed on the chart following each observation. It triggers a signal if either of its arms cuts the CUSUM path. This idea is quite elegant, although the V-mask has been found to be more difficult to implement than the one-sided CUSUM. In Section 2, we study the V-mask approach to monitoring a failure time mechanism, show its equivalence to the one-sided CUSUMs, and develop an alternative plotting mechanism based on monitoring bands that are simpler to use.

This work was motivated by the wish to provide real-time feedback to transplant centers given data reported to the Scientific Registry of Transplant Recipients (SRTR). For this purpose, we compared posttransplant outcomes at the center to those that would be expected from a model based on national data, where the expectations were risk adjusted to reflect the patient mix at the center under review. In this approach, the standard for comparison was obtained from a population model fitted to all centers combined. An alternative approach would use historical data for each center as the benchmark to define the expected outcomes, as suggested in Steiner et al. (2000, 2001) and Collett et al. (2009). This focuses on determining whether the center is performing better or worse than it has previously done. The use of historical benchmark can be satisfactory with very large centers or with the overall national picture, but it could be problematic for smaller centers, where the baseline rates (e.g., of 1-year patient survival) are rather poorly estimated (Kalbfleisch, 2009).

2. Method

2.1 *Notation*

In this section, we first describe an adjusted "national average failure rate," which is estimated by combining the outcomes from all of the transplant centers in the United States. Second, we consider individual centers and introduce a process to count the cumulative observed failures over time at each center. This is compared to a center-specific expected number of cumulative failures, which is obtained assuming that the outcome distribution of this center corresponds to that of the national average having adjusted for patient characteristics.

Let X represent the time from transplant to death, and suppose that we have a model for X based on transplantation data from all centers in the country. Given covariate vector Z_i for patient i measured at the time of transplant, a hazard function is defined as

$$\alpha_i(x) = \alpha(x; Z_i) = \lim_{\delta \to 0} P\{X \in (x, x+\delta) | X \geq x, Z_i\}/\delta, \quad (1)$$

which can be estimated through a failure time model. For example, we might have a (stratified) Cox model, an accelerated failure time model, or a parametric model to describe the national experience accounting, so much as possible, for covariates that influence outcomes.

Consider following a specific center in chronological time t beginning at $t = 0$ and suppose that patients receive transplants at times $S_1 < S_2 < \cdots$. In particular, subject i receives transplant at time S_i and subsequently fails at time T_i, so that the time to failure from transplant is $X_i = T_i - S_i$. Suppose that survival over a 1-year period is of interest, so that a *qualifying failure* occurs if $X_i \leq 1$. Other longer or shorter periods could also be considered. It is also assumed that, conditional on covariates Z_i, the null or "expected" distribution of X_i is known and defined by the hazard function $\alpha_i(x)$; in our case, $\alpha_i(x)$ is estimated based on the very large sample obtained by combining national experience of all transplant facilities. We suppose that the error in estimation of $\alpha_i(x)$ is small enough to be ignored.

Let $N_i^D(t)$ count the number of qualifying failures for subject i in $(0, t]$. Thus, $N_i^D(t)$ is 0 until a qualifying failure is observed, at which time it jumps to 1; if, on the other hand, a qualifying failure never occurs for subject i, $N_i^D(t)$ remains at 0 for all t. Thus,

$$N_i^D(t) = \begin{cases} I(T_i \leq t \leq S_i + 1) & \text{for } t \leq S_i + 1\ , \\ N_i^D(S_i + 1) & \text{for } t > S_i + 1. \end{cases}$$

Let $N^D(t) = \sum_{i=1}^{N_Q(t)} N_i^D(t)$ be the total observed number of qualifying failures in $(0, t]$ at the center, where $N_Q(t) = \sum_i I(S_i \leq t)$ denotes the number of transplants that have taken place in $(0, t]$. We define the "at risk" process for subject i as $Y_i(t) = I\{S_i < t \leq \min(T_i, S_i + 1)\}$.

We now suppose that the risk of a qualifying failure at this center is e^μ times the null or predicted rate $\alpha_i(x)$. Let the history for this center at t be given by $\mathcal{F}_{t^-} = \{N_Q(u), N_i^D(u), Y_i(u), Z_i,\ i = 1, \cdots, N_Q(t); 0 \leq u < t\}$ and define the intensity function of subject i at this center as

$$\begin{aligned} &E\{dN_i^D(t) | \mathcal{F}_{t^-}, \mu\} = e^\mu\, d\Lambda_i(t) \\ &= \begin{cases} Y_i(t) e^\mu \alpha_i(t - S_i) dt & \text{if } t > S_i; \\ 0 & \text{otherwise,} \end{cases} \end{aligned} \quad (2)$$

where α_i is defined in (1) and $d\Lambda_i(t)$ is being defined implicitly. When $\mu = 0$, national rates prevail and $E\{dN_i^D(t) | \mathcal{F}_{t^-}, \mu = 0\} = d\Lambda_i(t)$. In this case, $\Lambda_i(t) = \int_0^t d\Lambda_i(s)$ represents the cumulative intensity for individual i up to time t, and $A(t) = \sum_{i=1}^{N_Q(t)} \Lambda_i(t)$ denotes the overall cumulative intensity for the center up to t. Note that if $\mu = 0$, the death rates for patients at this center are identical to the expected or national rates; if $\mu > 0$ (or $\mu < 0$), the death rates in this center are higher (or lower) than the national rates.

We make the following notes: (i) Although we only include administrative censoring in this formulation, other independent censoring could be incorporated by suitable definition of $Y_i(t)$. (ii) We define the hazard $\alpha_i(x)$ for all $x > 0$ and restrict attention to qualifying failures through setting $Y_i(t) = 0$ once 1-year exposure is completed. Therefore, the proportional hazards assumption based on the constant relative risk e^μ for the center under review is only relevant for $0 < x < 1$. (iii) Finally, the choice of the proportional hazards model for center departures from the predicted rate is for convenience; other models, such as an accelerated failure time model or parametric model, could be used, but it would alter the formulation of the likelihood ratio and may increase the computational difficulty of the control limits.

2.2 *The O–E CUSUM with a V-Mask*

Based on the model (2), the likelihood of μ on data $\{N_Q(u), N_i^D(u), Y_i(u), 0 < u \leq t, i = 1, \cdots, N_Q(t)\}$ is proportional to $L(\mu) = \prod_{i=1}^{N_Q(t)} \exp\{\mu N_i^D(t) - e^\mu \Lambda_i(t)\}$.

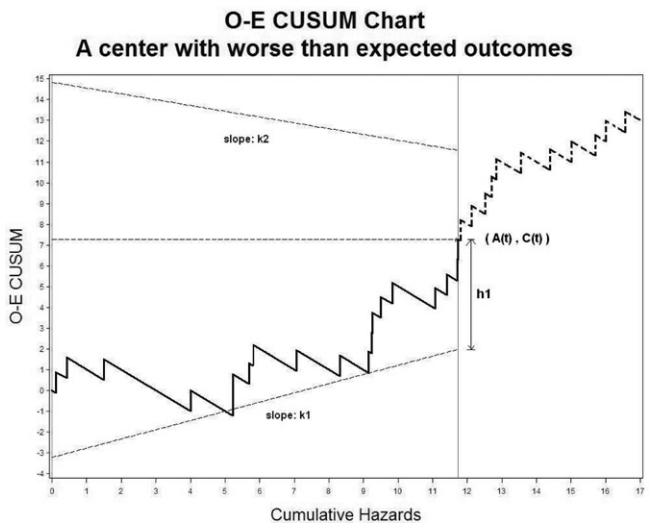

Figure 1. An O–E CUSUM with V-mask triggering "worse than expected" signal.

To construct the CUSUM, we consider a likelihood ratio test. The null hypothesis of interest is that the process is "in control" with relative risk 1 (H_0: $\mu = 0$). We consider simultaneously two alternative hypotheses: the process is "worse than expected" with a relative risk e^{θ_1} (H_-: $\mu = \theta_1$ with $\theta_1 > 0$), and the process is "better than expected" with a relative risk e^{θ_2} (H_+: $\mu = \theta_2$ with $\theta_2 < 0$).Here, θ_1 and θ_2 are predetermined constants.

The likelihood ratio of $\mu = \theta$ versus $\mu = 0$ for a specified center based on the data in $(s, t]$ with starting time $s \in (0, t]$ is $\mathrm{LR}(\theta; s, t) = \exp[\theta\{N^D(t) - N^D(s)\} - (e^\theta - 1)\{A(t) - A(s)\}]$. Therefore, the rejection region for H_- is $\log\{\mathrm{LR}(\theta_1; s, t)\} > a > 0$ (or $\log\{\mathrm{LR}(\theta_2; s, t)\} > b > 0$ for H_+). These two rejection regions can be rewritten as

$$C(s) < \{C(t) - h_1 - k_1 A(t)\} + k_1 A(s), \text{ for } H_-, \quad (3)$$

$$C(s) > \{C(t) + h_2 - k_2 A(t)\} + k_2 A(s), \text{ for } H_+, \quad (4)$$

where $C(t) = N^D(t) - A(t)$, $k_1 = (e^{\theta_1} - 1)/\theta_1 - 1 > 0$, $k_2 = (e^{\theta_2} - 1)/\theta_2 - 1 < 0$, $h_1 = a/\theta_1 > 0$, and $h_2 = -b/\theta_2 > 0$. Note that k_1 and k_2 are determined based on the target relative risk θ_1 and θ_2, whereas h_1 and h_2 can be adjusted to obtain desired properties (e.g., to achieve a certain false alarm rate over a given period of time). Here, we can view $N^D(t)$ as $O(t)$ and $A(t)$ as $E(t)$, as introduced in Section 1, so that $C(t) = O(t) - E(t)$.

Now consider a plot of $C(s)$ versus $A(s)$ for all $s \in (0, t]$ at a given t. The inequalities (3) and (4) correspond to straight-line boundaries (Figure 1) crossing the points $(A(t), C(t) - h_1)$ and $(A(t), C(t) + h_2)$ with slopes k_1 and k_2, respectively. These boundaries described the appropriate V-mask similar to that proposed by Barnard (1959) in the Gaussian case.

An alternative approach is to view the SPRT process in reverse time beginning with the "origin" $(A(t), C(t))$ at the current time t and looking backward at all previous times $s \le t$ (Wetherill, 1977). The same boundaries (3) and (4) can also be obtained from this approach.

We could plot the O–E CUSUM as $C(t)$ versus $A(t)$ or versus t. The former has the advantage of leading to the linear V-mask discussed earlier. In this plot, if either arm of the V-mask intersects the previous CUSUM path, a signal is recorded, suggesting a decrease (or increase) in the underlying failure rate from the nominal value. Thus, the O–E CUSUM can be implemented by applying the V-mask at each point in time until a signal occurs. If one continues a CUSUM indefinitely, whatever the true value of θ is, the CUSUM will eventually hit one of the boundaries and thus lead to rejection of the null hypothesis. Over any finite interval, however, there is a positive chance of no signal. Power and size are then of interest.

In the test outlined earlier, we plot $C(t)$ versus $A(t)$ and use the linear bounds. However, it is more natural to plot $C(t)$ versus t. In the next section, we respecify the CUSUM signals so that they can be implemented in a plot against t.

2.3 *Monitoring Bands*

The V-mask is generally viewed as a rather complicated presentation, which may be one reason why the one-sided CUSUMs discussed in the next section have been more widely used, at least in medical applications. In this section, we describe a novel way to present the O–E CUSUM chart to avoid the need of repeatedly applying the V-mask.

Consider the alternative hypothesis H_- at time t. From (3), let

$$M_1(t) = \inf_{s \le t}\{C(s) - k_1 A(s)\} + h_1 - \{C(t) - k_1 A(t)\}, \quad (5)$$

so that the chart signals at time t if $M_1(t) \le 0$, or it continues if $M_1(t) > 0$. In addition to the path $C(t)$, we can also plot $C(t) + M_1(t)$, graphically displaying the minimum distance of the CUSUM from the lower control arm of the V-mask at time t.

Similarly, we plot $C(t) - M_2(t)$ for "better than expected" detection, where $M_2(t) = \inf_{s \le t}\{-C(s) + k_2 A(s)\} + h_2 + \{C(t) - k_2 A(t)\}$. The CUSUM chart signals at time t if $M_2(t) \le 0$, or it continues if $M_2(t) > 0$. We refer to $C(t) + M_1(t)$ and $C(t) - M_2(t)$ as "monitoring bands," which now serve as control limits with the same signaling properties as the V-mask. These "monitoring bands" apply equally to a plot of $C(t)$ versus t as to a plot of $C(t)$ versus $A(t)$. Sample plots and detailed interpretations are given in Section 2.6.

It is worth noting that the computation of monitoring bands $M_1(t)$ and $M_2(t)$ is not so difficult as it might seem to be. For example, the infimum on the right side of (5) must occur before a jump point of $C(s)$. We only need to evaluate $\{C(s^-) - k_1 A(s^-)\} + h_1 - \{C(t) - k_1 A(t)\}$ at the failure times $s_1, s_2 \cdots t$, and select the minimum value as $M_1(t)$.

2.4 *The One-Sided CUSUM*

For comparison purposes, we discuss the one-sided CUSUM, which was introduced in the case of binary outcomes by Steiner et al. (2000) and modified to the present setting of continuous failure times by Biswas and Kalbfleisch (2008). The one-sided CUSUM is also based on an SPRT. For the alternative hypothesis of a relative risk e^θ, the one-sided CUSUM is defined by $G_{t+dt} = \max(0, G_t + dU_t)$ for $t > 0$, with $G_0 = 0$ and $dU_t = \theta dN^D(t) - (e^\theta - 1)dA(t)$.

If we are interested in detecting a relative risk of either e^{θ_1} ($\theta_1 > 0$) or e^{θ_2} ($\theta_2 < 0$), two one-sided CUSUMs can be performed simultaneously (Gandy et al., 2010), denoted as $G_t^{(1)}$ and $G_t^{(2)}$. The process $G_t^{(1)}$ remains at 0 until the first qualifying failure occurs, whereas $G_t^{(2)}$ immediately increases from 0. The $G_t^{(1)}$ CUSUM gives a signal of "worse than expected" when $G_t^{(1)}$ exceeds a predetermined control limit L_1 (> 0); and similarly, $G_t^{(2)}$ CUSUM signals "better than expected" when $G_t^{(2)}$ is greater than a predetermined control limit L_2 (> 0).

In contrast to the O–E CUSUM, the slope of any interval in the one-sided CUSUM is not directly interpretable as an estimated relative risk.

2.5 *Control Limits*

It is perhaps not so surprising that the O–E CUSUM with a V-mask is equivalent to the two one-sided CUSUMs with the usual horizontal control lines, because they are both derived from an SPRT. Both approaches lead to signals at the exact same time if the control lines and the parameters of the V-mask are suitably chosen. Specifically, with the choice $h_i = L_i/\theta_i$, $i = 1$ or 2, the O–E CUSUM V-mask designed to test $H_0 : \theta = 0$ versus $H_- : \theta = \theta_1 > 0$ and $H_+ : \theta = \theta_2 < 0$ has identical signal times to the simultaneous use of two one-sided CUSUMs constructed with regard to the same hypotheses. We show this equivalency in the Appendix.

Generally, we wish to choose a control limit so that there will tend to be a long waiting time until a signal occurs if the center failure rates are similar to the national average; at the same time, we wish to identify as quickly as possible the situation where the death rates are substantially higher (or lower) than the national average. The average run length (ARL) of a CUSUM is defined as the expected time to a signal. With the one-sided CUSUM $G_t^{(1)}$ and control limit L_1, the signal time is $\tau = \inf\{s : G_s^{(1)} \geq L_1\}$ and the ARL at a given relative risk e^θ is $\mathrm{ARL}(\theta) = E(\tau; \theta)$. One approach is to determine the control limit so as to attain a specified ARL when the process is operating at the null value; that is, we fix $E(\tau; \theta = 0)$.

In the one-sided CUSUM setting, Gandy et al. (2010) considered a timescale transformation $s = A(t)$. The modulated Poisson process $N^D(t)$ with intensity $A(t)$ is transformed to the new timescale s in which the event process $\tilde{N}(s)$ is a homogeneous Poisson process with rate 1. The log likelihood ratio up to time s is $\theta\tilde{N}(s) - (e^\theta - 1)s$, where $\tilde{N}(s) = N^D(A^{-1}(s))$ and $A^{-1}(s) = \inf\{t : A(t) > s\}$. Denote the signal time in the new timescale as $\tilde{\tau}$, so that $\tilde{\tau} = A(\tau)$ where τ is the signal time on the original timescale. They showed that the ARL in control on this new timescale, $E(\tilde{\tau}; 0)$, can be obtained analytically through constructing a Markov chain. This ARL is equal to the expected number of events until stopping on the original scale, $E(\tilde{\tau}) = E(N^D(\tau))$. In practice, one can calibrate L to obtain a desired ARL on a transformed timescale or, equivalently, expected number of failures until a false alarm on the original timescale. Since the one-sided CUSUM and O–E CUSUM with a V-mask both lead to signals at the same time when $h_i = L_i/\theta_i$, $i = 1$ or 2, we can also calibrate h_i in the O–E chart to obtain desired expected number of failures until a false alarm.

Biswas and Kalbfleisch (2008) conducted simulations to determine control limits. For a given center size, they set a false positive rate over a certain period, so that each center is subject to the same error rate if it operates at the national level. This yields control limits that are lower for smaller centers and higher for larger centers. We use a similar method of controlling type I error over a fixed period to obtain a control limit h for the O–E CUSUM; in the simulation, we categorize the results by the expected number of failures at a center. In the application of SRTR dataset, we use center size multiplied by national failure rate to approximate the number of expected failures and to determine the appropriate h. This approach subjects all centers regardless of size to a similar probability of a false positive.

2.6 *Some Examples of CUSUM Charts*

We consider liver transplant centers A and B followed over 3.5 years to illustrate the use and interpretation of CUSUM charts. For each center, the O–E CUSUM and two one-sided CUSUMs for 1-year posttransplant patient survival are presented. Similar charts could be constructed for other outcomes or other length of follow-up, such as 1-year graft survival or one-month survival.

In the O–E CUSUM chart, monitoring bands $C(t) + M_1(t)$ and $C(t) - M_2(t)$, chosen for testing alternatives of relative risk 2 and 0.5, respectively, are plotted along with the O–E CUSUM trajectory over time. $M_1(t)$ and $M_2(t)$ indicate how many additional and fewer failures at time t would have resulted in a signal. The values 2 and 0.5 as alternatives are chosen to represent differences in rates that would clearly be clinically important. These same values have been used in other presentations (e.g., Axelrod et al., 2009). For the one-sided charts, two one-sided CUSUMs are displayed on separate plots. We reflected the one-sided CUSUM versus the relative risk 0.5 and its control line through the X-axis in the presentation.

Center A. No signal of either "worse than expected" or "better than expected" was suggested in either CUSUM (Figure 2). The O–E chart (Figure 2.I) suggests that the outcomes of the center were similar to the national average over the 3.5 years. In July 2008, the CUSUM would have signaled "better than expected," had there been two fewer failures. The one-sided charts (Figure 2.II) show similar results that Center A performs at the national average level.

Center B. The failure rate at this center is close to the national average for the first year and a half, as suggested by the nearly horizontal plot line in the O–E chart (Figure 3.I). After that, the death rates were approximately twice the national average as illustrated by the O–E path having a slope close to the one for relative risk 2 in the legend. The CUSUM triggers a "worse than expected" signal in March 2008. Note that if the center had one more failure in November 2007, it would have triggered the signal then. As expected, the one-sided CUSUM chart (Figure 3.II) indicates a "worse than expected" signal at the same time.

It is worth noting that because we use national average rates as reference, an increasing trend, for example, could indicate either that the performance of the center has suddenly

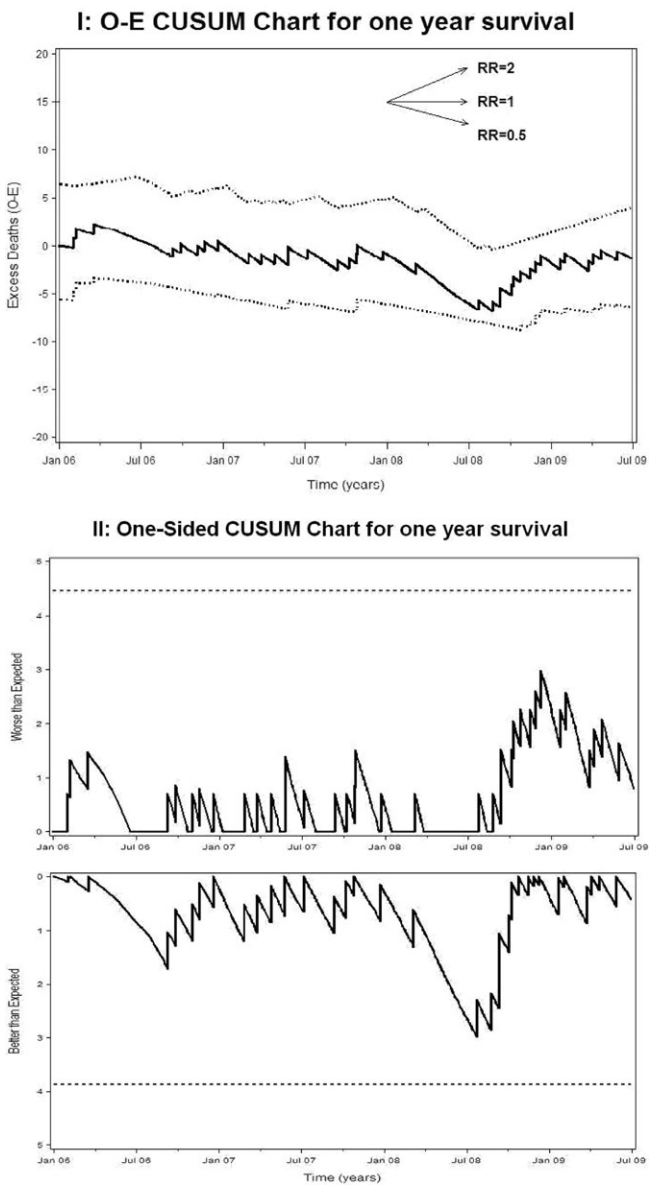

Figure 2. Center A, with 378 patients between January 01, 2006 and June 30, 2009.

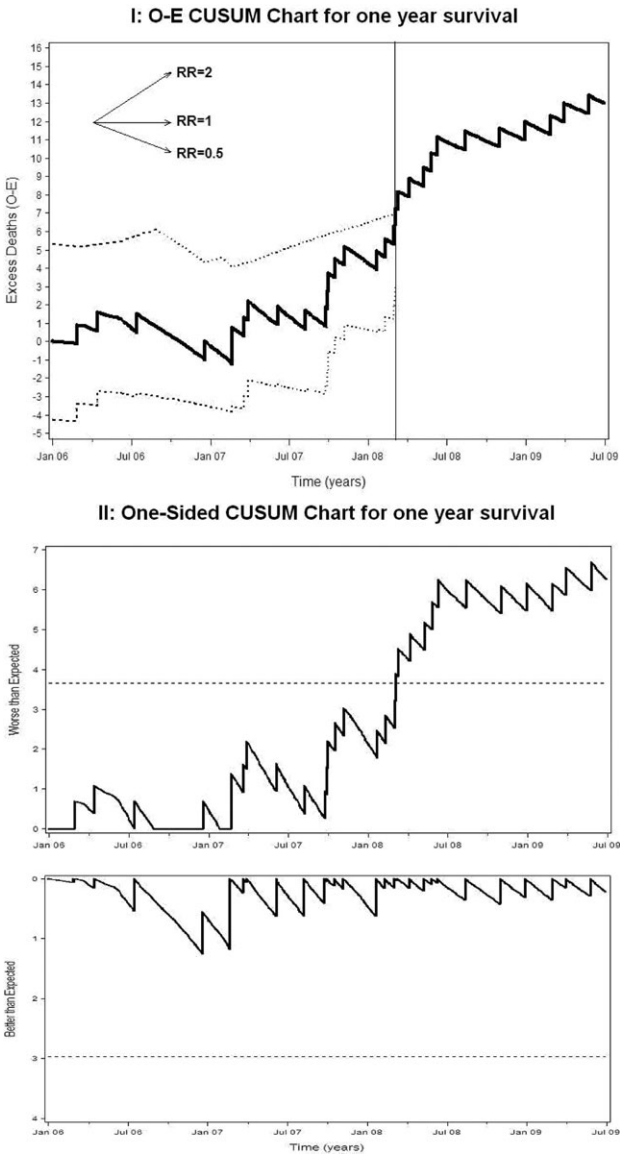

Figure 3. Center B, with 173 patients between January 01, 2006 and June 30, 2009.

changed to "worse than expected" or that it has consistently had "worse than expected" outcomes. When a center experiences a sudden change causing higher mortality rates, the CUSUM is expected to show a flat trajectory for a period of time followed by a substantially positive slope indicating such change, such as Center B in the example above. It then makes sense to look for an assignable cause associated with the time at which the change occurred. On the other hand, if the center has consistently had higher mortality rates compared to the national average, there would be no identifiable change point. In this situation, however, it is also desirable for the center to review its practice in light of the fact that its outcomes are poorer than one would expect based on the risk-adjusted national average outcomes.

2.7 *Head-Start*

When the CUSUM of a center leads to a "worse than expected" signal, it is appropriate for the center to examine its practice, especially changes in practice, to look for assignable causes, and to make adjustment as appropriate. Rather than resetting the CUSUM to zero, it is preferable to use a "head-start" by taking the plotting position somewhere less than the control limit (Lucas and Crosier, 1982). Gandy et al. (2010) discussed a head-start scheme in the one-sided CUSUMs. They reset the CUSUM to $L/2$ after a signal, and conducted a series of simulations to demonstrate the advantage of utilizing such head-start value. Collett et al. (2009) also used this head-start technique and argued the appropriateness of such resetting in monitoring transplant centers. The same idea could be used in an O–E CUSUM. For example, resetting the

CUSUM at $h_1/2$ below $C(t) + M_1(t)$ when a "worse than expected" signal occurs is equivalent to resetting the one-sided CUSUM to $L_1/2$.

3. Simulation Studies

3.1 *Control Limits*

We consider transplants arriving according to a homogeneous Poisson process and suppose that the posttransplant failure time distribution for the national average is exponential with rate λ_0, corresponding to a 1-year failure rate of $1 - e^{-\lambda_0} = 10\%$. As discussed before, the choice of a control limit for a center is affected by the size or the number of expected failures if the center failure rates are at the national average. To simulate centers that have expected failures within 1 year as 2, 5, 10, 15, and 20, Poisson processes are generated with rates 20, 50, 100, 150, and 200 transplants per year. Take $\theta_1 = -\theta_2 = \theta = \log(2)$ so that H_+ and H_- are symmetric hypotheses.

We chose parameter k based on the target relative risk e^{θ}, and chose h by controlling the rate of false signals to 8% over 3.5 years for each category of the expected number of failures per year. The choice of the 8% rate for the 3.5-year period gives a similar false positive rate to the standard 5% type I error rate over a 2.5-year period that has been used by the SRTR.

Simulation results confirm the equivalence of the one-sided CUSUM and the O–E CUSUM with respect to the signals that they generate, if $L_1 = h_1\theta_1$ and $L_2 = h_2\theta_2$. Table 1 gives the control limits of the O–E CUSUM obtained through controlling the type I error as described earlier. The column entitled "Power" specifies the probability that a center with relative risk 2 (or 0.5) would signal in a 3.5-year period. The ARLs in the table give the average number of follow-up years before the first signal occurs when the failure rate at the center is twice (or half) the national average. For example, if a center is expected to have five failures per year based on the national rates, but its true rate is twice that, there is a 92% probability that a "worse than expected" signal would be detected in the 3.5-year period, and on average, the first signal occurs after 1.71 years. The signal threshold h increases with the expected number of failures to maintain a constant probability of a false positive. As expected, when the expected number of failures increases, the power of CUSUMs increases.

3.2 *Sensitivity to Process Change in Relative Risk*

Of some particular interest is the behavior of the CUSUM when the center is initially experiencing failures at the overall (adjusted) national rate, but at a specific point in time, the rate changes substantially. To examine how sensitive the CUSUM is to sudden changes, we conducted simulations in two scenarios with a change point in the underlying risk.

In scenario 1, the process is under control with a relative risk 1 for subjects entering during the first year, and it changes to "worse than expected" with a relative risk 2 for subjects entering after year 1. This scenario mimics a systematic change in the quality of treatment that occurs at the time of transplant, such as the quality of the transplant surgical procedure. In scenario 2, the process operates at the national average level for the first year, and changes to "worse than expected" with a relative risk 2 for every subject that remains at risk or enters after year 1. This scenario reflects a sudden change of environment such as a change in the quality of care for all patients. In each case, the simulation evaluates the statistical power of the O–E CUSUM at the end of years 2, 2.5, 3, and 4. A signal counts in the power calculation only if it occurs after the change in rates at the end of year 1; if the chart signals before the end of year 1, we reset the CUSUM by applying the head-start described in Section 2.7 and then continue monitoring.

Table 2 shows that the CUSUM detects the sudden changes quickly, especially in centers with higher expected failures. After the change, the increase of cumulative failures is faster in scenario 2. Thus, as expected, the CUSUM is more powerful in detecting the scenario 2 type of change.

4. Case Studies

To demonstrate the use of the O–E CUSUM, we performed a retrospective analysis on 1-year posttransplant survival outcomes at liver transplant centers in the SRTR database. The cohort of patients receiving transplants between July 1, 2005 and December 31, 2008 was reviewed. Data included 11,861 liver transplants at 68 centers which ranged in size from 1 to 572 liver transplants over the 3.5-year period. We omitted 10 centers with fewer than eight transplants per year, for which the CUSUMs would be expected to yield little power.

The SRTR models for posttransplant survivals were utilized to represent the national rates and to compute the expected outcomes. The SRTR 1-year survival model for deceased donor transplants adjusted for 60 donor and recipient characteristics, whereas the model for living donor adjusted for eight donor and recipient characteristics. Because the models for deceased and living donors were quite different, SRTR computed expected outcomes for deceased and living donor cohorts separately using these two models. We do the same for the CUSUMs.

To specify control limits, we utilized the simulated values presented in Table 1. Thus, given the estimated expected number of failures at a center, we used linear interpolation to find an appropriate control limit h.

It is important to note that although we used a historical dataset for the purpose of demonstration, CUSUM charts can and should be used to monitor the center performance in real time; being able to effectively do this depends on prompt reporting of failures.

The number of signals and the average time to detect a signal for centers categorized by volume are summarized in Table 3. The O–E CUSUMs lead to relatively quick signals and for the most part, identify more quickly the same centers that are eventually identified as having results that are higher or lower than expected under the previous SRTR rules. Furthermore, if these charts were provided in real time (say quarterly) to the centers, they would have provided a simple graphical tool to identify when the center is experiencing relatively higher death rates and a clear indication of the potential for a signal as illustrated in Section 2.6.

It is worth noting if all centers perform at the national average level, we would expect to see 8% (about five signals out of 58 centers of interest) signaling on either direction. However, some centers may not operate at the null level during the time of interest; so as in this illustration, the test may detect more

Table 1
Control limits, power and ARL of the O–E CUSUM

Expected failures	Relative risk 2			Relative risk 0.5		
per year	h_1	Power	ARL	h_2	Power	ARL
2	4.08	0.70	2.98	3.00	0.42	4.60
5	5.34	0.92	1.71	4.36	0.71	3.04
10	6.36	1.00	1.05	5.50	0.91	2.04
15	6.81	1.00	0.77	6.10	0.98	1.56
20	7.25	1.00	0.61	6.46	0.99	1.27

Table 2
Statistical power of the CUSUM in scenario 1 where failure rates change for subjects entering after year 1, and scenario 2 where failure rates change for subjects at risk at year 1

Expected failures	Scenario 1				Scenario 2			
per year	Year 2	Year 2.5	Year 3	Year 4	Year 2	Year 2.5	Year 3	Year 4
2	0.08	0.20	0.34	0.56	0.21	0.34	0.47	0.67
5	0.13	0.37	0.59	0.83	0.38	0.60	0.74	0.90
10	0.24	0.60	0.81	0.96	0.64	0.83	0.92	0.99
15	0.34	0.73	0.91	0.99	0.80	0.94	0.98	1.00
20	0.42	0.85	0.97	1.00	0.89	0.98	0.99	1.00

Table 3
The number of centers signaled by the CUSUM (# of signals) and average time to signal (AVE) among signalled centers

Expected failures per year	Total # of centers	H_-: RR=2		H_+: RR=0.5	
		No. of signals	AVE	No. of signals	AVE
1–3	14	2	2.10	3	2.90
3–7	26	6	1.80	5	2.37
7–13	14	5	1.96	1	1.55
13–18	3	1	1.97	1	1.97
≥ 18	1	0	–	1	2.67

signals. In addition, the statistical power of each category in Table 1 shows that the test is more powerful in detecting the alternative hypothesis in larger centers (with more expected failures). This is consistent with what we see in Table 3.

5. Discussion

The usual one-sided CUSUM has the disadvantage of not giving a simple reading of the accumulating difference between observed and expected failures. For example, a horizontal path does not mean that the center is operating at the national average level, but rather that the center has a risk approximately half way between the national average and the target risk used in constructing the chart. In contrast, the O–E CUSUM gives a true reading as to whether or not the rate of deaths at a center is above or below the national average. The O–E CUSUM is easily plotted and its trends are easily interpreted; furthermore, when the monitoring bands are included, it provides simple rules for flagging.

Monitoring bands in O–E CUSUMs record the number of additional or fewer failures required for a signal. The one-sided CUSUM charts also provide such information, although in a somewhat disguised way. In the "worse than expected" one-sided CUSUM chart, the distance between CUSUM and the control line is proportional to the number of additional failures required for a signal at that time, with the constant of proportionality being the absolute value of the log of relative risk used in determining the chart.

Steiner and Jones (2010) proposed a risk-adjusted exponentially weighted moving average (EWMA) chart and claimed that its main advantage over a one-sided CUSUM is to provide an ongoing local estimate of the average score that is easier for clinical staff to interpret and understand. O–E CUSUM also provides such information, but in a simple chart based on the likelihood ratio.

Monitoring bands are similar to the Bollinger bands (Bollinger, 2002) used as a tool for technical evaluation of stock trading. Bollinger bands consist of a set of three curves drawn in relation to securities prices. The middle band is a measure of the intermediate-term trend, usually a simple moving average, that serves as the base for the upper band and lower band. The interval between the upper (or lower) and middle bands is determined by volatility, typically the standard deviation of the same data that were used for the average. Although somewhat different in purpose and construction, the Bollinger bands are used to graphically guide when appropriate actions (buying, holding, or selling) should be taken.

In constructing the CUSUM charts, we used a proportional hazards alternative. Other alternatives could be considered. Practitioners should be aware that a misspecified alternative would lead to reduced power and reduce the efficiency of the method. Also, the construction of the monitoring bands requires specification of alternative relative risk e^{θ_1} and e^{θ_2}. We chose $\theta_1 = \log(2) = -\theta_2$ in this article, which would represent important clinical differences. Other choices of θ_1 and θ_2 (e.g., $\theta_1 = \log(1.5) = -\theta_2$) could lead to different monitoring

bands and somewhat different operating characteristics. A systematic evaluation of the dependence of the ARL on the true relative risk e^{θ} and the specified alternatives would be of interest.

The national average failure rate is used in this article as the reference for evaluating each individual center. Alternatively, depending on one's interest, the historical performance of individual centers could also serve as the benchmark. In that case, a signal would indicate that the performance of the center has been improved or worsened compared to its own previous performance. Although this alternative way to set up a reference level has some appeal, one needs to be careful in interpretation. There is no guarantee on the quality of performance during reference period of time and the results would only show the comparison of the current performance relative to the historical performance for the particular center. For example, if a center has good performance during the reference period, the CUSUM could yield a "worse than expected" signal even though the center might in fact have normal performance levels compared to other centers. In addition, this approach can be problematic for smaller centers where there is a lot of inherent variation in the baseline period. Where possible, we believe that basing risk-adjusted charts on national outcomes, as we have discussed, provides a better approach to monitoring centers. Such plots indicate an overall propensity for the center to have higher rates of failure than the population as a whole. Abrupt changes in the slope of the CUSUM identify time points at which the rates within the center changed, and suggest the need of further explanation.

Acknowledgements

The SRTR is funded by contract 234-2005-37009C from the Health Resources and Services Administration (HRSA), U.S. Department of Health and Human Services. The views expressed herein are those of the authors and not necessarily those of the U.S. Government. This is a U.S. Government-sponsored work. There are no restrictions on its use.

References

Axelrod, D. A., Guidinger, M. K., Metzger, R. A., Wiesner, R. H., Webb, R. L., and Merion, R. M. (2006). Transplant center quality assessment using a continuously updatable risk-adjusted technique (CUSUM). *American Journal of Transplantation* **6,** 313–323.

Axelrod, D. A., Kalbfleisch, J. D., Sun, R. J., Guidinger, M. K., Biswas, P., Levine, G. N., Arrington, C. J., and Merion, R. M. (2009). Innovations in the assessment of transplant center performance: Implications for quality improvement. *American Journal of Transplantation* **9,** 959–969.

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

Biswas, P. and Kalbfleisch, J. D. (2008). A risk-adjusted CUSUM in continuous time based on the Cox model. *Statistics in Medicine* **27,** 3382–3406.

Bollinger, J. (2002). Bollinger on Bollinger bands. *McGraw Hill.*

Collett, D., Sibanda, N., Pioli, S., Bradley, A., and Rudge, C. (2009). The UK scheme for mandatory continuous monitoring of early transplant outcome in all kidney transplant centers. *Transplantation* **88,** 970–975.

Gandy, A., Kvaloy, J. T., Bottle, A., and Zhou, F. (2010). Risk-adjusted monitoring of time to event. *Biometrika* **97,** 375–388.

Kalbfleisch, J. D. (2009). Commentary on "The UK scheme for mandatory continuous monitoring of early transplant outcome in all kidney transplant centers" by Collett D, Sibanda N, Pioli S, Bradley A, and Rudge C. *Transplantation* **88,** 968–969.

Lucas, J. M. and Crosier, R. B. (1982). Fast initial response for CUSUM quality-control schemes: Give your CUSUM a head start. *Technometrics* **24,** 199–205.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika* **41,** 100–115.

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

Steiner, S., Cook, R., and Farewell, V. (2001). Risk adjusted monitoring of surgical outcomes. *Medical Decision Making* **21,** 163–169.

Steiner, S. and Jones, M. (2010). Risk adjusted survival time monitoring with an updating exponentially weighted moving average (EWMA) control chart. *Statistics in Medicine* **29,** 444–454.

Wetherill, G. B. (1977). *Sampling Inspection and Quality Control*, 2nd edition. London: Chapman & Hall.

Received June 2012. *Revised August* 2012
Accepted January 2013.

Appendix: The Equivalency Between the O-E CUSUM with a V-Mask and the Onesided CUSUM

With the choice $h_i = L_i/\theta_i,\ i = 1$ or 2, the O–E CUSUM with V-mask designed to test $H_0 : \theta = 0$ versus $H_- : \theta = \theta_1 > 0$ and $H_+ : \theta = \theta_2 < 0$ has identical hitting times to the simultaneous use of two one-sided CUSUMs constructed with regard to the same hypotheses.

Consider the path of one-sided CUSUM for "worse than expected" with parameters θ_1 and L_1. Consider an excursion beginning at s where $G_s^{(1)} > 0$ and $G_{s^-}^{(1)} = 0$. This excursion ends when the CUSUM reaches the control limit L_1 and triggers a signal or when it returns next to 0. If it returns to 0, it stays at 0 until the next failure when a new excursion begins. Suppose the original excursion begins at $s = 0$ and ends at time $\tau = \inf\{t > 0 : G_t^{(1)} = 0 \text{ or } G_t^{(1)} \geq L_1\}$, and let $J = I(G_\tau^{(1)} \geq L_1)$. If $J = 1$, for example, then

(i) $0 < \theta_1\{N^D(t) - N^D(s)\} - (e^{\theta_1} - 1)\{A(t) - A(s)\} < L_1, s < t < \tau$; and

(ii) $\theta_1\{N^D(\tau) - N^D(s)\} - (e^{\theta_1} - 1)\{A(\tau) - A(s)\} \geq L_1$.

If $J = 0$, then (ii) becomes (ii*) $\theta_1\{N^D(\tau) - N^D(s)\} - (e_1^{\theta} - 1)\{A(\tau) - A(s)\} = 0$.

It is easily seen that (i) implies that

$$\left\{\frac{e^{\theta_1} - 1}{\theta_1} - 1\right\}\{A(t) - A(s)\} < \{N^D(t) - A(t)\}$$
$$-\{N^D(s) - A(s)\} < \left\{\frac{e^{\theta_1} - 1}{\theta_1} - 1\right\}\{A(t) - A(s)\} + \frac{L_1}{\theta_1},$$

for $s < t < \tau$. This can be seen to be of the same form of the O–E CUSUM in (3). If we choose $h_1 = L_1/\theta_1$, the one-sided CUSUM does not signal on the interval $(0, \tau)$ if and only if the O–E CUSUM does not signal on the same interval. Similarly the two CUSUMs both signal at τ if the inequality (ii) holds. A similar argument shows an equivalence between the O–E CUSUM and the one-sided CUSUM for the test of "better than expected."