Evaluating Failure Outcomes with Applications to Transplant Facility Performance

by

Jie (Rena) Sun

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2012

Doctoral Committee:

Professor John D. Kalbfleisch, Co-Chair
Professor Douglas E. Schaubel, Co-Chair
Professor Robert M. Merion
Assistant Professor Min Zhang

ACKNOWLEDGEMENTS

This thesis would not have been possible without the guidance, encouragement and patience of my co-advisors Dr. Jack Kalbfleish and Dr. Doug Schaubel. I thank both of them for sharing their insights and knowledge with me. Besides the academic and research skills I learned from them, many transferrable skills such as structured creative thinking, effective writing and presentation have all benefited tremendously on my research, career and life. I feel privileged to have had the opportunity to work with both of them.

I am also deeply grateful to my committee members, Dr. Robert Merion and Dr. Min Zhang, for their suggestions and comments.

In addition, I would like to thank the Department of Biostatistics at the University of Michigan and Kidney Epidemiology and Cost Center for their assistance and financial support throughout my graduate studies.

As always, it is impossible to mention everybody who had an impact on this work. I would like to thank all of my colleagues who had supported me throughout the course of the thesis.

Finally my special thanks go to my wonderful parents, Baosheng Sun and Huifen Su, whose faith in me was the motivation to carry on with this work. I thank them for their encouragement, patience and understanding.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

LIST OF APPENDICES

Appendix

CHAPTER I

Introduction

In this thesis, I develop methods to evaluate mortality experience of medical facilities, with applications to transplant facility-specific post-transplant mortality and pre-transplant waitlist mortality. We aim to compare the center-specific outcomes with the standard practice while providing timely feedback to the centers.

In Chapter II, we introduce a risk-adjusted O-E (Observed-Expected) Cumulative Sum (CUSUM) chart along with monitoring bands as decision criterion, to monitor the post-transplant mortality in transplant programs. This can be used in place of a traditional but complicated V-mask and yields a more easily interpreted chart. The resulting plot provides bounds that allow for simultaneous monitoring of failure time outcomes with signals for 'worse than expected' or 'better than expected'. The plots are easily interpreted in that their slopes provide graphical estimates of relative risks and direct information on additional failures needed to trigger a signal. Appropriate rejection regions are obtained by controlling the false alarm rate (Type I error) over a period of given length.

In Chapter III, we discuss the construction of a weighted CUSUM to evaluate pre-transplant waitlist mortality of facilities in the context where transplantation is considered to be dependent censoring. This setting arises, for example, with patients

on the liver transplant waitlist. These 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. Waitlisted patients with higher MELD score have a higher risk of death and consequently are given higher priority to receive a liver transplant when available. Unless the time-dependent factors (such as MELD) are adjusted for in the pre-transplant death model, censoring (transplant) time is correlated with the patient's unobserved time of death. To evaluate the waitlist mortality of transplant centers, it is important to take this dependent censoring into consideration; failing to do so could yield biased results. We assume a 'standard' transplant practice through a transplant model, utilizing Inverse Probability Censoring Weights (IPCW) to construct a weighted CUSUM. We evaluate the properties of a weighted zero-mean process as the basis of the proposed weighted CUSUM. A rule of setting control limits is discussed. A case study on regional liver transplant waitlist mortality is carried out to demonstrate the use of the proposed weighted CUSUM.

In Chapter IV, we provide an explicit road map for using a Cox dependent censoring model in the IPCW approach, complete with details of implementation. The Cox IPCW method has not been widely adopted among practitioners, despite its flexibility and wide applicability. It is likely that the technical implementation, which seems tricky and challenging, is the main obstacle hindering its wide adoption. In addition to the software implementation details, we evaluate an alternative parametric IPCW approach to gain computational efficiency. Simulation studies and case study on the national liver transplant waitlist mortality are conducted to demonstrate the similarity in estimates between Cox IPCW and PWE IPCW, and the computational savings by the PWE IPCW as compared to the Cox IPCW.

In the last chapter, we discuss the future directions of our work.

CHAPTER II

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

2.1 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 et al. (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 one-year binary mortality outcomes using a cohort of transplanted patients at multiple centers. However, a built-in one-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 time scale, 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 time-scale 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 Figure 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 sug-

gested a CUSUM as a 'reversed' SPRT and showed that a pre-determined 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 compare post-transplant outcomes at the center to those that would be expected from a model based on national data, where the expectations are risk adjusted to reflect the patient mix at the center under review. In this approach, the standard for comparison is 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), Steiner et al. (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 one-year patient survival) are rather poorly estimated (Kalbfleisch, 2009).

2.2 Method

2.2.1 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

$$\alpha_i(x) = \alpha(x; Z_i) = \lim_{\delta \to 0} P\{X \in (x, x+\delta) | X \geq x, Z_i\}/\delta, \tag{2.1}$$

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 one-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,

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

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 $N_Q(t) = \sum_i I(S_i \le t)$ denotes the number of transplants that have taken place in $(0,t]$. We define the 'at risk' process for subject i as $Y_i(t) = I\{S_i < t \le \min(T_i, S_i + 1)\}$.

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 $\mathcal{F}_{t^-} = \{N_Q(u), N_i^D(u), Y_i(u), Z_i, i = 1, \cdots, N_Q(t); 0 \le u < t\}$ and define the intensity function of subject i at this center as

$$E\{dN_i^D(t)|\mathcal{F}_{t^-}, \mu\} = e^\mu d\Lambda_i(t) = \begin{cases} Y_i(t)e^\mu \alpha_i(t - S_i)dt & \text{if } t > S_i; \\ 0 & \text{otherwise,} \end{cases} \tag{2.2}$$

where α_i is defined in (2.1) 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 one-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.

2.2.2 The O-E CUSUM with a V-mask

Based on the model (2.2), 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 $L(\mu) = \prod_{i=1}^{N_Q(t)} \exp\{\mu N_i^D(t) - e^{\mu}\Lambda_i(t)\}$.

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 pre-determined constants.

The likelihood ratio of $\mu = \theta$ versus 0 for a center based on the data in $(s, t]$ with starting time $s \in (0, t]$ is $\mathrm{LR}(\theta; s, t) = \exp\left[\theta\{N^D(t) - N^D(s)\} - (e^{\theta} - 1)\{A(t) - A(s)\}\right]$. 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 re-written as

$$C(s) < \{C(t) - h_1 - k_1 A(t)\} + k_1 A(s), \text{ for } H_-, \tag{2.3}$$

$$C(s) > \{C(t) + h_2 - k_2 A(t)\} + k_2 A(s), \text{ for } H_+. \tag{2.4}$$

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 (2.3) and (2.4) correspond to straight-line boundaries (Figure 2.1) 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 $s \leq t$ (Wetherill, 1977). The same boundaries (2.3) and (2.4) 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 above. 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

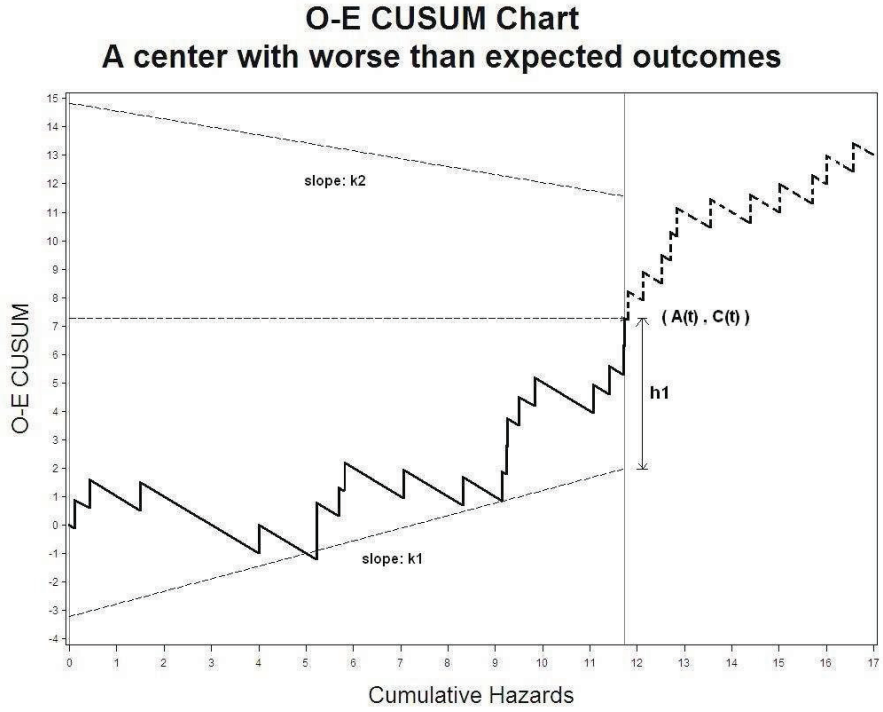

Figure 2.1: An O-E CUSUM with V-mask triggering 'worse than expected' signal.

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 above, 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 re-specify the CUSUM signals so that they can be implemented in a plot against t.

2.2.3 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 (2.3), let

$$M_1(t) = \inf_{s \le t}\{C(s) - k_1 A(s)\} + h_1 - \{C(t) - k_1 A(t)\}, \tag{2.5}$$

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 2.2.6.

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 (2.5) 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 $M_1(t)$.

2.2.4 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 a SPRT. For the alternative hypothesis of a rela-

tive 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.

2.2.5 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 $\text{ARL}(\theta) = E(\tau; \theta)$. 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 time-scale transformation $s = A(t)$. The modulated Poisson process $N^D(t)$ with intensity $A(t)$ is transformed to the new time-scale 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 time scale as $\tilde{\tau}$, so that $\tilde{\tau} = A(\tau)$ where τ is the signal time on the original time scale. They showed that the ARL in control on this new time scale, $E(\tilde{\tau}; 0)$, 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 time-scale or, equivalently, expected number of failures until a false alarm on the original time-scale. 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.

2.2.6 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 one-year post-transplant patient survival are presented. Similar charts could be constructed for other outcomes or other length of follow-up, such as one-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 2.2). The O-E chart (Figure 2.2.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 2 fewer failures. The one-sided charts (Figure 2.2.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 2.3.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 2.3.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 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.

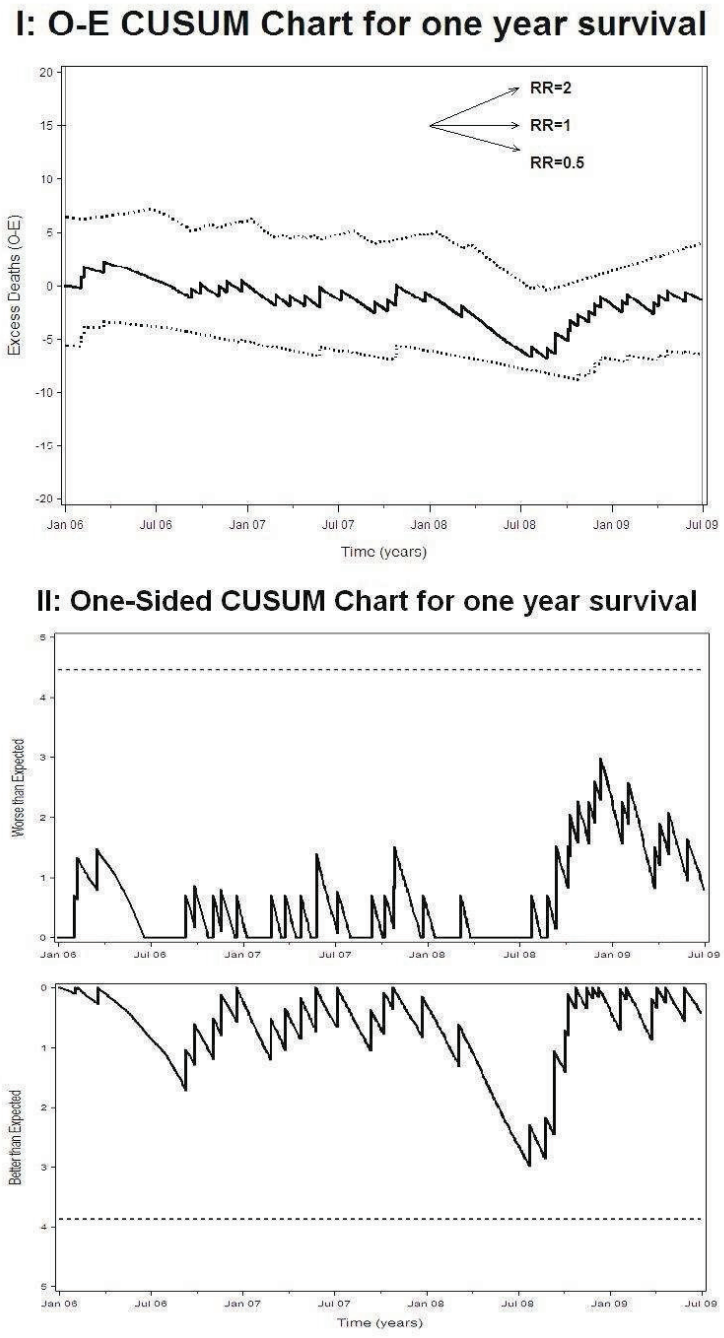

Figure 2.2: Center A, with 378 patients between January 01, 2006 and June 30, 2009

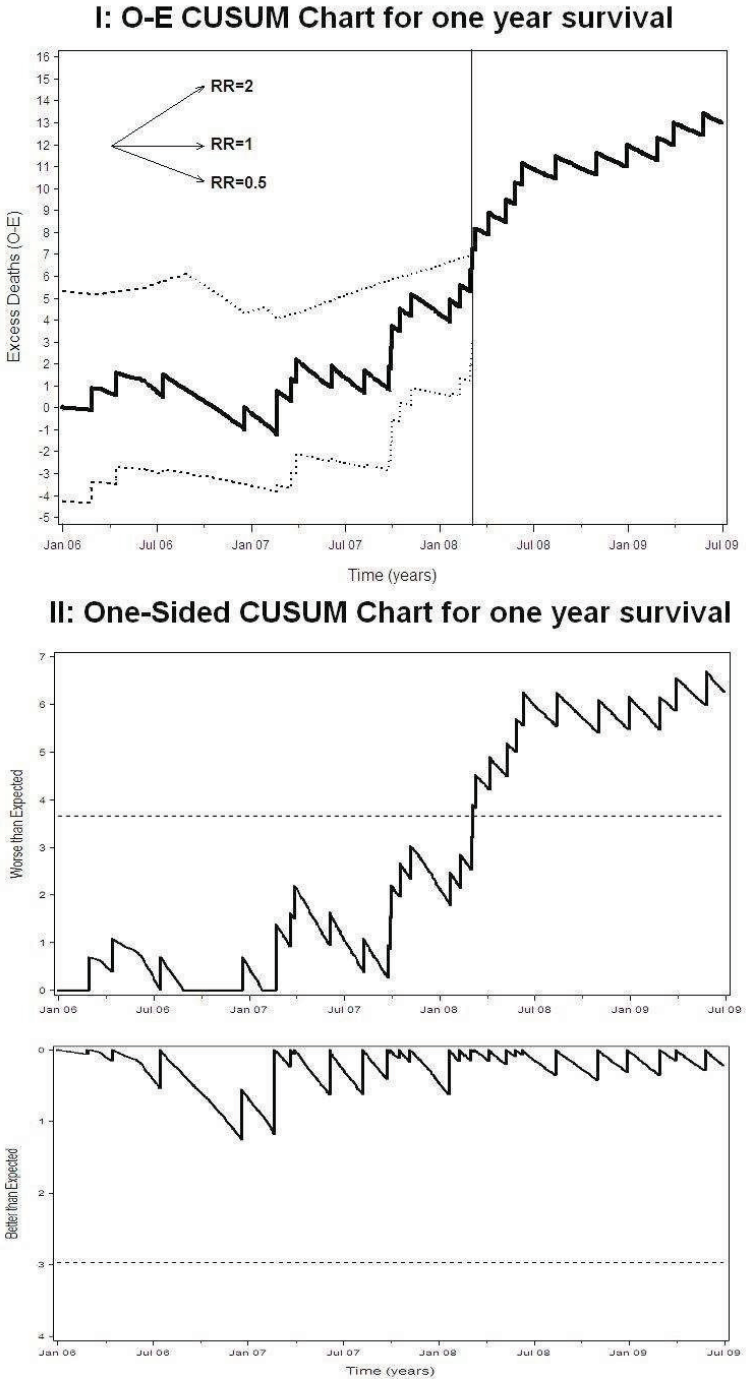

Figure 2.3: Center B, with 173 patients between January 01, 2006 and June 30, 2009

2.2.7 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 0, 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 $L_1/2$.

2.3 Simulation Studies

2.3.1 Control Limits

We consider transplants arriving according to a homogeneous Poisson process and suppose that the post-transplant failure time distribution for the national average is exponential with rate λ_0, corresponding to a one-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 one 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 con-

Table 2.1: Control limits, power and ARL of the O-E CUSUM.

Expected failures per year	Relative Risk 2			Relative Risk 0.5		
	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

trolling 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 2.1 gives the control limits of the O-E CUSUM obtained through controlling the Type I error as described above. 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 5 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.

2.3.2 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 re-set the CUSUM by applying the head-start described in Section 2.2.7 and then continue monitoring.

Table 2.2 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.

Table 2.2: 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

2.4 Case Studies

To demonstrate the use of the O-E CUSUM, we performed a retrospective analysis on one-year post-transplant 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 8 transplants per year, for which the CUSUMs would be expected to yield little power.

The SRTR models for post-transplant survivals were utilized to represent the national rates and to compute the expected outcomes. The SRTR one-year survival model for deceased donor transplants adjusts for 60 donor and recipient characteristics, whereas the model for living donor adjusts for 8 donor and recipient characteristics. Because the models for deceased and living donors are 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 2.1. 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 2.3. 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. Further, 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 2.2.6.

It is worth noting if all centers perform at the national average level, we would expect to see 8% (about 5 signals out of 58 centers of interest) signalling 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 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.

2.5 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

Table 2.3: The number of centers signalled 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	
		# of signals	AVE	# 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

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; further, 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 mis-specified 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 paper, 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.

CHAPTER III

Weighted Cumulative Sum (WCUSUM) to Monitor Medical Outcomes in the Presence of Dependent Censoring

3.1 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 clinicians by measuring the occurrence 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. (2000) and Steiner et al. (2001) developed a risk-adjusted one-sided CUSUM procedure based on the likelihood ratio in a logistic model for binary outcomes. They proposed a graphical method for identifying either a substantial or consistent change in risk-adjusted mortality. Axelrod et al. (2006) demonstrated the utility of the one-sided CUSUM method for tracking and analyzing one-year binary mortality outcomes using a cohort of transplanted patients at multiple centers. However, a built-in one-year lag is necessary in this approach. Biswas and Kalbfleisch (2008) developed a risk-adjusted one-sided CUSUM procedure constructed on a con-

tinuous 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 (SPRT) forms the basis of the one-sided CUSUM which examines whether there is evidence that could lead to rejection of the null hypothesis in favor of 'worse than expected' (or 'better than expected') performance at the center as compared to the reference national average mortality rates.

All these methods are developed based on the assumption of independent censoring. This could be violated in some cases, especially in medical settings where preventive approaches are applied on high-risk patients, and highly correlated dependent censoring may occur. For example, patients on the liver transplant waitlist are evaluated constantly to assess their current medical condition. One summary measure of time is Model for End-Stage Liver Disease score, or MELD score. 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 had the patient been left untransplanted. To evaluate the waitlist mortality of patients in transplant centers, it is important to take the dependent censoring factor (transplantation) into consideration; failing to do so yields biased results.

In this paper, we discuss a Weighted CUSUM (WCUSUM) to account for dependent censoring. Motivated by the waitlist mortality issue for liver transplant centers, we phrase the description of the method to address this case directly. Trans-

plant represents dependent censoring and mortality is the failure event. The method, however, could be adapted to monitor other datasets where dependent censoring is present.

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 national average performance, having adjusted for dependent censoring through the MELD score.

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 (Robins and Finkelstein, 2000). We describe the signalling rules for the WCUSUM. Simulation studies are conducted to demonstrate the properties of the weighted process. A case study is followed to illustrate the use of the proposed WCUSUM.

3.2 Notation

Assume patient i enters the cohort at calendar time S_i (e.g. time of initial listing on the transplant waitlist). Denote D_i as time to death since entry and C_i as time to transplant since entry. Let X_i be the observed event time since entry to either death or transplant whichever occurs first, $X_i = \min(D_i, C_i)$. Let T_i be the calendar time of the observed event, so that $T_i = S_i + X_i$. Let $Z_i(x), 0 \leq x \leq X_i$, be the set of time-dependent covariates (e.g. MELD scores) and let V_i be a set of baseline covariates.

Assume we have a population model on time to mortality since entry with a hazard function $\alpha_i(x) = \alpha(x; V_i)$ for subject i where $\alpha_i(x) = \lim_{\Delta \to 0} P\{D_i \in (x, x + \Delta) | D_i \geq x, V_i\}/\Delta$. Let $d\Lambda_i^*(t) = I(t > S_i)\alpha_i(t - S_i)dt$ define the hazard for subject i at calendar time t.

Now we build a process to count the qualifying failures at a particular center ϵ. Suppose that survival over a one-year period is of interest, so that at-risk indictor is $Y_i^*(t) = I\{S_i < t \leq \min(T_i, S_i + 1)\}$. Let $\delta_i = I(D_i = X_i)$ be the failure indicator. Let $N_i^*(t)$ count the number of qualifying failures in the chronological time interval $(0, t]$ for subject i:

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

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 number of qualifying failures in (0,t] for the center ϵ is $N^*(t) = \sum_{i \in \epsilon} N_i^*(t)$, where the summation is overall individuals i in this center ϵ.

3.3 Method

3.3.1 A Weighted Zero-Mean Process

In this section, we first state the key assumption on dependent censoring. We consider the independent censoring case and the usual zero-mean process. Then we take dependent censoring into account to construct a weighted zero-mean process. This is the foundation to the weighted CUSUM that is described in the following section.

Assume the cause-specific hazard for censoring is $\lambda_i^C(x|\bar{Z}_i(x), V_i) = \lim_{\Delta \to 0} P\{C_i \in$

$(x, x+\Delta)|D_i \geq C_i \geq x, \bar{Z}_i(x), V_i\}/\Delta$, where $\bar{Z}_i(x) = \{Z_i(s), 0 < s \leq x\}$. The key assumption is that

$$\lambda_i^C(x|\bar{Z}_i(x), V_i) = \lim_{\Delta \to 0} P\{C_i \in (x, x+\Delta)|C_i \geq x, \bar{Z}_i(y), V_i, D_i = y\}/\Delta, \tag{3.1}$$

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 knowledge of the future value of $D_i = y > x$ or the additional information on $\{Z_i(v), x < v \leq y\}$. Under this assumption, it follows that

$$P\{C_i > x|V_i, \bar{Z}_i(y), D_i = y\} = \exp\{-\int_0^x \lambda^C(u|\bar{Z}_i(u), V_i)du\} \tag{3.2}$$

for all $0 < x < y$. This assumption (3.1) and its consequence (3.2) are essential to the use of inverse weights and for the use of the process $Z_i(x)$ to fully correct for bias due to independent censoring (Robins and Finkelstein, 2000).

Let $\tilde{N}_i(t)$ represent the underlying failure counting process in the absence of dependent censoring so that $\tilde{N}_i(t) = I(S_i + D_i \leq t < S_i + 1)$ if $t \leq S_i + 1$ and $\tilde{N}_i(t) = \tilde{N}_i(S_i + 1)$ if $t > S_i + 1$. Similarly, let $\tilde{Y}_i(t)$ denote the underlying at-risk indicator in the absence of dependent censoring, $\tilde{Y}_i(t) = I\{S_i < t < \min(S_i + D_i, S_i + 1)\}$. It follows that

$$E(d\tilde{N}_i(t)|\tilde{Y}_i(t), V_i, S_i) = \tilde{Y}_i(t)\alpha_i(t - S_i)dt = \tilde{Y}_i(t)d\Lambda_i(t).$$

Without any censoring, the CUSUM at center ϵ could compare the observed number of failures $O(t) = \sum_{i \in \epsilon} N_i(t)$ with the expected number of failures $E(t) = \sum_{i \in \epsilon} \int_0^t \tilde{Y}_i(u)d\Lambda_i(u)$, and O(t)-E(t) is a zero-mean process if center ϵ has the same mortality rates as the reference population.

Now assuming the center ϵ still has the same mortality rates as the reference population but has dependent censoring, we aim to develop a zero-mean process analogue to O(t)-E(t) alone. Let $dM_i^*(t) = dN_i^*(t) - Y_i^*(t)d\Lambda_i^*(t) = Y_i^*(t)[d\tilde{N}_i(t) - d\Lambda_i^*(t)]$. Note that $Y_i^*(t) = \tilde{Y}_i(t)I(C_i > t - S_i)$ and

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

Under assumption (3.2), it follows that

$$E\{I(C_i > t - S_i)|\tilde{Y}_i(t), d\tilde{N}_i(t), \bar{Z}_i(t - S_i), S_i, V_i\} = \exp\left\{-\int_0^{t-S_i} \lambda_i^C(u|\bar{Z}_i(u), V_i)du\right\}.$$

So that

$$E[dM_i^*(t)] = E\left\{\exp\{-\int_0^{t-S_i} \lambda_i^C(u|\bar{Z}_i(u), V_i)du\}\tilde{Y}_i(t)[d\tilde{N}_i(t) - d\Lambda_i^*(t)]\right\}. \tag{3.3}$$

The expression (3.3) shows that the $M_i^*(t)$ process does not in general have mean zero. However, it also indicates how to obtain a zero-mean process.

Let $w_i^*(t) = w_i(t - S_i) = \exp\{\int_0^{t-S_i} \lambda_i^C(u|\bar{Z}_i(u), V_i)du\}$. It is now easy to see that

$$E[w_i^*(t)dM_i^*(t)|\tilde{Y}_i(t), V_i, S_i] = E\{\tilde{Y}_i(t)[d\tilde{N}_i(t) - d\Lambda_i^*(t)]|\tilde{Y}_i(t), V_i, S_i\} = 0. \tag{3.4}$$

This equation (3.4) 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 ϵ is $N^W(t) - A^W(t)$, where $N^W(t) = \sum_{i\in\epsilon} N_i^W(t)$ and $A^W(t) = \sum_{i\in\epsilon} A_i^W(t)$. In fact, we are replacing O(t) and E(t) above with estimates that adjusted for the dependent censoring.

In the independent censoring case, when all weights are equal to 1, this process reduces to the normal zero-mean Martingale, with

$$w_i^*(t)dM_i^*(t) = d\tilde{N}_i(t) - I(D_i \geq t - S_i)d\Lambda_i(t) = d\tilde{M}_i(t),$$

and $N^W(t) - A^W(t) = O(t) - E(t)$.

3.3.2 One-Sided Weighted CUSUM Chart

First let us revisit the one-sided CUSUM chart in the independent censoring case proposed by Biswas and Kalbfleisch (2008). At time t, consider testing $H_0 : \mu = 0$ versus $H_1 : \mu = \theta > 0$ ($e^\theta > 1$), where e^θ denotes relative risk of such process. The logarithm of the likelihood under relative risk e^θ, $\log L(t;\theta)$, is proportional to $\sum_i \{\theta N_i(t) - e^\theta A_i(t)\} = \theta N(t) - e^\theta A(t)$, where $N_i(t)$ counts the number of qualified failure for subject i up to time t, and $A_i(t)$ represents the cumulative hazards of this subject up to time t. So that the one-sided CUSUM G_t is defined by $G_{t+dt} = \max\{0, G_t + \theta dN(t) - (e^\theta - 1)dA(t)\}$, with $G_0 = 0$. This CUSUM can be designed to detect either a 'worse than expected' performance with $\theta > 0$ or 'better than expected' performance with $\theta < 0$. It triggers a signal if the process exceeds a pre-determined value.

With the presence of dependent censoring, we utilize weighted cumulative failures and weighted cumulative hazards defined in the last section in place of the ordinary values, use $dN^W(t) = \sum_i w_i^*(t) dN_i^*(t)$ and $dA^W(t) = \sum_i w_i^*(t) Y_i^*(t) d\Lambda_i^*(t)$ in place of $dN(t)$ and $dA(t)$. The one-sided Weighted CUSUM is

$$G^W_{t+dt} = \max\{0, G^W_t + \theta dN^W(t) - (e^\theta - 1)dA^W(t)\},$$

with $G^W_0 = 0$.

Weighted values still maintain the asymptotic properties as shown by previous research, but additional variation is introduced through the weights. In the next section, we quantify the variance of the weighted zero-mean process.

3.3.3 Variance of the Zero-Mean Process $N^W(t) - A^W(t)$

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 each individual i has mean zero under null hypothesis when a center has the same mortality rates as the reference population. We now investigate the variance of this process.

Consider the general case with true relative risk r, meaning that the mortality rates in the center ϵ are r times the rates of the population. A weighted zero-mean process for individual i would be $N_i^W(t) - rA_i^W(t) = \int_0^t w_i^*(u)dN_i^*(u) - r\int_0^t w_i^*(u)Y_i^*(u)d\Lambda_i^*(u)$. The variance of this process is

$$
\begin{aligned}
& \mathrm{Var}\left\{\int_0^t w_i^*(u)dN_i^*(u) - rw_i^*(u)Y_i^*(u)d\Lambda_i^*(u)\right\} \\
= \;& E\left\{\int_0^t w_i^*(u)dN_i^*(u) - rw_i^*(u)Y_i^*(u)d\Lambda_i^*(u)\right\}^2 \\
= \;& E\left\{\int_0^t [w_i^*(u)]^2 dN_i^*(u) - 2r\int_0^t w_i^*(u)dN_i^*(u)\int_0^t w_i^*(u)Y_i^*(u)d\Lambda_i^*(u)\right. \\
& \left. +r^2[\int_0^t w_i^*(u)Y_i^*(u)d\Lambda_i^*(u)]^2\right\}. \qquad (3.5)
\end{aligned}
$$

The second term in (3.5) has the expectation as follows

$$
\begin{aligned}
& E\{2r\int_0^t w_i^*(u)dN_i^*(u)\int_0^t w_i^*(u)Y_i^*(u)d\Lambda_i^*(u)\} \\
= \;& 2rE\{\int_0^t\int_0^t w_i^*(u)w_i^*(v)Y_i^*(u)Y_i^*(v)d\tilde{N}_i(u)d\Lambda_i^*(v)\} \\
= \;& 2rE\{\int_0^t w_i^*(u)Y_i^*(u)d\tilde{N}_i(u)\int_0^u w_i^*(v)d\Lambda_i^*(v)\} \\
& +2rE\{\int_0^t w_i^*(u)d\tilde{N}_i(u)\int_u^t w_i^*(v)Y_i^*(v)d\Lambda_i^*(v)\}, \qquad (3.6)
\end{aligned}
$$

with $Y_i^*(u)Y_i^*(v) = Y_i^*(u)$ for $v < u$ and $Y_i^*(u)Y_i^*(v) = Y_i^*(v)$ for $v > u$. Note that for the second term in (3.6) with $v > u$, when $Y_i^*(v) = 1$, $d\tilde{N}_i(u)$ has to be 0, because the fact that the subject is at risk for time v indicates it didn't fail at $u < v$. Similarly, when $d\tilde{N}_i(u) = 1$, it indicates that $Y_i^*(v) = 0$, meaning that if the

subject fails at time u, it is removed from the at risk set for time $v > u$. So that $E\{\int_0^t w_i^*(u)d\tilde{N}_i(u)\int_u^t w_i^*(v)Y_i^*(v)d\Lambda_i^*(v)\} = 0$. Therefore, (3.6) is

$$E\{2r\int_0^t w_i^*(u)dN_i^*(u)\int_0^t w_i^*(u)Y_i^*(u)d\Lambda_i^*(u)\} = 2rE\{\int_0^t\int_0^u w_i^*(u)w_i^*(v)Y_i^*(u)d\tilde{N}_i(u)d\Lambda_i^*(v)\}.$$

Now let us take a look at the third term in (3.5),

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

It is then obvious that, under the hypothesis of relative risk r and

$$E(Y_i^*(t)d\tilde{N}_i(t)|Y_i^*(t), V_i, r, S_i) = rY_i^*(t)d\Lambda_i^*(t),$$

the second and the third terms in (3.5) cancel. Thus (3.5) is

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

Under the null hypothesis of center having the same failure risk as the national average, or $r = 1$, we have $\mathrm{Var}\left\{\int_0^t w_i^*(u)dN_i^*(u) - w_i^*(u)Y_i^*(u)d\Lambda_i^*(u)\right\} = E\int_0^t [w_i^*(u)]^2 Y_i^*(u)d\Lambda_i^*(u)$. The variance of the process $N^W(t) - rA^W(t)$ under null accounting for all subjects at the center ϵ is then $\mathrm{Var}^W(t) = \mathrm{Var}\{N^W(t) - A^W(t)\} = \sum_{i\in\epsilon}\int_0^t [w_i^*(u)]^2 Y_i^*(u)d\Lambda_i^*(u)$.

In the special case of no dependent censoring, failure process is Poisson, weights reduce to 1, then the zero-mean process returns to the ordinary zero-mean process, and variance reduces to $r\int_0^t \tilde{Y}_i(u)d\Lambda_i(u) = rA_i(t)$ under the hypothesis of relative risk r.

3.3.4 Control Limits

A few different approaches have been discussed to set control limits for CUSUM processes. In the ordinary or independent censoring case, Gandy et al. (2010) utilize the expected number of observed events before stopping or the average run length in calendar time to calibrate control limits. Continuous time t is transformed to $\Lambda(t)$, which maps the counting process of observed failures to a homogeneous Poisson process with rate 1. In the dependent censoring scenario, the weighted counting process can no longer be mapped to a homogeneous Poisson process through the time transformation. Although the weighted expected number of failures recovers the underlying expected number of failures had there have been no dependent censoring, increased variance inflate the error rate α. Adopting an approach similar to Gandy et al. (2010), we can use both the weighted expected failures and the variance of the weighted zero-mean process under the null hypothesis to calibrate control limits. When the dependent censoring is positively correlated with death, the proportion of change in standard deviation of the weighted zero-mean process increases linearly with the proportion of change in control limit. We demonstrate this approach through simulation in the Appendix.

Biswas and Kalbfleisch (2008) and Sun and Kalbfleisch (2012) 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. For example, Biswas and Kalbfleisch (2008) uses a false positive rate of 8% over a 3.5 year period. This yields control limits that are lower for smaller centers and higher for larger centers. Based on our dataset and interest in monitoring, we choose to use a similar method of controlling Type I error over a fixed period to obtain a control limit L for the weighted CUSUM. Without

any knowledge of the mechanism for dependent censoring in the dataset, this can no longer be done via a simple simulation. We utilize resampling technique to calibrate control limits for a center of given size ψ and over a certain period of time. To do this, we require a reference population which forms the standard to which centers are to be compared. This reference population is subject to both failure and dependent censoring. We draw randomly and repeatedly samples of size ψ and construct WCUSUM. Then we choose the control limit so that a given population of the simulated WCUSUMs has a signal rate of α over the period of interest (e.g. 8% in 3.5 years).

3.3.5 IPCW Weights Calculation

Robins and Finkelstein (2000) has shown that under assumption (3.1) we can estimate the true hazards Λ_i with the presence of dependent censoring, using the inverse probability of censoring weights (IPCW) approach. Chapter 4 gives more details of the setup and implementation.

We assume a Cox model for the time to transplant with hazard function

$$\lambda^C(x|\bar{Z}_i(x), V_i, D_i > x) = \lambda_0^C(x)\exp\{\gamma^C Z_i(x) + \beta^C V_i\}, \tag{3.7}$$

where $\lambda_0^C(x)$ is an unspecified baseline hazard function, $\bar{Z}_i(x) = \{Z_i(s), 0 < s \leq x\}$ and V_i is a set of baseline covariates. For simplicity without loss of generality, we assume that the censoring rate at time x depends only on the most recent value $Z(x)$.

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

Under the model (3.7) and assumption (3.1), the conditional probability of not

receiving a transplant until time x for subject i where survival time exceeds x is,

$$K_i^V(x) = P\{C_i \geq x | D_i > x, \bar{Z}_i(x), V_i\} = \exp\{-\Lambda_i^C(x)\}, \tag{3.8}$$

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 further reduce the variation in the weights due to baseline heterogeneity while still get unbiased estimates for the marginal death model of interest, 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 (3.8). Stabilized weights are then $\hat{w}_{i2}(x) = \hat{K}_i^0(x)/\hat{K}_i^V(x)$. It has been shown that stabilized weights also give unbiased parameter estimates for the marginal death model, but with smaller variation. Therefore, stabilized weights are used to obtain the mortality hazards. The process of obtaining true hazards in the marginal death model with weights is presented in Appendix A, using the same approach that Robins and Finkelstein (2000) performed.

Note that the probability of having some large weights in this process is small, although it can sometimes happen. For example, the chance that a patient who is alive with large MELD score but has not received a transplant is very small.

In practice, one can also use a stratified Cox model or a parametric model (e.g. a piecewise exponential model) to obtain the weights. In our case, since transplant donors are strategically allocated within each OPO, instead of the entire national level, we utilize a stratified Cox model to estimate dependent censoring or transplant.

3.4 Simulation

3.4.1 Set-up

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 follows Bernoulli(p) 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 $Z_i(x) \sim \text{PP}(\mu e^{\gamma^D V_i})$. Suppose we are interested in one-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 (censoring).

Conditional on $Z_i(x)$ and V_i, we generate (cause-specific) censoring and mortality according to hazards 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. We choose an additive form for the conditional mortality model, to ensure that its marginal form taking expectation on $Z_i(x)$ is multiplicative (see Appendix) 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 is essential to generate a marginal mortality model in a proportional hazards format, so that Cox model can be used to estimate the mortality. The correlation between the transplant hazards and mortality hazards 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.

3.4.2 Variance of the Zero-Mean Process

In this section, we verify the variance calculation of the zero-mean process from equation (3.5). Consider a period of time in the equilibrium stage, say 1 year, and 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.

With relative risks 0.5, 1 and 2, Table 3.1 reports: the observed death rates; the dependent censoring rates; the Spearman rank correlation between latent death time and dependent censoring time; and the mean and the variance of $\mathrm{OE}_r^W = \mathrm{OE}_r^W(1) = N^W(1) - rA^W(1)$. In addition, it reports: the mean and standard deviation of what we refer to as the empirical variance,

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

the mean and standard deviation of the variance constructed in equation (3.5), $\widetilde{\mathrm{Var}} = \widetilde{\mathrm{Var}}\{OE_r^W(1)\} = \sum_i r\int_0^1 [w_i^*(u)]^2 Y_i^*(u)d\Lambda_i^*(u)$; and the mean and variance of the score statistic,

$$\mathrm{Score} = \mathrm{Score}\{\mathrm{OE}_r^W(1)\} = \frac{\mathrm{OE}_r^W(1)}{\sqrt{\widetilde{\mathrm{Var}}(\mathrm{OE}_r^W(1))}}.$$

Table 3.1: Confirmation of the expected variance and the zero-mean process.

				OE_r^W		$\widehat{\mathrm{Var}}$		$\widetilde{\mathrm{Var}}$		Score	
r	Death	Censoring	Corr.	mean	Var	mean	SD	mean	SD	mean	Var
0.5	11.1%	0	0	-0.27	55.4	55.4	6.8	55.7	2.5	-0.03	1.00
	8.6%	32.4%	0.13	-0.44	66.4	68.2	29.0	67.8	6.5	-0.04	0.98
1	20.7%	0	0	0.11	103.6	103.7	8.6	103.8	4.5	0.01	1.00
	16.3%	29.6%	0.17	-0.37	125.2	124.9	36.0	125.5	12.6	-0.03	1.00
2	36.3%	0	0	-0.14	179.0	181.2	10.7	181.5	8.2	-0.00	0.98
	29.6%	24.6%	0.22	0.34	220.0	216.8	39.3	216.1	21.8	0.04	1.06

Table 3.1 shows that the OE_r^W is a zero-mean process with the mean value close to

0 under all scenarios, and that $\widetilde{\text{Var}}$ and $\widehat{\text{Var}}$ are both valid estimates of the variance of OE_r^W, based on the close values they suggest at the mean level. However, $\widetilde{\text{Var}}$ from equation (3.5) possesses much smaller variation than $\widehat{\text{Var}}$ under all scenarios. Score statistic is constructed under each run, suggesting the same conclusions: that the OE_r^W process is mean zero and $\widetilde{\text{Var}}$ and the variance of OE_r^W agree closely.

3.4.3 Recovery of Underlying Failure Risks

We now 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 expected hazards using true IPCW weights and hazards under dependent censoring (Scenario 2). Note that we can never obtain the true weights or hazards in practice. To mimic the practical implementation, we also compare with the values 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 of censoring and mortality models.

We consider the following parameter setup: $\mu_0 = 100$, $p = 0.5$, $\mu = 3$, $\gamma^D = \log(2)$, $\lambda^D = 0.01$, $\lambda^C = 0.05$, $\gamma^C = \log(2)$, $\beta^D = 0.1$ and $\beta^C = \log(2)$. The simulation is conducted using 100 repetitions. The one-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 3.2: 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 failures	20.35	4.17	19.21	5.63	19.29	5.43
Expected failures	20.77	1.92	20.72	2.14	20.34	2.19
Variance	20.77	1.92	34.77	5.36	36.51	10.47

Table 3.2 shows that in both Scenario 2 and 3, weighted observed failures and weighted expected failures in the dependent censoring case recover the true value of underlying failures and hazards have the center had no such censoring. Note, however, that variance is inflated in the dependent censoring case due to 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.

3.5 Case Study

3.5.1 Data Description

As an example, we evaluate liver transplant waitlist mortality using the data obtained from the Scientific Registry of Transplant Recipients (SRTR). We consider a five-year cohort of patients from one of 11 regions in the U.S. waitlisted between January 1st, 2004 and December 31st, 2008. Patients recorded as Status 1 or 1A at baseline have acute liver failure at waitlisting and are not included in the analysis. In addition, we exclude patients listed in error, changed to kidney/pancreas transplants or with previous liver transplant history. Given that pediatric patients follows a different scheme of transplant, we only include adults of age 18 years above in the analysis. Two centers with fewer than 5 patients waitlisted over this five-year span are excluded. In the final working set, 3,314 patients from 7 centers and 5 Organ Procurement Organizations (OPOs) are included.

We measure baseline covariates gender, race, age, diagnosis categories, diabetes, previous malignancy indicator, Body Mass Index (BMI), blood type, and hospitalization and Intensive Care Unit (ICU) status. All these covariates are included in both the transplant (censoring) model and the mortality model.

Time dependent variables consist of the Model for End-Stage Liver Disease (MELD) score, inactive period, and sodium value. MELD is the scoring system used to prioritize patient on the liver wait list. It combines serum bilirubin, serum creatinine and the international normalized ratio for prothrombin time (INR). Allocation MELD score is used in practice as the main determining factor on liver deceased donor allocation. We record MELD as 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 1/1A. Assuming that a patient is being monitored sufficiently by the clinician, it would appear reasonable to believe that lack of a MELD update reflects the fact that the patient's MELD has not changed. This would imply that coding MELD score as a step-function (i.e. last-value carried-forward) would be appropriate. Sometimes, patients are temporarily removed from the waitlist for various reasons such as medical condition, refusing transplant, improved or deteriorated condition, or being inactive on the program for more than 2 years. Such patients are not supposed to receive offers of deceased donor livers when they are removed from or not on the waitlist. We set the inactive indicator as 1 to identify the period of removal and to capture this information in modeling. Sodium value is recorded as a continuous variable and is also included in the set of time dependent covariates. Alternative approaches of handling inactive time were used by Zhang and Schaubel (2011).

Death on the waitlist is the event of interest in our analysis. 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 of follow-up or received a living donor transplant, which typically is not predicted by MELD score.

The data from this region is considered as population data. We fit a Cox model of the equation (3.7) to the dependent censoring to obtain appropriate IPCW weights. In the censoring model, $Z_i(x)$ is the time-dependent MELD, inactive period and sodium level, with $Z_i(0)$ indicating the baseline values of these variables. V_i is the set of baseline covariates. The reader is referred to the case section in Chapter 4 for details of the baseline covariates.

We then conduct a weighted Cox death model stratified on centers with stabilized IPCW weights, using the same set of baseline covariates V_i and $Z_i(0)$. We used weighted Cox models to estimate hazards for death, controlling for time-dependent confounding variables (i.e. MELD, inactive period and sodium level). 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.

Resampling technique with replacement is used for 1,000 iterations to obtain control limits. Since the region of interest is the population or the reference, we sample N subjects from the entire region many times to choose an appropriate control limit for the facility, with N being the facility size. For example, for a facility with 300 patients arriving over a 5-year period, we randomly select 300 patients over the same 5-year period from the region and construct a weighted CUSUM. We repeat this process 1,000 times and calibrate a control limit L, so that the Type I error rate of the 5-year period is 10%. Now a WCUSUM for the facility of interest can be plotted with the same control limit L. Similarly, we repeat this process on facility sizes 100, 200 $\cdots$ 900, and 1,000 patients over the 5-year span. By controlling type I error rate at 10% for the entire 5-year period, we get control limit of each size summarized in Table 3.3.

3.5.2 Analysis and Results

We construct a WCUSUM in order to detect a relative risk of 2 for waitlist death rates at the center level, as compared to the overall regional data. Table 3.3 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.

Given the estimated expected number of failures at a center, we used linear interpolation based on values from Table 3.3 to find an appropriate control limit L. We apply the estimated control limits on the 7 centers in the selected region. No signal is presented in any center. Figure 3.1 demonstrates that the example center A with 472 patients over the 5 year period operates at the reference level for the first 4 years and has a spike in the number of deaths at the end of the fourth year, although the accumulation isn't enough to trigger a signal. Figure 3.2 shows center B with 1004 patients in the 5 year cohort has a large number of weighted failures observed around January 2006. We can see that although the actual number of failures are few, the weighted values are quite high which causes the spike of the WCUSUM. There are few high-risk patients that should have been transplanted but died on the waitlist at this center. Further investigation is suggested. After the spike at year 2006, the WCUSUM came back down around the zero line. This means that the sharp increase may just have been random variation.

3.6 Discussion

We assume that the information on MELD updates is accurate and that transplant (or censoring) model is correct, with no unmeasured confounders, so that the

Table 3.3: Control limits for Weighted CUSUM

Size	L	O^W	E^W	Var^W
100	5.08	16.95	16.90	29.35
200	5.98	33.71	33.78	59.76
300	6.76	50.42	50.43	88.00
400	7.28	68.36	68.14	120.97
500	7.47	84.97	85.27	150.57
600	7.73	101.59	101.45	176.98
700	7.92	117.98	118.29	205.92
800	8.10	135.23	135.73	239.10
900	8.15	152.47	152.44	268.13
1000	8.29	169.08	169.13	294.82

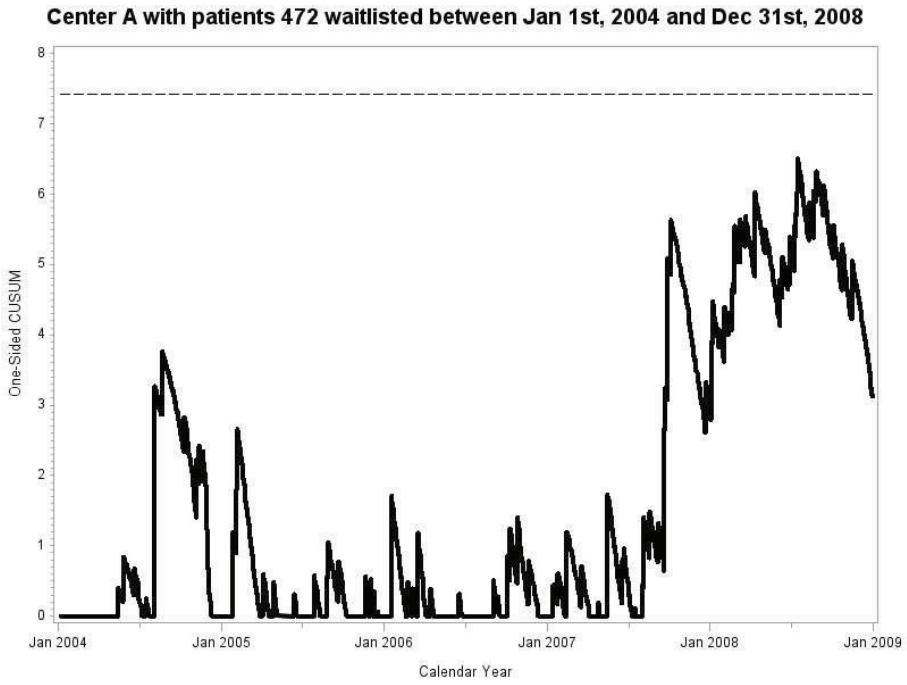

Figure 3.1: The weighted CUSUM of Center A for a 5-year period as compared to the standard practice of the region that Center A belongs to.

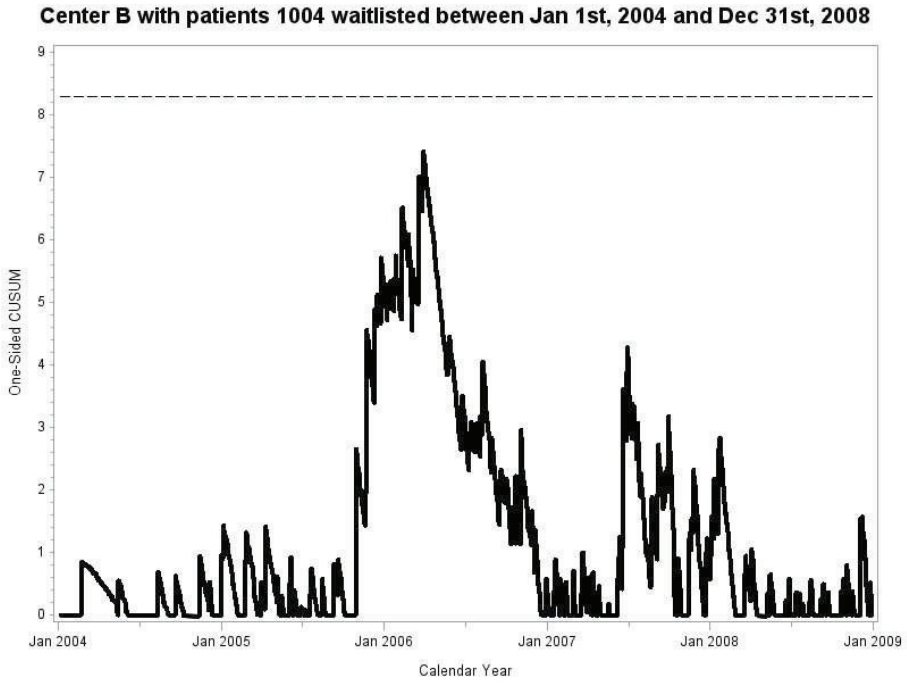

Figure 3.2: The weighted CUSUM of Center B for a 5-year period as compared to the standard practice of the region that Center B belongs to.

true hazards can be recovered by IPCW approach and the weighted process of the difference between cumulative observed number of failures and cumulative expected number of failures is a zero-mean process. We also assume that the Cox model for death is correct.

When 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 scheme is set nationally and should be strictly followed.

CHAPTER IV

Implementation of Inverse Probability Censoring Weighting using a Cox model and a Piecewise Exponential approach

4.1 Introduction

Time-to-event models are often used in analyzing biomedical data. However, in almost all application, the death time of interest may be censored, and the traditional independent censoring assumption is sometimes violated. This is especially true when a preventive approach is used. In studies collecting both longitudinal and survival information, time-dependent covariates are frequently related to both the event and dependent censoring. Ignoring the dependent censoring may introduce bias in estimating the failure hazards that would apply in the absence of censoring. For example, the receipt of a liver transplant constitutes dependent censoring in evaluating liver waitlist mortality. In this case, the Model for End-stage Liver Disease (MELD) score is highly predictive of both pre-transplant death and transplant time.

One way to estimate the underlying mortality model is through the Inverse Probability Censoring Weighting (IPCW) method (Robins and Rotnitzky, 1992; Robins and Finkelstein, 2000). The IPCW method first estimates weights based on the inverse probability that a surviving individual is uncensored via a time-dependent

censoring model. It then constructs an estimating equation based only on baseline covariates and weights from the censoring model.

Among regression methods for censored data, the Cox model is the most frequently used. The Cox model for death with IPCW weights has been shown to give consistent and unbiased estimates by Robins and Finkelstein (2000) and has been used in several works such as Schaubel et al. (2009) and Zhang and Schaubel (2011). However, the literature is not entirely clear on the implementation details. Several previous authors have also used a weighted pooled logistic model approximation (Hernán et al., 2000, 2002, 2006, 2008; Cole et al., 2005, 2007), which is asymptotically equivalent to a discrete Cox model and yields results close to the Cox model using exact times of events as illustrated by D'Agostino et al. (1990). However, this logistic approach treats each person-visit or person-day as an observation, and allow for a time-dependent intercept. Therefore, it expands the dataset into a much larger scale which results in considerable additional computational burden. In addition, to ensure consistent estimates in practice, this method requires to control the number of free parameters in the logisitc model using various means such as replacing intercepts with a linear term of a cubic splineHernán et al. (2000). It is worth noting that this approximate approach was first introduced to avoid the technical challenges before the time-varying weights were allowed in the Cox models by SAS 9.1 in 2004 (Hernán et al., 2000). Xiao et al. (2010) conducted simulations illustrating that the Cox death model yielded lower standard deviations of the treatment effect estimators than the pooled logistic regression approximation, less biased estimates in scenarios with more frequent events, and more accurate estimates for the indirect treatment effect. In summary, all evidence shows that a weighted Cox model for death should be preferred over the approximate approach.

A question remains as to what the best model would be for estimating censoring probabilities. Much methodological research has been done utilizing a Cox censoring model to obtain the IPCW weights (Robins and Finkelstein, 2000; Ghosh and Lin, 2002; Schaubel et al., 2009; Zhang and Schaubel, 2011), referred as 'Cox IPCW'; on the other hand, those who use a weighted pooled logistic death model opt for a pooled logistic censoring model to be consistent. Because the logistic censoring model allows for time-dependent intercepts and increases the number of parameters dramatically, the reduction of free parameters by combining intercepts is necessary to ensure model stability (Hernán et al., 2000).

Despite the obvious flexibility and wide applicability, the Cox IPCW approach has not been widely adopted among practitioners. Given that the Cox model features a non-parametric baseline hazard, it remains flexible while accounting for covariate effects through a parametric link. It is crucial but may be not obvious that the use of a Cox censoring model requires expanding the original analysis file into a larger dataset. Cox censoring model and Cox death model can both be accomplished using the standard Cox regression software. As far as we know, there is no report in the literature that describes how to implement this method. We also discuss a piece-wise exponential censoring model (PWE IPCW) as an alternative approach to Cox IPCW, and its advantage in reducing computation time when dependent censoring is heavy.

In this paper, we aim first to provide an explicit road map for using the Cox death model with Cox IPCW weights, complete with details of implementation. Then we consider PWE models to fit dependent censoring. Simulation demonstrates that PWE IPCW approach maintains most of the flexibility that the Cox IPCW offers. Guidance on PWE censoring model fitting is provided. Finally, we conduct

a case study based on liver waitlist mortality using data obtained from a national organ transplant registry to demonstrate the use of approaches and tricks mentioned throughout the paper.

4.2 Cox IPCW Approach

4.2.1 Notation

Denote D_i as time to death and C_i as time to dependent censoring for subject i. Let X_i be the observed time of death or censoring whichever occurs first, $X_i = \min(D_i, C_i)$. Let V_i be a set of baseline covariates and $Z_i(x), 0 \leq x \leq X_i$, be the set of time-dependent covariates. Let $Z_i \equiv Z_i(0)$ represents the baseline values of the time dependent covariates. Assume we have a population model for time to mortality with a hazard function,

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

In medical settings, it is very common that preventive treatments are prioritized to the patients with high risk of mortality. In that case, the patient censored is highly likely to die in the near future had the treatment has not been given. Ignoring the dependent censoring may introduce bias in estimating the failure hazard model.

4.2.2 Method

As shown by Robins and Finkelstein (2000), the IPCW approach correct for bias caused by dependent censoring that is attributable to a set of time dependent covariates $Z(x)$. The assumption underlying this approach is that the hazards of censoring at time x do not further depend on possibly unobserved death time D_i, or

$$\lambda^C(x|\bar{Z}_i(D_i), D_i, D_i > x) = \lambda^C(x|\bar{Z}_i(x), D_i > x), \tag{4.1}$$

where $\bar{Z}_i(x) = \{Z_i(s), 0 < s \leq x\}$, $\lambda^C(x|A) = \lim_{\Delta \to 0} P\{C_i \in [x, x+\Delta) | A, C_i \geq x\}/\Delta$ and C_i represents the censoring time. In fact, equation (4.1) says that given the true death time $D_i > x$ and the time-dependent covariates $Z_i(s)$ up to time D_i, the censoring rate depends only on the $\bar{Z}_i(x)$ and the fact that $D_i > x$, or that for an individual uncensored, the censoring rate at time x given the past and the covariates is unaffected by the future. This is referred to as the condition of 'no unmeasured counfounders' by Robins and Finkelstein (2000).

Assume that (4.1) holds, and that a Cox model holds for the time until censoring (transplant) with hazard function

$$\lambda^C(x|\bar{Z}_i(x), V_i, D_i > x) = \lambda_0^C(x) \exp\{\gamma^C Z_i(x) + \beta^C V_i\}, \tag{4.2}$$

where $\lambda_0^C(x)$ is an unspecified baseline hazard function and V_i is a set of baseline covariates. For notational convenience, we assume that the censoring rate at time x depends only on the most recent value $Z(x)$. Alternatively, one can build a stratified model to allow different baseline hazards across stratum: $\lambda_m^C(x|Z_i(x), V_i, D_i, D_i > x, m) = \lambda_{0m}^C(x) \exp\{\gamma^C Z_i(x) + \beta^C V_i\}$, where $\lambda_{0m}^C(x)$ is unspecified stratum-specific baseline hazard function. The basic unstabilized weights are defined as $w_i(x) = 1/K_i^V(x)$, where $K_i^V(x)$ represents the conditional probability of not receiving a transplant until time x for subject i where survival time exceeds x. That is,

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

where $\Lambda_i^C(x) = \int_0^x \exp\{\gamma^C Z_i(s) + \beta^C V_i\} d\Lambda_0^C(s)$.

Research (e.g. Robins et al. (2000)) has shown that in order to reduce the variation in the weights caused by baseline heterogeneity, stabilized weights should be considered. Robins and Finkelstein (2000) and Hernán et al. (2008) use $\hat{w}_i(x) = \hat{K}_i^0(x)/\hat{K}_i^V(x)$ with stabilizer $K_i^0(x)$ estimated by refitting the same model in (4.2)

with $Z_i(t)$ replaced with $Z_i(0)$. We use the same stabilizer, although other choices could be considered.

4.2.3 Software Implementation

In this section we describe a step-by-step procedure of technical implementation based on an example dataset. SAS code can be found in the Appendix. One can follow the same steps and implement the approach in R or other statistical software. Time-dependent covariate 'Z' is recorded in consecutive time intervals (t_1, t_2) for each subject indexed by 'id'. For convenience, it is assumed that the time-dependent Z remains constant in each time period and jumps to the next value when the period ends. Fixed baseline variables, 'age' and 'male', have the same value over time for each subject. 'Death' is the event of interest and 'transplant' is considered as dependent censoring. Table 4.1 presents a subset of the original dataset as an example.

To obtain the estimate of $K^V(x)$ for the denominator of stabilized weights, we fit model (4.2) with time-dependent Z, baseline covariates age and male, and transplant as the event using PHREG. We then output the linear predictor XBeta and cumulative baseline hazards using OUTPUT and BASELINE statements, and get incremental baseline hazards using the LAG function. If appropriate, a stratified censoring model can be fitted here with a STRATA statement, in which case the baseline hazards are recorded by stratum.

Now the critical step is to expand the dataset to all unique transplant times for each subject. This is essential to obtain the correct weights using a Cox IPCW approach. When a transplant occurs, the baseline hazards and the cumulative hazards jump, causing the weights to change. Viewed in this way, the weights are step functions and only change at the transplant times. Table 4.2 presents the expanded

Table 4.1: An example dataset.

id	t_1	t_2	Z	age	male	death	transplant
1	0	4	1	59	1	0	1
2	0	3	4	60	0	0	0
2	3	10	8	60	0	1	0
3	0	1	2	55	1	0	0
3	1	5	8	55	1	0	1
4	0	2	2	57	0	0	0
4	2	4	6	57	0	1	0

Table 4.2: The expanded dataset to cover all censoring times.

id	t_1	t_2	Z	age	male	death	transplant
1	0	4	1	59	1	0	1
2	0	3	4	60	0	0	0
2	3	4	8	60	0	0	0
2	4	5	8	60	0	0	0
2	5	10	8	60	0	1	0
3	0	1	2	55	1	0	0
3	1	4	8	55	1	0	0
3	4	5	8	55	1	0	1
4	0	2	2	57	0	0	0
4	2	4	6	57	0	1	0

Table 4.3: The contracted dataset to only include death times.

id	t_1	t_2	Z	age	male	death	transplant
1	0	4	1	59	1	0	1
2	3	4	8	60	0	0	0
2	5	10	8	60	0	1	0
3	1	4	8	55	1	0	0
4	2	4	6	57	0	1	0

dataset based on the original set in Table 4.1. If a stratified censoring model is assumed, one needs to expand the dataset to unique times of censoring within each stratum, as compared to an unstratified model with the expansion to all censoring times. As a result, a stratified censoring model tends to increase computational efficiency. For example, for the national dataset used in the case study, with 42,000+ patients and 20 records for each patient on average, a stratified censoring model based on 50 strata takes about 30 minutes while the unstratified censoring model

takes over 3.5 hours. Stratification is recommended for Cox IPCW censoring models when reasonable. Now, merging the newly expanded set with the data containing the baseline hazard increments and all covariates, we can obtain the estimates of the cumulative hazards via multiplying the cumulative baseline hazards by the exponetiated value of linear predictor Xbeta, or $\exp(\gamma^C Z_i(x) + \beta^C V_i)$. This yields the denominator of the stabilized weights.

We then refit the model with $Z_i(0)$ in replace of $Z_i(t)$ over time for all subjects to obtain $\hat{K}_i^0(x)$. Similar modeling and data manipulation steps as those for $\hat{K}_i^V(x)$ can be done.

Now we fit a Cox model for death using PHREG with a WEIGHT option to include the IPCW weights. Note that, the Cox model only takes the records at death times into account. To increase computational efficiency without altering the results, we use the subset with time periods that include death times. This technique results in a more significant time reduction for larger datasets and for datasets with heavy dependent censoring.

In summary, given the nonparametric nature of the baseline in Cox censoring model, the expansion of the dataset for weight calculations is essential. This results in substantial computational burden, especially when the dataset is large with many distinct censoring times or in the simulation studies where many iterations need to be carried out.

4.3 PWE IPCW Approach

4.3.1 Background

Obtaining IPCW weights via a Cox censoring model requires the data expansion to cover all unique censoring times in order to calculate weights correctly. A

parametric time-to-event model, in contrast, does not require such expansion. In this section, we explore a piece-wise exponential (PWE) censoring model for weights calculation (PWE IPCW) as an alternative.

Assume that a PWE model holds for the time until censoring (transplant) with hazard function for the kth interval

$$\lambda^C(x|Z_i(x), V_i, D_i > x, t_{k-1} < x \leq t_k) = \lambda_k^C \exp\{\gamma^C Z_i(x) + \beta^C V_i\}, \tag{4.3}$$

where λ_k^C represents the constant baseline hazard for the kth piece and $t_0 = 0$. This model requires pre-specification of the pieces or the cut-off knots. Along with γ^C and β^C, all λ_k^Cs need to be estimated. Therefore, $\hat{\Lambda}^C(x|\bar{Z}_i(x), D_i > x, V_i)$ increases linearly until $Z_i(x)$ or the hazard piece it lies in, $\hat{\lambda}_k^C$, changes, when it switches to a new slope to continue its accumulation. As we mentioned before, the Cox model for death only takes the records at death times into account; It is necessary to expand the dataset to all unique death times for the exact weights.

The difference in implementation between Cox IPCW and PWE IPCW gives each method computational advantage under different scenarios. In a dataset with many deaths but few censoring, Cox IPCW may run faster; on the other hand, in the settings with many censoring times and fewer deaths, the PWE IPCW approach tends to be computationally more efficient. In medical settings where preventive approaches are often used, the latter scenario with more censoring and fewer deaths is more common. Note that if the censoring percentage is quite low (e.g. 10%), then IPCW would usually not be required at all.

We propose the use of PWE model as an alternative to estimate censoring and IPCW weights, given its flexibility, ease of implementation, and potential gain of computational efficiency. Note that model (4.3) may be chosen for C_i either be-

cause it is believed to be the correct model, or because it is intended to be a close approximation to the (true) model given by (4.2).

4.3.2 Choice of Location and Number of Knots

Not knowing the actual shape of the baseline hazards, it may be challenging to determine appropriate number of pieces and locations of the cutoffs or knots for a PWE model.

Generally, two ways in determining the knots are commonly used. First and ideally, we allocate knots based on previous knowledge or theory. For example, following a heart transplantation, a patient faces an increasing hazard of death over the first ten days, while the body adapts to the new organ. The hazard then decreases with time as the patient recovers. In this case, we want to allocate more knots in the beginning and fewer towards the long term, to capture the main trends in the rate function. Another example, if it is known that transplants (dependent censoring) on the liver waitlist occur more frequently in the early stage of follow-up; as a consequence, making finer intervals at earlier follow-up times is recommended. Alternatively, without enough knowledge of the rate function, we suggest choosing knots based on the cumulative hazards for censoring estimated without covariates, or group the censoring events in equal number as pieces. Such a strategy helps to ensure sufficient data within each interval.

In the next section, we evaluate PWE IPCW method using pieces that are either equally spaced (on follow-up time) or that have equal number of events. Lawless and Zhan (1998) suggested that it is satisfactory to use piece-wise constant intensities with 4-10 pieces in most practical situations. Liu et al. (2012) conducted comprehensive simulation and recommended to include at least 6 pieces in the assumed

baseline rate function for a recurrent event model. We evaluate 4 and 6 pieces in our simulations with these recommendations as guidelines.

4.4 Simulation

In this section, we examine several common hazard distributions for censoring models, with the aim to compare the performance between PWE and Cox IPCW and to provide general guidance on choosing the pieces for PWE models.

We simulate samples with N subjects. For each subject i, we assume a baseline treatment covariate V_i $\sim$Bernoulli(p) and a time dependent covariate $Z_i(x)$ that follows a Poisson process on the follow-up time scale x, with rate depending on V_i and a predetermined baseline rate μ, $Z_i(x) \sim \text{PP}(\mu e^{\gamma^D V_i})$. Here γ^D is the coefficients of baseline covariates in the weighted Cox death model. Details are given in the following paragraph. Each subject is followed for five years and is censored at the end of fifth year if they have not experienced either a failure or a dependent censoring.

We model the censoring and mortality rates as $\lambda_i^C(x|V_i, Z_i(x)) = \lambda_0^C \exp\{\gamma^C V_i + \beta^C Z_i(x)\}$ and $\lambda_i^D(x|V_i, Z_i(x)) = \lambda_0^D \exp(\gamma^D V_i) + \beta^D Z_i(x)$. An additive form for the mortality model is chosen to ensure that its marginal form is multiplicative (see Appendix), $\lambda_i^D(x|V_i) = [\lambda_0^D - \mu(e^{-\beta^D x} - 1)]e^{\gamma^D V_i}$. It is advantagous to create a marginal mortality model in a proportional hazards format, so that Cox model can be used to estimate the mortality. The correlation between the censoring and mortality is mostly determined by the $Z_i(x)$ process. We use a Spearman rank correlation coefficient to measure the correlation between the death time and censoring time. In practice, however, we observe only one event among death, dependent censoring and independent censoring whichever occurs first.

We consider four scenarios with different shapes of baseline hazards: (I) constant

Table 4.4: Comparison among 4 baseline hazards, with censoring at ∼40%.

		Baseline Hazard							
		Constant		Piecewise unimodal		Weibull monotone $\lambda(t) = \alpha\gamma t^{\gamma-1}$			
		(I)$\lambda = 0.1$		(II)$\lambda = 0.06 \sim 0.14$		(III)$\alpha = 0.2$, $\gamma = 0.5$		(IV)$\alpha = 0.05$, $\gamma = 1.5$	
#	Weights	Est (bias)	SD	Est (bias)	SD	Est (bias)	SD	Est (bias)	SD
(1)	Cox	0.693 (0.001)	0.145	0.689 (-0.003)	0.154	0.675 (-0.017)	0.146	0.691 (-0.001)	0.134
(2)	PWE4no	0.694 (0.002)	0.143	0.689 (-0.003)	0.149	0.675 (-0.017)	0.144	0.692 (-0.000)	0.135
(3)	PWE4tm	0.691 (-0.001)	0.148	0.686 (-0.006)	0.157	0.667 (-0.025)	0.146	0.690 (-0.002)	0.143
(4)	PWE6no	0.694 (0.002)	0.141	0.689 (-0.003)	0.147	0.675 (-0.017)	0.147	0.691 (-0.001)	0.134
(5)	PWE6tm	0.695 (0.003)	0.146	0.688 (-0.004)	0.152	0.671 (-0.021)	0.144	0.691 (-0.001)	0.140

hazards; (II) unimodal piece-wise constant hazards equally spaced on the follow-up time; (III) Weibull monotone decreasing hazards $\lambda(t) = \alpha\gamma t^{\gamma-1}$ with $\gamma = 0.5$ and (IV) Weibull monotone increasing hazards $\lambda(t) = \alpha\gamma t^{\gamma-1}$ with $\gamma = 1.5$.

We set parameters as $p = 0.5$, $\mu = 3$, $\gamma^D = \log(2)$, $\lambda^D = 0.3$, $\gamma^C = \log(2)$, $\beta^D = 0.12$ and $\beta^C = \log(1.5)$. The parameters for baseline hazards are: (I) $\lambda = 0.1$; (II) $\lambda_1 = 0.08$, $\lambda_2 = 0.1$, $\lambda_3 = 0.12$, $\lambda_4 = 0.14$, $\lambda_5 = 0.1$, and $\lambda_6 = 0.06$; (III) $\alpha = 0.2$ and (IV) $\alpha = 0.05$. For these settings, censoring rates are all around 40% (38%-44%), and Spearman correlations between censoring time and death time are all approximately 0.2 (0.20-0.22).

Under each scenario, we evaluate parameter estimates of the mortality model for death using the following weights: (1) Cox IPCW weights; (2) PWE IPCW weights, 4 pieces with the equal number of censoring events; (3) PWE IPCW weights, 4 pieces with the equal intervals in the follow-up time; (4) PWE IPCW weights, 6 pieces with the equal number of censoring events, (5) PWE IPCW weights, 6 pieces with the equal intervals in the follow-up time. The results in Table 4.4 are based on 500 repetitions and a sample size 500.

All examples in the paper are carried out using SAS 9.3 (TS1M0) on a X64_7PRO platform (Windows) with dual CPU (Intel® Xeon® Processor X5570 @ 2.93 GHz) and 3GB RAM memory.

Table 4.5: Comparison among 4 baseline hazards, with censoring at ∼60%.

		Baseline Hazard							
		Constant		Piecewise unimodal		Weibull monotone $\lambda(t) = \alpha\gamma t^{\gamma-1}$			
		(I)$\lambda = 0.2$		(II)$\lambda = 0.16 \sim 0.24$		(III)$\alpha = 0.3$, $\gamma = 0.5$		(IV)$\alpha = 0.15$, $\gamma = 1.5$	
#	Weights	Est (bias)	SD	Est (bias)	SD	Est (bias)	SD	Est (bias)	SD
(1)	Cox	0.673 (-0.019)	0.201	0.670 (-0.022)	0.193	0.679 (-0.013)	0.196	0.672 (-0.020)	0.183
(2)	PWE4no	0.671 (-0.021)	0.199	0.674 (-0.018)	0.196	0.680 (-0.012)	0.193	0.671 (-0.021)	0.186
(3)	PWE4tm	0.663 (-0.029)	0.217	0.661 (-0.031)	0.208	0.665 (-0.027)	0.206	0.660 (-0.032)	0.201
(4)	PWE6no	0.670 (-0.022)	0.196	0.673 (-0.019)	0.193	0.679 (-0.013)	0.193	0.670 (-0.022)	0.182
(5)	PWE6tm	0.665 (-0.027)	0.212	0.667 (-0.025)	0.204	0.673 (-0.019)	0.198	0.665 (-0.027)	0.196

Table 4.4 shows that Cox IPCW and PWE IPCWs perform similarly in terms of both accuracy and efficiency (bias and standard deviation). Particularly, PWE with equal number of censoring events, Approach (2) and (4), gives results very close to the Cox IPCW. In addition, PWE with equal number of censoring events exhibits a small but consistent accuracy and efficiency gain comparing to PWE with equal time distance, given its smaller biases and smaller standard deviations. PWE with equal number of censoring events, therefore, is recommended. 4 or 6 pieces do not differentiate much in results. In most cases, PWE with 4 pieces is recommended, unless the follow-up period is long or the shape of baseline hazards is expected to be complicated, in which case, more pieces for PWE approach should be explored.

Now we increase the dependent censoring rates to 60%, and further compare the performance of these methods under different scenarios. We set parameters $\beta^D = 0.15$ and $\beta^C = \log(1.8)$, and the rest are the same as the ones used above. The four baseline hazards functions have parameters: (I) $\lambda = 0.2$; (II) $\lambda_1 = 0.18$, $\lambda_2 = 0.2$, $\lambda_3 = 0.22$, $\lambda_4 = 0.24$, $\lambda_5 = 0.2$, and $\lambda_6 = 0.16$; (III) $\alpha = 0.3$; and (IV) $\alpha = 0.15$, respectively. These values are 0.1 more than the values used in the previous setting. Now dependent censoring rates are around 60% (59%-61%), and Spearman correlations between censoring time and death time are approximately 0.2 (0.20-0.22).

Table 4.6: Average computation time of IPCW procedure (in seconds).

	Censoring rate 40%				Censoring rate 60%			
Method	(I)	(II)	(III)	(IV)	(I)	(II)	(III)	(IV)
Cox	24.3	23.2	20.6	20.1	15.0	15.7	16.3	15.1
PWE4no	13.6	13.9	13.7	13.2	5.8	6.3	6.0	6.1
PWE6no	15.0	16.2	10.5	14.6	6.6	6.9	6.7	6.3

Table 4.5 gives results similar to those in Table 4.4. Even with heavy dependent censoring around 60%, Cox IPCW and PWE IPCW approaches still perform well and give the estimates with bias smaller than 0.02 (or 3%). Similar as before, PWE with equal number of censoring events perform better than PWE with equal time distance, with smaller bias and standard deviation.

Our motivation to explore PWE IPCW as an alternative to Cox comes from the computational gain of the former approach when a large portion of the observations are dependently censored. We now compare the computational time of Cox IPCW and PWE IPCW with equal number of censoring events.

Table 4.6 shows the average computation time of estimating IPCW weights using these approaches, based on 10 runs and sample size $N = 1000$. The results show that PWE saves 50-70% of computation time as compared to Cox IPCW when dependent censoring rate is 40-60%. This time saving is important especially in a large dataset with heavy censoring. For example, in the case study dataset with 42,000 patients and 20 records per patient on average, PWE censoring model takes 2 minutes while Cox censoring model takes over 7 hours. A stratified model would reduce the difference to about 30 minutes, with increased computation time for PWE and decreased time for Cox censoring model.

4.5 Case Study

4.5.1 Data Description

As an example, we evaluate liver transplant waitlist mortality using the data obtained from the Scientific Registry of Transplant Recipients (SRTR), and compare the Cox IPCW and 4-piece PWE IPCW with equal number of censoring events approach. We consider a five-year cohort of patients waitlisted between January 1st, 2004 and December 31st, 2008. Patients recorded as Status 1 or 1A at baseline have acute liver failure at waitlisting and are not included in the analysis. In addition, we exclude patients listed in error, changed to kidney/pancreas transplants or with previous liver transplant history. Given that pediatric patients follows a different scheme of transplant, we only include adults of age 18 years above in the analysis. Among the 125 transplant centers, 25 centers with fewer than 10 patients listed per year are excluded. Similarly, among the 52 Organ Procurement Organizations (OPO), 2 with fewer than 10 patients listed per year are excluded. In the final working set, 42,021 patients from 100 centers and 50 OPOs are included.

We measure baseline covariates gender, race, age, diagnosis categories, diabetes, previous malignancy indicator, Body Mass Index (BMI), blood type, and hospitalization and Intensive Care Unit (ICU) status. All these covariates are included in both the transplant (censoring) model and the mortality model. In addition, 11 regions coded as binary indicators for each subject are also included.

Time dependent variables consist of the Model for End-Stage Liver Disease (MELD) score, inactive period, and sodium value. MELD is the scoring system used to prioritize patient on the liver wait list. It combines serum bilirubin, serum creatinine and the international normalized ratio for prothrombin time (INR). Allocation MELD score is used in practice as the main determining factor on liver deceased donor allo-

cation. We record MELD as 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 1/1A. Assuming that a patient is being monitored sufficiently by the clinician, it would appear reasonable to believe that lack of a MELD update reflects the fact that the patient's MELD has not changed. This would imply that coding MELD score as a step-function (i.e. last-value carried-forward) would be appropriate. Sometimes, patients are temporarily removed from the waitlist for various reasons such as medical condition, refusing transplant, improved or deteriorated condition, or being inactive on the program for more than 2 years. Such patients are not supposed to receive offers of deceased donor livers when they are removed from or not on the waitlist. We set the inactive indicator as 1 to identify the period of removal and to capture this information in modeling. Sodium value is recorded as a continuous variable and is also included in the set of time dependent covariates. Alternative approaches of handling inactive time were used by Zhang and Schaubel (2011).

Patients in need of a liver donor are often encouraged to be listed at multiple centers. In the five-year cohort, we have 39,680 patients listed at a single center, while 2,341 patients listed at multiple centers. Among the ones listed at multiple centers, 2,221 are listed at two centers, 111 are listed at three centers, and 9 are listed at four centers. When a patient is listed at multiple centers simultaneously, the first listing center is considered as his/her primary center; unless he/she is transferred to another center as an independent censoring, the new center would be appointed as the primary center and the patient is then treated as two people. We track all listing records for each patient and define dependent censoring if a patient receives a deceased donor transplant at any center.

Death on the waitlist is the event of interest in our analysis. 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 of follow-up or received a living donor transplant, which typically is not predicted by MELD score. We conduct a Cox death model stratified on center with Cox IPCW weights and PWE IPCW weights. In this case, we expect the baseline hazards decrease over time because patients are most likely getting transplants towards the beginning of waitlist period whenever qualified. We use piece-wised exponential censoring model with 4 pieces of equal number of transplants, as comparison to the Cox IPCW approach.

It is worth noting a computational trick for the Cox death model on any large dataset. Cox model only takes into the considerations the time points when a event or failure occurs. Therefore, a compressed dataset containing only records with death times yields the same results yet reduces computational time. This reduction in time is particularly significant in large datasets.

4.5.2 Results

Table 4.7 presents the parameter estimates of censoring model and death model using Cox IPCW and PWE IPCW in 4 pieces with equal number of events. The results of Cox censoring model are very close to those using PWE model. However, while Cox censoring model takes approximately 5 hours in computation, the PWE approach, on the other hand, only takes less than 4 minutes. The time reduction comes from two sources. First, although PHREG can conveniently deal with time dependent covariates, it increases computation time dramatically in large datasets. On the other hand, we implement PWE models in LIFEREG treating each record

as a piece of duration utilizing the memoryless property of exponential models. This reduced the computational time greatly as compared to PHREG procedure. Second, our dataset has approximately 50% of dependent censoring, which increases the burden in data expansion stage especially for Cox IPCW method where it is necessary to expand to all unique transplant times for all patients.

Both Cox and PWE censoring models give some extreme weights. Cox identified 31 subjects with maximum weights larger than 100; while PWE identifies 28. The same 28 patients have been suggested with extreme weights by both methods. PWE is slightly more stable in estimating weights for patients being in the at-risk set for a longer time with heavy tails. These 28 patients either have been in the at-risk set for longer than 2 years with medium MELD score or have been in the at-risk set for some time at a high MELD score 30+. Since 99.5% of subjects have maximum weights under 10, we use 10 as a cap for both Cox weights and PWE weights in the death model calculation.

Table 4.7 also shows that the death models using Cox IPCW weights and PWE IPCW weights yield very similar results. The expanded dataset of Cox IPCW is 7 times larger than that of PWE IPCW approach. With the computation trick to compress dataset to only include death records, the death models with both weights take the similar amount of time (14-16 minutes) in computation. As we expected, the higher the MELD score becomes, the more likely the patient would get a transplant and the more likely he or she would die. Since there is no patient with Status 1/1A at baseline, the parameter estimate for that covariate is 0.

Table 4.7: Censoring model and death model using Cox IPCW and PWE4 IPCW

	Censoring model						Death model					
	PWE4			Cox			PWE4 wts 10			Cox wts 10		
	Est	StErr	P-val.	Est	StErr	P-val.	Est	StErr	P-val.	Est	StErr	P-val.
Gender: Male	0.14	0.02	<0.01	0.14	0.02	<0.01	-0.02	0.03	0.44	-0.02	0.03	0.45
Race: White (ref)												
Black	-0.16	0.02	<0.01	-0.16	0.02	<0.01	0.03	0.06	0.57	0.02	0.06	0.66
Hispanic	-0.13	0.02	<0.01	-0.13	0.02	<0.01	-0.06	0.04	0.14	-0.06	0.04	0.12
Asian	-0.20	0.03	<0.01	-0.20	0.03	<0.01	-0.32	0.07	<0.01	-0.32	0.08	<0.01
Primary diagnosis: nonchron/cirr (ref)												
Chronic liver disease	0.07	0.03	0.01	0.07	0.03	0.01	0.04	0.06	0.48	0.04	0.06	0.54
Malignant neoplasm	0.07	0.03	0.01	0.06	0.03	0.03	0.57	0.09	<0.01	0.56	0.09	<0.01
Metastatic disease	0.06	0.05	0.22	0.06	0.05	0.22	0.25	0.11	0.02	0.24	0.11	0.03
HCV[1]	0.02	0.02	0.19	0.02	0.02	0.16	0.27	0.03	<0.01	0.27	0.03	<0.01
Etiology unknown	0.04	0.06	0.42	0.07	0.06	0.19	-0.02	0.11	0.88	-0.02	0.11	0.87
Other	-0.20	0.03	<0.01	-0.20	0.03	<0.01	0.09	0.06	0.16	0.09	0.06	0.14
Diabetes	-0.01	0.02	0.74	-0.01	0.02	0.73	0.12	0.03	<0.01	0.12	0.03	<0.01
Diabetes missing	-0.12	0.05	0.01	-0.12	0.05	0.01	0.01	0.08	0.91	0.01	0.08	0.89
Previous malignancy	-0.01	0.03	0.84	-0.01	0.03	0.62	-0.07	0.07	0.33	-0.09	0.07	0.24
BMI: 30 to 35 (ref)												
0 to 25	0.03	0.02	0.10	0.03	0.02	0.13	0.08	0.04	0.04	0.09	0.04	0.03
25 to 30	0.05	0.02	<0.01	0.05	0.02	<0.01	-0.06	0.04	0.11	-0.05	0.04	0.16
35+	-0.01	0.02	0.55	-0.01	0.02	0.53	0.08	0.04	0.08	0.08	0.04	0.07
Blood type: O (ref)												
A	0.11	0.02	<0.01	0.11	0.02	<0.01	0.06	0.03	0.05	0.06	0.03	0.03
B	0.35	0.02	<0.01	0.35	0.02	<0.01	-0.05	0.05	0.30	-0.05	0.05	0.31
AB	1.24	0.03	<0.01	1.24	0.03	<0.01	0.04	0.11	0.70	0.04	0.11	0.69
Hospitalization: Other conditions (ref)												
ICU [2]	0.00	0.04	0.94	0.03	0.04	0.48	1.28	0.08	<0.01	1.29	0.08	<0.01
not ICU	0.05	0.02	0.03	0.05	0.02	0.03	0.64	0.05	<0.01	0.65	0.05	<0.01
Age: 18 to 29 (ref)												
30 to 39	0.15	0.05	0.01	0.15	0.05	0.01	0.18	0.17	0.30	0.16	0.18	0.37
40 to 49	0.17	0.05	<0.01	0.17	0.05	<0.01	0.46	0.15	<0.01	0.43	0.15	0.01
50 to 54	0.16	0.05	<0.01	0.16	0.05	<0.01	0.71	0.15	<0.01	0.69	0.16	<0.01
55 to 59	0.16	0.05	<0.01	0.15	0.05	<0.01	0.75	0.15	<0.01	0.72	0.15	<0.01
60 to 64	0.20	0.05	<0.01	0.19	0.05	<0.01	0.90	0.16	<0.01	0.88	0.16	<0.01
65 to 69	0.17	0.05	<0.01	0.16	0.05	<0.01	1.08	0.16	<0.01	1.05	0.16	<0.01
70+	0.27	0.06	<0.01	0.25	0.06	<0.01	1.43	0.17	<0.01	1.41	0.17	<0.01
MELD score: 15 to 17(ref)												
6 to 8	-2.08	0.06	<0.01	-2.04	0.06	<0.01	-0.87	0.07	<0.01	-0.87	0.07	<0.01
9 to 11	-2.14	0.05	<0.01	-2.11	0.05	<0.01	-0.63	0.05	<0.01	-0.63	0.05	<0.01
12 to 14	-1.56	0.04	<0.01	-1.55	0.04	<0.01	-0.32	0.04	<0.01	-0.32	0.04	<0.01
18 to 20	0.92	0.03	<0.01	0.91	0.03	<0.01	0.41	0.05	<0.01	0.41	0.05	<0.01
21 to 23	1.70	0.03	<0.01	1.68	0.03	<0.01	0.55	0.06	<0.01	0.55	0.06	<0.01
24 to 26	2.37	0.03	<0.01	2.36	0.03	<0.01	1.06	0.08	<0.01	1.06	0.08	<0.01
27 to 29	2.97	0.03	<0.01	2.94	0.03	<0.01	1.83	0.09	<0.01	1.82	0.09	<0.01
30 to 32	3.44	0.04	<0.01	3.43	0.04	<0.01	2.19	0.11	<0.01	2.16	0.11	<0.01
33 to 35	3.68	0.04	<0.01	3.67	0.04	<0.01	2.51	0.10	<0.01	2.50	0.11	<0.01
36 to 39	3.77	0.04	<0.01	3.77	0.04	<0.01	2.81	0.11	<0.01	2.77	0.11	<0.01
40	3.88	0.04	<0.01	3.86	0.04	<0.01	3.43	0.10	<0.01	3.39	0.10	<0.01
Status 1/1A	4.91	0.11	<0.01	4.79	0.11	<0.01	0.00	-	-	0.00	-	-
Inactive	-2.30	0.07	<0.01	-2.23	0.07	<0.01	0.85	0.08	<0.01	0.85	0.08	<0.01
Serum sodium 138+ (ref)												
131-	0.10	0.02	<0.01	0.10	0.02	<0.01	0.83	0.05	<0.01	0.84	0.05	<0.01
132 to 137	0.05	0.02	<0.01	0.04	0.02	0.01	0.29	0.03	<0.01	0.29	0.03	<0.01
missing	-0.07	0.02	<0.01	-0.08	0.02	<0.01	0.21	0.04	<0.01	0.21	0.04	<0.01

1. HCV= hepatitis C.
2. ICU= intensive care unit.

4.6 Discussion

In the presence of dependent censoring, IPCW techniques can be useful to recover the true death hazards and are easy to implement. We detailed software implementation procedure, and describe techniques motivated by time reduction using stratification and compression of the final dataset for the death model. When the dataset is large or the dependent censoring portion is large, it is advantageous of using a PWE IPCW over Cox IPCW demonstrated via simulation and case study. In some extreme cases, PHREG may not run given the computer memory limitation, a PWE model can be used as an alternative.

We discussed some general guidance on how to choose pieces and knots for PWE censoring model.

CHAPTER V

Future Work

In the thesis, we first considered a risk-adjusted O-E CUSUM chart along with monitoring bands as decision criterion, to monitor the post-transplant mortality in transplant programs. The O-E CUSUM is easily plotted and its trends are easily interpreted; further, when the monitoring bands are included, it provides simple rules for flagging. In practice, a head-start technique can be used to provide a quick and sensitive detection after a signal, to ensure the problem causing the signal has been addressed properly. Further work can be done to delineate the average run length (ARL) among programs in the case study, incorporating a head-start after signal. The CUSUM described in this chapter, however, does not allow for dependent censoring, which is common especially in the medical setting. This motivated our work in the third chapter, where we developed a weighted CUSUM to account for the dependent censoring.

The construction of a weighted CUSUM based on the IPCW weights requires assumptions of accurate information, no unmeasured confounding, and correctness of the model. Given these assumptions, weighted O-E under null hypothesis is a zero-mean process with inflated variance. We derived the theoretical formula for the variance of this zero-mean process, which can be used for any point evaluation of

cumulative weighted O-E. For example, a score statistic can be constructed at any time point using accumulated information, to evaluate the null hypothesis based on a pre-determined Type I error rate. In our case, we are interested in the sequential usage of the accumulating information, for the purpose of providing timely feedback to centers. We discussed a resampling technique to obtain the control limit for a center at a given size and over a certain period of time. Further investigation on sampling can be done, such as comparing the results using sampling technique in the independent censoring scenario to the approach discussed in Chapter II. In addition, other ways of calibrating control limits, perhaps incorporating the weighted expected values and the variance of the weighted zero-mean process, are also of interest for future research.

In Chapter IV, we focused on the technical implementation of the Cox IPCW approach and compared the PWE approach with the Cox IPCW approach. While our censoring mechanism determines the simplicity or the monotone shape of the baseline hazard curve for the censoring model, PWE approach works well in giving similar results to the Cox IPCW model, while reducing computational time greatly. In practice, especially when the censoring mechanism is not as clearly defined as our case, the researcher needs to be careful in choosing the appropriate knots for the PWE approach. Much research has been done in this area. One may want to choose a simple and intuitive way to choose knots based on previous knowledge, if speed is of concern. Further investigation can be done in this area as well.

APPENDICES

APPENDIX A

Proof of Theorem in Chapter II

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'.

APPENDIX B

Cox model for death in Chapter III

We can estimate the true hazard in the mortality model using estimated inverse weights. Assuming a Cox PH mortality model $\lambda^D(x) = \lambda_0^D(x)\exp(\beta Z)$, where $\beta = (\beta_1, \cdots, \beta_p)$ and Z represents the vector $(Z(0), V(0))$ measured at baseline. Again, this mortality model might be obtained from an overall model or a stratified model. For demonstration, we use an overall model in our notation. The weighted Cox partial likelihood score function is

$$\begin{aligned} U(\beta) &= \sum_{i=1}^{n} \delta_i w_i(x_i)\{Z_i - \sum_j \frac{Y_j(x_i+S_i)w_j(x_i)Z_j\exp(\beta Z_j)}{\sum_k Y_k(x_i+S_i)w_k(x_i)\exp(\beta Z_k)}\} \\ &= \sum_{i=1}^{n} \int_0^x w_i(s)\{Z_i - \sum_j \frac{Y_j(s+S_i)w_j(s)Z_j\exp(\beta Z_j)}{\sum_k Y_k(s+S_i)w_k(s)\exp(\beta Z_k)}\} dN_i^*(s+S_i) \\ &= \sum_{i=1}^{n} \int_0^x w_i(s)\{Z_i - \bar{Z}_w(s;\beta)\} dN_i^*(s+S_i) \end{aligned}$$

$E(U(\beta)) = 0$. Use estimated stabilized weights $\hat{w}_i(x)$ for $w_i(x)$ and solve $U(\hat{\beta}) = 0$. $\hat{\beta}$ is a consistent estimator of β as shown by Robins and Finkelstein (2000).

APPENDIX C

Generating dependent censoring in Chapter III

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.

Let V_i represent the covariate of interest, e.g. treatment assignment, and suppose that V_i has a Bernoulli distribution with probability of success p. Generate Z_i (e.g. MELD score) as time dependent covariate for subject i based on his treatment assignment V_i. Let Z_i be a Poisson process with intensity $\mu e^{\gamma V_i}$. The time interval between successful jumps are i.i.d and follow $\exp\{\mu e^{\gamma V_i}\}$.

Assume a transplant model

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

and a conditional mortality model

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

Based on Jewell and Kalbfleisch (1996), the marginal survivor function for mor-

tality is

$$\begin{aligned} S(x|V_i) &= E_Z\{S(x|V_i, Z_i(x))\} = E_Z\{\exp(-\int_0^x [\lambda_0^D e^{\gamma V_i} + \beta^D Z_i(u)]du)\} \\ &= \exp\{-\lambda_0^D e^{\gamma V_i} x\} E_Z\{\exp \int_0^\infty \psi(u) Z_i(u) du\} \text{ where } \psi_t(u) = -\beta^D I(u < x) \\ &= \exp\{-\lambda_0^D e^{\gamma V_i} x\} \exp\{K_Z(\psi)\} \end{aligned}$$

Therefore,

$$\lambda^D(x|V_i) = -\frac{\partial \log S(x|V_i)}{\partial x} = \lambda_0^D e^{\gamma V_i} - \frac{\partial K_Z(\psi)}{\partial x}$$

and

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

So that $\partial K/\partial x = \mu e^{\gamma V_i}(e^{-\beta^D x} - 1)$ and the marginal mortality model given V_i is

$$\begin{aligned} \lambda^D(x|V_i) &= \lambda_0^D e^{\gamma V_i} - \mu e^{\gamma V_i}(e^{-\beta^D x} - 1) \\ &= [\lambda_0^D - \mu(e^{-\beta^D x} - 1)] e^{\gamma V_i} \\ &= \lambda_0^*(x) e^{\gamma V_i}, \end{aligned}$$

which is a Cox PH model.

APPENDIX D

Simulation studies to demonstrate alternative approach to choose control limit for WCUSUM in Chapter III

In this appendix, we demonstrate that in a weighted CUSUM, control limits increases linearly with the inflated variance measuring the zero-mean process as the dependent censoring rate increases, to maintain the same ARL or type I error rate over a certain period of time.

Assume a center with 100 patients per year for 3.5 years. One-year survival is of interest. The setup of censoring model and death model is the same as the rest of the paper, with parameters $p = 0.5$, $\mu = 3$.

We set death model parameters as $\lambda_1^D = 0.05$, $\gamma_1^D = \log(1.8)$, $\beta_1^D = 0.03$ (Pattern 1) and $\lambda_2^D = 0.015$, $\gamma_2^D = \log(2)$, $\beta_2^D = 0.05$ (Pattern 2) to compare two different shapes of baseline hazards.

We set parameters for censoring as $\lambda_1^C = 0$, $\gamma_1^C = \log(1.5)$, $\beta_1^C = \log(2)$; $\lambda_2^C = 0.03$, $\gamma_2^C = \log(1.5)$, $\beta_2^C = \log(2)$; $\lambda_3^C = 0.06$, $\gamma_3^C = \log(2)$, $\beta_3^C = \log(2)$; $\lambda_4^C = 0.1$, $\gamma_4^C = \log(1.5)$, $\beta_4^C = \log(1.5)$; $\lambda_5^C = 0.2$, $\gamma_5^C = \log(1)$, $\beta_5^C = \log(1.5)$, to obtain different rates of dependent censoring. 500 iterations are conducted on each scenario.

We stop each iteration as soon as the cumulative weighted expected number of failures reaches 40. StDev represents the square root of the variance for the weighted

Table D.1: Weighted CUSUM

	Pattern 1		Pattern 2	
Censor pattern	StDev	L	StDev	L
1 (no censoring)	6.1	5.07	6.1	5.10
2	6.7	6.68	6.9	6.98
3	7.1	7.25	7.5	8.08
4	7.5	7.92	7.9	8.85
5	7.7	8.80	8.2	9.50

zero-mean process. L is calibrated so that the weighted process has 5% type I error rate over the 3.5 year period. Power under hypothesis of relative risk 2 is also illustrated.

Table D.1 shows the linear trend.

APPENDIX E

Variance of the Weighted Zero-Mean process in Chapter IV

In this section, we show that the variance of the weighted zero-mean process increases in a stable rate in equilibrium stage. Assume patients arrive in a stable process (e.g. homogeneous Poisson process), and the number of patients at risk stays constant after the first year of accumulation (if we are interested in one-year survival).

Denote $B(t) = \sum_{i=1}^{n} \int_0^t [w_i^*(u)dN_i^*(u) - w_i^*(u)Y_i(u)d\Lambda_i(u)] = \sum_i \int_0^t w_i^*(u)dM_i^*(u)$, where $E\{w_i^*(u)dM_i^*(u)\} = 0$ for subject i and n is the number of patients up to time t at the center.

If relative risk of mortality is r, or the ratio of the observed number of deaths against the expected number of deaths is r, then

$$E\{\sum_i \int_0^t w_i^*(u)dN_i^*(u)|Y_i(u), w_i^*(u), V_i, S_i, r\} = r\sum_i \int_0^t w_i^*(u)Y_i(u)d\Lambda_i(u).$$

Generalize $B(t)$ to

$$B_r(t) = \sum_{i=1}^{n} \int_0^t [w_i^*(u)dN_i^*(u) - rw_i^*(u)Y_i(u)d\Lambda_i(u)] = \sum_i \int_0^t w_i^*(u)dM_i^*(u;r),$$

with $E\{w_i^*(u)dM_i^*(u;r)\} = 0$.

Assuming the expected risk of failure, the weights over time and the death process among all patients are i.i.d, according to Central Limit Theorem, we have

$$B_r(t)/\sqrt{n} \to_D N(0, \Sigma_r(t)), \tag{E.1}$$

and

$$\bar{\Sigma}_r(t) = \frac{1}{n}\sum_i \{\int_0^t w_i^*(u)dM_i^*(u;r)\}^2 = \frac{1}{n}\sum_i J_i(t;r) \to_p \Sigma_r(t), \text{ as } n \to \infty,$$

where

$$\begin{aligned} J_i(t;r) &= [\int_0^t w_i^*(u)dN_i^*(u) - \int_0^t r w_i^*(u)Y_i(u)d\Lambda_i(u)]^2 \\ &= \int_0^t [w_i^*(u)]^2 dN_i^*(u) - 2r\int_0^t w_i^*(u)dN_i^*(u)\int_0^t w_i^*(u)Y_i(u)d\Lambda_i(u) \\ &\quad + r^2[\int_0^t w_i^*(u)Y_i(u)d\Lambda_i(u)]^2. \end{aligned}$$

Note that although the expectation calculation of $\int_0^t w_i^*(u)dN_i^*(u)\int_0^t w_i^*(u)Y_i(u)d\Lambda_i(u)$ isn't trivial, if we assume $t \to \infty$ or there is no administrative censoring, equation above has a limiting value $E\{J_i(t;r)|w_i(u), V_i, S_i, 0 < u < t\} \to c_i$.

Then, $E\{\bar{\Sigma}_r\} = E\{\sum_i J_i(t)/n\} = \sum_i c_i(r)/n \to_p \bar{c_r}$, as $t \to \infty$, where $\bar{c_r}$ describes a population average.

This concludes that the variance of the weighted zero-mean process is stable in equilibrium stage when $t \to \infty$ (one year survival is the only independent censoring source), given patients' V_i, $Z_i(t)$ and relative risk. If patients arrive in a homogeneous Poisson process with rate λ: $E\{n\} = \lambda t$, then $\text{Var}\{B_r(t)\} \to \lambda t\bar{c_r} \doteq \sigma_r^2 t$.

We empirically verify that the standardized version of $B_r(t)$ or $B_r(t)/\text{SE}_r(t)$ has an approximate standard normal distribution, where

$$\{\text{SE}_r(t)\}^2 = \sum_i \hat{\text{Var}}\{\int_0^t w_i^*(u)dM_i^*(u;r)\} = \sum_i \{\int_0^t w_i^*(u)dN_i^*(u) - \int_0^t r w_i^*(u)Y_i(u)d\Lambda_i(u)\}^2,$$

with $E\{\int_0^t w_i^*(u)dM_i^*(u;r)\} = 0$.

We are interested in one-year outcomes. Therefore, after the first year of recruitment accumulation, the process reaches an equilibrium stage in which the distribution of $B_r(\Delta t) = B_r(t_2) - B_r(t_1)$ depends only on $\Delta t = t_2 - t_1$ for $t_2 > t_1 > 1$. We choose $t_2 = 2$ and $t_1 = 1$ in averaging the distribution of $B_r(\Delta t)/\text{SE}_r(\Delta t)$.

Table E.1: Standardized $B_r(t)$ between year 1 and year 2

r	death%	cen.%	corr	$B_r(\Delta t) \pm$ SE	$\text{Var}(\Delta t) \pm$ SE	$\frac{B_r(\Delta t)}{\sqrt{\text{Var}_r(\Delta t)}} \pm$ SE
2	22.8	36.2	0.22	-0.21±8.72	79.86±15.90	-0.09±0.98
1	12.3	41.4	0.16	0.14±6.62	44.61±13.12	-0.09±1.05
0.5	6.5	44.0	0.12	-0.28±5.05	22.03±11.39	-0.28±1.19

We consider the following parameter setup: $\mu_0 = 200$, $p = 0.5$, $\mu = 5$, $\gamma^D = \log(2)$, $\lambda^D = 0.01$, $\lambda^C = 0.01$, $\gamma^C = \log(2)$, $\beta^D = 0.05$ and $\beta^C = \log(2)$. The simulation is conducted with 500 repetitions.

In this Table, under relative risk $r = 2$, $r = 1$ and $r = 0.5$, we report observed dependent censoring rate, observed death rate, correlation between latent death time and dependent censoring time, and the normality properties of $B_r(\Delta t)/\text{SE}_r(\Delta t)$ between year 1 and year 2. The results confirm that the standardized zero-mean process at time t has mean 0 and variance 1.

BIBLIOGRAPHY

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

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

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

Biswas, P. and Kalbfleisch, J. (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.

Cole, S., Hernán, M., Anastos, K., Jamieson, B., and Robins, J. (2007). Determining the effect of highly active antiretroviral therapy on changes in human immunodeficiency virus type 1 RNA viral load using a marginal structural left-censored mean model. *Am J Epidemiol* **166**, 219–227.

Cole, S., Hernán, M., Margolick, J., Cohen, M., and Robins, J. (2005). Marginal structural models for estimating the effect of highly active antiretroviral therapy initiation on CD4 cell count. *Am J Epidemiol* **162**, 471–478.

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.

D'Agostino, R., Lee, M.-L., and Belanger, A. (1990). Relation of pooled logistic regression to time dependent Cox regression analysis: The Framingham Heart study. *Statistics in Medicine* **9**, 1501–1515.

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

Ghosh, D. and Lin, D. (2002). Marginal regression models for recurrent and terminal events. *Statistica Sinica* **12**, 663–688.

Hernán, M., Alonso, A., Logan, R., Grodstein, F., Michels, K., Willett, W., J.E., M., and J.M., R. (2008). Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology* **19**, 766–79.

Hernán, M., Brumback, B., and Robins, J. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology* **11**, 561–70.

Hernán, M., Brumback, B., and Robins, J. (2002). Estimating the causal effect of zidovudine on CD4 count with a marginal structural model for repeated measures. *Statistics in Medicine* **21**, 1689–1709.

Hernán, M., Lanoy, E., Costagliola, D.,and Robins, J. (2006). Comparison of dynamic treatment regimes via inverse probability weighting. *Basic Clin Pharmacol Toxicol* **98**, 237–42.

Jewell, N. and Kalbfleisch, J. (1996). Marker processes in survival analysis. *Lifetime Data Analysis* **2**, 15–29.

Kalbfleisch, J. (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.

Lawless, J. and Zhan, M. (1998). Analysis of interval-grouped recurrent-event data using piecewise constant rate functions. *The Canadian Journal of Statistics* **26**, 549–565.

Liu, D., Schaubel, D., and Kalbfleisch, J. (2012).Computationally efficient marginal models for clustered recurrent event data. *Biometrics* **68**, 779–788.

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

Page, E. (1954). Continuousinspection schemes. *Biometrika* **41**, 100–115.

Robins, J. and Finkelstein, D. (2000). Correcting for non-compliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics* **56**, 779–788.

Robins, J., Hernán, M., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11**, 550–560.

Robins, J. and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. *AIDS Epidemiology -Methodological Issues* pages 297–331.

Schaubel, D., Wolfe, R., Sima, C., and Merion, R. (2009). Estimating the effect of a time-dependent treatment by levels of an internal time-dependent covariate: Application to the contrast between liver wait-list and post-transplant mortality. *JASA* **104**, 49–59.

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

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. 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.

Sun, R.J. and Kalbfleisch, J.(2012). A risk-adjusted O-E CUSUM with monitoring bands for monitoring medical outcome. *To appear on Biometrics*.

Wetherill, G.B.(1977). *Sampling Inspection and Quality Control*. London,Chapman & Hall, Second Edition.

Xiao, Y., Abrahamowicz, M., and Moodie, E. (2010). Accuracy of conventional and marginal structural Cox model estimators: A simulation study. *The International Journal of Biostatistics* **6**, 13.

Zhang, M. and Schaubel, D. (2011). Estimating differences in restricted mean life time using observational data subject to dependent censoring. *Biometrics* **67**, 740–9.