Piecewise Constant Cross-Ratio Estimation for Association in Bivariate Survival Data with Application to Studying Markers of Menopausal Transition

Bin Nan, Xihong Lin, Lynda D. Lisabeth, and Siobán D. Harlow *

February 11, 2004

Abstract

A question of significant interest in female reproductive aging is to identify bleeding criteria for the menopausal transition. Although various bleeding criteria, or markers, have been proposed for the menopausal transition, their validity has not been adequately examined. The Tremin Trust data are collected from a long-term cohort study that followed a group of women throughout their whole reproductive life, and provide an unique opportunity for assessing the association between age at onset of a bleeding marker and age at onset of menopause. Formal statistical analysis of this dependence is challenging given the fact that both the marker event and menopause are subject to right censoring and their association depends on age at the marker event. We propose using cross-ratio to measure their dependence, which is assumed to be a piecewise constant function of age at onset of the marker event. Two estimation procedures using direct two-stage method and sequential two-stage method are proposed, while the latter is extended to allow for covariates in marginal survival functions. The proposed methods are applied to the analysis of the Tremin Trust data, and their performance is evaluated using simulations.

KEY WORDS: Bivariate survival; Clayton model; Cross-ratio function; Left truncation; Marker event; Menopausal transition; Two-stage method.

*Bin Nan is Assistant Professor and Xihong Lin is Professor, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109 (E-mail: bnan@umich.edu; xlin@umich.edu). Lynda D. Lisabeth is Research Investigator, Department of Neurology, University of Michigan, Ann Arbor, MI 48109 (E-mail: llisabet@med.umich.edu). Siobán D. Harlow is Associate Professor, Department of Epidemiology, University of Michigan, Ann Arbor, MI 48109 (E-mail: harlow@isr.umich.edu). The work of Nan and Lin was supported in part by U. S. National Cancer Institute Grant R01 CA76404. The authors thank David Oakes for helpful comments.

1 INTRODUCTION

1.1 The Tremin Trust Data

Considerable interest exists in developing a staging system for female reproductive aging (Mitchell, et al., 2000; Soules, et al., 2001). Such a system will help assess for a woman the need of contraception, the initiation of interventions such as bone density screening, and the approach of menopause. Reproductive life is commonly divided into the reproductive years and the transition years. The transition years include the early and late stages of the menopausal transition. Several bleeding pattern change criteria have been proposed as potential marker events for the early and late stages of the menopausal transition based on expert assessment of clinical studies (Soules, et al., 2001). For example, it has been suggested that age at onset of experiencing a menstrual cycle length at least 45 days might be a good marker for the early menopausal transition (Lisabeth, et al., 2003). However, the validity of these proposed bleeding markers and their associations with age at menopause have not been adequately examined, and formal statistical analysis can be complicated.

The Tremin Trust data were collected as part of the Menstrual and Reproductive Health Study (Treloar et al., 1967). This longitudinal cohort study followed participants throughout their reproductive life span. It provides an unique opportunity to investigate the process of female reproductive aging and the menopausal transition. The study sample consisted of white college students enrolled at the University of Minnesota. Data collection started in 1935 and enrolled a sample of 1,997 women over four years. Study participants were followed up to 40 years. Each woman was asked to use menstrual diary cards to record the days when bleeding was experienced. Menopause was defined as the final menstrual period (FMP), with the FMP confirmed after at least 12 months of amenorrhea. Only limited covariate information, e.g., age at menarche, was available.

Lisabeth et al. (2003) used a subset of the Tremin Trust data to study nine bleeding pattern change criteria for the early and late stages of the menopausal transition proposed by reproductive aging experts (Lisabeth, et al., 2003; Mitchell, et al., 2000; Soules, et al., 2001; Taffe and Dennerstein, 2001), such as age at onset of a 45 day cycle, age at onset of a 60 day cycle, etc. The subset consisted of 562 women in the original study cohort who were age 25 or less at enrollment, had information on age at menarche, and who were still participating in the study at age 35 which was used as the baseline age in their study.

To evaluate the validity of a proposed bleeding marker, one is interested in assessing the association between age at a marker event, defined as age at onset of a specific bleeding pattern change, and age at natural menopause, and studying how this association varies with age at the marker event. In this paper, we focus on age at which a woman first experienced a menstrual cycle at least 45 days in length, which has been proposed as a marker event for entry into the early menopausal transition stage. We call it the 45-day cycle marker.

Both time to a marker event and time to menopause were subject to right censoring in the Tremin Trust data. For each individual, the censoring time was the same for both events. Figure 1 shows the Kaplan-Meier estimates of the survival curves of time to menopause and time to the 45-day cycle marker in years. A total of 193 (34%) women were observed to experience natural menopause, and a total of 357 (64%) women were observed to experience a 45-day cycle marker. The median age at menopause was 51.7 years and the median age at the 45-day cycle marker was 42.7 years.

This problem hence can be formulated as estimating the dependence between censored bivariate survival times. The analysis, however, can be complicated by the observation in Lisabeth et al. (2003) that the dependence between these two event times varies with age at the 45-day cycle marker. Specifically, Lisabeth et al. (2003) explored the dependence between these two event times descriptively using interesting side-by-side box-plots given in Figure 2, which is reproduced from Lisabeth et al. (2003). They divided age into several intervals [35, 38), [38,40), and so on. For each age interval, a pair of box-plots were constructed to compare the distributions of age at menopause between two groups of women: for those who experienced a 45-day marker event within the age interval, their distribution of age at menopause was estimated by the Kaplan-Meier method and plotted in a gray box-plot; for those who had not yet experienced a 45-day marker event by the end of the age interval, their distribution of age at menopause was estimated by the Kaplan-Meier method and plotted in a white box-plot. The dependence of time to marker event and time to menopause can be easily visualized by comparing the paired box-plots in Figure 2. This comparison suggests that age at a 45-day cycle marker is weakly associated with age at menopause before age 40, but their association became much stronger after age 40. This indicates that the 45 day cycle marker might not be useful before age 40 but might be a good marker after age 40. In other words, the association between the two events varies with age at which the 45-day cycle marker occurs.

Figure 2 provides a convenient graphical tool to explore the dependence between those two events. However, Lisabeth et al. (2003) were not able to develop a formal statistical method to quantify this association displayed in Figure 2 and assess statistical significance of this association as a function of age at a 45-day marker event. We develop a statistical method to investigate this issue in this paper.

Our statistical model is motivated by our observation that the comparison between each pair of gray and white box-plots in Figure 2 can be characterized by the hazard ratio of time to menopause comparing women who experienced a 45-day cycle marker in a given age interval with women who have not experienced a 45-day cycle marker by the end of the age interval. As the width of each age interval approaches to zero, this hazard ratio becomes the cross-ratio (Clayton, 1978; Oakes, 1989) of the bivariate failure times: time to the marker event and time to menopause, which will be defined formally in Section 2. Hence quantitative analysis of the dependence observed in Figure 2 can be formulated by modelling cross-ratio as a function of age at marker event.

1.2 Statistical Background

There is a considerable recent literature on bivariate failure time analysis (Kalbfleisch and Prentice, Ch. 10, 2002). Most research in bivariate failure time analysis focuses on either nonparametric estimation of the joint survival function or estimation of regression coefficients in marginal models. The main interest in our application, however, is modelling the dependence between two event times (time to the marker event and time to menopause) as a function of one event time (time to marker event).

Several global dependence measures have been proposed for bivariate failure times, such as Kendall's τ and Spearman's correlation coefficient (Hougaard, Ch. 4, 2000), and average reciprocal cross-ratio (Fan, Hsu, and Prentice, 2000). However, these global measures are not always desirable, since they can mask important features of the data and do not address scientific questions of interest when the dependence of two event times is time-dependent and modelling such dependence is of major interest, as in the analysis of the Tremin Trust data.

Several local dependence measures have also been proposed for bivariate survival data. The cross-ratio function is of particular interest because of its attractive hazard ratio interpretation (Clayton, 1978; Oakes, 1989). When cross-ratio is constant, such as in Clayton model and Gamma frailty model, many estimating methods have been developed. See e.g. Clayton (1978), Oakes (1986b), Nielsen et al. (1992), Shih and Louis (1995), and Glidden (2000), among others. However, little work has been done to estimate the cross-ratio as a function of bivariate event times. In our application in the Tremin Trust data, we are particularly interested in estimating the cross-ratio as a function of age at a 45-day cycle marker event. Estimating cross-ratio as an arbitrary function of event times is very challenging. Hsu et al. (1999) discussed estimation of piecewise constant cross-ratios in a matched case-control setting using a pseudo-partial-likelihood function. However, their method relies heavily on the particular matched case-control design, and is not applicable to unmatched data which is the usual situation for cohort studies.

In this paper, we focus on modelling the dependence of bivariate survival data from cohort studies as a function of one event time using piecewise constants. When there is not covariate, in the discussion of Oakes (1986a), he conjectured that one might extend the constant cross-ratio model by assuming that the cross-ratio is a constant at each rectangle on the plane of bivariate failure times and that a separate distribution, subject to left truncation and right censoring, is defined within each rectangle. We first propose a direct two-stage method by building upon Oakes' proposal. A major limitation of this method is that it is difficult to incorporate covariates in marginal survival models. We then propose a sequential two-stage estimation procedure. A key advantage of this method is that it is applicable to both cases without and with covariates in marginal survival distributions. The basic idea of the sequential method is that one performs estimation sequentially under a modified Clayton model within each interval by using the estimated right boundary survival function from the previous interval as the left marginal survival function of the current interval.

The remainder of the paper is organized as follows. We first introduce in Section 2 our piecewise constant cross-ratio model. We derive in Section 3 the joint survival function of bivariate failure times under the piecewise constant cross-ratio model without/with covariates. We discuss in Section 4 the direct and the sequential two-stage estimation methods. We analyze the Tremin Trust data in Section 5 and conduct a simulation study in Section 6. Finally, we give conclusions in Section 7.

2 THE PIECEWISE CONSTANT CROSS-RATIO MODEL

We first assume there is no covariate and then extend the proposed model to allowing for covariates in marginal survival functions. Suppose that T_1 and T_2 are bivariate failure times. In the Tremin Trust data, T_1 is time to a 45-day cycle marker and T_2 is time to menopause.

Both are subject to right censoring. The cross-ratio (CR) function (Clayton, 1978; Oakes, 1989) for (T_1, T_2) is defined as

$$\theta(t_1, t_2) = \frac{\lambda_1(t_1|T_2 = t_2)}{\lambda_1(t_1|T_2 > t_2)} = \frac{\lambda_2(t_2|T_1 = t_1)}{\lambda_2(t_2|T_1 > t_1)} \ , \tag{1}$$

where $(T_1, T_2) \in [0, \infty) \times [0, \infty)$, and $\lambda_1(\cdot)$ and $\lambda_2(\cdot)$ are conditional hazard functions for T_1 and T_2, respectively. One can easily see that $\theta(t_1, t_2)$ is the relative risk of experiencing menopause at age t_2 comparing women who experiences the 45-day marker event at age t_1 with women who has not experienced the 45-day marker event by age t_1. The two event times are independent if $\theta(t_1, t_2) = 1$; positively correlated if $\theta(t_1, t_2) > 1$ and negatively correlated if $\theta(t_1, t_2) < 1$ (Kalbfleisch and Prentice, 2002).

Motivated by the Tremin Trust application, we assume $\theta(t_1, t_2)$ is a function of t_1, age at marker event,

$$\theta(t_1, t_2) = \theta(t_1). \tag{2}$$

A comparison of each pair of the side-by-side box-plots in Figure 2 is fully determined by cross-ratio $\theta(t_1)$. Specifically, from (1) and (2), one can easily see

$$\bar{F}(t_2|T_1 = t_1) \ = \ \exp\left\{-\int_0^{t_2} \lambda_2(s|T_1 = t_1)\ ds\right\} = \left\{\bar{F}(t_2|T_1 > t_1)\right\}^{\theta(t_1)}, \tag{3}$$

where $\bar{F}(\cdot|\cdot)$ denotes a conditional survival function. Each gray box-plot in Figure 2 represents $\bar{F}(t_2|T_1 = t_1)$ approximately (exactly when age intervals approach to zero), and the corresponding white box-plot represents $\bar{F}(t_2|T_1 > t_1)$. If $\theta(t_1) = 1$, the two survival functions in (3) are the same. If $\theta(t_1) > 1$, then $\bar{F}(t_2|T_1 > t_1) > \bar{F}(t_2|T_1 = t_1)$ for all t_2 and vice versa if $\theta(t_1) < 1$. Examination of Figure 2 suggests that the CR $\theta(t_1, t_2)$ is not a constant, and varies with marker time t_1 as in (2).

We estimate $\theta(t_1)$ nonparametrically by assuming that the cross-ratio $\theta(t_1)$ is a piecewise constant function of t_1. Let $[w_0, w_1)$, ..., $[w_{k-1}, w_k)$, ..., $[w_{K-1}, w_K)$ be a K finite partition of $[0, \infty)$, where $w_0 = 0$ and $w_K = \infty$. Suppose $\Pr(w_{k-1} \le T_1 < w_k) > 0$ for all k, $k = 1, \ldots, K$. We assume that

$$\theta(t_1, t_2) = \theta(t_1) = \theta_k, \quad \text{if } (t_1, t_2) \in A_k, \tag{4}$$

where θ_k is a constant and $A_k = [w_{k-1}, w_k) \times [0, \infty)$ is a strip corresponding to $t_1 \in [w_{k-1}, w_k)$. In practice, the follow-up time for any study is finite. Without loss of generality, we assume

$(T_1, T_2) \in [0, \tau_1] \times [0, \tau_2]$, and $w_K = \tau_1$. As an example, Figure 3(a) illustrates the piecewise constant cross-ratio model (4) for a partition of the support of T_1 into four intervals.

It is of significant practical interest to use the data to examine whether the piecewise constant cross-ratio assumption (4) is appropriate. Equation (3) provides a convenient way to graphically check this assumption. Specifically, simple calculations show that equation (4) can be rewritten as

$$\log[-\log\{\bar{F}(t_2|T_1 = t_1)\}] - \log[-\log\{\bar{F}(t_2|T_1 > t_1)\}] = \log(\theta_k), \tag{5}$$

which is constant for $w_{k-1} \leq t_1 < w_k$. It follows that we can group the data into K groups based on the data of t_1 similarly to that in Figure 1. Within each t_1 interval, plot the pair of estimated survival functions corresponding to the gray and white boxplots on a log(-log) scale and check whether they are parallel. This technique mimics the graphical method for checking the proportional hazards assumption in Cox model. We applied this graphical technique to The treminTrust data and found that the piecewise constant cross-ratio assumption is reasonable. For more detailed discussions, see Section 4.

The above discussion assumes no covariate. It is of interest to extend the proposed model to accommodate covariates in the marginal survival functions $\bar{F}_1(t_1)$ and $\bar{F}_2(t_2)$. In The treminTrust data, the marginal survival distributions of menopause and the 45 day cycle marker event are likely to be affected by age at menarche. To accommodate covariates, we assume the marginal distributions of T_1 and T_2 follow the Cox model as

$$\lambda_j(t_j|Z_j = z_j) = \lambda_{j0}(t_j)e^{z_j^T \beta_j}, \tag{6}$$

where $\lambda_j(t_j|Z_j)$ and $\lambda_{j0}(t_j)$ are the marginal hazards and the marginal baseline hazards for T_j, and Z_j are the covariate vectors associated with T_j, $j = 1, 2$.

3 CONSTRUCTION OF THE JOINT SURVIVAL FUNCTION

3.1 The Joint Survival Function Without Covariates

We are interested in constructing a bivariate survival function $\bar{F}(t_1, t_2|z_1, z_2)$ on the first quadrant of the plane $(t_1 > 0, t_2 > 0)$ using marginal survival functions $\bar{F}_1(t_1|z_1) = \bar{F}(t_1, 0|z_1)$ and $\bar{F}_2(t_2|z_2) = \bar{F}(0, t_2|z_2)$ and the piecewise constant cross-ratio model (4).

We first consider the case without covariates, and construct the bivariate survival function $\bar{F}(t_1, t_2)$ within each strip A_k. The piecewise constant cross-ratio model (4) assumes the cross-ratio $\theta(t_1, t_2)$ equals to the constant θ_k when $(t_1, t_2) \in A_k \equiv [w_{k-1}, w_k) \times [0, \infty)$. Following Clayton (1978) and Oakes (1986, 1989), equation (1) implies

$$f_{A_k}(t_1, t_2)\bar{F}_{A_k}(t_1, t_2) = \theta_k \frac{\partial \bar{F}_{A_k}(t_1, t_2)}{\partial t_1} \frac{\partial \bar{F}_{A_k}(t_1, t_2)}{\partial t_2} \ , \tag{7}$$

where $f_{A_k}(t_1, t_2)$ is the joint density function and $\bar{F}_{A_k}(t_1, t_2)$ the joint survival function of (T_1, T_2) given $(t_1, t_2) \in A_k$. We show in Appendix that the solution of partial differential equation (7) has the following form when $(t_1, t_2) \in A_k$,

$$\bar{F}_{A_k}(t_1, t_2; \theta_k) = \begin{cases} [(\theta_k - 1)\{a(t_1) + b(t_2)\}]^{-1/(\theta_k - 1)}, & \text{if } \theta_k \neq 1, \\ \exp[-\{a^*(t_1) + b^*(t_2)\}], & \text{if } \theta_k = 1; \end{cases} \tag{8}$$

where the univariate functions $a(\cdot)$, $b(\cdot)$, $a^*(\cdot)$, and $b^*(\cdot)$ can be determined by the left and bottom boundary conditions of A_k. We illustrate in Figure 3(b) the strips A_k and their left and bottom boundaries, denoted by L_k and B_k respectively. Then $a(\cdot)$, $b(\cdot)$, $a^*(\cdot)$, and $b^*(\cdot)$ can be determined by the marginal survival functions on the left and bottom boundaries L_k and B_k. We discuss the forms of $\bar{F}_{A_k}(t_1, t_2; \theta_k)$ for the following two cases of particular interest. Details can be found in Appendix.

The Classical Clayton Model

Let $K = 1$, i.e., θ is constant at any point (t_1, t_2) and the only strip A_1 becomes the whole first quadrant. The marginal survival functions at the two boundaries L and B become the marginal survival functions of T_1 and T_2: $\bar{F}(t_1, 0) = \bar{F}_1(t_1)$ and $\bar{F}(0, t_2) = \bar{F}_2(t_2)$. Then if $\theta \neq 1$, calculations in Appendix show that $\bar{F}(t_1, t_2)$ in equation (8) becomes Clayton's copula model (Clayton, 1978; Oakes, 1989) as, for $t_1, \ t_2 \geq 0$,

$$\bar{F}(t_1, t_2) = \left[\{\bar{F}_1(t_1)\}^{-(\theta-1)} + \{\bar{F}_2(t_2)\}^{-(\theta-1)} - 1\right]^{-1/(\theta-1)}. \tag{9}$$

When $\theta = 1$, $\bar{F}(t_1, t_2) = \bar{F}_1(t_1)\bar{F}_2(t_2)$, which means that T_1 and T_2 are independent.

The Piecewise Clayton Model with Left Truncation

Suppose there are K strips and the cross-ratio is a constant θ_k in the kth strip A_k, $k = 1, \cdots, K$. If the two boundary conditions $B_k : \bar{F}_{A_k}(t_1, 0)$ and $L_k : \bar{F}_{A_k}(w_{k-1}, t_2)$ are given, the joint survival function $\bar{F}_{A_k}(t_1, t_2)$ for $(t_1, t_2) \in A_k$ has the the following form when $\theta_k \neq 1$,

$$\begin{aligned} \bar{F}_{A_k}(t_1, t_2) \ = \ & \Big[\{\bar{F}_{A_k}(t_1, 0)\}^{-(\theta_k-1)} + \{\bar{F}_{A_k}(w_{k-1}, t_2)\}^{-(\theta_k-1)} \\ & - \{\bar{F}_{A_k}(w_{k-1}, 0)\}^{-(\theta_k-1)}\Big]^{-1/(\theta_k-1)} \end{aligned} \tag{10}$$

Let $t_1 = w_{k-1} + \tilde{t}_1$ and $t_2 = \tilde{t}_2$. Thus $(\tilde{t}_1, \tilde{t}_2) \in \tilde{A}_k \equiv [0, w_k - w_{k-1}) \times [0, \infty)$. Define

$$\begin{aligned} \tilde{F}_{\tilde{A}_k}(\tilde{t}_1, \tilde{t}_2) &= \Pr(T_1 > w_{k-1} + \tilde{t}_1, T_2 > \tilde{t}_2 | T_1 > w_{k-1}, T_2 > 0) \\ &= \bar{F}_{A_k}(w_{k-1} + \tilde{t}_1, \tilde{t}_2) / \bar{F}_{A_k}(w_{k-1}, 0) \ , \end{aligned} \tag{11}$$

which is the bivariate survival function for left truncated failure times T_1 and T_2 truncated at $(w_{k-1}, 0)$ (Kalbfleisch and Prentice, 2002). It follows that equation (10) can be rewritten as

$$\tilde{F}_{\tilde{A}_k}(\tilde{t}_1, \tilde{t}_2) = \left[\{\tilde{F}_{\tilde{A}_k}(\tilde{t}_1, 0)\}^{-(\theta_k - 1)} + \{\tilde{F}_{\tilde{A}_k}(0, \tilde{t}_2)\}^{-(\theta_k - 1)} - 1 \right]^{-1/(\theta_k - 1)} \ , \tag{12}$$

where $\tilde{F}_{\tilde{A}_k}(\tilde{t}_1, 0)$ and $\tilde{F}_{\tilde{A}_k}(0, \tilde{t}_2)$ are fully determined by the original two boundary conditions $\bar{F}_{A_k}(t_1, 0)$ and $\bar{F}_{A_k}(w_{k-1}, t_2)$ at the boundaries L_k and B_k. Hence the joint survival function for left truncated failure times has exactly the same form as the Clayton model (9) when $(\tilde{t}_1, \tilde{t}_2) \in \tilde{A}_k$ or equivalently $(t_1, t_2) \in A_k$. They are identical for $(t_1, t_2) \in A_1$ since $\bar{F}_{A_1}(0, 0) = 1$. When $\theta_k = 1$ and $(t_1, t_2) \in A_k$, we have $\bar{F}_{A_k}(t_1, t_2) = \bar{F}_{A_k}(t_1, 0)\bar{F}_{A_k}(w_{k-1}, t_2)/\bar{F}_{A_k}(w_{k-1}, 0)$ i.e., $\tilde{F}_{\tilde{A}_k}(\tilde{t}_1, \tilde{t}_2) = \tilde{F}_{\tilde{A}_k}(\tilde{t}_1, 0)\tilde{F}_{\tilde{A}_k}(0, \tilde{t}_2)$. This means that T_1 and T_2 are independent in A_k.

It should be noted that the two marginal survival functions $\bar{F}_1(t_1)$ and $\bar{F}_2(t_2)$ and the piecewise cross-ratio model (4) fully determine the bivariate survival function $F(t_1, t_2)$ on the whole support of (t_1, t_2): $t_1 > 0$, $t_2 > 0$. We only need to demonstrate that the bottom and left boundary conditions $B_k : \bar{F}_{A_k}(t_1, 0)$ and $L_k : \bar{F}_{A_k}(w_{k-1}, t_2)$ of each strip A_k are fully determined by the two marginal survival functions $\bar{F}_1(t_1)$ and $\bar{F}_2(t_2)$ and the cross-ratios θ_j, $j < k$, of the previous strips.

We refer to Figure 3(b) for the following discussion. First notice that the marginal survival function $\bar{F}_1(t_1)$ fully specifies the boundary conditions at $B_1, \cdots, B_K$. We only need to show how the boundary conditions at $L_1, \cdots, L_K$ are determined. Starting from the first strip A_1, the two marginal survival functions $\bar{F}_1(t_1)$ and $\bar{F}_2(t_2)$ specify the boundary conditions on L_1 and B_1. Given the cross-ratio parameter θ_1, the bivariate survival function in A_1 is $\bar{F}(t_1, t_2; \theta) = \bar{F}_{A_1}(t_1, t_2; \theta_1)$ which is specified by (12) with $\tilde{t}_1 = t_1$ and $\tilde{t}_2 = t_2$. Thus the right boundary of strip A_1 is fully determined. Notice that the right boundary of A_1 is the left boundary of A_2. It follows that the left boundary condition on L_2 of A_2 is fully determined. With the two known boundary conditions on L_2 and B_2 and the cross-ratio constant θ_2 in A_2, equation (12) specifies the bivariate survival function in A_2 as $\bar{F}(t_1, t_2; \theta) = \bar{F}_{A_2}(t_1, t_2; \theta_2)$. This now specifies the boundary condition on L_3. By applying this sequentially, the bivariate survival function $\bar{F}(t_1, t_2)$ is fully specified on the whole positive plane $t_1, t_2 > 0$. One can easily show that the resulting function $\bar{F}(t_1, t_2)$ constructed in such a piecewise fashion is a proper bivariate survival function as long as $\bar{F}_1(t_1)$ and $\bar{F}_2(t_2)$ are proper survival functions, i.e., $\bar{F}(t_1, t_2)$ is a non-increasing function satisfying $\bar{F}(0,0) = 1$ and $\bar{F}(\infty, \infty) = 0$. These calculations consist of the foundation of the sequential estimation method proposed in Section 4.

3.2 Extension to the Joint Survival Function with Covariates

The above results can be easily extended to construct the conditional bivariate survival function when the marginal survival functions of T_1 and T_2 depend on covariates Z_1 and Z_2 through the marginal Cox models in (6). Let $Z = (Z_1, Z_2)$ be the whole covariate vector. Consider the kth strip A_k. Suppose that the two boundary conditions $B_k : \bar{F}_{A_k}(t_1, 0|z)$ and $L_k : \bar{F}_{A_k}(w_{k-1}, t_2|z)$ are given. Define $\tilde{F}_{\tilde{A}_k}(\tilde{t}_1, \tilde{t}_2|z)$ similarly to (11) but conditional on $Z = z$. Similar calculations to Section 3.1 show that

$$\begin{aligned}
\tilde{F}_{\tilde{A}_k}(\tilde{t}_1, \tilde{t}_2|z) &= \bar{F}_{A_k}(w_{k-1} + \tilde{t}_1, \tilde{t}_2|z)/\bar{F}_{A_k}(w_{k-1}, 0|z) \ , & (13)\\
&= \left[\{\tilde{F}_{\tilde{A}_k}(\tilde{t}_1, 0|z)\}^{-(\theta_k-1)} + \{\tilde{F}_{\tilde{A}_k}(0, \tilde{t}_2|z)\}^{-(\theta_k-1)} - 1\right]^{-1/(\theta_k-1)} \ , & (14)
\end{aligned}$$

which takes the same form as (12) except that the two boundary conditions $\bar{F}_{A_k}(t_1, 0|z)$ and $\bar{F}_{A_k}(w_{k-1}, t_2|z)$ depend on covariates.

Now examine these boundary conditions $\bar{F}_{A_k}(t_1, 0|z)$ and $\bar{F}_{A_k}(w_{k-1}, t_2|z)$. Since the two marginal survival function $\bar{F}_1(t_1|z_1) = \bar{F}(t_1, 0|z)$ follows Cox model (6), simple calculation shows that the left truncated marginal survival function $\tilde{F}_{\tilde{A}_k}(\tilde{t}_1, 0|z)$ at the bottom boundary B_k of the kth strip A_k also follows Cox model. However, although the marginal survival function $\bar{F}_2(t_2|z_2) = \bar{F}(0, t_2|z)$ follows Cox model, the left truncated marginal survival function $\bar{F}_{\tilde{A}_k}(0, \tilde{t}_2|z)$ on the left boundary L_k of A_k generally does not follow Cox model any more for $k > 1$. Specifically, we can show that $\tilde{F}_{\tilde{A}_k}(0, \tilde{t}_2|z)$ takes the form

$$\begin{aligned}
\bar{F}_{\tilde{A}_k}(0, \tilde{t}_2|z) &= \bar{F}_{A_k}(w_{k-1}, t_2|z)/\bar{F}_{A_k}(w_{k-1}, 0|z)\\
&= \Big[1 + \{\bar{F}_{A_{k-1}}(w_{k-2}, t_2|z)/\bar{F}_{A_{k-1}}(w_{k-1}, 0|z)\}^{-(\theta_{k-1}-1)}\\
&\qquad - \{\bar{F}_{A_{k-1}}(w_{k-2}, 0|z)/\bar{F}_{A_{k-1}}(w_{k-1}, 0|z)\}^{-(\theta_{k-1}-1)}\Big]^{-1/(\theta_{k-1}-1)}, \quad (15)
\end{aligned}$$

which obviously does not have the form of Cox model when $\theta_j \neq 1$ for some $j < k$. This result complicates the estimation procedure in the presence of covariates. However, it should be noted that from (15), using similar arguments as those in Section 3.1 one can show that the function $\bar{F}(t_1, t_2|z)$ is a fully specified proper survival function determined by the cross-ratio model (4) and the two marginal Cox models (6).

4 PIECEWISE CONSTANT CROSS-RATIO ESTIMATION

We discuss two methods in this section for estimating cross-ratio parameters $\theta = (\theta_1, \cdots, \theta_K)^T$: the direct two-stage method and the sequential two-stage method. The direct two-stage method is simple and is applicable when there is no covariate. However, the estimated survival function might not be a proper survival function in finite samples, and it is not applicable in presence of covariates. The sequential method is slightly more complicated but is applicable no matter covariates are involved or not, and the estimated survival function is always proper.

4.1 The Direct Two-Stage Method

We first consider the case without covariates. In the discussions of Oakes (1986a), he conjectured that one might define a separate distribution for T_1 and T_2 in each strip A_k, $k = 1, \ldots, K$, subject to left truncation at its left boundary and right censoring at its right boundary, and estimate the cross-ratio θ_k using the left truncated and right censored data in A_k. We fully develop this idea in this section using the results in Section 3 and term it as the direct two-stage method.

Specifically, for each strip A_k, we first construct a left truncated and right censored data set. Since the resulting left truncated data follow Clayton model (12), we adopt the two-stage method of Shih and Louis (1995) by estimating the two boundary conditions $\tilde{F}_{\tilde{A}_k}(\tilde{t}_1, 0)$ and $\tilde{F}_{\tilde{A}_k}(0, \tilde{t}_2)$ nonparametrically at the first stage and then estimating the cross-ratio θ_k using the likelihood specified by (12) at the second stage.

Suppose that there are n subjects. We introduce subscript i to indicate subject i. Let Δ_{ji} be the censoring indicator and $Y_{ji} = \min(T_{ji}, C_{ji})$ be the observed time for the jth failure time of subject i, where T_{ji} is the jth failure time and C_{ji} is the jth censoring time, $i = 1, \ldots, n$, $(j = 1, 2)$. For example, in The treminTrust data, T_{1i} is age at a 45-day cycle marker event and T_{2i} is age at menopause of subject i.

For each strip A_k, we construct a data set using the original cohort data by left truncating Y_{1i} at w_{k-1} and right censoring Y_{1i} at w_k. Specifically, we consider a subset of n_k subjects whose $Y_{1i} \geq w_{k-1}$. Let the new survival time be $\tilde{T}_{1i} = T_{1i} - w_{k-1}$, the new observed time be $\tilde{Y}_{1i} = \min(Y_{1i} - w_{k-1}, w_k - w_{k-1})$ and the new censoring indicator be $\tilde{\Delta}_{1i} = 1$ if $\tilde{Y}_{1i} = \tilde{T}_{1i}$ and 0 otherwise. All the variables associated with the second failure time T_2 remain the same, i.e., $\tilde{Y}_{2i} = Y_{2i}$ and $\tilde{\Delta}_{2i} = \Delta_{2i}$ for all $i = 1, \ldots, n_k$. Denote by $\mathcal{D}_k = \{\tilde{Y}_{1i}, \tilde{\Delta}_{1i}, \tilde{Y}_{2i}, \tilde{\Delta}_{2i}; i = 1, \cdots, n_k\}$ the resulting left truncated and right censored data set that will be used for estimation in

A_k. For The treminTrust data, $\mathcal{D}_k$ contains the subset of women who had not experienced a 45-day cycle marker event by age w_{k-1}. For those who had not experienced the marker event by age w_k, their times to the marker event, T_1, are censored at w_k. The second failure time T_2 is time to menopause.

At the first stage, we use the data set $\mathcal{D}_k$ to estimate the two marginal survival functions on the left and bottom boundaries L_k and B_k of the strip A_k. They can be written as

$$\tilde{\bar{F}}_{\tilde{A}_k}(\tilde{t}_1, 0) = \frac{\bar{F}_{A_k}(t_1, 0)}{\bar{F}_{A_k}(w_{k-1}, 0)}, \quad \tilde{\bar{F}}_{\tilde{A}_k}(0, \tilde{t}_2) = \frac{\bar{F}_{A_k}(w_{k-1}, t_2)}{\bar{F}_{A_k}(w_{k-1}, 0)}. \tag{16}$$

It can be easily seen that $\tilde{\bar{F}}_{\tilde{A}_k}(\tilde{t}_1, 0)$ and $\tilde{\bar{F}}_{\tilde{A}_k}(0, \tilde{t}_2)$ can be estimated nonparametrically simply by Kaplan-Meier estimates using data $(\tilde{Y}_{1i}, \tilde{\Delta}_{1i})$ and $(\tilde{Y}_{2i}, \tilde{\Delta}_{2i})$ respectively (Kalbfleisch and Prentice, 2002). It is worth noting that the Kaplan-Meier estimator of $\tilde{\bar{F}}_{\tilde{A}_k}(\tilde{t}_1, 0)$ using $(\tilde{Y}_{1i}, \tilde{\Delta}_{1i})$ can be equivalently calculated from the Kaplan-Meier estimator of the marginal survival function $\bar{F}_1(t_1) = \bar{F}(t_1, 0)$ using all the data (Y_{1i}, Δ_{1i}) in the original cohort.

At the second stage, we use the data set $\mathcal{D}_k$ to construct and maximize the likelihood function specified using bivariate survival function (12) with respect to θ_k by treating the two estimated marginal survival functions $\widehat{\tilde{\bar{F}}}_{\tilde{A}_k}(\tilde{t}_1, 0)$ and $\widehat{\tilde{\bar{F}}}_{\tilde{A}_k}(0, \tilde{t}_2)$ as fixed. Following Shih and Louis (1995), one can easily show that the likelihood function involving θ_k can be written as

$$\begin{aligned} L(\theta_k) \quad = \quad & \prod_{i=1}^{n_k} \left\{ \frac{\partial^2 H(u_i, v_i; \theta_k)}{\partial u_i \partial v_i} \right\}^{\tilde{\Delta}_{1i} \tilde{\Delta}_{2i}} \left\{ \frac{-\partial H(u_i, v_i; \theta_k)}{\partial u_i} \right\}^{\tilde{\Delta}_{1i}(1-\tilde{\Delta}_{2i})} \\ & \times \left\{ \frac{-\partial H(u_i, v_i; \theta_k)}{\partial v_i} \right\}^{(1-\tilde{\Delta}_{1i})\tilde{\Delta}_{2i}} \{H(u_i, v_i; \theta_k)\}^{(1-\tilde{\Delta}_{1i})(1-\tilde{\Delta}_{2i})} \end{aligned} \tag{17}$$

where $H(u, v; \theta) = \left\{u^{-(\theta-1)} - v^{-(\theta-1)} - 1\right\}^{-1/(\theta-1)}$, $u_i = \widehat{\tilde{\bar{F}}}_{\tilde{A}_k}(\tilde{Y}_{1i}, 0)$ and $v_i = \widehat{\tilde{\bar{F}}}_{\tilde{A}_k}(0, \tilde{Y}_{2i})$. The standard error of the resulting maximum likelihood estimator $\widehat{\theta}_k$ can be estimated using the method similarly to that described in Shih and Louis (1995).

The direct two-stage method is easy to implement. However, it has two major limitations. First, the estimated bivariate survival function $\widehat{\bar{F}}(t_1, t_2)$ might not be a proper survival function in finite samples since it views the survival functions on A_k, $k = 1, \ldots, K$, as separate functions and might have positive jumps at the boundaries L_k. Second, it is not straightforward to be extended to the case with covariates.

With the classical Clayton model where cross-ratio θ is constant on $[0, \infty) \times [0, \infty)$, Glidden (2000) extended the two-stage method of Shih and Louis (1995) by allowing for covariates in the marginal survival functions $\bar{F}_1(t_1)$ and $\bar{F}_2(t_2)$ using Cox model (6). Specifically, one can simply fit Cox models to estimate the two marginal conditional survival functions at the first stage and perform the same maximum likelihood calculations for θ at the second stage. Under piecewise constant cross-ratio assumption (4), when covariates are involved in the marginal survival functions $\bar{F}_1(t_1|z)$ and $\bar{F}_2(t_2|z)$ under Cox model (6), the bivariate survival function of $\tilde{T}_{1i}$ and $\tilde{T}_{2i}$ conditional on covariates still follows Clayton model (12). However, for each strip A_k, $k > 1$, the marginal survival function $\tilde{F}_{\tilde{A}_k}(0, \tilde{t}_2|z)$ on the left boundary L_k of A_k in (12) does not follow the Cox model any more. Its form is more complicated as seen in (15), and its components depend on the two marginal survival functions $\bar{F}_1(t_1)$ and $\bar{F}_2(t_2)$ and all θ_j's, $j = 1, \ldots, k$. Hence, if one uses the direct two-stage method by simply fitting the marginal Cox models using data $(\tilde{Y}_{1i}, \tilde{\Delta}_{1i}, Z_{1i})$ and $(\tilde{Y}_{2i}, \tilde{\Delta}_{2i}, Z_{2i})$ at the first stage, then the MLE estimator of θ_k at the second stage would be inconsistent.

4.2 The Sequential Two-Stage Method

We propose in this section a sequential two-stage method to overcome the two major limitations of the direct two-stage method. This method ensures the estimated bivariate survival function is a proper survival function and accommodates covariates in the marginal survival functions under Cox model (6).

The sequential two-stage method differs from the direct two-stage method by calculating the left and bottom boundary conditions at each strip sequentially from the first strip A_1 to the last strip A_K. Unlike the direct two-stage method, which treats $\bar{F}(t_1, t_2)$ separately in each strip A_k, the sequential two-stage method views $\bar{F}(t_1, t_2)$ as one whole survival function in $[0, \infty) \times [0, \infty)$ under the piecewise constant cross-ratio model (4) and specifies $\bar{F}(t_1, t_2)$ sequentially using the procedure layed out at the end of Section 3.1. Hence $\bar{F}(t_1, t_2)$ is seamless at the left boundary of A_k's, i.e., the right boundary condition of A_k and the left boundary condition of A_{k+1} are identical.

The main idea of the sequential two-stage method can be easily illustrated by referring to Figure 3(b). We first estimate the two marginal survival functions $\bar{F}_1(t_1)$ and $\bar{F}_2(t_2)$ or $\bar{F}_1(t_1|Z_1)$ and $\bar{F}_2(t_2|Z_2)$, which specify the boundary conditions on L_1 and all B_k's, $k = 1, \cdots, K$. We start from the first strip A_1 whose boundary conditions L_1 and B_1 are given. Estimation proceeds sequentially from strip A_1 to strip A_k. Within each strip A_k, given the left and bottom boundary conditions L_k and B_k, we estimate θ_k and $\bar{F}_{A_k}(t_1, t_2)$ using the left truncated and right censored data $\mathcal{D}_k$ under Clayton model (12) or (14). The survival function on the right boundary of A_k is then available and specifies the left boundary condition L_{k+1} of the next strip A_{k+1}. One hence can estimate the left boundary condition L_k of each strip sequentially and obtain estimators of θ_k and $\bar{F}_{A_k}(t_1, t_2)$ sequentially.

We describe below in detail the sequential two-stage method. For simplicity, We focus on the case with covariates. The method for the case without covariates is the same except that the two marginal survival functions $\bar{F}_1(t_1) = \bar{F}(t_1, 0)$ and $\bar{F}_2(t_2) = \bar{F}(0, t_2)$ are estimated using Kaplan-Meier method.

Step 1. Estimate the two marginal survival functions. We fit the marginal Cox models (6) using $(Y_{1i}, \Delta_{1i}, Z_{1i})$ and $(Y_{2i}, \Delta_{2i}, Z_{2i})$ respectively, $i = 1, \cdots, n$, and calculate the estimators of marginal survival functions $\bar{F}_1(t_1|z) = \bar{F}(t_1, 0|z)$ and $\bar{F}_2(t_2|z) = \bar{F}(0, t_2|z)$, where the marginal baseline hazards are estimated using Breslow estimator and the regression coefficients β_1 and β_2 are estimated using the partial likelihood method. These calculations provide us the estimators of the left boundary condition $\bar{F}_{A_1}(0, t_2|z)$ on L_1 of strip A_1 and all the bottom boundary conditions $\bar{F}_{A_k}(t_1, 0|z)$ on B_k of strips A_k, $k = 1, \ldots, K$.

Step 2. Estimation of θ_1 and $\bar{F}_{A_1}(t_1, t_2|z)$. Use the right-censored data $\mathcal{D}_1$ and treat the estimators of the left boundary condition $\bar{F}_{A_1}(0, t_2|z)$ on L_1 and the right boundary condition $\bar{F}_{A_1}(t_1, 0|z)$ on B_1 obtained from Step 1 as known. Calculate the maximum likelihood estimator of θ_1 by maximizing the likelihood (17). Then estimate the bivariate survival function $\bar{F}_{A_1}(t_1, t_2|z)$ from equation (14).

Step 3. Estimation of θ_k for $k > 1$. Suppose one obtains the estimators of θ_{k-1} and the bivariate survival function $\bar{F}_{A_{k-1}}(t_1, t_2|z)$ in strip A_{k-1}.

(i) Use the estimator of bivariate survival function $\bar{F}_{A_{k-1}}(t_1, t_2|z)$ to compute the estimator of marginal survival function on the right boundary of A_{k-1}, denoted as $\hat{\bar{\bar{F}}}_{A_{k-1}}(w_{k-1}, t_2|z)$. This gives the estimator of the survival function on the left boundary L_k of strip A_k, i.e., $\hat{\bar{\bar{F}}}_{A_{k-1}}(w_{k-1}, t_2|z) = \hat{\bar{\bar{F}}}_{A_k}(w_{k-1}, t_2|z)$. Convert $\hat{\bar{\bar{F}}}_{A_k}(w_{k-1}, t_2|z)$ to the marginal survival function for the left truncated failure time $\tilde{T}_{2i}$ using equation (13).

(ii) Plug the estimators of the left marginal survival function $\tilde{F}_{\tilde{A}_k}(w_k, \tilde{t}_2|z)$ obtained in (i) and the bottom marginal survival function $\tilde{F}_{\tilde{A}_k}(\tilde{t}_1, 0|z)$ obtained in Step 1 into the likelihood function (17) using data $\mathcal{D}_k$ and treat them as known. Maximize (17) with respect to θ_k. The standard error estimator of $\widehat{\theta}_k$ depends on the variability of all the estimated marginal survival functions and cross-ratios θ_j's for all$j < k$ of the previous strips, and can thus be very complicated. We hence use bootstrap to calculate the standard error estimator of $\widehat{\theta}_k$.

Step 4. Let $k = k + 1$, and iterate Step 3 until $k = K$.

As discussed at the beginning of Section 4.2, comparing to the direct method, the main advantages of the sequential method are that it ensures the estimated survival function to be proper and is able to accommodate covariates in the marginal survival functions. It is also worth noting that in the absence of covariates, if the piecewise cross-ratio model (4) holds, the estimators of θ_k's under the sequential method might be more efficient than those given by the direct method. The direct method on the other hand might be more robust against model misspecification of piecewise constant cross-ratios, since estimation of the left margin of A_k by the direct method does not depend on the estimators of the previous strips.

A brief discussion of the asymptotic properties of the direct and the sequential two stage estimators is given in the Appendix. The estimators $\hat{\theta}_k$, $k = 1, \ldots, K$, are consistent and each has an asymptotically normal distribution. Their bootstrap standard deviation estimators are also consistent.

5 ANALYSIS OF THE TREMIN TRUST DATA

We applied both direct and sequential two-stage methods to the analysis of The treminTrust data discussed in Section 1.1. Our main interest was to estimate the piecewise constant cross-ratios for assessing the association between time to the 45-day cycle marker event and time to menopause. Our secondary interest was to estimate the survival distribution of menopause given a woman's age at her 45-day cycle marker event. In our analysis, time zero was defined as age 35, T_1 was time to the 45-day cycle marker event and T_2 was time to menopause. Both event times were subject to censoring. The detailed descriptive statistics of the data can be found in Section 1.1.

We first ignored covariates by assuming the marginal distributions of time to the 45 day cycle marker and time to menopause did not depend on covariates, and analyzed the data under piecewise constant cross-ratio model (4). We began with considering two different partitions for the time axis of the 45-day cycle marker to investigate how sensitive the results were with respect to different partitions. The first partition consisted of one-year intervals from age 35 to 49 and the interval of age 50 and above. The second partition was exactly the same as that used by Lisabeth et al. (2003) in Figure 2, which consisted of 8 intervals including the interval of 50+ that was not shown in the figure. We found both partitions gave similar estimators of the cross-ratios θ's. To save some space, we only present here the results using the second partition which corresponds to the partition used in Figure 2.

We next applied the graphical technique discussed in Section 2 to examine the piecewise cross-ratio assumption. Specifically, using the results in equation (5), we plotted in Figure 4 the paired log(-log) conditional survival functions corresponding to the paired gray and white boxplots in Figure 1. The results in Figure 4 suggest that the curves are roughly parallel within each pair. This provides empirical evidence that the piecewise constant cross-ratio assumption (4) is appropriate for the Tremin Trust data.

We then applied both the direct two-stage method and the sequential two-stage method proposed in Section 4 to estimating the cross-ratios θ_k. The results are presented in Table 1. For the direct two-stage method, both the model-based standard errors and the bootstrap standard errors were calculated. For the sequential two-stage method, only the bootstrap standard errors were calculated. The bootstrap standard error estimators were obtained based on 1000 bootstrap samples. The results in Table 1 suggest that the cross-ratio estimates are very close for both methods.

For the two age intervals of experiencing a 45 day cycle marker event before age 40, the cross-ratio estimates were similar and were both a little less than 1, suggesting a weak negative association between time to the 45-day cycle marker and time to menopause before age 40. For the three intervals of T_1 between 40 and 45, the cross-ratio estimates were similar and were close to 2, suggesting a strong positive association between the two events. For the two intervals of T_1 between 46 and 49, the cross-ratio estimates were both larger than 3, suggesting a very strong positive association between the two events. For the interval of $T_1 \geq 50$, the cross-ratio estimate was reduced to 1.57, suggesting a weak positive association between the two events. Under the direct method, the model based standard error estimates of θ_k's of Shih and Louis (1995) were smaller than their bootstrap counterparts. This discrepancy might be due to instability of the estimates and fine partitions. The bootstrap standard error estimates using the direct and the sequential methods, however, were similar.

The above discussions suggest that we might consider a wider partition by assuming the cross-ratio $\theta(t_1)$ to be piecewise constants in four intervals of T_1: 35-39, 40-45, 46-49 and 50+. The resulting cross-ratio estimates are likely to be more stable. The results for this partition calculated using both direct and sequential methods are also presented in Table 1. Under the direct method, both the model based standard error estimates and the bootstrap standard error estimates now agreed well with each other. The cross-ratio estimates and their standard errors using both methods were also similar.

To examine whether it was appropriate to group the original 8 intervals into the wider 4 intervals, we performed formal statistical tests. Specifically, we were interested in testing whether the cross-ratios corresponding to the 8 interval partition were the same within each of the final four intervals, i. e., we considered three null hypotheses (a) $H_0 : \theta_1 = \theta_2$; (b) $H_0 : \theta_3 = \theta_4 = \theta_5$; and (c) $H_0 : \theta_6 = \theta_7$, and performed a test for each H_0. Under each H_0, we used the point estimates of corresponding θ's and their bootstrap covariance estimates, and performed a chi-square test for the specified contrast. For example, to test (a) $H_0 : \theta_1 = \theta_2$, we calculated the contrast $T = (-1, 1)(\widehat{\theta}_1, \widehat{\theta}_2)'$ and $\text{var}(T) = (-1, 1)\text{cov}\{\widehat{\theta}_1, \widehat{\theta}_2)\}(-1, 1)'$. We then used the test statistic $T^2/\text{var}(T)$, which follows a chi-square distribution with one degree of freedom asymptotically. Note that the chi-square test for the second null hypothesis (b) was based on a two degree-of-freedom test. The p-values for these three null hypotheses were 0.90, 0.91, 0.54 using the direct method and 0.93, 0.91, 0.29 using the sequential method. They suggested that the four interval partition was appropriate and it was reasonable to assume a constant cross-ratio within each of the final four intervals.

We now interpret the cross-ratio estimates. Our results suggest that time to the 45-day cycle marker event and time to menopause are weakly negatively associated (CR=0.80, 95% CI=(0.60, 1.00)) if a woman experiences the 45 day cycle marker event before age 40. Hence the 45 day cycle marker is not very useful for assessing menopause age if it occurs before age 40. Between age 40 and 49, the association of the two event times becomes positive and strong with the strongest association observed between age 46 and 49 (CR=2.17, 95% CI=(1.47, 2.87) if $40 \leq T_1 \leq 45$; CR=4.11, 95% CI=(2.41,5.81) if $46 \leq T_1 \leq 49$). These cross-ratios can be interpreted as the relative risks. For example, CR=2.17 means the risk of experiencing menopause at any given age for a woman who experiences the 45-day cycle marker event at any time $t_1 \in [40, 45)$ is 2.17 times higher than a woman who experiences the 45-day cycle marker event after age t_1. This strong positive association indicates if a woman experiences the 45-day cycle marker between age 40 and 49, the earlier she experiences the marker event, the earlier she will experience menopause. Hence the 45-day cycle marker event is useful for assessing menopause age in this age interval. After age 50, the estimated cross-ratio declines towards 1 and is not statistically significant. This indicates that the positive association of the two event times diminishes after age 50, and the 45-day cycle marker after age 50 is not particularly useful for assessing age at onset of menopause.

Our secondary interest was to estimate the survival function for time to menopause (T_2) given age at the 45-day cycle marker event (T_1), i.e., we were interested in estimating $\bar{F}(T_2|T_1 = t_1)$. This estimation is of both clinical and a woman's own interests. For example, if a woman first experiences a 45-day cycle at a certain age, say 45, she would like to know her expected median age of menopause. This information is also helpful for clinicians to evaluate a woman's need for continuing contraception and the appropriateness of initiating interventions such as bone density screening.

We used equation (3) to estimate the conditional survival function $\bar{F}(T_2|T_1 = t_1)$, since the conditional survival function $\bar{F}(t_2|T_1 > t_1)$ could be estimated either nonparametrically using left truncated data or using model (10). We computed the estimates of the conditional survival function $\bar{F}(t_2|T_1 = t_1)$ at $t_1 = 36, 39, 42, 45, 48, 51$, where the cross-ratios were estimated using the sequential two-stage method, and the conditional survival function $\bar{F}(t_2|T_1 > t_1)$ was estimated nonparametrically using the Kaplan-Meier method. The estimated survival curves are plotted in Figure 5, and several estimated percentiles are presented in Table 2. For example, if a woman experiences the 45-day cycle marker at age 36, 39, 42, 45, 48, or 51, her median age of menopause is expected to be 52.2, 52.2, 50.2, 51.1, 51.7, or 53.9. One can easily see from Figure 5 that the survival distribution of age at onset of menopause is similar for women experiencing the 45 day cycle marker before age 40, e.g., comparing the $T_1 = 36$ and 39 curves. This result is consistent with the weak cross-ratio estimates before age 40. Women who experience the 45 day cycle marker before age 40 are likely to have a later onset of menopause than women who experience the 45 day cycle marker between 40 and 49. Among women who experience the 45 day cycle marker between 40 and 49, a later onset of the marker event would imply a later onset of menopause.

We next incorporated the covariate, age at menarche, in our analysis. Specifically, both event times, time to the 45-day cycle marker and time to menopause were assumed to marginally follow the Cox proportional hazards model (6), while the cross-ratios were assumed to follow the piecewise constant model (4). Only the sequential two-stage method was applicable in the presence of covariates.

Similarly to the analysis without covariate, we assumed $\theta(t_1)$ be piecewise constants in four intervals: 36-39, 40-45, 46-49, and 50+. The cross-ratio estimates and their standard error estimates are also given in Table 1. The results were rather similar to those without covariate.

This was not surprising for the Tremin Trust data since the association of age at menarche with age at menopause was of borderline significance (relative risk =0.89 and p-value=0.038), and age at menarche was not significantly associated with age at 45-day cycle marker event (relative risk =0.94 and p-value=0.17). The estimated conditional survival distributions of age at menopause given age at the 45-day cycle marker were similar to those in Figure 5. Due to space limitation, these conditional survival curves are not presented here. However, we present in Table 2 the estimated percentiles of age at menopause given a series of values of age at the 45-day cycle marker and the median age of menarche that is 12 years old. One can see that the results are similar to those without covariate. The interpretation of the results are similar to those without covariate and thus are omitted.

6 SIMULATION STUDIES

We conducted simulation studies to examine the finite sample performance of the direct and the sequential two-stage methods without/with covariates. We first considered the case without covariates. The marginal distributions of T_1 and T_2 were specified as unit exponential, and the distributions of the two censoring times were assumed to be independent and uniformly distributed over $(0, 2.3)$ as in Shih and Louis (1995) and Glidden (2000).

To mimic the analysis results of the Tremin Trust data, we assumed the cross-ratio $\theta(t_1)$ was piecewise constants over four intervals: $\theta(t_1) = 0.9$ when $t_1 \in [0, 0.25)$, $\theta = 2.0$ when $t_1 \in [0.25, 0.5)$, $\theta = 4.0$ when $t_1 \in [0.5, 0.75)$, and $\theta = 1.5$ when $t_1 > 0.75$. To generate bivariate random variables (T_1, T_2) following the piecewise Clayton model (10), we first generated two independent uniform $(0, 1)$ variables U_1 and U_2. We then set $T_1 = -\log U_1$ and calculated T_2 by solving the equation $\bar{F}(T_2|T_1) = U_2$ for T_2, where the conditional distribution $\bar{F}(t_2|t_1)$ was derived sequentially from the joint survival function $\bar{F}(t_1, t_2)$ in (10) and the two exponential margins as k increases from 1 to 4. The resulting random variables (T_1, T_2) followed the bivariate survival function $F(t_1, t_2)$ in (10). We set the sample size $n = 500$, which is close to the sample size in The treminTrust Data set. A total of 100 replications were conducted.

For each simulated data set, we analyzed the data using both direct and sequential two-stage methods. Under the direct method, the nonparametric estimates of the marginal survival functions $\bar{F}_1(t_1)$ and $\bar{F}_2(t_2)$ were obtained using Kaplan-Meier method. The standard errors of the cross-ratio estimates were calculated using both the model based method (Shih and Louis, 1995) and the bootstrap method based on 100 bootstrap samples. Under the sequential method, only the bootstrap standard errors were calculated. For both methods, the empirical 95% coverage probabilities using the bootstrap standard errors were calculated. The simulation results are reported in Table 3.

The results in Table 3 suggest that both direct and sequential methods perform very well and their performances are similar. The biases of the cross-ratio estimates using both methods were less than 3%. Under the direct method, both the model-based SEs and the bootstrap SEs agreed well with the empirical SEs. Under the sequential method, the bootstrap SEs were also close to the empirical SEs. The coverage probabilities were close to the normal value. These indicate that the bootstrap method performed well in estimating standard errors.

We next simulated the case where covariates were present in the two marginal survival functions under Cox model (6). Specifically, we assumed $\bar{F}_1(t_1|z_1) = \exp(\beta_1 z_1)$ and $\bar{F}_1(t_2|z_2) = \exp(\beta_2 z_2)$, where $\beta_1 = \beta_2 = 1$ and Z_1 and Z_2 were simulated from Uniform $(0, 1)$ and Normal(1,1), respectively. We chose the same setting for θ_k, $k = 1, \ldots, 4$. The results are presented in Table 4. Each simulated data set was analyzed using the naive direct method and the sequential two-stage method. The naive direct method used the left-truncated and right censored data $\mathcal{D}_k$ and assumed the left and bottom margins L_k and B_k of strip A_k followed Cox models. The results in Table 4 show that the sequential method performs well and the bias of cross-ratio estimates is minimal. However, the cross-ratio estimates given by naive direct method had considerable bias when the true value of cross-ratio is large.

7 DISCUSSION

We consider in this paper modelling the dependence between two event times using a piecewise constant cross-ratio model, where the cross-ratio is assumed to be a step function of one of the two failure times. This model is motivated by research in female reproductive aging, where it is of interest to model the dependence of time to a bleeding pattern based marker event and time to menopause as a function of time to the marker event. We propose two estimation procedures: the direct two-stage method and the sequential two-stage method. The former is applicable to the case without covariates, while the latter is applicable to both cases with/without covariates. Our simulation results suggest both methods work well in the absence of covariates and the sequential method also performs well when covariates are present.

Our analysis of the Tremin Trust data indicates that the onset of a 45-day cycle marker is not statistically significantly associated with age at onset of menopause before age 40 and after age 50. However, it is strongly positively associated with age at onset of menopause between age 40 and 50, and the association is the strongest between age 45 and 50. This analysis supports the descriptive results of Lisabeth et al. (2003). Women who experience the 45 day cycle marker before age 40 might represent a group of women who have different menstrual histories and will experience a prolonged period of transition from late reproductive life to menopause. The onset of a 45 day cycle marker is a good candidate marker for entering into the early stage of menopausal transition if it is experienced between age 40 and 50. It is of scientific interest to apply the proposed method to the analysis of the other proposed bleeding markers.

We consider in this paper two types of two-stage methods. A key advantage of both methods is that they are computationally easy. Several alternative methods might be investigated in future research. Instead of estimating θ_k individually in each strip, one approach is to modify the second stage of the two stage method by simultaneously estimating all θ_k's using all the observed data from the likelihood constructed by the bivariate survival function (10) for all k. This method might yield more efficient estimators of θ_k's. However, a main challenge is that the form of the marginal hazard function on the left margin L_k is complicated and depends on the cross-ratios θ_j, $j < k$. The joint estimation would be computationally difficult. Another approach is to use the semiparametric maximum likelihood method where the likelihood is specified using (10). One simultaneously estimates the cross-ratios θ_k's and the marginal survival functions $\bar{F}_1(t_1)$ and $\bar{F}_2(t_2)$ or their Cox model version with covariates (6), where the baseline marginal hazards are estimated nonparametrically. Although this method might yield semiparametrically efficient cross-ratio estimators, its computation is even more difficult. More future research is needed.

APPENDIX A: THE SOLUTION TO EQUATION (7)

The General Solution

We show the results for a more general situation where θ is constant when $(t_1, t_2) \in A \equiv [u_1, u_2) \times [v_1, v_2)$. The strip A_k can be viewed as a special rectangle A with $u_1 = w_{k-1}$, $u_2 = w_k$, $v_1 = 0$, and $v_2 = \infty$. For simplicity, we drop the subscription A_k from survival functions and replace θ_k by θ. Let $h(t_1, t_2) = -\log\bar{F}(t_1, t_2)$. Suppose θ is constant when $(t_1, t_2) \in A$. Then equation (7) becomes the following second order partial differential equation:

$$\frac{\partial^2 h}{\partial t_1 \partial t_2} + (\theta - 1)\frac{\partial h}{\partial t_1}\frac{\partial h}{\partial t_2} = 0 \ ,$$

which is equivalent to

$$\exp\{(\theta-1)h\}\frac{\partial^2 h}{\partial t_1 \partial t_2} + (\theta - 1)\exp\{(\theta-1)h\}\frac{\partial h}{\partial t_1}\frac{\partial h}{\partial t_2} = 0 \ .$$

When $\theta \neq 1$, the above equation is equivalent to

$$\frac{\partial^2}{\partial t_1 \partial t_2}\Big[(\theta-1)^{-1}\exp\{(\theta-1)h\}\Big] = 0 \ .$$

Thus we have

$$(\theta-1)^{-1}\exp\{(\theta-1)h\} = a(t_1) + b(t_2)$$

for arbitrary functions $a(\cdot)$ and $b(\cdot)$. Hence

$$\bar{F}(t_1, t_2) = [(\theta-1)\{a(t_1) + b(t_2)\}]^{-1/(\theta-1)}.$$

It is trivial to show that when $\theta = 1$, the solution to the equation $\partial^2 h/\partial t_1 \partial t_2 = 0$ is $h(t_1, t_2) = a^*(t_1) + b^*(t_2)$ for arbitrary functions $a^*(\cdot)$ and $b^*(\cdot)$. Hence

$$\bar{F}(t_1, t_2) = \exp[-\{a^*(t_1) + b^*(t_2)\}].$$

The Classical Clayton Model

Let $u_1 = v_1 = 0$ and $u_2 = v_2 = \infty$. We have $\bar{F}(0,0) = 1$. Given the two boundary conditions $\bar{F}(t_1, 0) = \bar{F}_1(t_1)$ and $\bar{F}(0, t_2) = \bar{F}_2(t_2)$, if $\theta \neq 1$, from equation (8) we have

$$\begin{aligned} a(t_1) + b(0) &= (\theta-1)^{-1}\bar{F}_1(t_1)^{-(\theta-1)}, \\ a(0) + b(t_2) &= (\theta-1)^{-1}\bar{F}_2(t_2)^{-(\theta-1)}, \\ a(0) + b(0) &= (\theta-1)^{-1}. \end{aligned}$$

Thus we have $(\theta-1)\{a(t_1) + b(t_2)\} = \bar{F}_1(t_1)^{-(\theta-1)} + \bar{F}_2(t_2)^{-(\theta-1)} - 1$, and we then obtain equation (9). When $\theta = 1$, from equation (8) we have

$$a^*(t_1) + b^*(0) = -\ \log\bar{F}_1(t_1), \quad a^*(0) + b^*(t_2) = -\ \log\bar{F}_2(t_2), \quad a^*(0) + b^*(0) = 0.$$

Thus $a^*(t_1) + b^*(t_2) = -\ \log\{\bar{F}_1(t_1)\bar{F}_2(t_2)\}$, and we then have $\bar{F}(t_1, t_2) = \bar{F}_1(t_1)\bar{F}_2(t_2)$.

The Clayton Model with Left Truncation

Suppose that $u_1,\ v_1 > 0$ and $\bar{F}(t_1, v_1)$ and $\bar{F}(u_1, t_2)$ are known. We only show the case when $\theta \neq 1$. The proof of the case with $\theta = 1$ is similar. When $(t_1, t_2) \in A$, from equation (8) we have

$$\begin{aligned} a(t_1) + b(v_1) &= (\theta - 1)^{-1}\bar{F}(t_1, v_1)^{-(\theta-1)}, \\ a(u_1) + b(t_2) &= (\theta - 1)^{-1}\bar{F}(u_1, t_2)^{-(\theta-1)}, \\ a(u_1) + b(v_1) &= (\theta - 1)^{-1}\bar{F}(u_1, v_1)^{-(\theta-1)}. \end{aligned}$$

Thus

$$(\theta - 1)\{a(t_1) + b(t_2)\} = \bar{F}(t_1, v_1)^{-(\theta-1)} + \bar{F}(u_1, t_2)^{-(\theta-1)} - \bar{F}(u_1, v_1)^{-(\theta-1)}\ ,$$

and we then obtain equation (10) with $u_1 = w_{k-1}$ and $v_1 = 0$. The Clayton model for left truncated data in equation (12) follows easily from equation (10) using the left truncated survival function in equation (11).

APPENDIX B: ASYMPTOTIC PROPERTIES

Since $\Pr(w_{k-1} \leq T_1 < w_k) > 0$ for all k $(k = 1, \ldots, K)$, the number of subjects n_k used for estimating each θ_k goes to infinity proportionally as the sample size $n \to \infty$. For each strip A_k, the estimates of the two marginal survival functions are root-n consistent in the support region $(t_1, t_2) \in [0, \tau_1) \times [0, \tau_2)$ with $P(T_1 > \tau_1, C_1 > \tau_1) > 0$ and $P(T_2 > \tau_2, C_2 > \tau_2) > 0$ which can be shown sequentially, thus as in Shih and Louis (1995) and Glidden (2000) we have that $\sqrt{n}(\hat{\theta}_k - \theta_k)$ converges weakly to a mean zero normally distributed random variable under their regularity conditions and model assumption (4). More rigorous proof of consistency for two-stage method in Clayton model can be found in Hu (1998). It can further be shown using the functional delta method that $\sqrt{n}(\hat{\theta}_k - \theta_k)$ and the bootstrapped $\sqrt{n}(\hat{\theta}_k^* - \hat{\theta}_k)$ given observed data converge weakly to random variables following the same normal distribution.

REFERENCES

Clayton, D. G. (1978), "A Model for Association in Bivariate Life Tables and Its Application in Epidemiological Studies of Familial Tendency in Chronic Disease Incidence," *Biometrika*, 65, 141-151.

Cox, D. R. (1972), "Regression Models and Life-Tables (with discussion)," *Journal of the Royal Statistical Society B*, 34, 187-220.

Fan, J., Hsu, L., and Prentice, R. L. (2000), "Dependence Estimation Over a Finite Bivariate Failure Time Region," *Lifetime Data Analysis*, 6, 343-355.

Glidden, D. V. (2000), "A Two-Stage Estimator of the Dependence Parameter for the Clayton-Oakes Model," *Lifetime Data Analysis,* 6, 141-156.

Hougaard, P. (2000), *Analysis of Multivariate Survival Data,* New York: Springer-Verlag.

Hsu, L., Prentice, R. L., Zhao, L. P., and Fan J. J. (1999), "On Dependence Estimation Using Correlated Failure Time Data from Case-Control Family Studies," *Biometrika,* 86, 743-753.

Hu, H-L (1998), *Large Sample Theory for Pseudo-Maximum Likelihood Estimates in Semiparametric Models*, Ph.D. Dissertation, University of Washington.

Kalbfleisch, J. D. and Prentice, R. L. (2002), *The Statistical Analysis of Failure Time Data*, 2nd Ed., Wiley.

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,* to appear.

Nielsen, G. G., Gill, R. D., Andersen, P. K., and Srensen, T. I. A. (1992), "A Counting Process Approach to Maximum Likelihood Estimation in Frailty Models," *Scandinavian Journal of Statistics,* 19, 25-43.

Oakes, D. (1986a), "A Model for Bivariate Survival Data," in *Modern Statistical Methods in Chronic Disease Epidemiology*, eds. Moolgavkar, S. H. and Prentice, R. L., Wiley.

Oakes, D. (1986b), "Semiparametric Inference in a Model for Association in Bivariate Survival Data," *Biometrika,* 73, 353-361.

Oakes, D. (1989), "Bivariate Survival Models Induced by Frailties," *Journal of the American Statistical Association*, 84, 487-493.

Shih, J. H. and Louis, T. A. (1995), "Inference on the Association Parameter in Copula Models for Bivariate Survival Data," *Biometrics,* 51, 1384-1399.

Taffe, J. and Dennerstein, L. (2001),“ Menstrual Patterns Leading to Final Menstrual Period,” *Menopause*, 9, 32-40.

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.

Table 1: Two-stage estimates of the piecewise constant cross-ratios for time to the 45-day cycle marker event (T_1) and time to menopause in the Tremin Trust data. The bootstrap standard errors were calculated using 1000 bootstrap samples. The labels are $\hat{\theta}$: point estimate; SE_{SL}: model-based standard error estimate in Shih and Louis (1995); SE_{BS}: bootstrap standard error estimate.

	Direct method		Sequential method		
	Without covariate		Without covariate		With covariate
T_1	$\hat{\theta}$ ($\mathrm{SE}_{SL}, \mathrm{SE}_{BS}$)	$\hat{\theta}$ ($\mathrm{SE}_{SL}, \mathrm{SE}_{BS}$)	$\hat{\theta}$ (SE_{BS})	$\hat{\theta}$ (SE_{BS})	$\hat{\theta}$ (SE_{BS})
35-37	0.80 (0.14, 0.15)	0.80 (0.12, 0.11)	0.80 (0.15)	0.80 (0.10)	0.81 (0.11)
38-39	0.83 (0.20, 0.17)		0.82 (0.18)		
40-41	2.44 (0.50, 0.70)	2.19 (0.27, 0.33)	2.48 (0.74)	2.17 (0.35)	2.17 (0.34)
42-43	1.98 (0.47, 0.71)		2.04 (0.77)		
44-45	2.18 (0.39, 0.63)		2.12 (0.70)		
46-47	4.48 (1.19, 1.92)	3.64 (0.71, 0.78)	5.72 (1.95)	4.11 (0.85)	4.41 (0.94)
48-49	3.18 (0.81, 0.97)		3.28 (1.10)		
50+	1.57 (0.40, 0.55)	1.57 (0.40, 0.55)	1.39 (0.38)	1.37 (0.41)	1.47 (0.47)

Table 2: Estimated percentiles of survival probability for age at menopause given age at the 45-day cycle marker.

	Without covariate					With Covariate				
						Age at menarche = 12				
Age at marker	90%	75%	50%	25%	10%	90%	75%	50%	25%	10%
36	47.9	49.9	52.2	54.2	55.5	47.9	49.9	52.2	54.1	55.3
39	47.9	49.9	52.2	54.1	55.3	47.8	49.8	52.1	54.1	55.3
42	47.2	48.5	50.2	51.8	52.8	46.4	48.1	49.9	51.6	52.8
45	48.8	49.5	51.1	52.2	53.0	47.4	48.9	50.8	52.2	53.3
48	50.5	51.4	51.7	52.2	52.7	48.2	49.7	51.4	52.2	53.0
51	53.1	53.2	53.9	55.2	55.5	52.8	53.8	54.7	55.5	56.2

Table 3: Simulation results for the cross-ratio estimates without covariates based on 100 replications. The true values are $\theta = (0.9, 2.0, 4.0, 1.5)$ when t_1 is in intervals $[0, 0.25)$, $[0.25, 0.5)$, $[0.5, 0.75)$ and above 0.75, The sample size is 500. The labels are $\hat{\theta}$: point estimate average; SE_{SL}: average of the standard error estimates using Shih and Louis (1995); SE_{BS}: average of the bootstrap standard error estimates using 100 bootstrap samples; SE_{E}: empirical standard error; CP: empirical 95% coverage probability.

	Oakes's Method					Sequential Method			
θ	$\hat{\theta}$	SE_{SL}	SE_{BS}	SE_{E}	CP	$\hat{\theta}$	SE_{BS}	SE_{E}	CP
0.9	0.92	0.12	0.12	0.12	0.93	0.92	0.12	0.12	0.93
2.0	2.04	0.29	0.29	0.31	0.95	2.04	0.29	0.31	0.93
4.0	4.09	0.70	0.73	0.66	0.97	4.09	0.71	0.65	0.96
1.5	1.51	0.25	0.25	0.25	0.93	1.51	0.25	0.25	0.94

Table 4: Simulation results base on 100 replications for the cross-ratio estimates with covariates assuming the two marginal survival distributions follow the Cox model. The sample size is 500. The true values are $\theta = (0.9, 2.0, 4.0, 1.5)$ when t_1 is in intervals $[0, 0.25)$, $[0.25, 0.5)$, $[0.5, 0.75)$ and above 0.75, respectively. The labels are $\hat{\theta}$: average of the point estimates; SE_{E}: empirical standard error.

	Direct Method		Sequential Method	
θ	$\hat{\theta}$	SE_{E}	$\hat{\theta}$	SE_{E}
0.9	0.89	0.08	0.89	0.08
2.0	2.13	0.22	2.02	0.27
4.0	2.35	0.33	3.98	0.68
1.5	1.52	0.31	1.51	0.30

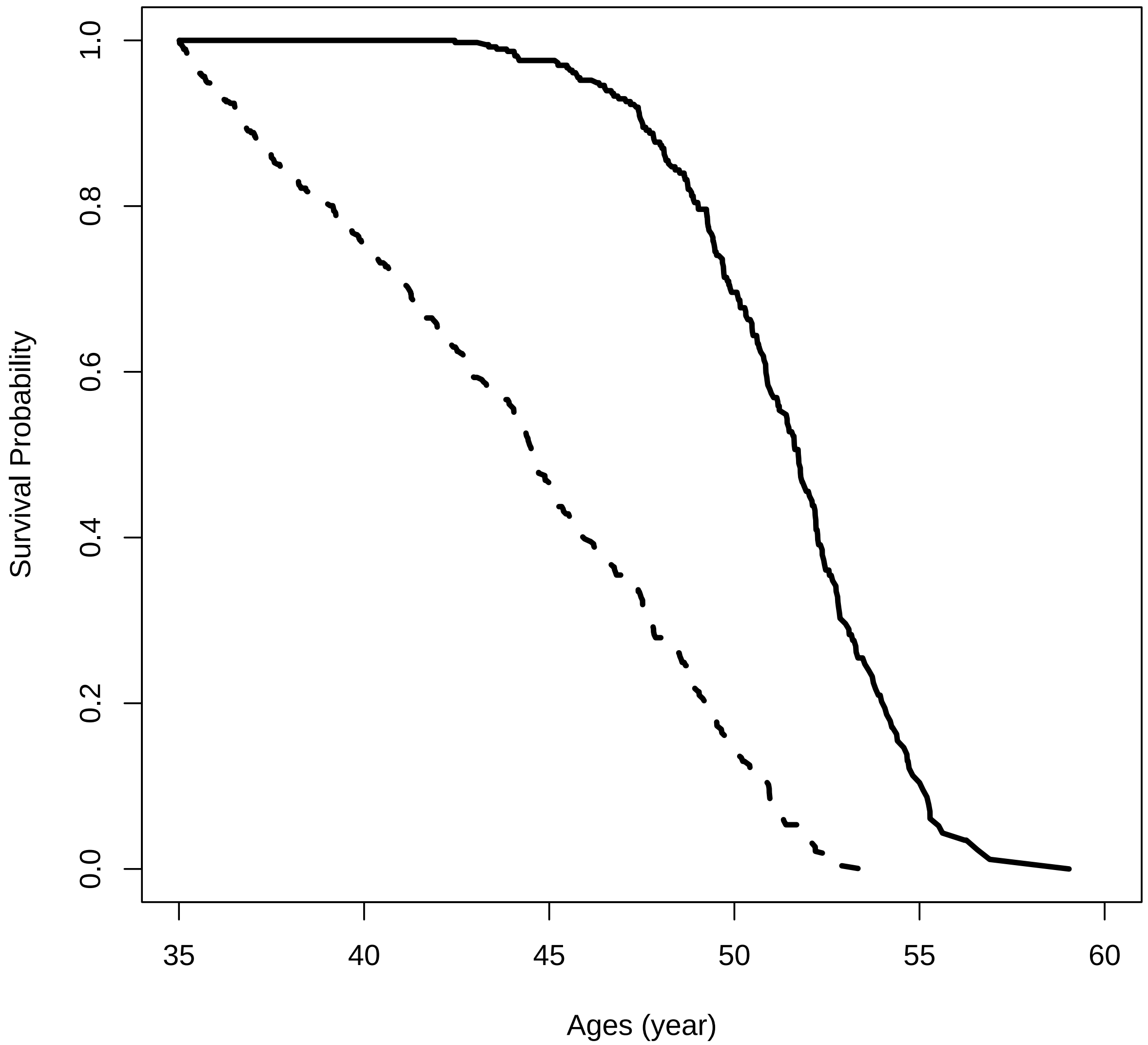

Figure 1: Kaplan-Meier estimators for time to menopause (solid line), time to the 45-day cycle marker (dash line), and time to censoring (dot line).

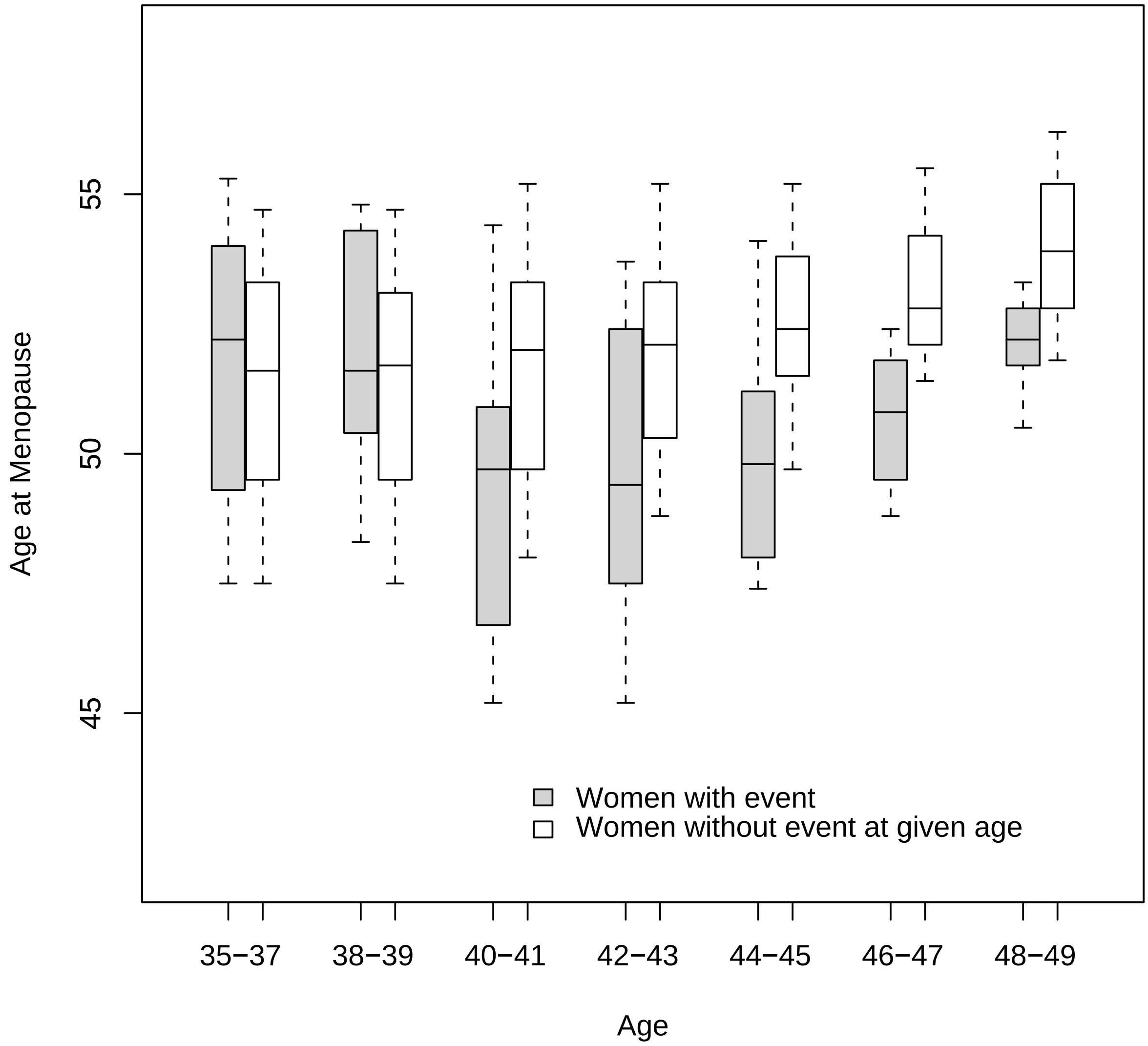

Figure 2: Paired box-plots of estimated survival probabilities comparing women experiencing the 45-day cycle marker event at a given age interval and those having not experienced the marker event by the end of the age interval.

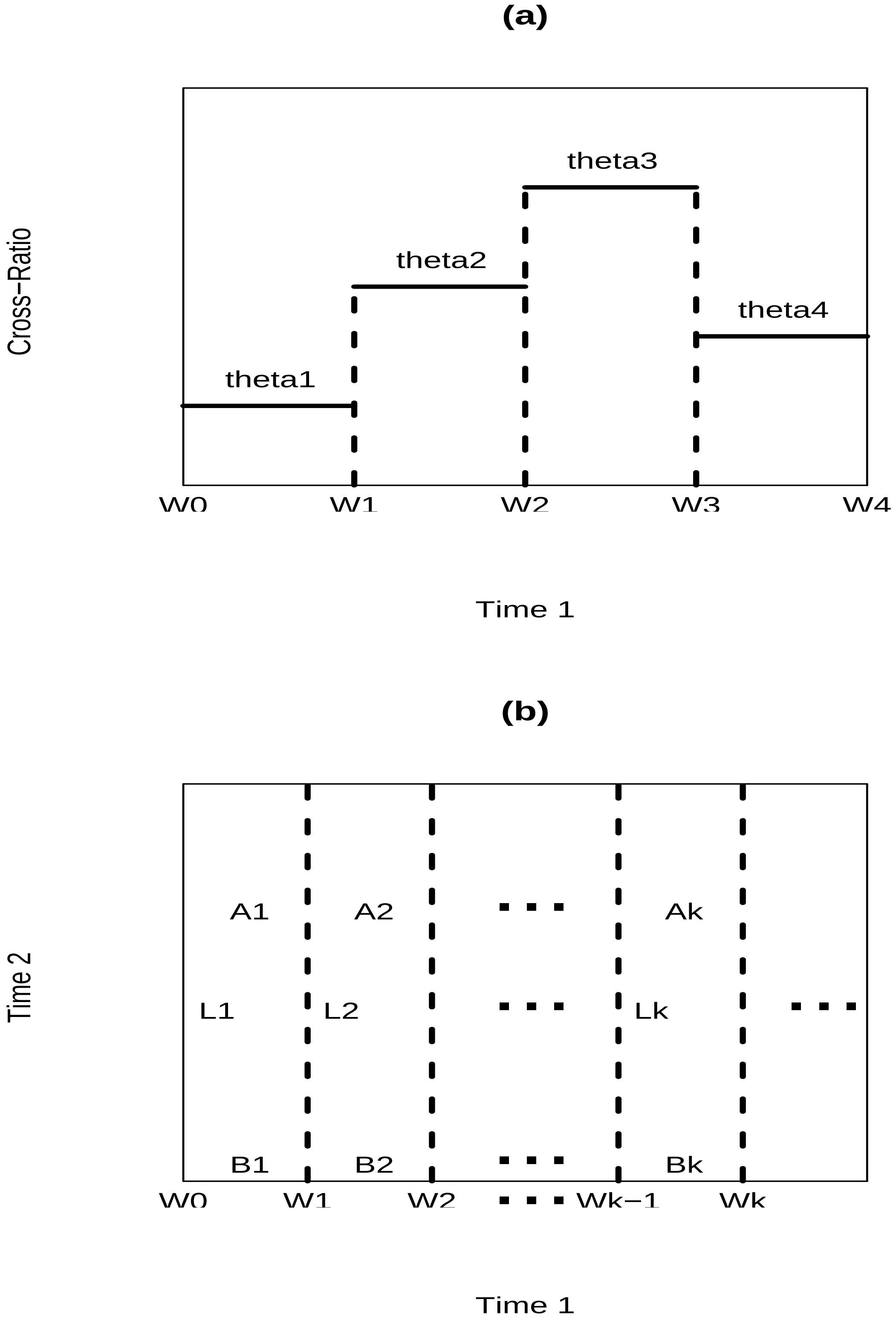

Figure 3: (a) Illustration of the piecewise constant cross-ratio model; (b) Illustration of the partition of T_1 with strips A_k and their left and bottom boundaries L_k and B_k.

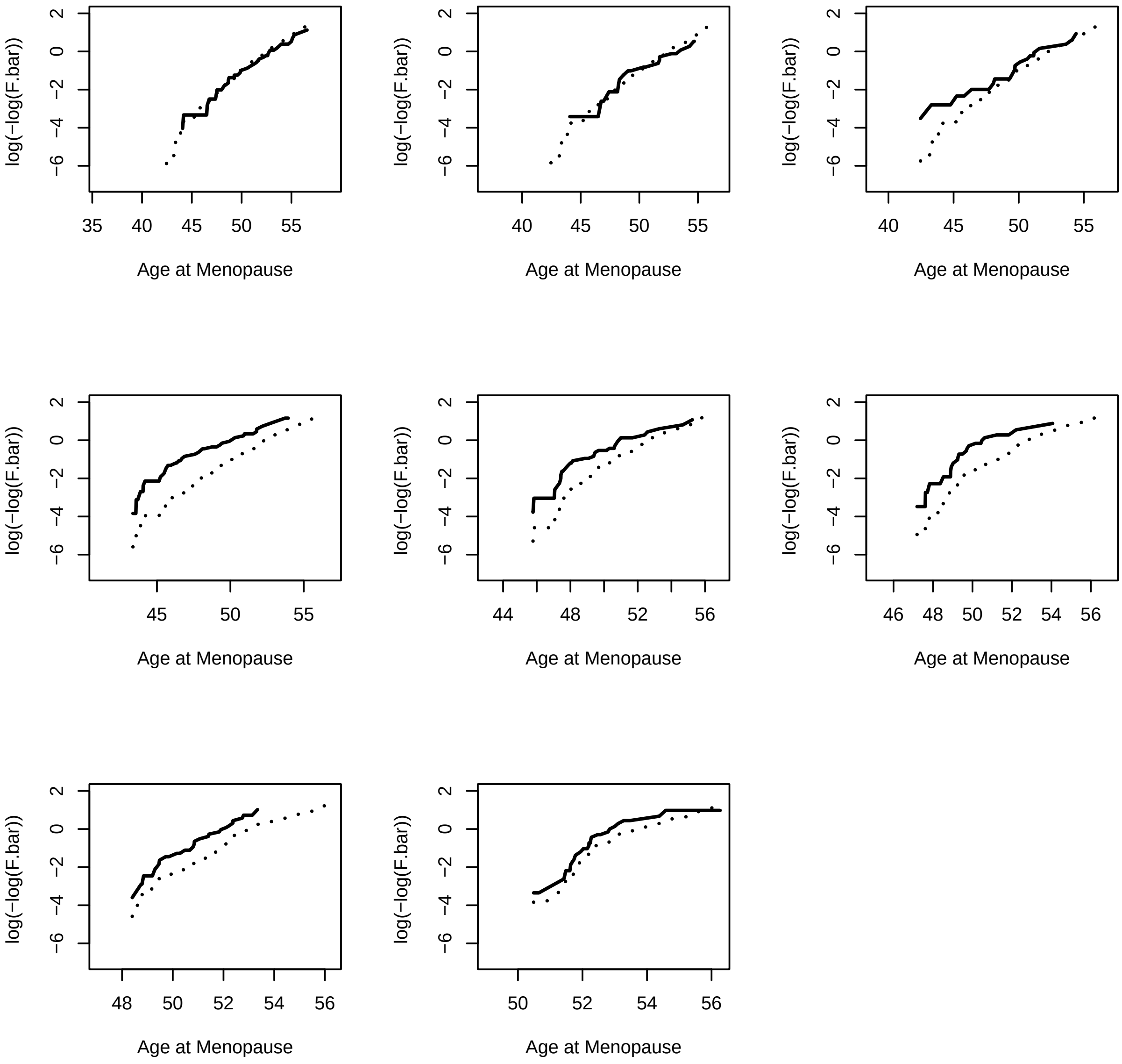

Figure 4: Paired log(-log) survival plots for checking the assumption of piecewise constant cross-ratios. In each plot, the solid line represents $\log[-\log\{\bar{F}(t_2|T_1 = t_1)\}]$ for data with $T_1 \in (t_1 - 1, t_1 + 1)$ except the first interval which is $T_1 \in (t_1 - 1.5, t_1 + 1.5)$; dot line represents $\log[-\log\{\bar{F}(t_2|T_1 > t_1)\}]$ for data with $Y_1 > t_1$. From left to right and top to bottom, $t_1 = 36.5, 38, 40, 42, 44, 46, 48,$ and 50.

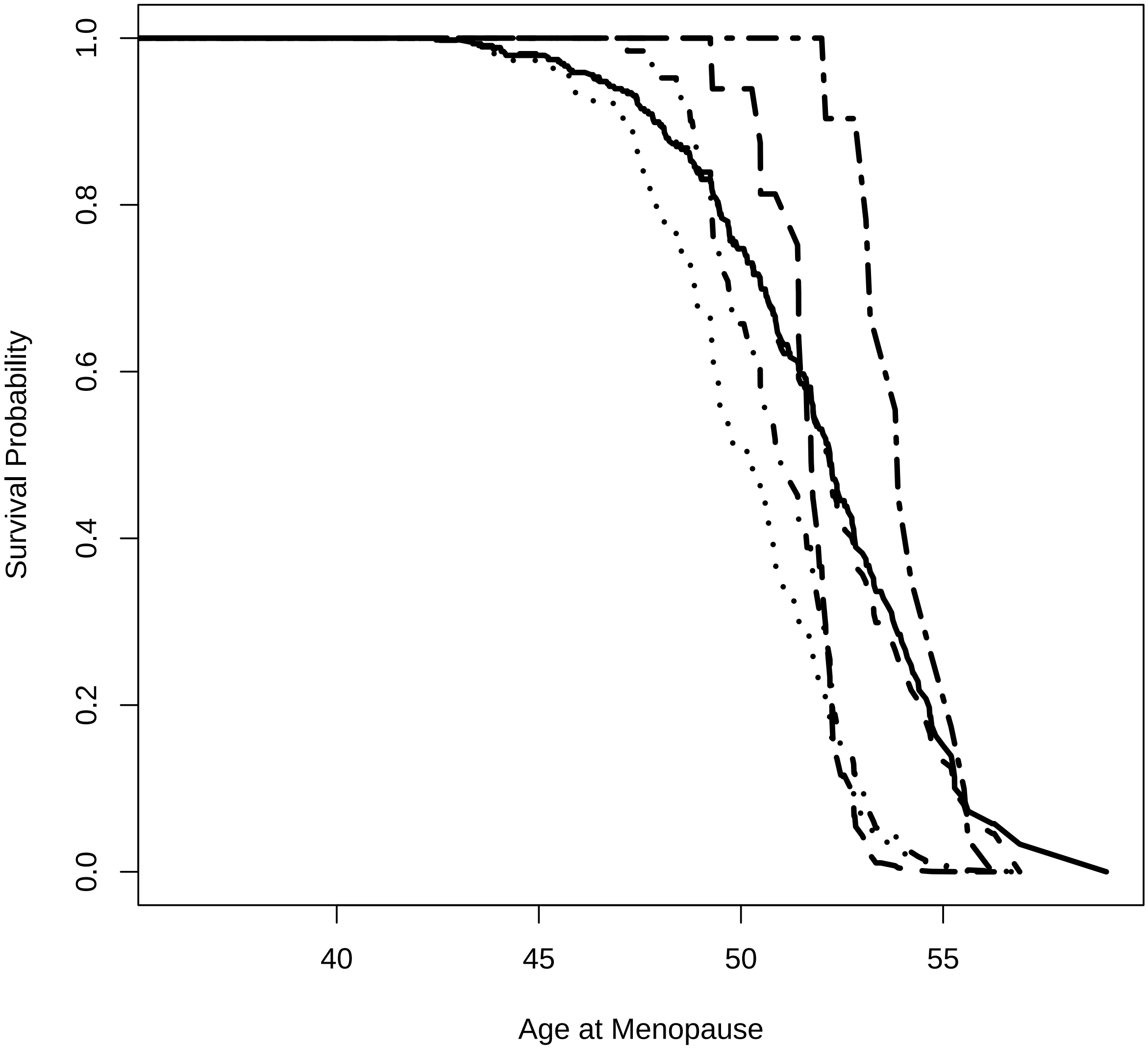

Figure 5: Estimated survival functions for time to menopause given age at the 45-day cycle marker event: —— Age 36; – – – Age 39; · · · Age 42; – · – Age 45; — — Age 48; — · — Age 51.