University of Michigan School of Public Health
The University of Michigan Department of Biostatistics Working Paper Series

Year 2003 *Paper* 16

A Varying-Coefficient Cox Model for the Effect of Age at a Marker Event on Age at Menopause

Bin Nan* Xihong Lin†
Lynda D. Lisabeth‡ Sioban D. Harlow**

*University of Michigan, bnan@umich.edu
†University of Michigan, xlin@umich.edu
‡Universtiy of Michigan, llisabet@umich.edu
**University of Michigan, harlow@umich.edu

http://biostats.bepress.com/umichbiostat/paper16

A Varying-Coefficient Cox Model for the Effect of Age at a Marker Event on Age at Menopause

Bin Nan, Xihong Lin, Lynda D. Lisabeth, and Sioban D. Harlow

Abstract

. It is of recent interest in reproductive health research to investigate the validity of a marker event for the onset of menopausal transition and to estimate age at menopause using age at the marker event. We propose a varying coefficient Cox model to investigate the association between age at a marker event, denned as a specific bleeding pattern change, and age at menopause, where both events are subject to censoring and their association varies with age at the marker event. Estimation proceeds using the regression spline method. The proposed method is applied to the Tremin Trust Data to evaluate the association between age at onset of the 60-day menstrual cycle and age at menopause. The performance of the proposed method is evaluated using a simulation study.

A Varying-Coefficient Cox Model for the Effect of Age at a Marker Event on Age at Menopause

Bin Nan[1], Xihong Lin[2], Lynda D. Lisabeth[3], and Siobán D. Harlow[4]

[1,2] Department of Biostatistics
[3] Department of Neurology
[4] Department of Epidemiology
University of Michigan, Ann Arbor, MI 48109
[1] *email*: bnan@umich.edu

September 8, 2003

Summary. It is of recent interest in reproductive health research to investigate the validity of a marker event for the onset of menopausal transition and to estimate age at menopause using age at the marker event. We propose a varying coefficient Cox model to investigate the association between age at a marker event, defined as a specific bleeding pattern change, and age at menopause, where both events are subject to censoring and their association varies with age at the marker event. Estimation proceeds using the regression spline method. The proposed method is applied to the Tremin Trust Data to evaluate the association between age at onset of the 60-day menstrual cycle and age at menopause. The performance of the proposed method is evaluated using a simulation study.

Key Words: *B*-splines; Cox regression; Generalized cross validation; Marker events; Non-parametric regression; Survival Analysis; Time-dependent covariates.

1 Introduction

It is of recent interest in female reproductive aging research to identify marker events for the onset of the menopausal transition, and to investigate their use for estimating age at menopause. Menopause is defined as the final menstrual period (FMP), with the FMP confirmed after at least 12 months of amenorrhea. Although several marker events based on menstrual bleeding criteria have been proposed (Mitchell, et al., 2000; Soules, et al., 2001; Taffe and Dennerstein, 2001), there is a lack of appropriate statistical models to formally evaluate their validity due to the complex nature of the data. Specifically both age at onset of a marker event and age at menopause are subject to censoring, and their relationship is complicated and varies with age at onset of the marker event.

This paper is motivated by the analysis of the Tremin Trust data. This data set provides a unique opportunity to evaluate the association between age at menopause and ages at onset of the marker events proposed by reproductive health experts based on bleeding criteria (Treloar, et al., 1967). The study enrolled 1997 white college students at the University of Minnesota between 1935 and 1939 and followed them up to 40 years through their reproductive life. The study participants were asked to use menstrual diary cards to record the days when bleeding was experienced. Only limited covariate information was available in the data.

Lisabeth et al. (2003) analyzed a subset of 562 women from the original Tremin Trust cohort, who were age 25 or less at enrollment and still participating in the study at age 35. They performed descriptive analyses to examine the associations between ages at onset of several proposed marker events, e.g., the age a woman first experienced a menstrual cycle at least 60-days in length, and age at natural menopause, which was defined as the age a woman experienced the final menstrual period. Their preliminary analysis results suggest that a 60-day cycle might be a useful marker for predicting age at menopause. To illustrate

the main issue of the analysis and the main idea of the proposed approach, we focus in this paper on investigating the association between age at onset of the 60-day cycle marker event and age at menopause.

Figure 1 provides the Kaplan-Meier curves of age at the 60-day cycle marker event and age at menopause. A total of 282 women experienced the 60-day cycle marker event, and 280 (50%) women were censored for the marker event. The median age at the 60-day cycle marker was 48.7 years. A total of 193 women experienced natural menopause, and 369 (66%) women were censored for menopause. The median age at menopause was 51.7 years. There were 9 women who experienced menopause without having the 60-day cycle marker, and 271 women who were censored for both the 60-day cycle and menopause events. Note that these 271 women who were censored for the marker event were part of the 369 women who were censored for menopause and their censoring times for those two events were the same.

As the first step to explore the relationship between age at the 60-day cycle marker and age at menopause, we restricted ourselves to the 282 women who had an observed 60-day cycle marker event and classified them into several groups based on their ages at onset of the 60-day marker as [35, 40), [40,43), and so on. For each marker age group, we calculated the quartiles of age at menopause using the Kaplan-Meier method and displayed these estimated quartiles using a boxplot. These boxplots are given in Figure 2. The number of women in each marker age group is given above the corresponding boxplot. Figure 2 shows that the relationship between age at the 60-day cycle marker and age at menopause is complicated and varies with age at the 60-day cycle marker. For example, women who had experienced the 60-day cycle marker before age 40 had a median age at menopause of 53; while for women who had experienced the 60-day cycle marker between age 40 and 43, the median age at menopause dropped quickly to 47, with median menopause age subsequently increasing with age at marker event.

The first scientific interest is to quantify the association between age at the 60-day marker and age at menopause using a statistical model. The second scientific interest, especially for clinicians and women themselves, is to estimate the distribution of age at menopause given age at onset of the 60-day cycle marker. For example, if a woman first experiences a 60-day cycle at age 40, she would like to know from her physician her expected median age of menopause. From a clinical point of view, this would be a very useful piece of information for helping determine a woman's need for continued contraception and the likelihood of initiating interventions such as bone density screening.

The development of such a statistical model however has the following challenges. First, the graphical analysis, such as Figure 2, was informative but only descriptive. Further, it was restricted to the 282 women who had observed the 60-day cycle marker. However, the remaining 280 women who had been censored for the 60-day cycle marker event also contained some information about age at menopause, and one would like to include these women in the analysis. Second, we need a flexible model to allow the association between age at the 60-day cycle marker and age at menopause to vary with age at the 60-day cycle marker.

Several approaches have been proposed for modeling intermediate marker events. Crowley and Hu (1977) analyzed the Stanford heart transplant data using the Cox partial likelihood method and treated transplant status, which was an intermediate marker event, as a time dependent covariate. Lefkopoulou and Zelen (1995) and Nam and Zelen (2001) studied the same model from a different angle, which leads to a contingency table interpretation. For an overview of the existing methods handling intermediate marker events, see Kalbfleisch and Prentice (2002, Section 6.4). All of these authors assumed a constant regression coefficient for modeling the effect of the intermediate marker event. The results in Figure 2 however suggest that this assumption is not appropriate for the Tremin Trust data. In particular, if

we model age at onset of the 60-day marker event as a time-dependent covariate, we need to allow its regression coefficient to vary with age at the marker event. We hence consider a varying coefficient model.

Hastie and Tibshirani (1993) proposed general varying-coefficient models. In the Cox model setting, it is commonly assumed in such models that the regression coefficient $\beta(\cdot)$ is a function of the follow-up time, e.g., see Murphy and Sen (1991), Marzec and Marzec (1997), Cai and Sun (2003), among others. Since our purpose is to evaluate the effect of age at the 60-day cycle marker event on age at menopause, as demonstrated in Figure 2, it is natural and biologically more interpretable to assume that, the regression coefficient $\beta(\cdot)$ of the time dependent covariate, which indicates the onset of the marker event, be a function of age at the marker event, instead of a function of the follow-up time.

The remainder of the paper is organized as follows. We introduce in Section 2 a varying coefficient Cox model for age at menopause, where age at onset of the 60-day cycle marker is a time dependent binary covariate and its coefficient is assumed to be a smooth function of the marker event age. We discuss in Section 3 an estimation procedure using regression splines. We analyze in Section 4 the Tremin Trust data, and conduct a simulation study in Section 5 to evaluate the performance of the proposed method, followed by concluding remarks in Section 6.

2 The Varying-Coefficient Model

Suppose the data consist of n subjects. Let Y_i be the observed time to the event of interest, which is defined as the minimum of the survival time T_i, e.g., age at menopause, and the censoring time C_i for the ith subject ($i = 1, \cdots, n$). We assume independent censoring. Let Δ_i be a censoring indicator, which takes value 1 if a failure is observed and 0 otherwise. Let $Z_i(t)$ be a binary time dependent variable to indicate the onset of a marker event.

Assuming $\lambda_0(t)$ is the baseline hazard and $\lambda_i\{t|Z_i(t)\}$ is the hazard rate of the survival time to the endpoint event at t given $Z_i(t)$. A standard Cox model with a time dependent covariate has the following form:

$$\lambda_i\{t|Z_i(t)\} = \lambda_0(t)\exp\{\beta Z_i(t)\} \ . \tag{1}$$

It is common to use (1) to model the effect of an intermediate marker event by defining $Z_i(t)$ as a binary indicator which takes value 0 and changes to 1 after the marker event happens (Crowley and Hu, 1977; Kalbfleisch and Prentice, 2002).

In the Tremin Trust data, t is age, time to the endpoint event is age at menopause, and time to the marker event is age at the first occurrence of the 60-day cycle marker event. Model (1) assumes the log relative risk comparing subjects who have experienced the marker event with subjects who never experienced the marker event is constant β, and is irrelevant to the age they have experienced the marker event. The discussions in Section 1 suggest that the association between age at menopause and age at the marker event varies with age at the occurrence of the 60-day cycle marker event in the Tremin Trust data. We hence propose to extend model (1) to allow the regression coefficient β to be a function of age at the marker event.

Let S_i be the true age at the 60-day marker event for woman i. We have

$$Z_i(t) = \begin{cases} 1 & \text{if } t \geq S_i \\ 0 & \text{if } t < S_i. \end{cases} \tag{2}$$

Equivalently, $Z_i(t) = I[t \geq S_i]$, where $I(\cdot)$ is an indicator function. This means that $Z_i(t)$ is zero and jumps to 1 if a woman experiences the marker event at age S_i. We extend model (1) to allow the association between age at menopause and age at the marker event to depend on age at the marker event $S_i = s$ as

$$\lambda\{t|Z_i(t)\} \ = \ \lambda_0(t)\exp\{\beta(s)Z_i(t)\} = \begin{cases} \lambda_0(t)\exp\{\beta(s)\} & \text{if } t \geq s \\ \lambda_0(t) & \text{if } t < s, \end{cases} \tag{3}$$

where $\beta(s)$ is an unknown smooth function.

We now compare the interpretations of model (1) and model (3). In both models, the baseline hazard $\lambda_0(t)$ is the hazard for women who never experienced the marker event. Suppose a woman experiences the marker event at s. Before the marker event happens, her hazard of menopause is the baseline hazard $\lambda_0(t)$. Once the marker event happens at age $S_i = s$, her hazard of menopause changes to the baseline hazard $\lambda_0(t)$ multiplied by a factor $\exp\{\beta(s)\}$ as $t \geq s$, and her survival function starts to diverge if $\beta(s) \neq 0$, from those who have not yet observed any marker event at time s. This means that we assume the hazards are proportional after the onset of the marker event. Under the constant relative risk Cox model (1), this proportional factor is constant and free of the age at onset of the marker event s, e.g., $\exp\{\beta(s)\} = \exp(\beta)$. Under the varying-coefficient Cox model (3), the proportional factor varies with age at onset of the marker event s.

The difference between model (1) and model (3) is more easily illustrated using Figure 3(a) and 3(b) on the log hazard scale by contrasting two subjects who have experienced the marker event at time 1 and time 2 respectively. Under the constant coefficient Cox model (1), the first subject's hazard is the baseline hazard before time 1 (s_1) and changes from the baseline hazard since time 1 by an amount of β, while the second subject's hazard is the baseline hazard before time 2 (s_2) and changes from the baseline hazard since time 2 by the same amount β. Under the varying-coefficient Cox model (3), both women's hazards also change at time 1 and time 2 respectively, but by different constants $\beta(s_1)$ and $\beta(s_2)$ respectively. Note that the lines are all parallel and reflect the proportional hazards assumption.

It should be noted that $Z_i(t)$ is always observable at any t during the observed follow-up time period, even though both age at marker event and age at menopause are subject to censoring and their true values are unknown. Specifically, if the marker event is observed for

a woman, $Z_i(t)$ is fully observed at any t and is a step function, which takes value 0 before the marker event time and 1 afterwards, during the follow-up. If the marker event is not observed for a woman during the follow-up time, i.e., age of the marker event S_i is censored and the true value of S_i is not observed, then $Z_i(t) = 0$ during the observed follow-up time, and her age at menopause will either be censored at the same time or is observed before S_i.

If baseline covariates $\boldsymbol{X}_i$ are available, model (3) can be easily extended to incorporate baseline covariates $\boldsymbol{X}_i$ as

$$\lambda_i\{t|Z_i(t), \boldsymbol{X}_i\} = \lambda_0(t)\exp\{\beta(s)Z_i(t) + \boldsymbol{\gamma}'\boldsymbol{X}_i\} \ , \tag{4}$$

where $\boldsymbol{X}_i$ is age at menarche in the Tremin Trust data. For extension of (4) to accommodate multiple marker events, see Discussion. Since model (3) is a special case of model (4), we shall focus on model (4) in this paper.

3 The Estimation Procedure

3.1 Estimation Using B-splines

We consider estimation of the nonparametric function $\beta(s)$ using the regression spline method by approximating $\beta(s)$ using the natural cubic B-spline basis. Let K be the number of interior knots. Knot locations are usually chosen such that there are roughly equal numbers of observed data points between any two adjacent knots. This can be done by placing the K knots using the $100j/(K+1)$ $(j = 1, \cdots, K)$ percentiles of the observed marker event times. We discuss in Section 3.2 estimation of the number of knots K using GCV.

Since a natural spline is constrained to be linear beyond the two boundary knots, the function $\beta(s)$ can be parameterized using $K+2$ natural cubic B-spline basis functions $B_k(s)$ $(k = 1 \ldots, K+2)$ as

$$\beta(s) = \sum_{k=1}^{K+2} \theta_k B_k(s). \tag{5}$$

Replacing $\beta(s)$ by its B-spline approximation in equation (5), model (4) can be written as

$$\lambda\{t|Z_i(t), \boldsymbol{X}_i\} = \lambda_0(t)\exp\{\boldsymbol{\theta}'\widetilde{\boldsymbol{Z}}_i(t) + \boldsymbol{\gamma}'\boldsymbol{X}_i\} \ , \tag{6}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{K+2})'$ and $\widetilde{\boldsymbol{Z}}_i(t) = \{B_1(s)Z_i(t), \ldots, B_{K+2}(s)Z_i(t)\}'$. Note that $\widetilde{\boldsymbol{Z}}_i(t)$ is always observable during follow-up because $Z_i(t)$ is fully observed during follow-up. Specifically, if the marker event is observed at S_i for the ith woman during follow-up, then $\widetilde{\boldsymbol{Z}}_i(t) = \mathbf{0}$ if $t < S_i$ and $\widetilde{\boldsymbol{Z}}_i(t) = \{B_1(S_i), \ldots, B_{K+2}(S_i)\}'$ if $t \geq S_i$. If the marker event is not observed, i.e., S_i is censored and the true S_i is not available, then $\widetilde{\boldsymbol{Z}}_i(t) = \mathbf{0}$ for any observed follow-up time t. This is because in the latter case, even though the true S_i is unknown and thus $B_k(S_i)$ cannot be evaluated, we have $Z_i(t) = 0$ and hence $\widetilde{\boldsymbol{Z}}_i(t) = \mathbf{0}$. It follows that $\boldsymbol{\theta}$ can be estimated and hence $\beta(s)$ is identifiable and can be estimated without difficulty.

Now model (6) becomes a standard Cox proportional hazards model with the time dependent covariate vector $\widetilde{\boldsymbol{Z}}(t)$ and the baseline covariate vector $\boldsymbol{X}_i$. Estimation of the parameter vectors $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ can be obtained by maximizing the following log partial likelihood function

$$\ell(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \sum_{i=1}^{n} \int \left[\boldsymbol{\theta}'\widetilde{\boldsymbol{Z}}_i(t) + \boldsymbol{\gamma}'\boldsymbol{X}_i - \log \sum_{j=1}^{n} I(Y_j \geq t)\exp\left\{\boldsymbol{\theta}'\widetilde{\boldsymbol{Z}}_j(t) + \boldsymbol{\gamma}'\boldsymbol{X}_j\right\} \right] dN_i(t) \ , \tag{7}$$

where Y_i is the minimum of censoring time C_i and time to menopause T_i and $N_i(\cdot)$ is the counting process of menopause event for subject i.

The maximum partial likelihood estimators of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ can be calculated using any statistical package that does Cox regression, e.g., SAS PROC PHREG. Note that as discussed above, if the marker event is censored for subject i, i.e. S_i is not observable, then $Z_i(t) = 0$ at any observed follow-up time t, and we can assign S_i any value for numerical implementation when fitting the Cox model (6). Denote by $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\gamma}})$ the maximum partial likelihood estimators of $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ and $\mathrm{cov}(\widehat{\boldsymbol{\theta}})$ and $\mathrm{cov}(\widehat{\boldsymbol{\gamma}})$ their covariance estimators. The nonparametric function

$\beta(s)$ can be estimated by

$$\widehat{\beta}(s) = \sum_{k=1}^{K+2} \widehat{\theta}_k B_k(s) \ . \tag{8}$$

The pointwise confidence interval for $\widehat{\beta}(s)$ can be estimated using its variance estimator $\text{var}\{\widehat{\beta}(s)\} = \boldsymbol{B}(s)'\text{cov}(\widehat{\boldsymbol{\theta}})\boldsymbol{B}(s)$, where $\boldsymbol{B}(s) = \{B_1(s), \cdots, B_{K+2}(s)\}'$. Note that for two arbitrary women who have experienced the marker event at age s_1 and s_2 ($s_2 > s_1$), their log relative risk after age s_2 can be easily calculated as $\widehat{\beta}(s_2) - \widehat{\beta}(s_1)$ and its variance can be easily calculated as $\{\boldsymbol{B}(s_2) - \boldsymbol{B}(s_1)\}'\text{cov}(\widehat{\boldsymbol{\theta}})\{\boldsymbol{B}(s_2) - \boldsymbol{B}(s_1)\}$.

As discussed in the Introduction Section, it is of both clinical interest and a woman's own interest to estimate age at menopause if a woman has experienced the 60-day marker event at a certain age. This can be done by estimating the survival function of age at menopause given a specified age at onset of the 60-day marker event s using model (4). We first estimate the baseline cumulative hazard function $\Lambda_0(t)$ using Breslow estimator,

$$\widehat{\Lambda}_0(t) = \int_0^t \left[\sum_{i=1}^n I(Y_i \geq u)\exp\{\widehat{\beta}(S_i)Z_i(u) + \widehat{\boldsymbol{\gamma}}'\boldsymbol{X}_i\}\right]^{-1} \left\{\sum_{i=1}^n dN_i(u)\right\}. \tag{9}$$

Then the survival function for menopause given the age at the marker event $S = s$ and the covariates $\boldsymbol{X} = \boldsymbol{x}$ can be estimated by

$$\widehat{F}(t|s, \boldsymbol{x}) = \exp\left\{-\int_0^t \exp[\widehat{\beta}(s)Z(u) + \widehat{\boldsymbol{\gamma}}'\boldsymbol{x}] \ d\hat{\Lambda}_0(u)\right\}, \tag{10}$$

where $Z(u) = I(u \geq S)$.

3.2 Estimation of the Number of Knots

An advantage of the use of a regression spline for estimating the nonparametric function $\beta(s)$ is its computational simplicity. However this method requires estimation of the number of knots. For uncensored data, cross-validation (CV) and generalized cross validation (GCV)

are commonly used to choose the number of knots, see, e.g. Hastie and Tibshirani (1990). For survival data, O'Sullivan (1988) proposed CV and GCV for choosing the smoothing parameter for the smoothing spline estimator assuming that the baseline cumulative hazard function $\Lambda_0(t) = \int_0^t \lambda_0(u)du$ is known. We extend O'Sullivan's method to choose the number of knots in the regression spline setting and account for the fact that $\Lambda_0(t)$ is unknown and is estimated.

We first consider the case when $\Lambda_0(t)$ is known. Following O'Sullivan (1988), under model (6), for a given number of knots K, if $\Lambda_0(t)$ is a known function, the likelihood function of $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ is available and can be maximized using an iterated reweighted least square algorithm. If the estimators of $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ at the lth iteration are $(\widehat{\boldsymbol{\theta}}_{(l)}, \widehat{\boldsymbol{\gamma}}_{(l)})$, the working weight w_i and the working dependent variable y_i for subject i can be written as

$$\begin{aligned} w_i &= \frac{1}{2}\Lambda_0(Y_i)\exp\{\widehat{\boldsymbol{\theta}}'_{(l)}\widetilde{\boldsymbol{Z}}_i(Y_i) + \widehat{\boldsymbol{\gamma}}'_{(l)}\boldsymbol{X}_i\} \ , \\ y_i &= \widehat{\boldsymbol{\theta}}'_{(l)}\widetilde{\boldsymbol{Z}}_i(Y_i) + \widehat{\boldsymbol{\gamma}}'_{(l)}\boldsymbol{X}_i + \Delta_i/(2w_i) - 1 \ , \end{aligned}$$

where $\Delta_i = I(T_i \leq C_i)$ is the censoring indicator. One calculates $(\widehat{\boldsymbol{\theta}}_{(l+1)}, \widehat{\boldsymbol{\gamma}}_{(l+1)})$ by minimizing $\sum_{i=1}^n w_i\{y_i - \boldsymbol{\theta}'\widetilde{\boldsymbol{Z}}_i(Y_i) - \boldsymbol{\gamma}'\boldsymbol{X}_i\}^2$. Denote by $\widetilde{\boldsymbol{X}}_i = \{\widetilde{\boldsymbol{Z}}_i(Y_i), \boldsymbol{X}'_i\}'$ and $\widetilde{\mathbf{X}} = \left(\widetilde{\boldsymbol{X}}'_1, \cdots, \widetilde{\boldsymbol{X}}'_n\right)'$. Denote by $\widehat{\boldsymbol{y}} = (\widehat{y}_1, \cdots, \widehat{y}_n)'$, $\widehat{\mathbf{W}} = \mathrm{diag}(\widehat{w}_1, \cdots, \widehat{w}_n)$ and $\widehat{\boldsymbol{f}} = (\widehat{f}_1, \ldots, \widehat{f}_n)'$ the working dependent variable, the working weight matrix, and the predicted value vector at convergence. Then $\widehat{\boldsymbol{f}}$ can be calculated as $\widehat{\boldsymbol{f}} = \widetilde{\mathbf{X}}(\widetilde{\mathbf{X}}'\widehat{\mathbf{W}}\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}'\widehat{\mathbf{W}}\widehat{\boldsymbol{y}} = \widehat{\mathbf{H}}\widehat{\boldsymbol{y}}$, where $\widehat{\mathbf{H}}$ is the linearized hat matrix. The generalized cross-validation (GCV), which is a function of the number of knots K, is given by

$$GCV(K) = \frac{\sum_{i=1}^n w_i(\widehat{y}_i - \widehat{f}_i)^2}{(1-\bar{h})^2} \ , \tag{11}$$

where $\bar{h}$ is the average of the diagonal elements of $\widehat{\mathbf{H}}$, the so-called mean leverage.

We now consider the case when the baseline hazard $\Lambda_0(t)$ is unknown and is estimated by the Breslow estimator (9). O'Sullivan (1988) suggested to calculate the Breslow estimator of

$\Lambda_0(\cdot)$ for each K and plug it into (11) as if it were known. However, this plug-in procedure ignores the fact that different choices of K give different baseline hazard estimators of $\Lambda_0(t)$, but the above procedure assumes the same true baseline hazard is used for different K. We hence propose a modified procedure to account for this.

First, a series of Cox models as in (6) are fitted for a range of the number of interior knots K. We used 1 to 20 in the analysis of Tremin Trust data. For each choice of K, the cumulative baseline hazard function estimator $\widehat{\Lambda}_0(t;K)$ and the B-spline estimator $\widehat{\beta}(s;K)$ are calculated. They are then plugged into equation (11) to calculate GCV(K). Note that different baseline hazard estimators are used for different K at this step. We then select K that minimizes GCV(K), call it K_* and obtain the corresponding baseline hazard estimator $\widehat{\Lambda}_0(t;K_*)$. At the next step, we replace the true $\Lambda_0(t)$ by this estimated $\widehat{\Lambda}_0(t;K_*)$ and treat it as fixed and known. Then recalculate the GCV statistic (11) using the above least square procedure for each choice of K and select a new K that minimizes GCV(K). Note here a common $\widehat{\Lambda}_0(t;K_*)$ is used to calculate GCV for different choices of K. The procedure is repeated until the chosen K_* at the current step is the same as the K_* at the previous step. The cross-validation (CV) statistic can be calculated similarly.

4 The Analysis of the Tremin Trust Data

We applied the proposed varying coefficient Cox model to the analysis of the Tremin Trust data. The goals of our study were to investigate the relationship between age at menopause and age at the 60-day cycle marker event, and to estimate the distribution of age at menopause given any particular age at onset of the 60-day cycle marker. The data were described in detail in Section 1. The data used in our analysis were the same as that used in Lisabeth, et al. (2003) and consisted of 562 women who were 25 years or younger at enrollment and still participated in the study at age 35. We used age 35 as the time origin

in our analysis. The median age at enrollment was 19 years, the median age at menarche was 12 years and ranged from 9 to 18 years, and the length of follow-up ranged from 9 to 39 years with median of 27 years.

For each woman, the data set contained the minimum of the age at menopause and the censoring age, i.e., the observed menopause age; a censoring indicator for age at menopause; a 60-day cycle marker event indicator; the age at the marker event if it occurred during the follow-up time; and age at menarche. For descriptive Kaplan-Meier analysis results of age at menopause and age at the 60-day cycle marker event, see Section 1 and Figure 2.

We considered the semiparametric varying coefficient Cox model (4) by letting $Z_i(t)$ be a time-dependent binary indicator for the onset of the 60-day cycle marker event and the baseline covariate X_i be age at menarche. We fit model (4) using the B-spline method via the Cox model (6). The method of Therneau and Grambsch (2000) was used to expand the data set for the time dependent covariate $Z_i(t)$. For subject i, if the 60-day cycle marker event was observed, i.e. S_i was observed, then the i-th record was expanded into two records: the first one had the starting time 0 and the stopping time S_i with $Z_i(t) = 0$ and $\Delta_i = 0$; the second one had the starting time S_i and the stopping time $Y_i = \min(T_i, C_i)$ with $Z_i(t) = 1$ and the observed censoring indicator Δ_i. If the 60-day cycle marker event was not observed, i.e. S_i was censored, the starting time was zero and the stopping time was Y_i, and $Z_i(t) = 0$ and Δ_i was the observed censoring indicator Δ_i. Then the risk set at time t was defined as the set of all the records with t falling in between each pair of their starting times and stopping times. The baseline covariate X_i kept the same value during the data expansion.

Among the 562 women, 282 experienced the 60-day cycle marker during the follow-up. The observed marker times were used to determine the knot allocations and generate the natural cubic B-spline basis functions $B_k(s)$ used for estimating $\beta(s)$. The extreme values of the observed marker times were used as the two boundary knots. The number and the

locations of the interior knots were determined using the GCV method described in Section 3.2. The optimal number of interior knots was estimated as $K_{optimal} = 8$. The spline estimator of $\beta(s)$ and its 95% point-wise confidence interval are plotted in Figure 4. For illustrative purpose, we also considered approximating $\beta(s)$ using piecewise constants as $\beta(s) = \sum_{k=1}^{K+1} \beta_j I[s_{k-1} < S_i \leq s_k]$, where $\{s_0, s_1, \cdots, s_{K+1}\}$ is the set of knots including the boundary knots, and fit $\lambda\{t|Z_i(t), X_i\} = \lambda_0(t)\exp\Big\{ \sum_{k=1}^{K+1} \beta_j I[s_{k-1} < S_i \leq s_k]Z_i(t) + \gamma X_i \Big\}$. The piecewise constant estimator of $\beta(s)$ using the age intervals $[35, 38)$, $[38, 40)$, etc., is superimposed in Figure 4. We can see that the B-spline estimate and the piece-wise constant estimate of $\beta(s)$ agree well with each other.

The results in Figure 4 suggest that the 60-day cycle marker is strongly associated with age at menopause, and its effect varies with age at the 60-day cycle marker event. But when age at marker event is close to 35, the estimated $\beta(s)$ does not significantly differ from zero which implies that having a marker around age 35 is uninformative about age at menopause. The curve is mainly positive and increases before age 44 and then starts to decrease. This indicates that before age 44, the association between age at menopause and age at the 60-day cycle marker becomes stronger as age increases. Among women who first experience the 60-day cycle before 44, as age at onset of the 60-day cycle increases, she is likely to have menopause more quickly. For example, consider two women: the first woman experiences the 60-day cycle at age 39 and the second woman experiences the 60-day cycle at age 42. Then relative risk of menopause at any age after age 42 for the second woman is $\exp\{\widehat{\beta}(42) - \widehat{\beta}(39)\}$ =exp(4.1-2.2)=6.7 times higher than the first woman (p-value < 0.0001).

The estimated $\beta(s)$ curve starts decreasing after age 44. This indicates that after age 44, the association between age at menopause and age at the 60-day cycle marker becomes weaker as age increases. Among women who first experience the 60-day cycle after 44, as age at onset of the 60-day cycle increases, a woman is likely to have menopause at a later

age. For example, consider two women: the first woman experiences the 60-day cycle at age 48 and the second woman experiences the 60-day cycle at age 51. Then relative risk of menopause at any age after age 48 for the second woman is $\exp\{\widehat{\beta}(51) - \widehat{\beta}(48)\}$=exp(1.9-3.2)=0.27 times lower than the first woman (p-value < 0.0001). In other words, the relative risk of menopause at any age after age 51 for the first woman is 1/0.27=3.7 times higher of the second woman.

The estimated log relative risk for age at menarche was -0.16 ($RR = 0.85$) for a one year increment (p-value = 0.01). This means that a younger age at menarche has a significant effect on advancing the expected age at menopause. We also found that the effect of age at the 60-day cycle marker was independent of age at menarche. Particularly the estimated curves of $\beta(s)$ were almost identical with and without adjusting for age at menarche.

The survival probabilities of age at menopause were calculated using equation (10) for several selected ages at the 60-day cycle marker event given age of menarche equaled to 12, which was the median age of menarche. The estimated survival curves are plotted in Figure 5(a) and the estimated corresponding percentiles are summarized in Table 1. For example, if a woman experiences the 60-day cycle marker at age 36, 39, 42, 45, 48, or 51, her median age of menopause is expected to be 54.4, 51.9, 47.5, 49.3, 50.6 or 53.1. These results are consistent with the pattern of the estimated $\beta(s)$ curve in Figure 4. For a woman who experiences the 60-day cycle marker before age 44, the later she experiences the marker event, the earlier she is likely to experience menopause. For a woman who experiences the 60-day cycle marker after age 44, the later she experiences the marker event, the later she is likely to experience menopause.

These results are biologically meaningful. Women who are observed to have a 60-day cycle before age 40 may belong to a subgroup of women who cycle infrequently, e.g. women with polycystic ovarian disease, and for whom the pattern of change in menstrual bleeding

with age may differ from other women. Additional research on this subgroup of women is needed.

Another interesting and probably more intuitive piece of information for both clinicians and midlife women is the number of years from the onset of marker event to menopause. The percentiles of this quantity can be easily calculated by subtracting age at marker event from the corresponding estimated percentiles for age at menopause, which are also given in Table 1. The median number of years for a woman who experiences the 60-day cycle marker at age 36, 39, 42, 45, 48, or 51, is expected to be 18.4, 12.9, 5.5, 4.3, 2.6, or 2.1. The survival curves for menopause after the onset of marker event are plotted in Figure 5(b).

5 The Simulation Study

We conducted a simulation study to evaluate the performance of the natural cubic B-spline estimator for $\beta(s)$ in model (3). The follow-up time was restricted from 0 to 1. To roughly mimic the shape of the estimated $\widehat{\beta}(s)$ for the 60-day cycle marker event in Figure 4, we assumed that true $\beta(s) = 3\sin(\pi s)$. The age at the marker event S was generated from a Weibull distribution with shape parameter 2 and scale parameter 1. The age at menopause T was generated from the model $\lambda\{t|Z(t)\} = \lambda_0(t)\exp\{\beta(s)Z(t)\}$, where $Z(t) = I(t \geq S)$ and the baseline hazard $\lambda_0(t) = 0.5t^2$, which corresponds to the hazard of a Weibull distribution with shape parameter 2 and scale parameter 4. The censoring time C was generated by $C = U \cdot I(U \leq 1) + I(U > 1)$, where $U \sim \text{Uniform}(0, 2)$. Thus the observed time $Y = \min(T, C)$ was within the interval $[0, 1]$. The censoring percentage was about 70%. We assumed a sample size of $n = 500$ in each simulated data set.

To reduce the computational burden, we chose the optimal number of interior knots in spline estimating $\beta(\cdot)$ by minimizing the mean square error of $\hat{\beta}(\cdot)$ defined as MSE $= \sum_{j=1}^{J}\left\{\hat{\beta}(t_j) - \beta(t_j)\right\}^2$, where t_j, $j = 1, \ldots, J$, are equally spaced grid points in (0,1). We

used $J = 1000$. The two boundary knots were chosen to be 0 and 1, respectively.

We performed 100 simulations and analyzed each simulated data set using the varying coefficient model (3) using B-splines by fitting the Cox model (6). The estimated optimal numbers of interior knots varied from 1 to 6 with the average number of estimated knots equal to 1.6. The average of the 100 estimated $\widehat{\beta}(s)$ and the true curve $\beta(s)$ are plotted in Figure 6. The 95% pointwise confidence intervals for $\widehat{\beta}(s)$ using the empirical standard errors and the average of the 100 estimated standard errors are also plotted. Figure 6 suggests that the pointwise biases of the B-spline estimator $\widehat{\beta}(\cdot)$ are close to zero, and the pointwise model based SEs of $\widehat{\beta}(s)$ agree well with their empirical counterparts, except for the boundary.

6 Discussion

We have proposed in this paper a varying-coefficient Cox model to investigate the association between time to an intermediate marker event and time to a primary endpoint event, where the coefficient of the time dependent marker indicator is assumed to be a nonparametric function of time at the marker event, and baseline covariate effects are modeled parametrically. We estimate the nonparametric regression function using B-splines which can be easily formulated into a standard Cox model and fitted using the standard partial likelihood method. We estimate the number of knots using a modification of O'Sullivan (1988)'s GCV method. Our simulation results suggest the proposed method works well in finite samples.

The large sample theory for the partial likelihood based regression spline estimator $\widehat{\beta}(s)$ is beyond the scope of this paper. For discussions of such spline estimators in linear regression settings, see Zhou et al. (1998) and Huang (2003). An extension of their results to the Cox model setting requires further research. Our simulation results provide empirical evidence that similar results are likely to hold for Cox regression with varying-coefficients. If the parametric parameter $\boldsymbol{\gamma}$ in the semiparametric varying coefficient model (4) is of primary

interest, the results of Huang (1999) might be extended to our model.

For simplicity, we have focused in this paper on a single marker event. Several other bleeding criteria based markers were also considered by Lisabeth, et al. (2003), e.g., a 90-day cycle marker or a skipped cycle marker. Model (4) can easily be extended to incorporate multiple marker events with varying coefficients. Specifically, suppose there are two marker events. Let S_1 be the time to marker 1 and S_2 the time to maker 2. Let $Z_1(t)$ and $Z_2(t)$ be the marker indicators at time t for these two markers, respectively. Let $\boldsymbol{X}$ be the baseline covariate vector. Then we can model the hazard function for the time to the final event as

$$\begin{aligned}\lambda\{t|Z_1(t),Z_2(t),\boldsymbol{X}\} &= \lambda_0(t)\exp\{\beta_1(s_1)Z_1(t)+\beta_2(s_2)Z_2(t)+\boldsymbol{\gamma}'\boldsymbol{X}\} \\ &= \begin{cases} \lambda_0(t)\exp\{\boldsymbol{\gamma}'\boldsymbol{X}\} & \text{if } t<\min(s_1,s_2) \\ \lambda_0(t)\exp\{\beta_1(s_1)+\boldsymbol{\gamma}'\boldsymbol{X}\} & \text{if } s_1\le t<s_2 \\ \lambda_0(t)\exp\{\beta_2(s_2)+\boldsymbol{\gamma}'\boldsymbol{X}\} & \text{if } s_2\le t<s_1 \\ \lambda_0(t)\exp\{\beta_1(s_1)+\beta_2(s_2)+\boldsymbol{\gamma}'\boldsymbol{X}\} & \text{if } t\ge\max(s_1,s_2). \end{cases} \end{aligned} \tag{12}$$

Notice that for each subject, only one of the middle two conditions in (12) can happen. Similar to the single marker situation, $\beta_1(\cdot)$ and $\beta_2(\cdot)$ can be estimated using the natural cubic B-splines based on the partial likelihood method.

Interactions between a marker event and baseline covariates can also be modeled similarly. Let X be a scalar for illustrative purpose. Then model (4) can be extended to

$$\lambda\{t|Z(t),X\} = \lambda_0(t)\exp\{\beta(s)Z(t)+\gamma X+\phi(s)Z(t)X\}\ , \tag{13}$$

where $\phi(\cdot)$ is the interaction effect. Model (13) does not introduce any extra technical difficulty in terms of estimation when spline methods are used to estimate $\beta(\cdot)$ and $\phi(\cdot)$.

Acknowledgment

The work of Nan and Lin was supported in part by U. S. National Cancer Institute Grant R01 CA76404. The work of Lisabeth and Harlow was supported in part by U. S. National Institute of Aging Grant AG021543.

References

Cai, Z. and Sun, Y. (2003). Local linear estimation for time-dependent coefficients in Cox's regression models. *Scandinavian Journal of Statistics* **30,** 93-111.

Crowley, J. and Hu, M. (1977). Covariance analysis of heart transplant survival data. *Journal of the American Statistical Association* **72,** 27-36.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models.* Chapman & Hall.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society, Series B* **55,** 757-796.

Huang, J. (1999). Efficient estimation of the partly linear additive Cox model. *Annals of Statistics* **27,** 1536-1563.

Huang, J. H. (2003). Local asymptotics for polynomial spline regression. *Annals of Statistics*, to appear.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd edition. Hoboken, New Jersey: John Wiley & Sons.

Lefkopoulou, M. and Zelen, M. (1995). Intermediate clinical events, surrogate markers and survival. *Lifetime Data Analysis* **1,** 73-85.

Lisabeth, L. D., Harlow, S. D., Gillespie, B., Lin, X., and Sowers, M. F. (2003). Staging reproductive aging: a comparison of proposed bleeding criteria for the menopausal transition. *Menopause,* in press.

Marzec, L. and Marzec, P. (1997). On fitting Cox's regression model with time-dependent coefficients. *Biometrika* **84,** 901-908.

Mitchell, E. S., Woods, N. F., and Mariella A. (2000). Three stages of the menopausal transition from the Seattle Midlife Women's Health Study: toward a more precise definition. *Menopause* **7**, 334-349.

Murphy, S. A. and Sen, P. K. (1991). Time-dependent coefficients in a Cox-type regression model. *Stochastic Processes and Their Applications* **39,** 153-180.

Nam, C. M. and Zelen, M. (2001). Comparing the survival of two groups with an intermediate clinical events. *Lifetime Data Analysis* **7,** 5-19.

O'Sullivan, F. (1988). Nonparametric estimation of relative risk using splines and cross-validation. *SIAM Journal on Scientific and Statistical Computing* **9,** 531-542.

Soules, M. R., Sherman, S., Parrott, E. P., Rebar R., Santoro, N., Utian, W., Woods, N. (2001). Executive summary: stages of reproductive aging workshop. *Fertil Steril* **76**, 874-878.

Taffe, J. and Dennerstein, L. (2001). Menstrual patterns leading to final menstrual period. *Menopause* **9**, 32-40.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model.* Springer-Verlag New York, Inc.

Treloar, A. E., Boynton, R. E., Behn, B. G., and Brown, B. W. (1967). Variation of the human menstrual cycle through reproductive life. *International Journal of Fertility* **12,** 77-126.

Zhou, S., Shen, X., and Wolfe, D. A. (1998). Local asymptotics for regression splines and confidence regions. *Annals of Statistics* **26,** 1760-1782.

Table 1: Estimated percentiles for the survival probabilities of age at menopause given age at onset of the 60-day cycle marker. For each marker event age, the upper entries are the estimated percentiles of chronological age at menopause; and the lower entries are the estimated percentiles of the length in years from the onset of marker event to menopause.

Age at the marker event	Estimated percentiles				
(in years)	90%	75%	50%	25%	10%
36	50.6	52.5	54.4	55.3	56.2
	14.6	16.5	18.4	19.3	20.2
39	47.5	49.8	51.9	53.1	54.3
	8.5	10.8	12.9	14.1	15.3
42	43.3	45.2	47.5	49.2	50.3
	1.3	3.2	5.5	7.2	8.3
45	46.3	47.6	49.3	50.6	51.7
	1.3	2.6	4.3	5.6	6.7
48	48.6	49.3	50.6	51.7	52.4
	0.6	1.3	2.6	3.7	4.4
51	51.5	52.2	53.1	54.4	54.9
	0.5	1.2	2.1	3.4	3.9

List of Illustrations

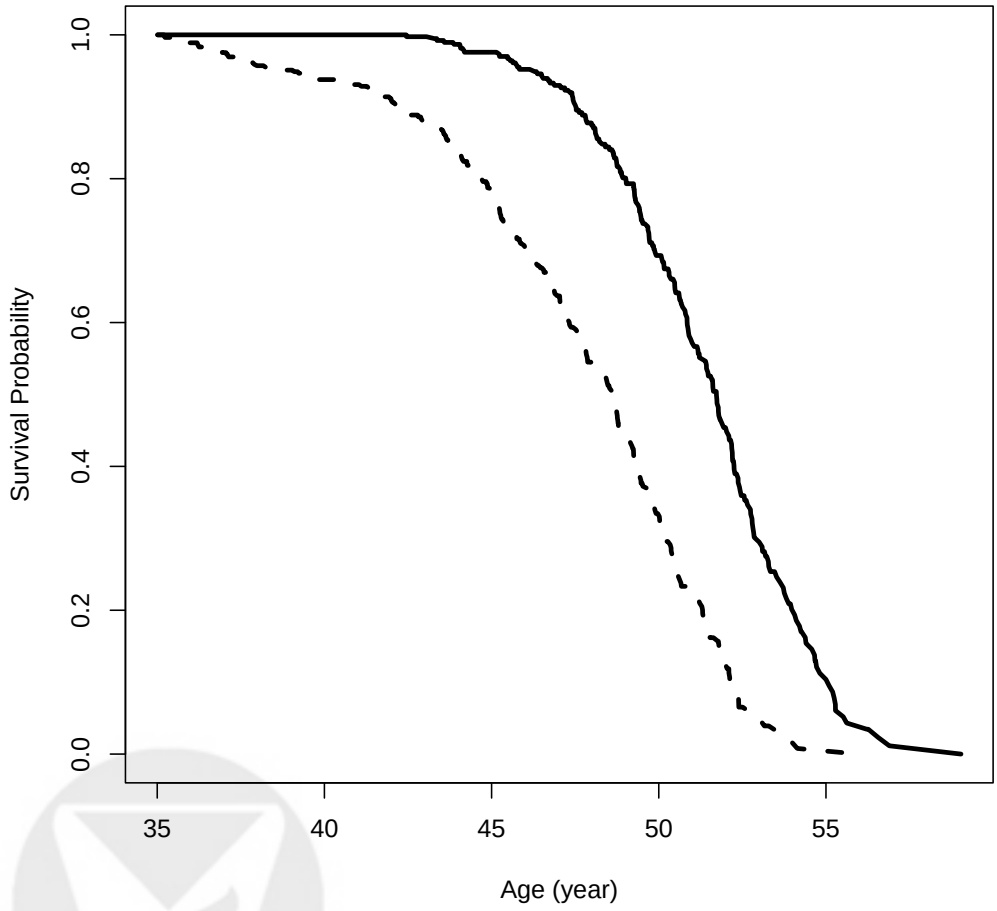

Figure 1:

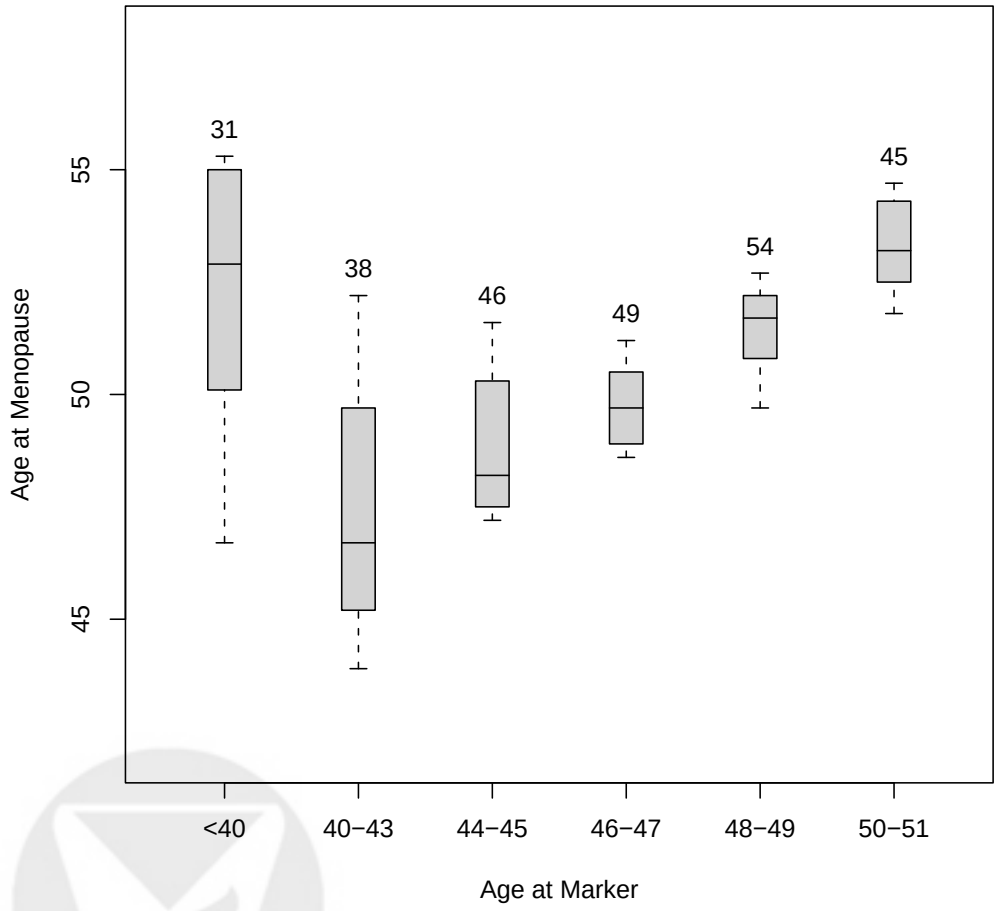

Figure 2:

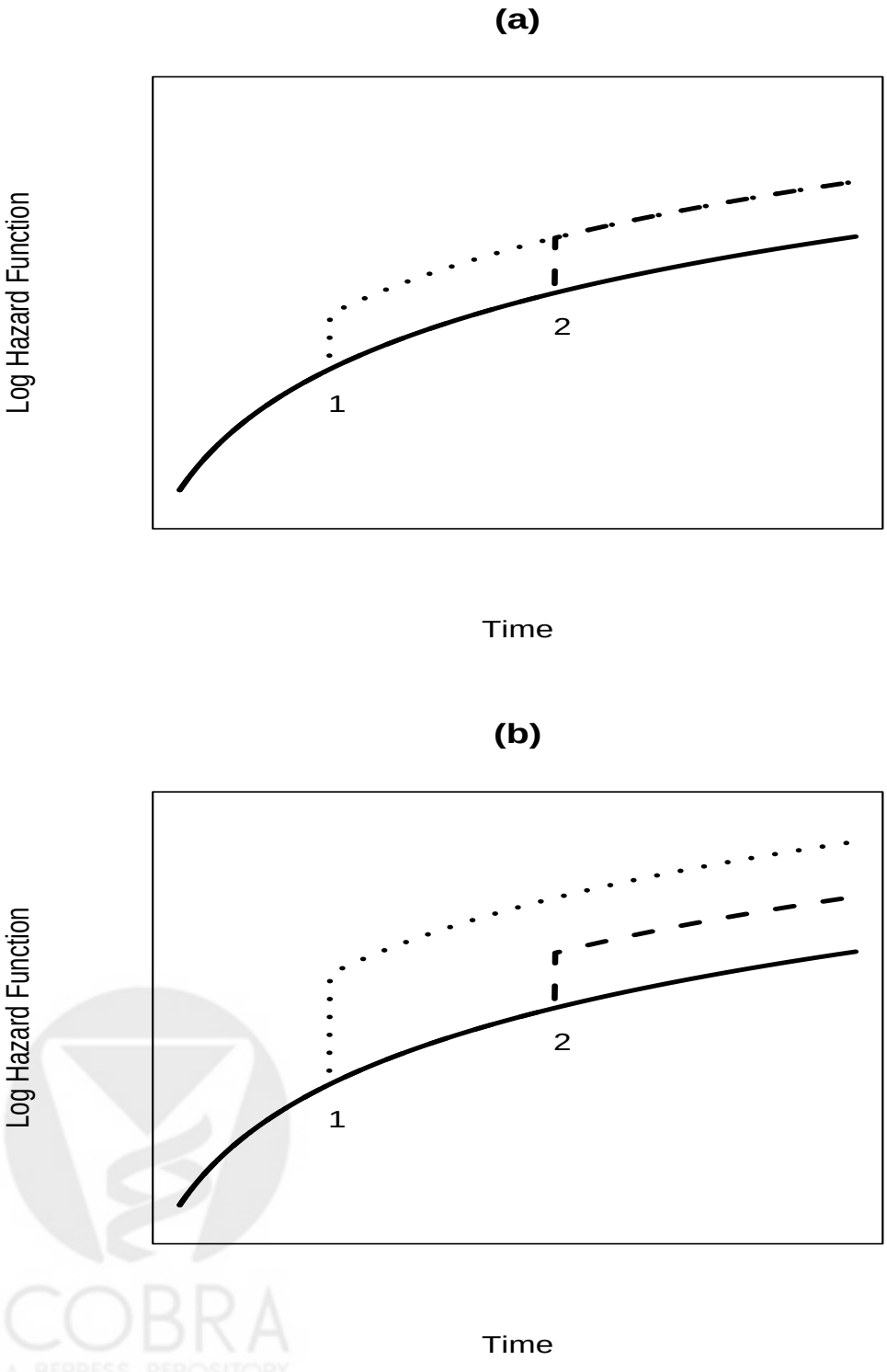

Figure 3:

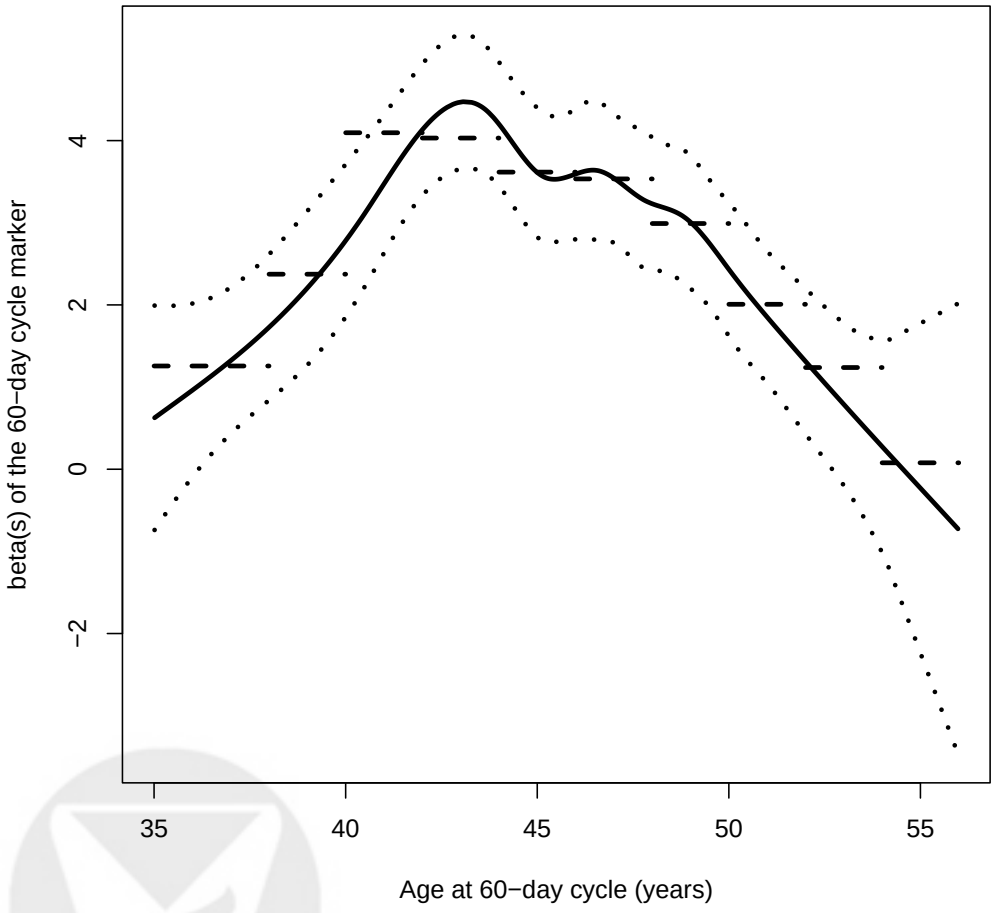

Figure 4:

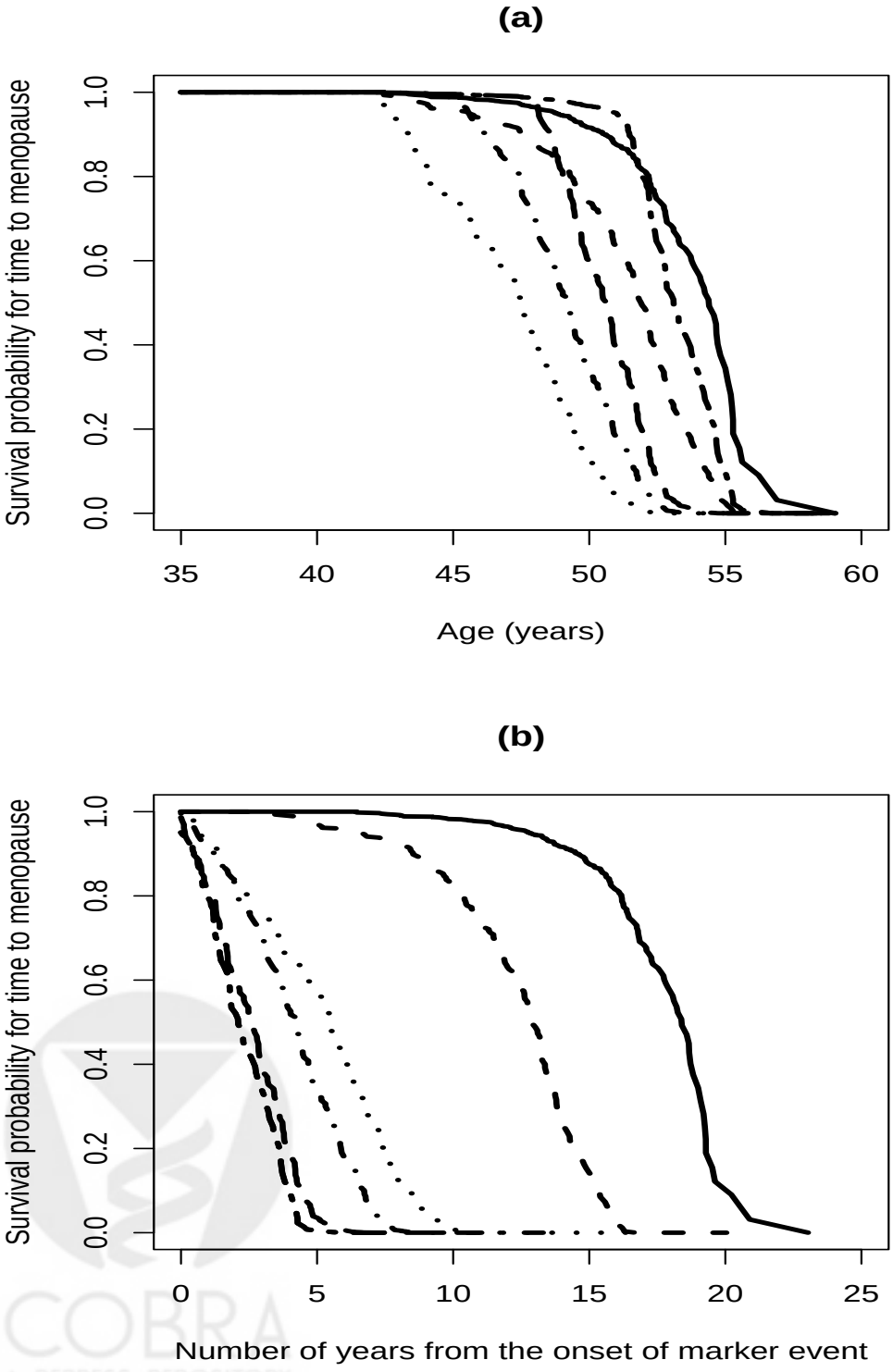

Figure 5:

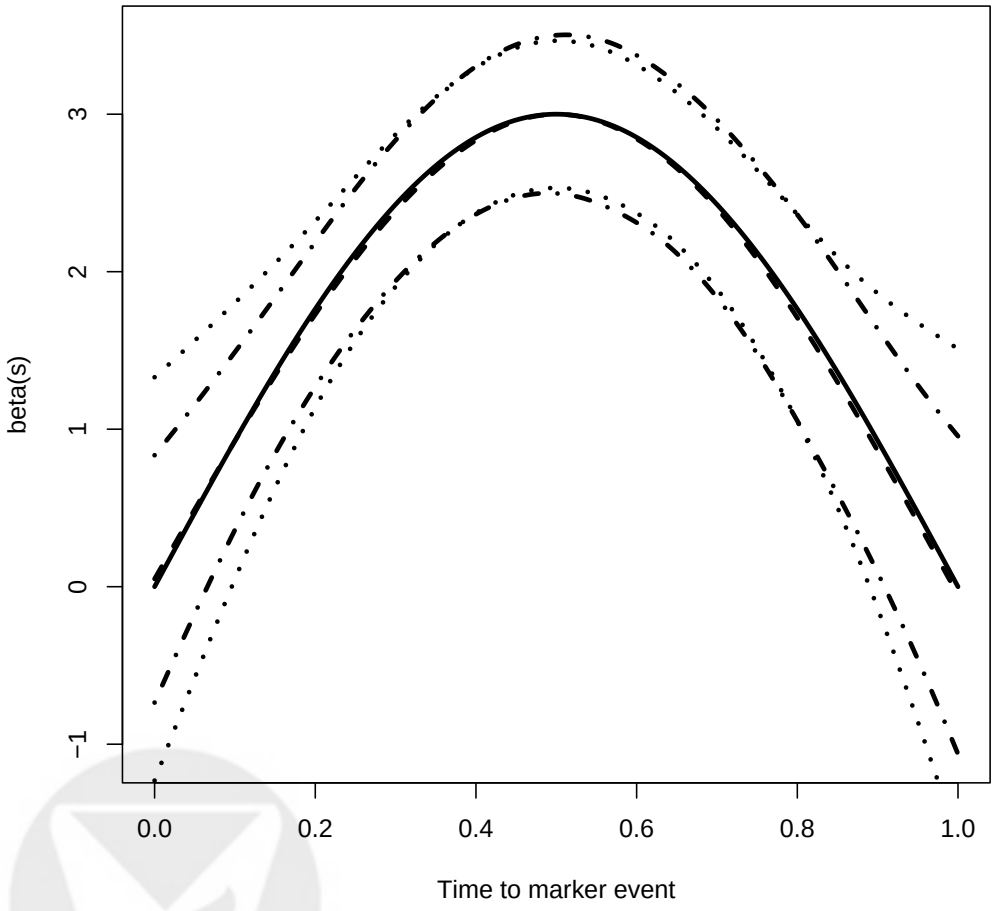

Figure 6: