Statistics
in Medicine

Research Article

Received 13 March 2013, Accepted 14 February 2014 Published online 13 March 2014 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6139

A weighted cumulative sum (WCUSUM) to monitor medical outcomes with dependent censoring

Rena Jie Sun, John D. Kalbfleisch*† and Douglas E. Schaubel

We develop a weighted cumulative sum (WCUSUM) to evaluate and monitor pre-transplant waitlist mortality of facilities in the context where transplantation is considered to be dependent censoring. Waitlist patients are evaluated multiple times in order to update their current medical condition as reflected in a time-dependent variable called the Model for End-Stage Liver Disease (MELD) score. Higher MELD scores are indicative of higher pre-transplant death risk. Moreover, under the current liver allocation system, patients with higher MELD scores receive higher priority for liver transplantation. To evaluate the waitlist mortality of transplant centers, it is important to take this dependent censoring into consideration. We assume a 'standard' transplant practice through a transplant model and utilize inverse probability censoring weights to construct a WCUSUM. We evaluate the properties of a weighted zero-mean process as the basis of the proposed WCUSUM. We then discuss a resampling technique to obtain control limits. The proposed WCUSUM is illustrated through the analysis of national transplant registry data. Copyright © 2014 John Wiley & Sons, Ltd.

Keywords: cumulative sum (CUSUM); dependent censoring; inverse probability weights; failure time data; quality control; quality improvement; resampling; control limits; risk adjustment

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 have been suggested to monitor the performance of medical providers by measuring the rate of deaths or other outcomes after a surgical procedure. This approach enables early detection of an unacceptable number of deaths, for example, and can help with timely identification and correction of problems.

Steiner *et al.* [1, 2] developed a risk-adjusted one-sided CUSUM procedure based on the likelihood ratio in a logistic model for binary outcomes. The authors proposed a graphical method for identifying either a substantial or consistent change in risk-adjusted mortality. Axelrod *et al.* [3, 4] demonstrated the utility of the one-sided CUSUM method for tracking and 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 [5] developed a risk-adjusted one-sided CUSUM procedure constructed on a continuous time scale to monitor transplant survival outcomes sequentially by incorporating exposure and failures as soon as they occur. They compared the observed number of deaths at a given center to the expected number of deaths at that center assuming that the center has the same adjusted death rates as the overall national average. A sequential probability ratio test forms the basis of the one-sided CUSUM, which examines whether there is evidence that could lead to rejection of the null hypothesis that the center's rates are the same as the expected value in the overall national experience in favor of 'worse than expected' (or 'better than expected') performance.

Each method listed in the preceding paragraph was developed based on the assumption of independent censoring. This would be violated in many cases, especially in medical settings where preventive

Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, MI, U.S.A.
**Correspondence to: John D. Kalbfleisch, Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, MI, U.S.A.*
†E-mail: jdkalbfl@umich.edu

approaches are applied selectively on high-risk patients. For example, in such settings, pre-treatment mortality is dependently censored by the receipt of treatment. In particular, the methods we propose in this article were motivated by the evaluation of mortality among patients waitlisted for liver transplantation. Patients on the liver transplant waitlist are regularly evaluated to assess their current medical condition. One particularly important summary measure is the Model for End-Stage Liver Disease (MELD) score, computed as a log-linear combination of serum creatinine, bilirubin, and international normalized ratio (INR) of prothrombin time. Waitlisted patients with higher MELD score have a higher risk of death and consequently are given priority to receive liver transplants when available. The 'censoring time', through receiving a transplant, is therefore correlated with the patients' unobserved time of death on the waitlist that would have occurred had the patient been left untransplanted. To evaluate transplant center-specific waitlist mortality, it is important to take the dependent censoring (transplantation) into consideration; failing to do so may yield substantially biased results.

In this article, we discuss a weighted CUSUM (WCUSUM) to account for dependent censoring. Motivated by the waitlist mortality issue for liver transplant centers, we present the method to address this case directly. However, the proposed methods could be adapted to monitor other data sets where dependent censoring is present. For example, the methods would be generally useful in dealing with transplant data and evaluating the survival on the waitlist, at least in any transplant situation in which the chance of transplant depends on a time-dependent risk factor for mortality. Another application could be in patients with prostate cancer where some interventions depend on the current value of a measurement of the prostate-specific antigen. In order to investigate survival without the intervention, one could use these methods. This could be useful in assessing overall survival in a particular region of the country or in comparing treatment centers.

We assume all centers follow a standard liver transplantation guideline on donor allocation, which can be described by a transplant model. We then make use of inverse weights in order to obtain adjusted CUSUMs (or WCUSUMs) that take account of the dependent censoring, where the weights are determined by the time-dependent MELD scores and their relationship to transplant. The resulting WCUSUMs are designed to compare the waitlist mortality at a center to the overall national average, having adjusted for patient mix and dependent censoring through the MELD score. We utilize a resampling technique to obtain control limits for the WCUSUM.

In the following sections, we introduce some basic notation before constructing a WCUSUM, where the weights and the hazard of death are obtained using an inverse probability of censoring approach [6]. We then describe the signaling rules for the WCUSUM, illustrated by a case study on waitlisted liver transplant mortality.

2. A weighted Observed - Expected (O-E) CUSUM

2.1. Notation

Assume patient i from a given facility $\mathcal{F}$ of interest enters the cohort at calendar time S_i (e.g., time of initial listing on the transplant waitlist). Let D_i denote the time to death since entry and C_i the time to transplant since entry. Let $X_i = \min(D_i, C_i)$ be the observed event time since entry to either death or transplant, whichever occurs first. The calendar time of the observed event is $T_i = S_i + X_i$. Finally, $Z_i(x), 0 \leqslant x \leqslant X_i$, denotes the set of time-dependent covariates (e.g., MELD scores), and V_i is a set of baseline covariates measured at the time of entry. Typically, V_i includes $Z_i(0)$.

Assume we have a population model on time to mortality (in the absence of censoring) since entry with a hazard function $\alpha_i(x) = \alpha(x; V_i)$ for subject i, where

$$\alpha_i(x) = \lim_{\Delta \to 0} P\{D_i \in (x, x+\Delta) | D_i \geqslant x, V_i\}/\Delta. \tag{1}$$

Let $d\Lambda_i(t) = I(t > S_i)\alpha_i(t - S_i)dt$ define the hazard or failure intensity function for subject i at calendar time t.

Suppose that survival over a 1-year period is of interest and refer to a death that occurs at time $D_i \leqslant 1$ as a 'qualifying death'. Note that 1-year survival is often used as a critical endpoint in assessing transplant centers, but the methods could be used to assess mortality over other windows of time, such as 3-year survival, with minor adjustment. Individual i is at risk of a qualifying death at time t if $Y_i^*(t) = 1$, where $Y_i^*(t) = I\{S_i < t \leqslant \min(T_i, S_i + 1)\}$. Let $\delta_i = I(D_i = X_i)$ be the failure indicator and $N_i^*(t)$ count the observed number of qualifying failures in the chronological time interval $(0, t]$ for subject i:

$$N_i^*(t) = \begin{cases} 0 & t \leqslant S_i; \\ \delta_i I\{T_i \leqslant t \leqslant S_i + 1\} & S_i < t \leqslant S_i + 1; \\ N_i^*(S_i + 1) & t > S_i + 1. \end{cases}$$

Note that $N_i^*(t)$ is either 0 or 1. It takes the value 1 if the ith individual enters at a time $S_i < t$ and has a qualifying failure before time t. The observed number of qualifying failures in $(0, t]$ for the center $\mathcal{F}$ is $N^*(t) = \sum_{i \in \mathcal{F}} N_i^*(t)$.

2.2. *Definition and properties of the weighted O-E CUSUM*

In this section, we first state the key assumption regarding the dependent censoring mechanism. We then consider the case of independent censoring and the unweighted O-E CUSUM. Taking the dependent censoring into account, we construct a weighted O-E CUSUM that provides an estimate of the underlying true CUSUM. We show the weighted O-E CUSUM has mean zero and evaluate its variance function. The components of the O-E CUSUM form the foundation to the weighted one-sided CUSUM that is described in the following section.

Assume the cause-specific hazard for censoring is $\lambda_i^C(x|\bar{Z}_i(x), V_i) = \lim_{\Delta \to 0} P\{C_i \in (x, x + \Delta)|D_i \geqslant x, C_i \geqslant x, \bar{Z}_i(x), V_i\}/\Delta$, where $\bar{Z}_i(x) = \{Z_i(s), 0 < s \leqslant x\}$. The key assumption needed to construct the WCUSUM is

$$\lambda_i^C(x|\bar{Z}_i(x), V_i) = \lim_{\Delta \to 0} P\{C_i \in (x, x + \Delta)|C_i \geqslant x, \bar{Z}_i(y), V_i, D_i = y\}/\Delta, \tag{2}$$

for all $y > x$. It says that all information about the rate of dependent censoring at time x is contained in $\bar{Z}_i(x)$ and the fact that the individual is surviving and uncensored at time x. This rate is not changed by the additional knowledge of the future value of $D_i = y > x$ or the additional information on $\{Z_i(v), x < v \leqslant y\}$. Under this assumption, it follows that

$$P\{C_i > x|V_i, \bar{Z}_i(y), D_i = y\} = \exp\left\{-\int_0^x \lambda_i^C(u|\bar{Z}_i(u), V_i)du\right\} \tag{3}$$

for all $0 < x < y$. This assumption (2) and its consequence (3) are essential for the inverse weights arising from the process $Z_i(x)$ to fully correct for bias due to dependent censoring [6].

Let $N_i(t)$ represent the underlying counting process of qualifying failures in the absence of dependent censoring, so that $N_i(t) = I(S_i + D_i \leqslant t < S_i + 1)$ if $t \leqslant S_i + 1$ and $N_i(t) = N_i(S_i + 1)$ if $t > S_i + 1$. Similarly, let $Y_i(t)$ denote the underlying at-risk indicator in the absence of dependent censoring, $Y_i(t) = I\{S_i < t < \min(S_i + D_i, S_i + 1)\}$. It follows that $E(dN_i(t)|Y_i(t), V_i, S_i) = Y_i(t)\alpha_i(t - S_i)dt = Y_i(t)d\Lambda_i(t)$.

Without any censoring, the O-E CUSUM at center $\mathcal{F}$ would compare the observed number of failures $O(t) = N(t) = \sum_{i \in \mathcal{F}} N_i(t)$ with the expected number of failures $E(t) = A(t) = \sum_{i \in \mathcal{F}} \int_0^t Y_i(u)d\Lambda_i(u)$, and $O(t) - E(t)$ is a zero-mean process if center $\mathcal{F}$ has the same mortality rates as the reference population. A plot of $O(t) - E(t)$ versus t provides a tracking of the outcomes from this facility as compared to the reference rates described by $d\Lambda_i(t), i = 1, 2, \ldots$. When this plot trends upwards (downwards), the observed failure rates in this center are higher (lower) than those in the reference population, and so these plots provide a useful descriptive tool. Further discussion can be found in [7] or [8]. See also [9].

Consider now the situation where the center $\mathcal{F}$ has the same mortality rates as the reference population but is subject to dependent censoring as described earlier. We aim to develop a process analogous to $O(t) - E(t)$ that is adjusted for the dependent censoring but retains the zero-mean property when the death rates in the facility $\mathcal{F}$ correspond to the reference rates. Let $dM_i^*(t) = dN_i^*(t) - Y_i^*(t)d\Lambda_i(t) = Y_i^*(t)[dN_i(t) - d\Lambda_i(t)]$. Note that $Y_i^*(t) = Y_i(t)I(C_i > t - S_i)$ so that

$$\begin{aligned} E[dM_i^*(t)] &= E\left\{E\{dM_i^*(t)|Y_i(t), dN_i(t), \bar{Z}_i(t - S_i), S_i, V_i\}\right\} \\ &= E\left\{E\{I(C_i > t - S_i)|Y_i(t), dN_i(t), \bar{Z}_i(t - S_i), S_i, V_i\}Y_i(t)[dN_i(t) - d\Lambda_i(t)]\right\}. \end{aligned}$$

Under assumption (3), it follows that $E\{I(C_i > t - S_i)|Y_i(t), dN_i(t), \bar{Z}_i(t - S_i), S_i, V_i\} = P\{C_i > t - S_i|D_i > t - S_i, \bar{Z}_i(t - S_i), S_i, V_i\} = \exp\left\{-\int_0^{t-S_i} \lambda_i^C(u|\bar{Z}_i(u), V_i)du\right\}$, so that

$$E[dM_i^*(t)] = E\left[\left\{Y_i(t)[dN_i(t) - d\Lambda_i(t)]\exp\left\{-\int_0^{t-S_i}\lambda_i^C(u|\bar{Z}_i(u), V_i)du\right\}\right\}\right]. \quad (4)$$

The expression (4) shows that the $M_i^*(t)$ process does not in general have mean zero under the reference distribution. However, (4) also indicates how to obtain a zero-mean process.

First, we define the weights in chronological time so that $w_i^*(t) = w_i(t - S_i) = \exp\left\{\int_0^{t-S_i}\lambda_i^C(u|\bar{Z}_i(u), V_i)du\right\}$. It is now easy to see that

$$E[w_i^*(t)dM_i^*(t)|Y_i(t), V_i, S_i] = E[\{Y_i(t)[dN_i(t) - d\Lambda_i(t)]|Y_i(t), V_i, S_i\}] = 0. \quad (5)$$

This equation (5) shows that the difference between the weighted cumulative observed failures $N_i^W(t) = \int_0^t w_i^*(u)dN_i^*(u)$ and the weighted cumulative hazards $A_i^W(t) = \int_0^t w_i^*(u)Y_i^*(u)d\Lambda_i(u)$ is a zero-mean process, for any subject i.

Thus, the weighted zero-mean process for center $\mathcal{F}$ is $O^W(t) - E^W(t)$, where $O^W(t) = \sum_{i\in\mathcal{F}} N_i^W(t)$ and $E^W(t) = \sum_{i\in\mathcal{F}} A_i^W(t)$. In fact, we are replacing $O(t)$ and $E(t)$ above with estimates that are adjusted for the dependent censoring. We refer to $O^W - E^W$ as the weighted O-E CUSUM. In the case of no dependent censoring, when all weights are equal to 1, this process reduces to the usual zero-mean Martingale, with $w_i^*(t)dM_i^*(t) = dN_i(t) - Y_i(t)d\Lambda_i(t) = dM_i(t)$ and $O^W(t) - E^W(t) = O(t) - E(t)$.

The process $O^W(t) - E^W(t)$ has a jump discontinuity at any time t where a qualifying failure is observed to occur. If it is the ith individual who fails, the size of the jump is $w_i^*(t)$, which is the inverse of the probability that individual i would be untransplanted at time t given survival to that time. The compensating process, $E^W(t)$, makes the same adjustment to the conditional probabilities of failure among those at risk at time t. As before, when this process trends up (down), the observed failure rate in center $\mathcal{F}$ is higher (lower) than the rate in the general population.

The variance of the process $O^W(t) - E^W(t)$ under the reference distribution and accounting for all subjects at the center $\mathcal{F}$ is

$$\mathrm{Var}^W(t) = \mathrm{Var}\left\{O^W(t) - E^W(t)\right\} = \sum_{i\in\mathcal{F}} E\left[\int_0^t \{w_i^*(u)\}^2 Y_i^*(u)d\Lambda_i(u)\right],$$

which is derived in the Appendix. The derivation of this result depends on a novel partition of the relevant integrals that leads to this relatively very simple result. As compared to the $O(t) - E(t)$ process, the weighted process, $O^W(t) - E^W(t)$, has an additional source of variation introduced by the weights. We also consider a more general case in which the facility of interest has relative risk, r, so that each patient has a hazard $r\alpha_i(x)$. In this case, the corresponding mean zero process is $O^W(t) - rE^W(t)$, which has variance function

$$\mathrm{Var}_r^W(t) = \sum_{i\in\mathcal{F}} rE\left[\int_0^t \{w_i^*(u)\}^2 Y_i^*(u)d\Lambda_i(u)\right]. \quad (6)$$

The variance of the processes is relatively easily calculated and could be used to assist in the construction of appropriate stopping rules for the CUSUM. Further, the statistic $\{O^W(t) - E^W(t)\}/\sqrt{\mathrm{Var}^W(t)}$ provides a test statistic at time t that could be compared to the $N(0, 1)$ distribution. We will not pursue this approach further in this article, as the resampling approach discussed in Section 3.2 provides a simpler and more easily implemented approach for determining stopping rules.

3. One-sided WCUSUM chart

3.1. The one-sided CUSUM chart

Biswas and Kalbfleisch [5] proposed a one-sided CUSUM chart, applicable if censoring is independent. At time t, the null hypothesis is taken to be that the rates of qualifying deaths in $\mathcal{F}$ correspond to those in the general or reference population. Thus, for individual i, the hazard function is $r\alpha_i(x)$ with $r = 1$. It is convenient to write $r = \exp(\mu)$, so that the hull hypothesis is $H_0 : \mu = 0$. To construct the CUSUM, they consider the alternative, worse than expected with a relative risk $r = e^\theta > 1$.

Thus, the alternative hypothesis is $H_1 : \mu = \theta$ for some suitably chosen $\theta > 0$ that typically represents a relative risk of clinical importance. In this paper, as in another similar work, we consider $\theta = \log 2$ for the worse than expected case. In [5], it is shown that the logarithm of the likelihood ratio of $\mu = \theta$ versus $\mu = 0$ is proportional to $\sum_i [\theta N_i(t) - \{e^\theta A_i(t)\}] = \theta O(t) - \{e^\theta - 1\}E(t)$. From this, they develop a one-sided CUSUM based on a sequential likelihood ratio test. This approach involves plotting the function G_t, which can be most easily defined in terms of its increments by $G_{t+dt} = \max\{0, G_t + \theta dO(t) - (e^\theta - 1)dE(t)\}$, with $G_0 = 0$.

The process G_t remains at 0 until the first qualifying failure occurs when it has a jump discontinuity of size θ. The process then drifts downward according to the term $(e^\theta - 1)dE(t)$ until it reaches 0 or until the next qualifying failure, whichever occurs first. If another failure occurs, it jumps again by θ and then continues the downward trend. If it reaches 0, this constitutes a renewal, and the process remains there until the next qualifying failure occurs. For quality monitoring purposes, the one-sided CUSUM registers a signal of worse than expected when G_t exceeds a predetermined control limit $L > 0$. Refer to Figure 1 for an example chart designed to detect worse than expected signals, although this chart is revised to accommodate dependent censoring. Note that a one-sided CUSUM can be designed to help detect either a worse than expected performance with $\theta > 0$ or a better than expected performance with $\theta < 0$; see [7, 10].

With the presence of dependent censoring, we utilize weighted cumulative failures and weighted cumulative hazards defined in the last section. Thus, in place of $O(t)$ and $E(t)$, we use $O^W(t)$ and $E^W(t)$. The one-sided WCUSUM $G^W(t)$ is defined in terms of its increments by

$$G^W_{t+dt} = \max\left\{0, G^W_t + \theta dO^W(t) - (e^\theta - 1)dE^W(t)\right\}, \tag{7}$$

with $G^W_0 = 0$. This gives rise to a plot that is very similar to the case discussed earlier with no dependent censoring. In this case, however, the one-sided WCUSUM for worse than expected jumps by $\theta / P(C_i > t - S_i) = \theta w^*_i(t)$, where i represents the subject experiencing a qualifying failure at time t. When there is no failure, the one-sided WCUSUM trends down by $(e^\theta - 1)$ times the accumulating hazards scaled by the $w^*_i(t)$'s. In essence, an individual who is observed to have a qualifying failure, but for whom the chance of being untransplanted at the time of failure is small, yields much larger jump in the process G^W_t than a similar individual who fails, but who had a large chance of continuing to be untransplanted at the time of failure. The one-sided WCUSUM for better than expected also follows closely the corresponding CUSUMs in the case of no dependent censoring, as discussed in [7, 10].

The development of the O-E and one-sided WCUSUMs is straightforward. As in the case of no dependent censoring, their implementation requires the definition of suitable control limits in order to detect signals in a desirable way. In the next section, we consider this problem of specifying control limits L for the one-sided WCUSUM for detecting worse than expected. It is also possible to develop monitoring bands for the O-E WCUSUMs along the same lines as introduced in [7] in the case of no dependent censoring. We illustrate both the one-sided WCUSUM and the O-E WCUSUM in monitoring liver transplant centers in Section 4.

3.2. Obtaining control limits by resampling

Biswas and Kalbfleisch [5] and Sun and Kalbfleisch [7] conducted simulations to determine suitable 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 has failure rates that correspond exactly to the reference or national rates. For example, Biswas and Kalbfleisch [5] used a false-positive rate of 8% over a 3.5-year period; these values were chosen to be comparable to the flagging rates in use by the Scientific Registry of Transplant Recipients (SRTR) in monitoring transplant centers. This approach of controlling the false-positive rates for all centers yields control limits that are lower for smaller centers and higher for larger centers. Simulations were performed using a Poisson process for subject arrivals, with each patient having an exponential failure distribution at the reference failure rate (of the national average). In this way, appropriate limits could be obtained for each center according to its size (in terms of average number of arrivals per year). It should be noted that these limits are not adjusted for the particular patient characteristics at the center, although simulations suggested that moderate variation in patient characteristics did not change much the false-positive rate. In the situation with dependent censoring and inverse weights, a similar simulation approach could be used. In addition to an assumption about the failure model, however, one would also need to model the time-dependent model for the dependent censoring

mechanism and so generate patient characteristics and a time-dependent risk score as well as a suitable censoring model. This seemed very complicated to execute, and the sensitivity to model assumptions would always be a concern. It therefore seemed better to seek a more empirical approach that would require fewer assumptions.

Gandy *et al.* [10] considered an alternative approach to selecting control limit in the independent censoring case. They defined a revised time scale, $s = E(t)$, and noted that, in terms of this time scale, the counting process of qualifying failures, $O^{\#}(s) = O(E^{-1}(s))$, is a homogeneous Poisson process with rate 1. They showed that the average run length (ARL) in control on this new time scale 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 if the center's death rates correspond to the reference or national norm. In practice, one can calibrate L to obtain a desired ARL on the transformed time scale; this is equivalent to setting L to correspond to the level needed to achieve a certain average number of qualifying failures at the signal on the original time scale. This approach would have the same limits apply to all centers with the disadvantage that small centers would be subjected to a smaller risk of a false-positive signal over any given period, whereas the largest facilities would have a much higher probability of a false positive in a given interval. The choice between criteria is a policy decision. In the case of dependent censoring, however, there is no time scale that would map the process $O^W(t)$ into a Poisson process. This approach does not apply in the dependent censoring case.

In order to circumvent these problems, we utilize a resampling technique to calibrate control limits for a center of given size. In effect, this approach could be used any time that the reference is determined by a large national or regional population and whether particular facilities have outcomes that are outside of the national norm is of interest. The idea is to draw patients at random from the population to repeatedly compose a center of given size m and, for each such draw, obtain the corresponding one-sided WCUSUM. In the repetitions, this gives a realistic picture of the natural variation in the population of interest. The approach has the decided advantage that the time-dependent censoring is automatically included in the variation. Various rules could be used to determine appropriate control limits for the center of interest (i.e., of size m). For example, we could choose the limit L so that the resampled or simulated patients would lead to a signal at a given proportion of the time over an interval of specified length. This is analogous to the approach taken in [5, 7]. Alternatively, we could choose the control limit L in order to fix the average time until a signal occurs for a given center size, assuming the reference population stably spans over a long enough period. If we are interested in 1-year mortality outcomes, a 1-year lead-in period to reach equilibrium is important in calibrating the control limit L and in constructing the WCUSUM. The WCUSUM can then be operated continuously under the same L and still maintain a comparable type I error rate.

We develop the first of these ideas more specifically. Suppose that the criterion we wish to implement is that there would be an 8% chance of a false-positive signal over a 3.5-year period. We set this criterion when the process is in equilibrium. To do this, we would simulate the process with new arrivals starting at time 0 and extending over a 4.5-year period. If the qualifying failure must occur within 1 year since entry, the G_t process would be in equilibrium by the beginning of year 1. We first estimate the weights and true hazards by constructing a dependent censoring model and a weighted death model using the population data. Consider, for example, a center that admits 30 patients per year. To evaluate this, we select at random and with replacement a sample of size 135 from the population from the patients who arrive during a 4.5-year period. For this sample, we construct a one-sided WCUSUM, which we begin to observe at year 1 and follow for an additional 3.5 years and record the maximum value, $G^{\max} = \max\{G_t : 1 < t \leqslant 4.5\}$ that this WCUSUM achieves. Over a large number of B repetitions (e.g., $B = 1000$), we find $G_i^{\max}, i = 1, \ldots, B$ and choose L so that 8% of these B runs have a maximum one-sided WCUSUM value larger than L; that is , 8% gives a false-positive signal. This approach can be repeated for various facility sizes.

This resampling technique can also be used in the case of independent censoring and should give results similar to those obtained by the approach in [5, 7].

4. Case study

4.1. Data description

We consider mortality rates for liver transplant patients who have been waitlisted for a liver transplant using data obtained from the SRTR. We consider a 3.5-year cohort of patients waitlisted between

1 January 2005 and 30 June 2008 from one of 11 regions in the USA. In this example, the region is being considered as the population in our model. Patients recorded as status 1 or 1A at the time of waitlisting have acute liver failure at waitlisting and are not included in the analysis. In addition, we exclude patients who were waitlisted in error, who changed to kidney/pancreas transplants, or who had a previous liver transplant. Given that pediatric patients follow a different scheme of transplant, we only include adults of age at least 18 years at the time of waitlisting in the analysis. Two centers with fewer than five patients waitlisted over this 3.5-year span are excluded. In the final data set, 2578 patients from *a single region with* seven centers and five organ procurement organizations are included.

The following baseline covariates are considered: gender, race, age, diagnosis categories, diabetes, previous malignancy indicator, body mass index, blood type, and hospitalization and intensive care unit status. All these covariates as well as baseline MELD score (see succeeding discussions) and sodium value are included in both the transplant (censoring) model and the mortality model. Further discussion of the data and models can be found in [11].

Time-dependent variables consist of the MELD score, inactive period, and sodium value. MELD is a function of measurements on serum bilirubin, serum creatinine, and the INR for prothrombin time, and allocation MELD score is used in practice as the main determining factor in the allocation of livers from deceased donors. For analyses reported here, we record MELD using 12 binary indicators for whether the score is in 6-8, 9-11, 12-14, 15-17 (as the reference level), 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39, 40+, and status 1 or 1A. Assuming that a patient is being monitored sufficiently by the clinician, it is reasonable to assume that the lack of a MELD score update implies that, to a reasonable approximation, the patient's MELD score has not changed. MELD is updated frequently when a patient's score is increasing because it is important to the center to make this known to increase the chance of a transplant, and MELD only rarely decreases. This suggests that coding MELD score as a step function (i.e., last-value carried-forward) would be appropriate. Waitlisted patients are sometimes declared inactive and so temporarily removed from the waitlist; this may be due to a temporary sickness or other event that makes transplant impossible for a period. Inactive patients should not receive transplant offers. Inactive status is tracked with a time-dependent indicator variable that takes the value 1 if the patient is inactive. Alternative approaches of handling inactive time were used by Zhang and Schaubel [12]. The inactive indicator is also included in the set of time-dependent covariates along with MELD and sodium value in the dependent censoring model. Patients are sometimes permanently removed from the waitlist for reasons such as medical condition, refusing transplant, improved or deteriorated condition, or being inactive on the program for more than 2 years. Without over-assumption, we treat removed patients as right truncation.

We consider death within 1 year since waitlisting as the outcome of primary interest in our analysis, although other periods could clearly be considered as well. A patient is considered as dependently censored if he or she experienced any type of deceased donor transplant or died during a deceased donor transplant procedure. A patient is independently censored if he or she is lost to follow-up, removed from the waitlist, or received a living donor transplant, which typically is not predicted by MELD score.

4.2. *Modeling and control limits*

We first generate a suitable population model based on the totality of the regional data for both the dependent censoring process (transplant) and mortality. We will then use these models to construct WCUSUMs for each facility under consideration. Appropriate control limits for the one-sided CUSUMs are obtained by resampling the population and defining levels that, on resampling, give rise to prespecified operating characteristics. It should be noted that all centers within a region share the same rules for transplantation (depending on MELD) and so share the same censoring mechanism. This also means that the relatively large sample will allow fairly precise estimation of the censoring distribution and associated with weights.

In order to develop a suitable model for the transplant (dependent censoring), we consider a time-dependent Cox model with hazard function,

$$\lambda^C(x|\bar{Z}_i(x), V_i, D_i > x) = \lambda_0^C(x)\exp\{\gamma^C Z_i(x) + \beta^C V_i\}, \tag{8}$$

where $\lambda_0^C(x)$ is an unspecified baseline hazard function, $\bar{Z}_i(x) = \{Z_i(s), 0 < s \leqslant x\}$, and V_i is a set of baseline covariates. In the censoring model, $Z_i(x)$ is a vector of time-dependent covariates including MELD, inactive period, and sodium level for subject i, with $Z_i(0)$ indicating the baseline values of these variables. V_i is the set of baseline covariates and includes $Z_i(0)$. The rate of transplantation is taken to

depend only on the current value (most recent measurement) of Z_i, which corresponds to the policy that utilizes MELD score as the main determinant of priority for transplantation.

Fitting the model to the dependent censoring data using standard techniques, we obtain estimates $\hat{\gamma}^C$, $\hat{\beta}^C$, and $\hat{\Lambda}_0^C(x)$, where the last is an estimate of $\Lambda_0^C(x) = \int_0^x \lambda_0^C(u)du$.

With the model for censoring developed, estimation of mortality rates on the waitlist is the second part of the population model needed to implement the WCUSUM techniques. As above, we assume that the hazard of death in the absence of censoring is again a Cox model but conditioned only on the baseline covariates, V_i. Thus, the model for the hazard $\alpha_i(x)$ in (1) is

$$\alpha_i(x) = \lambda_0(x)\exp(\beta V_i). \tag{9}$$

Appropriate estimation of the parameters in (9) requires appropriately accounting for the dependent censoring, which we do through an analysis, stratified on centers, where the individuals in the risk sets are appropriately weighted with stabilized inverse probability censoring weights (IPCW) weights. (Stabilized weights are briefly discussed in Section 4.4.) The weights are being used to control for the confounding variable of censoring. Because these confounders are controlled by the weights rather than by inclusion as covariates in the Cox models, this approach avoids the problem that such confounders could also be intermediate on the causal pathway to the outcome of death. Once estimates of the regression parameter β are obtained from the model stratified on center, an estimate of an appropriate population level baseline hazard function is obtained by using a nonstratified model of the form (9) with $\exp(\beta V_i)$ taken as an offset.

With appropriate models for censoring and mortality now in place, we construct CUSUM charts for each of the facilities in the study. In fact, to use the CUSUMs prospectively, we would use the estimates to monitor new events as they occur on the weight list in each facility in the region and so use the methods for prospective monitoring of the mortality rates in the centers. For the purpose of this illustration, however, we develop CUSUMs over the period 1 January 2005 to 30 June 2008. We focus attention on 1-year mortality and begin monitoring each facility 1 year earlier in January 2004. If the facility is in accordance with the overall average performance in mortality and transplant, the 1-year period in advance should lead to a process that is approximately in equilibrium. We construct WCUSUMs for each center. The one-sided WCUSUM defined in (7) requires the specification of a suitable target relative risk for the alternative hypothesis, and we select $e^\theta = 2$. Thus, the WCUSUM is being constructed so as to be particularly sensitive to a relative risk of 2 for waitlist death rates at the center level, as compared to the overall regional data.

In order to obtain an appropriate control limit for a center of interest, which admits on average m patients per year, we use a resampling approach as described in Section 3.2. We select an appropriate criterion and for this follow the recommendation of Biswas and Kalbfleisch [5] to obtain a type I error rate of 8% over a 3.5-year period. For facility size m, we select at random $4.5m$ patients from those entering the entire population over the 4.5-year period from 1 January 2004 to 30 June 2008 and, for this selection, construct a one-sided WCUSUM beginning in January 2005 and extending for 3.5 years. This process is repeated 1000 times, and the control limit L is chosen so that 8% of these 1000 WCUSUMs would yield a signal. The results of this process for facility sizes $m = 30, 50, 100, 150, 200$ are summarized in Table I.

4.3. Analysis results

Table I shows that as size increases, the control limit increases, and the weighted expected number of failures and the variance of the weighted zero-mean process increase linearly. The weighted observed number of failures and the weighted expected number of failures are very close.

Table I. Control limits for WCUSUM.

Size (per year)	L	O^W	E^W	Var^W
30	5.66	15.36	15.48	26.82
50	6.35	25.03	25.19	43.28
100	7.23	50.64	50.80	85.79
150	8.10	75.98	76.17	132.03
200	8.33	101.60	101.80	174.66

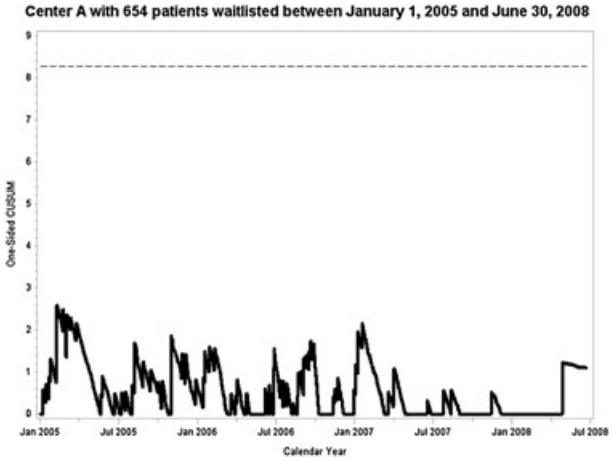

Figure 1. The weighted one-sided CUSUM, with control limit, of center A over a 3.5-year period.

Figure 2. The weighted one-sided CUSUM, with control limit, of center B over a 3.5-year period.

Given the estimated expected number of failures at a center, we used linear interpolation based on values from Table I to find an appropriate control limit L. We apply the estimated control limits on the 7 centers in the selected region and one center is signaled during the period of interest. Figure 1 demonstrates that the example Center A with 654 patients over the 3.5-year period operates at the reference level for the entire period. Figure 2 shows Center B with 368 patients in the 3.5 year cohort. It has several failures observed around the end of September 2006 in a short period and continues with a number of observed failures, which leads to the signal at the beginning of July 2007.

In practice, the signal in Figure 2 should lead to a review by the center for possible issues in process that might account for this signal and the apparent very high death rates in effect. Following this signal and possible remedial action, it would be normal to restart the WCUSUM before continuing. There is advantage to using a 'head start', whereby the CUSUM would begin at a position $L/2$ instead of 0. This has the effect of increasing the chance of an additional signal especially if the process is currently out of control as the CUSUM suggests. Head starts are discussed in [10, 13].

As noted earlier, an alternative presentation of CUSUMs was discussed in [7] in which the O-E CUSUM is plotted with monitoring bands to indicate when the CUSUM yields a signal. Figures 3 and 4 present the weighted version $O^W - E^W$ along with the appropriate monitoring bands for CUSUMs for the same two facilities as in Figures 1 and 2. The dark path is the CUSUM and the broken path is the monitoring band for a worse than expected signal. Figure 3 does not signal and Figure 4 signals at the same time as the one-sided CUSUM in Figure 2. Note that the advantage of these plots is that when the CUSUM trends up or down, it indicates that the rate of failure in the center is respectively higher or lower than that in the general population.

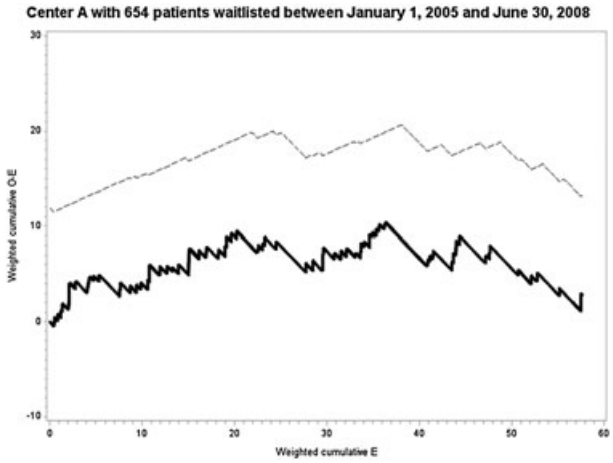

Figure 3. The weighted O-E CUSUM for center A. The dotted line is the monitoring bound for a worse than expected flag.

Figure 4. The weighted O-E CUSUM for center B. The dotted line is the monitoring bound for a worse than expected flag.

4.4. Unstabilized and stabilized IPCW weights

Under assumption (2), Robins and Finkelstein [6] have shown that we can estimate the true hazards Λ_i in the presence of dependent censoring, using the IPCW approach.

We assume a Cox model for the time to transplant (or dependent censoring) with hazard function (8). Under this model and the assumption (2), the conditional probability of not receiving a transplant until time x for subject i whose survival time exceeds x is,

$$K_i^V(x) = P\{C_i \geqslant x | D_i > x, \bar{Z}_i(x), V_i\} = \exp\left\{-\Lambda_i^C(x)\right\}, \tag{10}$$

where $\Lambda_i^C(x) = \int_0^x \exp\{\gamma^C Z_i(s) + \beta^C V_i\} d\Lambda_0^C(s)$. This is estimated as $\hat{K}_i^V(x)$ by replacing γ^C, β^C and Λ_0^C with their estimated values. The commonly used (unstabilized) weights are defined as $\hat{w}_{i1}(x) = 1/\hat{K}_i^V(x)$.

To reduce the variation in the weights, we can stabilize the weights by including a numerator $\hat{K}_i^0(x)$ obtained by using $Z_i(0)$ in place of $Z_i(s)$ in (10). Stabilized weights are then $\hat{w}_{i2}(x) = \hat{K}_i^0(x)/\hat{K}_i^V(x)$. It can be seen that the stabilized weights also give unbiased estimating equations for the parameters of the marginal death model, and these are often used to reduce the variability of the estimates as we have done in Section 4.2.

5. Simulations of $O^W - E^W$

In this section, we describe an approach that we use to simulate data that satisfy the dependent censoring models considered in this paper. These simulations are then used to evaluate the properties of the $O^W - E^W$ process.

Assume patients arrive at a given center according to a homogeneous Poisson process with rate μ_0 patients per year. We refer to μ_0 as the facility size. For each patient i, assume a baseline covariate V_i that is Bernoulli(p) variable, and a time-dependent covariate $Z_i(x)$ that follows a Poisson process on the follow-up time x with rate depending on V_i; specifically, we assume $Z_i(x) \sim PP(\mu e^{\gamma^D V_i})$, where $PP($ denotes Poisson process. Suppose we are interested in 1-year mortality. Patients are followed for one year from entry and are censored at one year if they have not experienced either a failure or a transplant.

Conditional on $Z_i(x)$ and V_i, we generate (cause-specific) censoring and mortality according to hazard functions $\lambda_i^C(x|V_i, Z_i(x)) = \lambda_0^C \exp\{\gamma^C V_i + \beta^C Z_i(x)\}$ and $\lambda_i^D(x|V_i, Z_i(x)) = \lambda_0^D \exp(\gamma^D V_i) + \beta^D Z_i(x)$, respectively. As shown in the Appendix in Section A.2, the additive form for the conditional mortality model yields a marginal form, after taking an expectation over $Z_i(x)$, that is multiplicative with structure $\lambda_i^D(x|V_i) = [\lambda_0^D - \mu(e^{-\beta^D x} - 1)]e^{\gamma^D V_i}$. This step in the simulation allows us to generate a marginal mortality model within the proportional hazards class. This means that an ordinary Cox model can be used to estimate the relative risks and baseline hazard. The degree of correlation between the transplant hazard and mortality hazard is determined by the $Z_i(x)$ process. We use a Spearman rank correlation coefficient to measure the correlation between the latent death time and transplant time. In practice, we observe only one event among death, transplant and independent censoring whichever occurs first.

We first conduct some simulations to verify the variance formulas given in (6). We consider the following parameter setup: $\mu_0 = 500$, $p = 0.5$, $\mu = 5$, $\gamma^D = \log(2)$, $\lambda^D = 0.01$, $\gamma^C = \log(1.5)$, $\beta^D = 0.06$ and $\beta^C = \log(2)$. The simulation is conducted using 1000 repetitions.

For relative risks $r = 0.5$, 1 and 2, Table II reports the observed death rates, the dependent censoring rates, and the Spearman rank correlation between latent death time and dependent censoring time. In addition, Table II reports: the mean and standard deviation of the *empirical variance* of $O^W(1) - rE^W(1)$,

$$\widehat{\mathrm{Var}} = \widehat{\mathrm{Var}}\left\{O^W(1) - rE^W(1)\right\} = \sum_i \left\{\int_0^1 w_i^*(u)dN_i^*(u) - rw_i^*(u)Y_i^*(u)d\Lambda_i(u)\right\}^2;$$

and the mean and standard deviation of the variance from equation (6),

$$\widetilde{\mathrm{Var}} = \widetilde{\mathrm{Var}}\left\{O^W(1) - rE^W(1)\right\} = \sum_i r\int_0^1 [w_i^*(u)]^2 Y_i^*(u)d\Lambda_i(u).$$

It is expected that both variance estimators would have the same mean with $\widetilde{\mathrm{Var}}$ having the smaller standard deviation. Note that in all of these calculations, we are taking the weights $w_i^*(u)$ and the intensity functions $\Lambda_i(u)$ are taken as given, their estimates in practice being based on a large population sample.

Table II. Empirical verification of the properties of $O^W(1) - rE^W(1)$ under simulations.

				OE_r^W		$\widehat{\mathrm{Var}}$		$\widetilde{\mathrm{Var}}$	
r	Death (%)	Censoring (%)	Corr.	Mean	Var	Mean	SD	Mean	SD
0.5	11.1	0	0	−0.27	55.4	55.4	6.8	55.7	2.5
	8.6	32.4	0.13	−0.44	66.4	68.2	29.0	67.8	6.5
1	20.7	0	0	0.11	103.6	103.7	8.6	103.8	4.5
	16.3	29.6	0.17	−0.37	125.2	124.9	36.0	125.5	12.6
2	36.3	0	0	−0.14	179.0	181.2	10.7	181.5	8.2
	29.6	24.6	0.22	0.34	220.0	216.8	39.3	216.1	21.8

Table III. Recovery of underlying failures and risks in the case of dependent censoring.

	Scenario 1 (indep)		Scenario 2 (dep)		Scenario 3 (dep)	
	Mean	SD	Mean	SD	Mean	SD
Observed (O or O^W)	20.35	4.17	19.21	5.63	19.29	5.43
Expected (E or E^W)	20.77	1.92	20.72	2.14	20.34	2.19
Variance of $O-E$ or O^W-E^W	20.77	1.92	34.77	5.36	36.51	10.47

Table II verifies that the $O^W(1) - rE^W(1)$ has mean value close to 0 under all scenarios and that $\widetilde{\text{Var}}$ and $\widehat{\text{Var}}$ are both valid estimates of its variance. However, $\widetilde{\text{Var}}$ has much smaller variation than $\widehat{\text{Var}}$ under all scenarios.

In Table III, we compare the number of observed failures and the number of expected failures in the independent censoring case (scenario 1) with the weighted observed failures and weighted number of expected failures under dependent censoring (scenario 2). Again, the weights and the hazards are assumed known or estimated precisely from a large sample. Note that we can never obtain the true weights or hazards in practice. To mimic the practical implementation, we also compare with the results obtained from the estimated weights and hazards (scenario 3), where we generate a separate large sample (or population) with 5000 subjects and run IPCW analysis to obtain the parameter estimates for the censoring and mortality models. We consider the following parametric settings: $\mu_0 = 100$, $p = 0.5$, $\mu = 3$, $\gamma^D = \log(2)$, $\lambda^D = 0.01$, $\lambda^C = 0.05$, $\gamma^C = \log(2)$, $\beta^D = 0.1$, and $\beta^C = \log(2)$. The simulation is conducted using 100 repetitions. The 1-year cohort has 13.9% deaths and 39.3% dependent censoring, while the latent death rate is 20.6%. Spearman rank correlation between latent death time and dependent censoring time is 0.18.

Table III shows that in both scenarios 2 and 3, the mean of the weighted observed failures and weighted expected failures in the dependent censoring case are close to those in the case of no censoring. Note, however, that variance of $O^W - E^W$ is inflated in the dependent censoring case because of the additional uncertainty introduced by the weights. Weighted values using estimated weights and estimated hazards in scenario 3 agree closely with those obtained using true weights and true hazards in scenario 2.

6. Discussion

The construction of a WCUSUM with IPCW weights requires assumptions of accurate information, no unmeasured confounding, and correctness of the censoring model. Given these assumptions, the weighted O-E process under the null hypothesis is a zero-mean process but with inflated variance as compared to the CUSUM with no dependent censoring.

When the dependent censoring model is misspecified, a WCUSUM might give a variety of results depending on the actual censoring pattern. It is important to have a correct dependent censoring model. In our case, this does not present a problem because the transplant priorities are set nationally to depend on current MELD score and should be strictly followed within each region depending primarily on the MELD score. If the censoring varied across centers, it would still be possible to carry out separate estimations within each center and utilize the corresponding censoring distribution in the WCUSUM. There would be additional uncertainty in this case because of the potential error in the weights because of the smaller sample, and one might wish to take that into account in specifying control limits.

We view the use of CUSUMs as primarily a quality improvement tool, and so they are used to provide quick feedback to the medical providers so that they can monitor outcomes and review for possible problems when a signal occurs. These methods could also be used by an oversight organization to monitor facilities under its purview. The control limits we obtained were with a view to the quality improvement use, and control limits in the context of an oversight organization should depend on the purpose of the flagging and the actions to be taken. For example, very different flagging criteria would be used to suggest what centers might profit from an audit of procedures versus a situation in which the signal would lead to financial penalties or censure.

We presented a resampling technique to obtain the control limit for a center of a given size and over a certain period. In doing this, it was essentially assumed that each center draws its patients and covariates from the population, so that the future distribution of profiles in each center would mirror that in the overall population. For our example, Figure 5 gives side-by-side box plots of the distributions of the

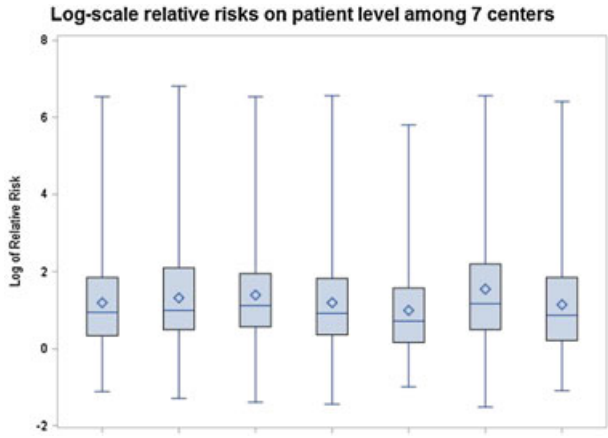

Figure 5. Side-by-side box plots of the estimated relative risks ($\exp(\beta' V_i)$) of patients in the seven centers.

estimated relative risks for the seven centers. These suggest that there are no large differences in risk profile and so provide some support for our approach. If these distributions were quite different, and it seemed more reasonable to assume that each center would have risk distributions in the future that looked like those in the past, then we could stratify the bootstrap approach so that bootstrap data for each center would be obtained by sampling nearest neighbors. A lead-in period to reach equilibrium is suggested. Our impression is that the control limits are not terribly sensitive to the risk profile, but a systematic examination of this would be useful. The sampling approach to determining control limits should also be investigated further because it provides a very simple empirical approach to this problem. It would be useful, for example, to compare the results from this approach in the independent censoring case to those obtained from the approach in [5, 10].

Appendix A

A.1. Variance of $O^W(t) - rE^W(t)$ with relative risk r

We begin by examining the process

$$N_i^W(t) - A_i^W(t) = \int_0^t w_i^*(u) dN_i^*(u) - \int_0^t w_i^*(u) Y_i^*(u) d\Lambda_i(u)$$

for a given individual i. We have already seen that this process has mean zero under the null hypothesis that individual i has the same mortality rate as an individual with the same covariates in the reference population. We now investigate the variance of this process.

Consider the more general case where the individual has a relative risk r, meaning that the mortality rate for this individual is a constant r times the mortality rate of a similar individual from the population. The corresponding weighted zero-mean process is

$$N_i^W(t) - rA_i^W(t) = \int_0^t w_i^*(u) dN_i^*(u) - r\int_0^t w_i^*(u) Y_i^*(u) d\Lambda_i(u). \tag{A.1}$$

The variance of this process can be evaluated through the following steps:

$$\begin{aligned}\mathrm{Var}\left(N_i^W(t) - rA_i^W(t)\right) &= E\left\{\int_0^t w_i^*(u) dN_i^*(u) - r w_i^*(u) Y_i^*(u) d\Lambda_i(u)\right\}^2 \\ &= E\left\{\int_0^t [w_i^*(u)]^2 dN_i^*(u)\right\} - 2rE\left\{\int_0^t w_i^*(u) dN_i^*(u) \int_0^t w_i^*(v) Y_i^*(v) d\Lambda_i(v)\right\} \\ &\quad + r^2 E\left\{\left[\int_0^t w_i^*(u) Y_i^*(u) d\Lambda_i(u)\right]^2\right\}.\end{aligned} \tag{A.2}$$

Statistics in Medicine

The second term in (A.2) can be written as

$$2rE\left\{\int_0^t w_i^*(u)Y_i^*(u)dN_i(u)\int_0^u w_i^*(v)d\Lambda_i(v)\right\}+2rE\left\{\int_0^t w_i^*(u)dN_i(u)\int_u^t w_i^*(v)Y_i^*(v)d\Lambda_i(v)\right\}, \tag{A.3}$$

where we have used $Y_i^*(u)Y_i^*(v) = Y_i^*(u)$ for $v < u$. Note that the second term in (A.3), which is an integral over the range $v > u$, must be 0 because if $Y_i^*(v) = 1$, then $dN_i(u) = 0$ for all $u < v$, and on the other hand, if $dN_i(u) = 1$, then $Y_i^*(v) = 0$ for all $v > u$. Under the hypothesis of a relative risk of r, it follows that

$$E(Y_i^*(t)dN_i(t)|Y_i^*(t), V_i, r, S_i) = rY_i^*(t)d\Lambda_i(t).$$

Therefore, (A.3) reduces to

$$\begin{aligned}2r^2E\left\{\int_0^t\int_0^u w_i^*(u)w_i^*(v)Y_i^*(u)dN_i(u)d\Lambda_i(v)\right. &= 2r^2E\left\{\int_0^t\int_0^u w_i^*(u)w_i^*(v)Y_i^*(u)d\Lambda_i(u)d\Lambda_i(v)\right\}\\ &= r^2E\left\{\left[\int_0^t w_i^*(u)Y_i^*(u)d\Lambda_i(u)\right]^2\right\}.\end{aligned}$$

Thus, the second and third terms in (A.2) cancel and

$$\begin{aligned}\mathrm{Var}\left\{N_i^W(t) - rA_i^W(t)\right\} &= E\int_0^t [w_i^*(u)]^2dN_i^*(u)\\ &= rE\int_0^t [w_i^*(u)]^2Y_i^*(u)d\Lambda_i(u).\end{aligned}$$

Consider now a center $\mathcal{F}$ in which each individual has a relative risk r compared to the overall average mortality rate. In this case, $O^W(t) = \sum_{i\in\mathcal{F}} N_i^W(t)$ and $E^W(t) = \sum_{i\in\mathcal{F}} A_i^W(t)$, and as individuals are independent,

$$\mathrm{Var}\left\{O^W(t) - rE^W(t)\right\} = r\sum_{i\in\mathcal{F}} E\int_0^t [w_i^*(u)]^2Y_i^*(u)d\Lambda_i(u).$$

In the special case of no dependent censoring, the variance reduces to $r\int_0^t Y_i(u)d\Lambda_i(u) = rA_i(t)$, which is the usual martingale result. The special case $r = 1$, corresponding to the usual null hypothesis, is of special interest.

A.2. Dependent censoring simulation background

For the simulations in Section 5, we utilized a joint model for the mortality and censoring mechanisms as outlined in this section. In particular, we show that for an additive conditional mortality model given V_i and $Z_i(x)$, the expectation of the hazards on $Z_i(x)$ results in a multiplicative form, which can then be analyzed using a standard Cox PH model. Thus, this approach leads to both the time-dependent Cox model for the dependent censoring and the Cox model with fixed baseline covariates for the mortality model. The derivation follows that in [14].

Let V_i represent the baseline covariate of interest, for example, treatment assignment, and suppose that V_i has a Bernoulli distribution with probability of success p. The time-dependent covariate, $Z_i(t)$ (e.g., MELD score) for subject i is generated as a Poisson process with intensity $\mu e^{\gamma V_i}$ depending on the value of V_i. The time-dependent model for transplant or dependent censoring has hazard function

$$\lambda_i^C(x|Z_i(x), V_i) = \lambda_0^C\exp\left\{\beta^C Z_i(x) + \gamma^C V_i\right\}.$$

In an analogous way, the model for mortality given $Z_i(t)$ has hazard function

$$\lambda_i^D(x|Z_i(x), V_i) = \lambda_0^D\exp(\gamma V_i) + \beta^D Z_i(x).$$

Thus, the time-dependent covariate $Z_i(t)$ induces a correlation between the death time and the censoring time for the ith individual.

Following the derivations in [14], the marginal survivor function for mortality can be calculated as

$$\begin{aligned} S(x|V_i) &= E\{S(x|V_i, Z_i(x))|V_i\} = E\left\{\exp\left(-\int_0^x \left[\lambda_0^D e^{\gamma V_i} + \beta^D Z_i(u)\right] du\right)|V_i\right\} \\ &= \exp\left\{-\lambda_0^D e^{\gamma V_i} x\right\} E\left\{\exp\int_0^\infty \psi(u) Z_i(u) du | V_i\right\}, \\ &= \exp\left\{-\lambda_0^D e^{\gamma V_i} x\right\} \exp\{K_Z(\psi)\}, \end{aligned}$$

where $\psi_t(u) = -\beta^D I(u < x)$ and

$$\begin{aligned} K_Z(\psi) &= -\int_0^x \mu e^{\gamma V_i} ds + \int_0^x \mu e^{\gamma V_i} \exp\left\{\int_s^x \psi(v) dv\right\} ds \\ &= -\mu e^{\gamma V_i} x + \mu e^{\gamma V_i} \int_0^x e^{-\beta^D (x-s)} ds \\ &= \mu e^{\gamma V_i} \left(\frac{1}{\beta^D} - \frac{e^{-\beta^D x}}{\beta^D} - x\right). \end{aligned}$$

Therefore, the marginal hazard for mortality is

$$\begin{aligned} \lambda^D(x|V_i) &= -\frac{\partial \log S(x|V_i)}{\partial x} = \lambda_0^D e^{\gamma V_i} - \frac{\partial K_Z(\psi)}{\partial x} \\ &= \left[\lambda_0^D - \mu\left(e^{-\beta^D x} - 1\right)\right] e^{\gamma V_i} \\ &= \lambda_0^*(x) e^{\gamma V_i}, \end{aligned}$$

which is a standard Cox model with fixed covariates and baseline hazard, $\lambda_0^*(x) = \lambda_0^D - \mu(e^{-\beta^D x} - 1)$.

Acknowledgements

We would like to thank the Kidney Epidemiology and Cost Center at the University of Michigan for its support of this research. We also thank Dr Min Zhang and Dr Robert Merion for their valuable input to this work and the Associate Editor and referees, whose comments and questions helped greatly to improve the manuscript. The data for this study were made available by the SRTR, which is funded by contract from the Health Resources and Services Administration (HRSA), US Department of Health and Human Services.

References

1. Steiner S, Cook R, Farewell V, Treasure T. Monitoring surgical performance using risk-adjusted cumulative sum charts. *Biostatistics* 2000; **1**:441–452.
2. Steiner S, Cook R, Farewell V. Risk adjusted monitoring of surgical outcomes. *Medical Decision Making* 2001; **21**:163–169.
3. Axelrod D, Guidinger MK, Metzger RA, Wiesner RH, Webb RL, Merion RM. Transplant center quality assessment using a continuously updatable risk-adjusted technique (CUSUM). *American Journal of Transplantation* 2006; **6**:313–323.
4. Axelrod DA, Kalbfleisch JD, Sun RJ, Guidinger MK, Biswas P, Levine GN, Arrington CJ, Merion RM. Innovations in the assessment of transplant center performance: implications for quality improvement. *American Journal of Transplantation* 2009; **9**:959–969.
5. Biswas P, Kalbfleisch JD. A risk-adjusted CUSUM in continuous time based on the Cox model. *Statistics in Medicine* 2008; **27**:3382–3406.
6. Robins JM, Finkelstein D. Correcting for non-compliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics* 2000; **56**:779–788.
7. Sun RJ, Kalbfleisch JD. A risk-adjusted O-E CUSUM with monitoring bands for monitoring medical outcome. *Biometrics* 2013. DOI: 10.1111/j.1541-0420.2012.01822.x.
8. Collett D, Sibanda N, Pioli S, Bradley A, Rudge C. The UK scheme for mandatory continuous monitoring of early transplant outcome in all kidney transplant centers. *Transplantation* 2009; **88**:970–975.
9. Kalbfleisch JD. 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* 2009; **88**:968–969.
10. Gandy A, Kvaloy JT, Bottle A, Zhou F. Risk-adjusted monitoring of time to event. *Biometrika* 2010; **97**:375–388.

Statistics in Medicine

11. Dickinson DM, Shearon TH, O'Keefe J, Wong H-H, Berg CL, Rosendale JD, Delmonico FL, Webb RL, Wolfe RA. SRTR center-specific reporting tools: posttransplant outcomes. *American Journal of Transplantation* 2006; **6**:1198–1211.
12. Zhang M, Schaubel DE. Estimating differences in restricted mean lifetime using observational data subject to dependent censoring. *Biometrics* 2011; **67**:740–749.
13. Lucas J, Crosier R. Fast initial response for CUSUM quality-control schemes: give you CUSUM a head start. *Technometrics* 1982; **24**:199–205.
14. Jewell NP, Kalbfleisch JD. Marker processes in survival analysis. *Lifetime Data Analysis* 1996; **2**:15–29.