Published in final edited form as:

Biometrics. 2017 September ; 73(3): 835–845. doi:10.1111/biom.12655.

Proportional Hazards Model with a Change Point for Clustered Event Data

Yu Deng^{1,*}, Donglin Zeng^{1,**}, Jinying Zhao^{2,***}, and Jianwen Cai^{1,****}

¹Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina, U.S.A.

²Department of Epidemiology, Tulane University, New Orleans, Louisiana, U.S.A.

Summary

In many epidemiology studies, family data with survival endpoints are collected to investigate the association between risk factors and disease incidence. Sometimes the risk of the disease may change when a certain risk factor exceeds a certain threshold. Finding this threshold value could be important for disease risk prediction and diseases prevention. In this work, we propose a change-point proportional hazards model for clustered event data. The model incorporates the unknown threshold of a continuous variable as a change point in the regression. The marginal pseudo-partial likelihood functions are maximized for estimating the regression coefficients and the unknown change point. We develop a supremum test based on robust score statistics to test the existence of the change point. The inference for the change point is based on the m out of n bootstrap. We establish the consistency and asymptotic distributions of the proposed estimators. The finite-sample performance of the proposed method is demonstrated via extensive simulation studies. Finally, the Strong Heart Family Study dataset is analyzed to illustrate the methods.

Keywords

Change point; Clustered event; m out of n bootstrap; Proportional hazards model

1. Introduction

The change point models have been widely applied in clinical research to decide the subgroup of participants who have a much higher risk for specific diseases. Change point effects have been observed in many medical studies for different traits, such as fasting plasma glucose in the Australian Diabetes Obesity and Lifestyle Study (AusDiab) (Tapp et al., 2006), midthigh muscle cross-sectional area in the COPD Study (Marquis et al., 2002), and leukocyte telomere length in the Strong Heart Family Study (SHFS) (Zhao et al., 2014). In these studies, risk of disease changes when a continuous risk factor passes a threshold

*yuliadeng@gmail.com

**dzeng@bios.unc.edu

***jzhao5@tulane.edu

****cai@bios.unc.edu

Supplementary Materials

Web Appendices referenced in Section 1 and Section 3 and an R code for implementing the proposed methods are available with this paper at the Biometrics website on Wiley Online Library.

value. For example, Zhao et al. (2014) investigated the association between leukocyte telomere length (LTL) and diabetes incidence in the SHFS. SHFS is a longitudinal family-based cohort study of cardiovascular disease, type 2 diabetes and their risk factors among American Indians residing in Oklahoma, Arizona and South/North Dakota. The authors found that participants with shorter LTL (lower quartile) have nearly two-fold increased risk for developing incident diabetes compared to those with longer LTL. Such a change point for LTL and diabetes incidence was also observed by Willeit et al. (2014). It is well-known that telomere length shortens progressively with each cell division until it reaches a threshold value beyond which cells enter into senescence or die, a phenomenon called “Hayick limit”. Even though the change point observed in these studies is consistent with the theory of “Hayick limit”, the precise change point location in LTL remains to be determined. Finding this threshold value is helpful to identify at-risk individuals and risk prediction. Thus, it is of great interest to develop a rigorous and comprehensive framework to conduct the change point analysis for survival data subject to censoring.

The change point analysis has been studied in the univariate Cox proportional hazards model. The Cox proportional hazards model (Cox, 1972) was widely used to estimate the association between disease incidence and potential risk factors. Different change point models in the Cox proportional hazards model are proposed for various purposes. Liang et al. (1990), Luo (1996), and Pons (2002) discussed the change point at an unknown time for the lag effect of the covariates. Gandy et al. (2005), Gandy and Jensen (2005), and Jensen and Lütkebohmert (2008) considered the Cox model with a smooth change in the regression coefficient. They assumed that the slopes are different for the covariates above and below the change points. Another class of models assumes a non-smooth “jump” effect at an unknown threshold of a covariate (Luo and Boyett, 1997; Pons, 2003; Kosorok and Song, 2007). Here, we focus on the change point analysis based on a non-smooth “jump” effect of a covariate. Maximum partial likelihood methods were proposed to estimate the change point and regression coefficients in this type of models. Luo and Boyett (1997) applied a two-step procedure to estimate the change point and proved the consistency of a resulting estimator. Later, Pons (2003) proved that this estimator asymptotically follows a composite Poisson process. Kosorok and Song (2007) generalized this estimator to transformation models and established the asymptotic properties of this class of models, which includes the Cox model as a special case. The change point analysis proposed for the univariate case cannot be applied directly to clustered survival data, because the proposed methods did not take into account the correlation between subjects within the same cluster.

In this paper, we focus on developing a Cox-type marginal hazards model (Lee et al., 1992) with a change point in a covariate for clustered survival data. The Cox marginal hazards model uses a pseudo-likelihood approach with a working independence assumption, while adjusting for the correlation by a sandwich estimate of the covariance matrix. The marginal hazards model is useful when the focus is on making inferences on the population average effect of risk factors on failure time. One major difficulty for the change point analysis in the Cox marginal hazards model is the complicated asymptotic distribution of the change point estimator for clustered data. With univariate survival data, Pons (2003) proved that the change point estimator asymptotically follows a composite Poisson process which depends on the change point locations across all the subjects. However, the existing theory for the

univariate Cox model cannot be applied directly to the change point analysis in the Cox marginal hazards model. The asymptotic distribution of the change point estimator for the clustered data is a weighted mixture summation of the composite Poisson process based on the cluster sizes and the joint distribution of the covariates. Considering the varying cluster size and all the possible situations of the covariate passing the true threshold across every member within each cluster, we prove that the asymptotic distribution of the proposed change point estimator follows a more complicated composite Poisson process.

The structure of this paper is as follows. In Section 2, we describe the estimation method based on a two-step procedure. We then provide an inference method based on m out of n bootstrap, and a testing procedure for the existence of a change point. In Section 3, we establish the consistency, convergence rates and asymptotic distributions of the proposed estimators. Simulation studies evaluating the small sample performance of the method are presented in Section 4. In Section 5, data from the Strong Heart Family Study are analyzed using our approach. The details of the proofs are given in the Supplementary Materials.

2. Methods

2.1 Model and Parameter Estimation

Consider n independent and identically distributed (i.i.d) clusters with the i th cluster containing K_i subjects ($i = 1, \dots, n$). For the j th subject in the i th cluster, $j = 1, \dots, K_i$, let $\tilde{T}_{ij}$ be the survival time, X_{ij} denote a one-dimensional continuous covariate whose effect on the response may have a change point, and $\mathbf{Z}_{ij}(t)$ denote other potentially time-dependent covariates whose effects could be different before or after X_{ij} passes the change point. In other words, the proportional hazards model with a change point assumes that the hazard rate function for $\tilde{T}_{ij}$ given $\mathbf{W}_{ij}(t) \equiv (X_{ij}, \mathbf{Z}_{ij}^T(t))^T$ takes a form

$$\lambda(t|\mathbf{W}_{ij}) = \lambda_0(t) \exp \{ \boldsymbol{\beta}_1^T \mathbf{Z}_{ij}(t) + \beta_2 I(X_{ij} > \zeta) + \boldsymbol{\beta}_3^T \mathbf{Z}_{ij}(t) I(X_{ij} > \zeta) \},$$

where $\lambda_0(t)$ is an unknown baseline function, ζ is the unknown change point for which the covariate X_{ij} has different effects for $X_{ij} \leq \zeta$ and $X_{ij} > \zeta$, and $\boldsymbol{\beta} \equiv (\boldsymbol{\beta}_1^T, \beta_2, \boldsymbol{\beta}_3^T)^T$ is a vector of $2J + 1$ unknown parameters with $J = \dim(\mathbf{Z}_{ij}(t))$. Therefore, the proposed model implies that the effect of $\mathbf{Z}_{ij}$ is $\boldsymbol{\beta}_1$ when $X_{ij} \leq \zeta$, and it becomes $(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_3)$ when $X_{ij} > \zeta$. Furthermore, the hazard ratio between $X_{ij} > \zeta$ and $X_{ij} \leq \zeta$ is $\exp \{ \beta_2 + \boldsymbol{\beta}_3^T \mathbf{Z}_{ij}(t) \}$ for given $\mathbf{Z}_{ij}(t)$.

If we define $r_\theta\{\mathbf{W}_{ij}(t)\} \equiv \boldsymbol{\beta}_1^T \mathbf{Z}_{ij}(t) + \beta_2 I(X_{ij} > \zeta) + \boldsymbol{\beta}_3^T \mathbf{Z}_{ij}(t) I(X_{ij} > \zeta)$ and $\boldsymbol{\theta} \equiv (\zeta, \boldsymbol{\beta}^T)^T$, then a marginal pseudo-partial likelihood function for n clusters with right censoring can be formulated as

$$L(\boldsymbol{\theta}) = \prod_{i=1}^n \prod_{j=1}^{K_i} \left(\frac{\exp[r_\theta\{\mathbf{W}_{ij}(T_{ij})\}]}{\sum_{l=1}^n \sum_{k=1}^{K_l} I(T_{lk} \geq T_{ij}) \exp[r_\theta\{\mathbf{W}_{lk}(T_{ij})\}]} \right)^{\Delta_{ij}},$$

where $T_{ij} = \min(\tilde{T}_{ij}, C_{ij})$ with C_{ij} being the censoring time assumed to be independent of $\tilde{T}_{ij}$ given the covariates W_{ij} , and $\delta_{ij} = I(\tilde{T}_{ij} \leq C_{ij})$ is the failure indicator.

To estimate the model parameters, we propose to maximize the logarithm of the pseudo-likelihood function, which is defined as $l_n(\zeta, \beta) \equiv \log \{L(\theta)\}$. Computationally, we adopt the following two-step procedure for maximization. For any fixed value of ζ in a pre-specified range $[\zeta_1, \zeta_2]$, we maximize the logarithm of the pseudo-likelihood function via the Newton-Raphson method, which yields the global maximum due to the strict concavity of $l_n(\zeta, \beta)$ for the given ζ . We thus obtain the profile function for ζ . In the second step, we apply a grid-search algorithm to find the optimal estimator for ζ . It is possible to have multiple ζ reaching the same maximum value, because the profile function of ζ is a step function. To retain the unique value of ζ , we choose the smallest one as our estimate of ζ . Thus, $(\hat{\zeta}, \hat{\beta}) = \arg \max_{\zeta \in [\zeta_1, \zeta_2], \beta} l_n(\zeta, \beta)$. In addition, the cumulative baseline hazard function $\Lambda_0(t)$ is estimated by the Breslow-type estimator, which is given in the following form:

$$\hat{\Lambda}_0(t) = \sum_{i=1}^n \sum_{j=1}^{K_i} \frac{I(T_{ij} \geq t) \Delta_{ij}}{\sum_{l=1}^n \sum_{k=1}^{K_l} I(T_{lk} \geq T_{ij}) \exp[r_{\hat{\theta}}\{\mathbf{W}_{lk}(T_{ij})\}]}$$

2.2 Inference for ζ and β

To make inference for ζ and β , we utilize the asymptotic results which will be given in Section 3. In that section, we show that $\hat{\zeta}$ and $\hat{\beta}$ are asymptotically independent and the asymptotic distribution of $\hat{\beta}$ remains the same regardless whether ζ is known or not. Thus, the inference for β can be carried out in a similar manner as the marginal proportional hazard model for clustered survival data, treating $\zeta = \hat{\zeta}$ as fixed (c.f. Lee et al., 1992). However, the inference for $\hat{\zeta}$ is challenging due to the intractable asymptotic distribution shown in Section 3. The bootstrap approach is commonly applied to generate the empirical distributions of the estimators with complicate asymptotic distributions (Efron and Tibshirani, 1994). The usual bootstrap approach is to draw a sample of n with replacement from the dataset of n samples. Efron and Tibshirani (1986) demonstrated its performance in generating standard errors and confidence intervals under regular conditions. However, the usual bootstrap approach produces inconsistent estimators in some non-standard problems. Dömbgen (1993) and Shao (1994) demonstrated the failure of the usual bootstrap in non-differentiable objective functions or non-smooth statistics. In addition, Shao (1994) proposed a remedy of such situation by sampling a ratio of the size of the original dataset. Given the dataset of size n , the m out of n bootstrap approach is defined as sampling with replacement of size m , where $m \rightarrow \infty$, and $m/n \rightarrow 0$. Similar concepts are also proposed by Bickel et al. (2012) and Politis and Romano (1999). Such method is widely used in non-standard problems, such as non-differentiable objective functions (Huang et al., 1996; Chakraborty et al., 2013) and non- $n^{-1/2}$ asymptotics (Abrevaya and Huang, 2005; Sen et al., 2010). In addition, Xu et al. (2014) proved the consistency of the m out of n bootstrap in the case of the Cox proportional hazards model with a change point. Xu et al. (2014) applied m out of n bootstrap based on some fixed values of m , which are $n^{4/5}$, $n^{9/10}$, and $n^{14/15}$. Based

on this group of pre-specified m , the poor coverage rates have been reported in their simulations. Here, we will adopt a data-driven approach to select the optimal m .

For the m out of n bootstrap, several data-driven approaches for choosing m have been proposed (Hall et al., 1995; Lee, 1999; Cheung et al., 2005; Bickel and Sakov, 2005; Bickel and Sakov, 2008). Among them, Bickel and Sakov (2008) proposed a method to select m for extrema functions. Based on their approach, the desired m is selected from a sequence of possible re-sampling sample sizes. The rule is to select the maximum sample size that achieves the minimum distance defined on supremum norm between two empirical distributions, which are based on any two adjacent re-sampling sample sizes. Thus, the selected m can achieve the stable empirical distributions of the proposed estimator. Hence, we adapt this algorithm to select m in the following way.

1. Construct a sequence of the re-sampling sample sizes $m_j = \lfloor j \times \frac{n}{q} \rfloor$, where $j = q, q-1, \dots, 1$, n/q is the interval between two adjacent re-sampling sample sizes, and $\lfloor a \rfloor$ is the largest integer no larger than a .
2. For the m_j out of n bootstrap, the empirical cumulative distribution function for the change point estimator is constructed as follows:

$$R_{m_j}(x, \hat{\zeta}) = \frac{1}{B} \sum_{b=1}^B I \left\{ m_j \left(\hat{\zeta}_{m_j}^{(b)} - \hat{\zeta} \right) \leq x \right\},$$

where $\hat{\zeta}$ is the change point estimator based on the full dataset, $\hat{\zeta}_{m_j}^{(b)}$ is the change point estimator based on the dataset with m_j samples in the b th replication, $b = 1, 2, \dots, B$, and B is the total number of bootstrap replications.

3. The m will be selected as the maximum value which minimizes the supremum difference between two adjacent empirical cumulative distributions in the m_j sequence.

$$m = \max \arg \min_{m_j} \sup_x \left| R_{m_j}(x, \hat{\zeta}) - R_{m_{j+1}}(x, \hat{\zeta}) \right|$$

Based on the selected m , the m out of n bootstrap is to draw m samples with replacement out of the overall n samples. The standard error of the proposed estimator is estimated by the sample standard deviation based on B replicates divided by n/m . In addition, the equal-tailed

95% confidence intervals are generated as $\left[\hat{\zeta} - \frac{Q_{\hat{\zeta}, 0.95}}{n/m}, \hat{\zeta} + \frac{Q_{\hat{\zeta}, 0.95}}{n/m} \right]$, where $Q_{\hat{\zeta}, 0.95}$ is the

95th quantile of the absolute value $\left| \hat{\zeta} - \hat{\zeta}_m^{(b)} \right|$ for the replicate $b = 1, 2, \dots, B$. Both the standard error estimator and the confidence interval are adjusted by n/m , which corrects the over-estimated variance and wide confidence intervals based on the m out of n bootstrap.

2.3 Hypothesis Testing for the Change Point

In practice, one important question is whether the change point exists. The null hypothesis is specified as $H_0: \beta_2 = 0, \beta_3^T = 0$ in our proposed model. However, the change point is not identifiable given both β_2 and β_3 are zero, because the estimation of the change point relies on either β_2 or β_3 unequal to zero. To handle it, in general, there are two testing methods in the change point method literatures, which are the maximum efficiency robust tests (MERT) (Gastwirth, 1966, Gastwirth, 1985) and the supremum (SUP) tests (Davies, 1977, Davies, 1987, Kosorok and Song, 2007). Zucker et al. (2013) conducted extensive simulations to compare these two approaches. Based on their simulation results, the SUP tests are more powerful under different scenarios. Here, we adopt the SUP type of test but rely on robust score statistics for the clustered survival time. Specifically, our test statistic is

$$\text{SUP}_k = \sup_{\zeta \in [\zeta_1, \dots, \zeta_k]} \mathbf{U}(\zeta)^T \boldsymbol{\Sigma}(\zeta)^{-1} \mathbf{U}(\zeta),$$

$$\text{where } \mathbf{U}(\zeta) = \frac{\partial l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \text{ and } \boldsymbol{\Sigma}(\zeta) = \sum_{i=1}^n \sum_{j=1}^{K_i} \sum_{l=1}^{K_i} \mathbf{H}_{ij}(\zeta) \mathbf{H}_{il}(\zeta)^T,$$

$$\mathbf{H}_{ij}(\zeta) = \left\{ \tilde{\mathbf{Z}}_{ij}(T_{ij}) - \frac{S_n^{(1)}(T_{ij}; \zeta, \boldsymbol{\beta})}{S_n^{(0)}(T_{ij}; \zeta, \boldsymbol{\beta})} \right\} - \sum_{s=1}^n \sum_{l=1}^{K_i} \frac{\Delta_{sl} \exp\{\boldsymbol{\beta}^T \tilde{\mathbf{Z}}_{ij}(T_{sl})\}}{n S_n^{(0)}(T_{sl}; \zeta, \boldsymbol{\beta})} \left\{ \tilde{\mathbf{Z}}_{ij}(T_{sl}) - \frac{S_n^{(1)}(T_{sl}; \zeta, \boldsymbol{\beta})}{S_n^{(0)}(T_{sl}; \zeta, \boldsymbol{\beta})} \right\},$$

$S_n^{(r)}(t; \zeta, \boldsymbol{\beta}) = \frac{1}{n} \left(\sum_{i=1}^n \sum_{j=1}^{K_i} Y_{ij}(t) \tilde{\mathbf{Z}}_{ij}^{\otimes r}(t; \zeta) \exp[r_{\theta}\{\mathbf{W}_{ij}(t)\}] \right)$, $Y_{ij}(t) = \mathbb{I}(T_{ij} \geq t)$, and $\tilde{\mathbf{Z}}_{ij}(t; \zeta) = (\mathbf{Z}_{ij}(t), \mathbb{I}(X_{ij} > \zeta), \mathbf{Z}_{ij}(t) \mathbb{I}(X_{ij} > \zeta))$ for $r = 0, 1$. For a column vector $\mathbf{Z}$, $\mathbf{Z}^{\otimes 0}$ refers to the scalar 1, $\mathbf{Z}^{\otimes 1}$ refers to the vector $\mathbf{Z}$, and $\mathbf{Z}^{\otimes 2}$ refers to the matrix $\mathbf{Z}\mathbf{Z}^T$. Davies (1987) proved that the asymptotic distribution of such supremum test statistics does not follow a standard chi-squared distribution. Hence, we applied permutations under the null hypothesis to generate the critical value for the supremum test. Under the null hypothesis, there is no change point effect on the response. Thus, we randomly shuffle the covariate X_{ij} for sufficient times. Then, we obtain the permutation distribution of the proposed test statistics. We reject the null hypothesis at a significance level of α if SUP_k is larger than the upper α -quantile of the permutation distribution.

3. Asymptotic Results

In this section, we establish the consistency and asymptotic distributions of the estimators for the change point, the regression parameters and the cumulative baseline hazard function. The following conditions are needed to establish the asymptotic properties of the estimators.

- (C.1) The density of X_{ij} is assumed to be strictly positive, bounded and continuous in a neighborhood of ζ_0 , denoted by $\mathcal{V}_0$.
- (C.2) For any ζ in $\mathcal{V}_0$, the information matrix $I(\boldsymbol{\theta}) = \int_0^\tau \mathbf{v}(t; \zeta, \boldsymbol{\beta}) s^{(0)}(t; \zeta, \boldsymbol{\beta}) \lambda_0(t) dt$ is positive definite, where $\mathbf{v}(t; \zeta, \boldsymbol{\beta}) = s^{(2)}(t; \zeta, \boldsymbol{\beta}) / s^{(0)}(t; \zeta, \boldsymbol{\beta}) - [s^{(1)}(t; \zeta, \boldsymbol{\beta}) / s^{(0)}(t; \zeta, \boldsymbol{\beta})]^2$.

$$\zeta, \beta]^{2 \otimes 2}, s^{(r)}(t; \zeta, \beta) = E \left(\sum_{j=1}^{K_i} Y_{ij}(t) \tilde{\mathbf{Z}}_{ij}^{\otimes r}(t; \zeta) \exp[r_{\theta} \{ \mathbf{W}_{ij}(t) \}] \right), \text{ and } r = 0,$$

1, 2. In addition, $\lambda_{\min} \left(\int_0^{\tau} E \left[Y_{ij}(t) \{1, \mathbf{Z}_{ij}(t)\}^{\otimes 2} \middle| X_{ij} = \zeta_0 \right] d\Lambda_0(t) \right) > 0$, where $\lambda_{\min}(\mathbf{A})$ is the smallest eigenvalue of any square matrix $\mathbf{A}$.

(C.3) There exists a convex and bounded neighborhood Θ of θ_0 such that for $k = 0, 1, 2$, and $r = 1, 2$, $\sup_{\zeta \in [\zeta_1, \zeta_2]} E \{ \sup_{t \in [0, \tau]} \sup_{\theta \in \Theta} (\|\mathbf{Z}_{ij}(t)\|^k \exp[r_{\theta} \{ \mathbf{W}_{ij}(t) \}])^\tau X_{ij} = \zeta \} < \infty$.

(C.4) The random process $\sup_{t \in [0, \tau]} \sup_{\theta \in \Theta} \|S_n^{(r)}(t; \zeta, \beta) - s^{(r)}(t; \zeta, \beta)\|$ converges almost surely to zero, where $s^{(r)}(t; \zeta, \beta) < \infty$, and $r = 0, 1, 2$. When $r = 0$, $s^{(0)}(t; \zeta, \beta)$ is bounded away from zero.

(C.5) $\sup_{t \in [0, \tau]} \lambda_0(t) < \infty$, and $P(Y_{ij}(t) = 1) > 0$ for all $t \in [0, \tau]$.

(C.6) $P(K_i = k_0) = 1$, where $1 \leq k_0 < \infty$.

(C.1) and (C.2) are needed for the identifiability of the change point and regression coefficients. (C.2) holds if $\mathbf{Z} = (\mathbf{Z}_{11}, \mathbf{Z}_{12}, \dots, \mathbf{Z}_{nK_n})$ has a full rank given $\mathbf{X} = (X_{11}, X_{12}, \dots, X_{nK_n})^T$. (C.3) shows that $s^{(r)}(t; \zeta, \beta)$ is bounded on Θ for $t \in [0, \tau]$, and it holds if all the covariates are bounded. (C.4) guarantees that $S_n^{(r)}(t; \zeta, \beta)$ converges almost surely to $s^{(r)}(t; \zeta, \beta)$. (C.5) shows that $\lambda_0(t)$ is bounded and the at risk probability is non-zero for $t \in [0, \tau]$. (C.6) assumes that all cluster sizes are bounded.

Our first two theorems establish the consistency and convergence rates of the estimators.

Theorem 1

Under conditions (C.1)–(C.6), $\hat{\theta}$ converges in probability to θ_0 in the neighborhood Θ as $n \rightarrow \infty$.

In the proof of Theorem 1, we first show that $G_n(\theta) = n^{-1} \{l_n(\theta) - l_n(\theta_0)\}$ converges uniformly to $G(\theta)$ in probability, where $G(\theta)$ is defined in the Supplementary Materials. Next, we verify that $G(\theta)$ is a strictly concave function in a neighborhood of θ_0 . From the uniform convergence of $G_n(\theta)$ to $G(\theta)$, it gives $\liminf G(\hat{\theta}) = G(\theta_0)$ with probability one. Since $G(\theta)$ has the unique maximum θ_0 in $\mathcal{V}_0$, we conclude that $\hat{\theta}$ should converge to θ_0 in probability.

Theorem 2

Under conditions (C.1)–(C.6),

$$\lim_{A \rightarrow \infty} \lim_{n \rightarrow \infty} P_0 \left(n|\hat{\zeta} - \zeta_0| > A \right) = 0,$$

$$\lim_{A \rightarrow \infty} \lim_{n \rightarrow \infty} P_0 \left(n^{1/2} \|\hat{\beta} - \beta_0\| > A \right) = 0.$$

Theorem 2 implies that the convergence rates for $\hat{\zeta}$ and $\hat{\beta}$ are $1/n$ and $1/\sqrt{n}$, respectively. These rates will be used to derive the asymptotic distributions of the estimators in Theorem 3.

Let $\theta_{n,u} = (\zeta_{n,u}, \beta_{n,u}^T)^T$, $\zeta_{n,u} = \zeta_0 + n^{-1}u_1$, and $\beta_{n,u} = \beta_0 + n^{-1/2}u_2$, where u_1 and u_2 satisfy that $(\|u_1\| + \|u_2\|^2)^{1/2} \leq n^{1/2}\epsilon$. To obtain the asymptotic distributions of the estimators, we first need the expansions of $\{l_n(\theta_{n,u}) - l_n(\theta_0)\}$. In Theorem 3, we prove that

$l_n(\theta_{n,u}) - l_n(\theta_0) = Q_n(u_1) + u_2^T \tilde{l}_n - \frac{1}{2} u_2^T I(\theta_0) u_2 + o_p(1)$, where $Q_n(u_1)$ and $\tilde{l}_n$ are defined as

$$Q_n(u_1) = \sum_{i=1}^n \sum_{j=1}^{K_i} \Delta_{ij} \left[\left\{ \beta_{20} + \beta_{30}^T \mathbf{Z}_{ij}(T_{ij}) \right\} \{I(\zeta_0 \geq X_{ij}) - I(\zeta_{n,u} \geq X_{ij})\} \right. \\ \left. - \frac{S_n^{(0)}(T_{ij}; \zeta_{n,u}, \beta_0) - S_n^{(0)}(T_{ij}; \zeta_0, \beta_0)}{S_n^{(0)}(T_{ij}; \zeta_0, \beta_0)} \right],$$

$$\tilde{l}_n = n^{-1/2} \sum_{i=1}^n \sum_{j=1}^{K_i} \int_0^\tau \left(\tilde{\mathbf{Z}}_{ij}(t, \zeta_0) - \frac{S_n^{(1)}(t; \zeta_0, \beta_0)}{S_n^{(0)}(t; \zeta_0, \beta_0)} \right) dM_{ij}(t).$$

For the cluster with m subjects, we define the set

$A_{m1}^+ = \{K_i = m, \text{ only one } X_{ij} > \zeta_0, \text{ all the other } X_{i1}, \dots, X_{ij-1}, X_{ij+1}, \dots, X_{im} \leq \zeta_0\}$, where $m = 1, \dots, K$, and $K = \max(K_i)$ is the maximum cluster size. We further define the element of A_{m1}^+ as $A_{m1}^{k+} = \{X_{ik} > \zeta_0, X_{i1}, \dots, X_{ik-1}, X_{ik+1}, \dots, X_{im} \leq \zeta_0\}$, where $k = 1, 2, \dots, m$. Similarly, A_{m1}^- and A_{m1}^{k-} are defined for the situations when only one $X_{ij} < \zeta_0$. Let $V_{mk,l}^+$ and $V_{mk,l}^-$ be independent sequences of identically and independently distributed random variables with the characteristic functions

$$E\{\exp(it V_{mk,l}^+)\} = E\left\{\exp\left(it q_s \eta_{ns,ik}^{(1)}\right) \middle| K_i = m, A_{m1}^{k+}, X_{ik} = \zeta_0^+\right\},$$

$$E\{\exp(it V_{mk,l}^-)\} = E\left\{\exp\left(it q_s \eta_{ns,ik}^{(1)}\right) \middle| K_i = m, A_{m1}^{k-}, X_{ik} = \zeta_0^-\right\},$$

where $l \geq 1$, q_s is an arbitrary constant, and

$$\eta_{ns,ik}^{(1)} = -\Delta_{ik} \{\beta_{20} + \beta_{30}^T \mathbf{Z}_{ik}(T_{ik})\} - \int_0^\tau Y_{ik}(t) \exp\{\beta_{10}^T \mathbf{Z}_{ik}(t)\} [1 - \exp\{\beta_{20} + \beta_{30}^T \mathbf{Z}_{ik}(t)\}] d\Lambda_0(t)$$

. We further denote v_{mk}^+ and v_{mk}^- to be the real jump processes such that $v_{mk}^+ = 0$ on R^- and $v_{mk}^- = 0$ on R^+ . We further denote $v_{mk}^+(s)$ to be a Poisson variable with mean rate $sp(m)f_{X_{ik}}(\zeta_0^+)P(A_{m1}^{k+}|K_i=m, A_{m1}^+)$, and $v_{mk}^-(s)$ to be a Poisson variable with mean rate $sp(m)f_{X_{ik}}(\zeta_0^-)P(A_{m1}^{k-}|K_i=m, A_{m1}^-)$, where $p(m)$ is probability of the cluster with m

subjects, $P(A_{m1}^{k+}|K_i=m, A_{m1}^+)$ is the conditional probability of A_{m1}^{k+} given A_{m1}^+ , and $f_{X_{ik}}(\zeta_0^+)$ is marginal density function of X_{ij} at ζ_0^+ . Similarly, we define $P(A_{m1}^{k-}|K_i=m, A_{m1}^-)$ and $f_{X_{ik}}(\zeta_0^-)$ for X_{ik} at ζ_0^- . Let $Q(s) \equiv Q^+(s) - Q^-(s)$, where

$$Q^+(s) \equiv \sum_{m=1}^K \sum_{k=1}^m \sum_{0 \leq l \leq v_{mk}^+(s)} V_{mk,l}^+ \text{ and } Q^-(s) \equiv \sum_{m=1}^K \sum_{k=1}^m \sum_{0 \leq l \leq v_{mk}^-(s)} V_{mk,l}^-.$$

Thus, we further establish the following Theorem 3.

Theorem 3

Under conditions (C.1)–(C.6), $n(\hat{\zeta} - \zeta_0)$ and $n^{1/2}(\hat{\beta} - \beta_0)$ are asymptotically independent. Furthermore, $n(\hat{\zeta} - \zeta_0)$ converges in distribution to $\arg \max Q(s)$. $n^{1/2}(\hat{\beta} - \beta_0)$ converges weakly to a Gaussian variable $N(0, I(\beta_0)^{-1} \Sigma(\beta_0) I(\beta_0)^{-1})$.

Note that Q_n is defined as a random variable on the space of right-continuous functions with left-hand limits equipped with the Skorohod topology. The major challenge of Theorem 3 is to prove that the process Q_n converges weakly to Q on the space $D[0, \infty)$. We show that the characteristic function of Q_n converges to the characteristic function of Q by considering all the possible situations of cluster sizes and the allocation of X_{ij} relative to ζ_0 within each cluster.

Because we have proved that the change point can be estimated more accurately with a faster convergence rate of $1/n$, the inference for this estimator remains the same as if the change point is known. In other words, the standard asymptotic results as given in Spiekerman and Lin (1998) still apply. That is, the Breslow-type estimator for the cumulative baseline hazard function $\hat{\Lambda}_0(t)$ is consistent and asymptotically normal.

4. Simulation Studies

We conducted simulation studies to evaluate the performance of our proposed method. Our first set of studies was designed to assess the bias of the estimators and the coverage rate of the confidence interval. We considered one covariate $Z \sim N(1, 4)$ and one change point variable $X \sim \text{Uniform}(0, 2)$ with the true change point at 0.75 or 1. We generated the marginal survival times $\tilde{T}_{ij}$ under the proportional hazards model $\Lambda(t|X, Z) = t \exp\{\beta_1 Z + \beta_2 I(X > \zeta) + \beta_3 Z I(X > \zeta)\}$, where $(\beta_1, \beta_2, \beta_3) = (-1, -1.5, 2)$. The censoring time follows $\text{Uniform}(0, 80)$ and the censoring rate is 10%. The correlated failure times were generated in the same way as in Cai and Shen (2000), which is a multivariate extension of the Clayton and Cuzick (1985) method. The conditional cumulative density function of the survival time for the j th subject in the i th cluster is

$$F_i(\tilde{T}_{ij}|\tilde{T}_{i1}, \dots, \tilde{T}_{i(j-1)}) = 1 - \left\{ \sum_{h=1}^j S_{ih}(\tilde{T}_{ih})^{-1/\gamma} - (j-1) \right\} \left\{ \sum_{h=1}^j S_{ih}(\tilde{T}_{ih})^{1/\gamma} - (j-2) \right\}^{\gamma+j-1},$$

where $S_{ij}(t) = P(\tilde{T}_{ij} > t)$ is the marginal survival function, γ indicates the degree of dependence between $\tilde{T}_{ij}$ and $\tilde{T}_{ih}(h = 1, \dots, j-1)$. The Kendall's tau coefficient can be

expressed as $\tau_K = \frac{1}{2\gamma+1}$, where $\gamma = 0.25$ or 1.5 indicates strong or moderate positive dependence within each cluster. We considered both the small cluster sizes with 2 or 2–5 subjects and the large cluster size with 20 subjects. The number of clusters is 100 or 200. The searching range of the change point is $[0.5, 1.5]$. To select m for each simulation, we considered q to be 5 or 10. The number of grids is 500 for the small cluster size, and 1000 for the large cluster size. For example, if the number of grids is 500, then we would search

through all the points $0.5 + j \frac{1}{500-1}$ for $j = 0, 1, \dots, 499$. All results are based on 500 replications and each m out of n bootstrap consists of 150 replicates.

In Table 1, the proposed method provides approximately unbiased estimates for the change point $\zeta = 0.75$, and the m out of n bootstrap generates proper coverage rates. When the cluster size and/or the number of clusters increase, the bias of the change point estimate and the variance estimates decrease. For the m out of n bootstrap (results not shown), the choices of m are not influenced by the dependence (moderate vs high dependence) within the clusters. However, the choice of m increases as the number of clusters increases. The results also show that the estimates for the regression coefficients β are approximately unbiased and the confidence intervals using normal approximation generate proper coverage rates for both highly and moderately correlated clusters. The finite sample performance of the change point estimator is not sensitive to the magnitude of the correlation within the clusters. However, the cluster sizes have a substantial impact on the performance of the change point estimator. In Table 2, the proposed methods draw the same conclusion for the change point $\zeta = 1$ as Table 1. Comparing Table 1 and Table 2, the results show that the finite sample performance of the change point estimator is not very sensitive to the change point location.

Table 3 shows that the Breslow-type estimator provides approximately unbiased estimates for the cumulative baseline hazard function at failure time 1 and 2. The confidence intervals using normal approximation generate proper coverage rates when the cluster size and/or the number of clusters increase.

Our second set of simulation studies were aimed at comparing type I error and power of the SUP₁, SUP₃, and SUP₁₁ tests under varying scenarios. We examine the performance of these tests with the highly/moderately correlated clusters of size 2 or 2 to 5 with 100 clusters, and clusters of size 20 with 50 clusters. We set the true change point to be 1 or 0.75, the grid for the SUP₁ test to be 1, the grids for the SUP₃ test to be $\{0.5, 1, 1.5\}$, and the grids for the SUP₁₁ test to be $\{0.5, 0.6, 0.7, \dots, 1.4, 1.5\}$. Thus, the SUP₁ test is the optimal test if the true change point is the same as the pre-assumed change point 1. The regression coefficients (β_{20}, β_{30}) are set to $(0, 0)$ for type I error, and $(0.2, -0.35)$, $(0.2, -0.27)$, or $(0.2, -0.15)$ for power under the cluster size 2, 2 to 5 and 20, respectively. The results for type I error and power are based on 10000 and 1000 replicates, respectively. All the other specifications are the same as the first set of simulations.

Table 4 shows that type I errors of all three tests are close to 0.05, regardless of where the true change point is. For the power, the performance of the supremum tests is determined by the number of grids and the minimum distance between the grids and the true change point. The minimum distance is calculated as the smallest absolute difference between the true change point and the grids. For example, when the true change point is 1, the minimum distances for all three tests are 0. In this case, the SUP_1 test is the optimal test with the highest power, while the SUP_{11} test has the lowest power. This finding is expected because the SUP_1 test is only evaluated once, while the SUP_{11} test is evaluated on more grids. When the true value is 0.75, these tests have different minimum distances. In this case, the SUP_{11} test is the most powerful test among the three tests because it has the smallest minimum distance. The SUP_1 test has a slightly higher power than SUP_3 , since both tests have the same minimum distance and the SUP_3 test is evaluated on a larger set. Consequently, the power of the supremum test increases if the minimum distance decreases. Given the same minimum distance, the tests based on a smaller set of grids have a slightly higher power.

5. Analysis of Strong Heart Study Data

The SHFS recruited 3665 American Indians (aged 15 and older) from 94 extended families in three geographic areas: Arizona, Oklahoma, and Dakota. Each participant attended clinical and physical examinations at baseline (2001–2003) and 5-year follow-up (2006–2009). There are 2315 participants free of diabetes at baseline, among whom 292 developed incident diabetes by the end of 5-year follow-up (median survival time=5.4 years). Zhao et al. (2014) used a trial-and-error approach and observed that those individuals with LTL less than the 25th percentile had a significantly higher risk of developing new diabetes than the other individuals. Here, we took a more systematic approach to identify the change point in LTL for diabetes incidence.

We included LTL with an unknown change point to be estimated, gender, age, body mass index (BMI) ($<25 \text{ kg/m}^2$, $25-29.9 \text{ kg/m}^2$, and $\geq 30 \text{ kg/m}^2$), fasting glucose, total triglycerides, and their interactions with the dichotomized LTL (long vs. short) as predictors in the Cox marginal hazards model. First of all, we applied the proposed supremum test with the robust score statistics to verify the existence of the change point. We set the grids for the supremum test to be $\{0.5, 0.9, 1.3\}$, which correspond to the lower 5% quantile, median, and upper 5% quantile of LTL, respectively. The p -value is 0.002, which is highly significant. This indicates the existence of a change point in LTL for diabetes incidence. We next applied the two-step procedure to estimate the change point and the m out of n bootstrap with 500 replicates to generate the 95% confidence interval of the change point. The range for the grid search is usually specified to be a wide range of X 's support to ensure the inclusion of the change point. In some situation, they may be specified based on biological background. In our analysis, we used the 1th and 99th quantiles as ζ_1 and ζ_2 to form the range for the grid search. The estimated change point is 0.870 and its 95% confidence interval is [0.834, 0.907]. Only the interaction between the change point of LTL and total triglycerides is statistically significant (p -value = 0.036). We removed the non-significant interaction terms and presented the final model as Model 1 in Table 5. The marginal test for the effect of total triglycerides among the participants with LTL larger than the change point is highly significant with p -value < 0.001 . For this group of participants, the increase in the level of

total triglycerides results in an increase in the risk of developing incident diabetes. In contrast, the marginal effect of total triglycerides among the participants with LTL less than the change point is not significant (p -value = 0.583). The hazard ratio of diabetes for shorter LTL ($< \zeta$) compared to longer LTL given the mean total triglycerides (147 mg/dL) is 2.476 [1.866, 3.285]. We verified proportional hazard assumptions for all covariates in Figure 1. For categorical variables (leukocyte telomere length, gender, and BMI), we generated plots of log of negative log of survival functions versus time, which show parallel trends between different levels for each covariate. For continuous variables (age, fasting glucose, and total triglycerides), the scattered plots show that the Schoenfeld residuals based on Model 1 in Table 5 are evenly distributed on both sides of the reference line, suggesting that the proportional hazards assumptions are satisfied for all predictors.

As mentioned before, Zhao et al. (2014) used a trial-and-error approach to find the change point. After trying different cutpoints, they located the change point somewhere near the first quartile (0.872). Their results are presented under Model 2 in Table 5. Although the ad-hoc estimate of the change point is very close to our estimate, their approach did not reveal a statistically significant interaction between the change point and the total triglycerides. Thus, it could not differentiate the effect of total triglycerides on developing incident diabetes among the short and long LTL participants. Based on this ad-hoc estimate, total triglycerides did not have a significant effect on developing incident diabetes for both short and long LTL participants (p -value = 0.136). In addition, the ad-hoc method cannot provide a confidence interval for the change point estimate. In contrast, our approach can estimate the change point and corresponding 95% CI.

6. Discussion

Change point effects are commonly seen in regression problems. Although a number of approaches have been developed to estimate the change point in linear regression and the univariate Cox model, no research has been done for clustered survival data. In this paper, we developed for the first time a two-step approach to estimate the change point and a testing procedure to verify the existence of a change point for clustered survival data. We adopted an adaptive m out of n bootstrap to construct the confidence interval and provided an easy way to determine the appropriate m . We proved the asymptotic properties of the proposed change point estimator. As shown in our simulation studies, the estimator is approximately unbiased and its confidence interval has a good coverage rate.

The motivation of this paper is to estimate the change point of the leukocyte telomere length for Type II diabetes. As was mentioned in the Introduction, the Hayflick limit phenomenon is the reason to assume a “jump” effect in the leukocyte telomere length for Type II diabetes. However, in some situations, there may exist a smooth change in regression coefficients. Such models for smooth change in regression coefficients were developed by Gandy et al. (2005), Gandy and Jensen (2005), and Jensen and Lütkebohmert (2008). In biomedical research, either the “jump” model or the “smooth” model could be plausible, depending on the underlying biological mechanism. Our estimation approach can be extended to handle the smooth model. However, the asymptotic properties of the smooth model will be very

different from the “jump” effect model. For example, the convergence rate in the smooth model is no longer $1/n$.

In this paper, we considered the change point analysis in the Cox-type marginal hazards model with a common baseline hazard function for clustered event data. Our method can be readily extended to incorporate non-homogeneous baseline hazard functions in studies where the baseline hazards are different for different members in a cluster or for different disease types. For the inference procedure, we can adopt the modified pseudo-partial likelihood function proposed in Wei et al. (1989). The asymptotic properties need to be modified to reflect non-homogeneous baseline hazard functions in the model. The limiting distribution of the change point estimator will follow a different compound Poisson process.

The tests for the null hypothesis $H_0 : \beta_2 = 0, \beta_3^T = 0$ are applied to verify the existence of the change point. Once the existence of the change point is established, we can fit the Cox-type marginal hazards model and test the significance of the interaction terms. We can apply the score test with the robust covariance estimator to test the null hypothesis $H_0 : \beta_3^T = 0$. This test can be used to determine whether the effects of other risk factors are different before and after the change point of the exposure variable.

We applied our methods to estimate the change point of LTL for diabetes incidence in the SHFS. Because telomere length is genetically determined (Zhu et al., 2013), it is likely that the change point is racial or ethnic specific. Thus, it will be of interest to investigate the change point of LTL in other ethnic groups. In addition, the change point of LTL is disease-specific. The estimated change point for LTL may be different for diabetes from that for other diseases, such as carotid atherosclerosis. We can apply our methods to identify the change point of LTL for other diseases in future studies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was partially supported by the National Institutes of Health grants (P01CA142538, R01ES021900 and R01DK091369) and the National Center for Research Resources grant (UL1 RR025747). The authors thank the Strong Heart Study participants, Indian Health Service facilities, and participating tribal communities for their extraordinary cooperation and involvement, which has contributed to the success of the Strong Heart Study.

References

- Abrevaya J, Huang J. On the bootstrap of the maximum score estimator. *Econometrica*. 2005; 73:1175–1204.
- Bickel, PJ., Götze, F., van Zwet, WR. Resampling fewer than n observations: gains, losses, and remedies for losses. Springer; 2012.
- Bickel PJ, Sakov A. On the choice of m in the m out of n bootstrap and its application to confidence bounds for extreme percentiles. 2005 *Unpublished manuscript*.
- Bickel PJ, Sakov A. On the choice of m in the m out of n bootstrap and confidence bounds for extrema. *Statistica Sinica*. 2008; 18:967–985.

- Cai J, Shen Y. Permutation tests for comparing marginal survival functions with clustered failure time data. *Statistics in medicine*. 2000; 19:2963–2973. [PubMed: 11042626]
- Chakraborty B, Laber EB, Zhao Y. Inference for optimal dynamic treatment regimes using an adaptive m-out-of-n bootstrap scheme. *Biometrics*. 2013; 69:714–723. [PubMed: 23845276]
- Cheung K, Lee SM, Young GA. Iterating the m out of n bootstrap in nonregular smooth function models. *Statistica Sinica*. 2005; 15:945.
- Clayton D, Cuzick J. Multivariate generalizations of the proportional hazards model. *Journal of the Royal Statistical Society. Series A (General)*. 1985:82–117.
- Cox DR. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*. 1972; 34:187–220.
- Davies RB. Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*. 1977; 64:247–254.
- Davies RB. Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*. 1987; 74:33–43.
- Dümbgen L. On nondifferentiable functions and the bootstrap. *Probability Theory and Related Fields*. 1993; 95:125–140.
- Efron B, Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical science*. 1986:54–75.
- Efron, B., Tibshirani, RJ. An introduction to the bootstrap. CRC press; 1994.
- Gandy A, Jensen U. On goodness-of-fit tests for aalen's additive risk model. *Scandinavian Journal of Statistics*. 2005; 32:425–445.
- Gandy A, Jensen U, Lütkebohmert C. A cox model with a change-point applied to an actuarial problem. *Brazilian Journal of Probability and Statistics*. 2005; 19:93–109.
- Gastwirth JL. On robust procedures. *Journal of the American Statistical Association*. 1966; 61:929–948.
- Gastwirth JL. The use of maximin efficiency robust tests in combining contingency tables and survival analysis. *Journal of the American Statistical Association*. 1985; 80:380–384.
- Hall P, Horowitz JL, Jing B-Y. On blocking rules for the bootstrap with dependent data. *Biometrika*. 1995; 82:561–574.
- Huang J, Sen P, Shao J. Bootstrapping a sample quantile when the density has a jump. *Statistica Sinica*. 1996; 6:299–309.
- Jensen U, Lütkebohmert C. A cox-type regression model with change-points in the covariates. Lifetime data analysis. 2008; 14:267–285. [PubMed: 18224437]
- Kosorok MR, Song R. Inference under right censoring for transformation models with a change-point based on a covariate threshold. *The Annals of Statistics*. 2007; 35:957–989.
- Lee, E., Wei, L., Amato, D., Leurgans, S. Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In: Klein, J., Goel, P., editors. *Survival Analysis: State of the Art*. Vol. 211. Netherlands: Springer; 1992. p. 237-247. of *Nato Science*
- Lee SM. On a class of m out of n bootstrap confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 1999; 61:901–911.
- Liang KY, Self SG, Liu XH. The Cox proportional hazards model with change point: an epidemiologic application. *Biometrics*. 1990; 46:783–793. [PubMed: 2242414]
- Luo X. The asymptotic distribution of mle of treatment lag threshold. *Journal of statistical planning and inference*. 1996; 53:33–61.
- Luo X, Boyett JM. Estimations of a threshold parameter in cox regression. *Communications in Statistics - Theory and Methods*. 1997; 26:2329–2346.
- Marquis K, Debigaré R, Lacasse Y, LeBlanc P, Jobin J, Carrier G, Maltais F. Midthigh muscle cross-sectional area is a better predictor of mortality than body mass index in patients with chronic obstructive pulmonary disease. *American Journal of Respiratory and Critical Care Medicine*. 2002; 166:809–813. [PubMed: 12231489]
- Politis, D., Romano, J. Subsampling. New York: Springer; 1999.
- Pons O. Estimation in a cox regression model with a change-point at an unknown time. *Statistics: A Journal of Theoretical and Applied Statistics*. 2002; 36:101–124.

- Pons O. Estimation in a Cox regression model with a change-point according to a threshold in a covariate. *The Annals of Statistics*. 2003; 31:442–463.
- Sen B, Banerjee M, Woodroffe M, et al. Inconsistency of bootstrap: The grenander estimator. *The Annals of Statistics*. 2010; 38:1953–1977.
- Shao J. Bootstrap sample size in nonregular cases. *Proceedings of the American Mathematical Society*. 1994; 122:1251–1262.
- Spiekerman CF, Lin D. Marginal regression models for multivariate failure time data. *Journal of the American Statistical Association*. 1998; 93:1164–1175.
- Tapp R, Zimmet P, Harper C, de Courten M, McCarty D, Balkau B, Taylor H, Welborn T, Shaw J, Group AS, et al. Diagnostic thresholds for diabetes: the association of retinopathy and albuminuria with glycaemia. *Diabetes research and clinical practice*. 2006; 73:315–321. [PubMed: 16644057]
- Wei L-J, Lin DY, Weissfeld L. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American statistical association*. 1989; 84:1065–1073.
- Willeit P, Raschenberger J, Heydon EE, Tsimikas S, Haun M, Mayr A, Weger S, Witztum JL, Butterworth AS, Willeit J, et al. Leucocyte telomere length and risk of type 2 diabetes mellitus: new prospective cohort study and literature-based meta-analysis. *PloS one*. 2014; 9:e112483. [PubMed: 25390655]
- Xu G, Sen B, Ying Z. Bootstrapping a change-point Cox model for survival data. *Electronic Journal of Statistics*. 2014; 8:1345–1379. [PubMed: 25400719]
- Zhao J, Zhu Y, Lin J, Matsuguchi T, Blackburn E, Zhang Y, Cole SA, Best LG, Lee ET, Howard BV. Short leukocyte telomere length predicts risk of diabetes in american indians: the strong heart family study. *Diabetes*. 2014; 63:354–362. [PubMed: 23949319]
- Zhu Y, Voruganti VS, Lin J, Matsuguchi T, Blackburn E, Best LG, Lee ET, MacCluer JW, Cole SA, Zhao J. Qtl mapping of leukocyte telomere length in american indians: the strong heart family study. *Aging (Albany NY)*. 2013; 5:704. [PubMed: 24036517]
- Zucker DM, Agami S, Spiegelman D. Testing for a Changepoint in the Cox Survival Regression Model. *Journal of Statistical Theory and Practice*. 2013; 7:360–380.

Figure 1.

Diagnostic Plots. The log of negative log of survival functions versus time are plotted for leukocyte telomere length, gender, and BMI. Schoenfeld residuals are plotted for age, fasting glucose, and total triglycerides.

Table 1 Simulation Results for the Change Point and Regression Parameters ($\zeta = 0.75$).

Correlation	Cluster Size	# of Clusters	Bias ($\hat{\zeta}$) ($\times 10^{-3}$)	SSD ($\hat{\zeta}$) ($\times 10^{-3}$)	95% CI($\hat{\zeta}$)	Length($\hat{\zeta}$) ($\times 10^{-2}$)	Parameters	Bias	SSD	SEE	95% CI
High	2	100	-7.41	1.57	0.95	8.56	β_1	-0.023	0.109	0.099	0.924
							β_2	-0.035	0.212	0.202	0.920
							β_3	0.043	0.174	0.164	0.942
		200	-3.32	0.84	0.94	4.27	β_1	-0.005	0.077	0.069	0.928
							β_2	-0.011	0.147	0.143	0.950
							β_3	0.013	0.129	0.115	0.922
	2-5	100	-3.71	0.90	0.97	4.80	β_1	-0.013	0.094	0.085	0.920
							β_2	-0.031	0.178	0.163	0.908
							β_3	0.027	0.169	0.149	0.926
		200	-1.25	0.47	0.96	2.45	β_1	-0.010	0.065	0.062	0.936
							β_2	-0.017	0.120	0.117	0.944
							β_3	0.018	0.113	0.108	0.936
Moderate	2	100	-0.44	0.16	0.94	0.80	β_1	-0.013	0.071	0.067	0.932
							β_2	-0.023	0.112	0.108	0.942
							β_3	0.026	0.137	0.129	0.924
		200	-0.19	0.09	0.92	0.41	β_1	-0.004	0.049	0.048	0.954
							β_2	-0.007	0.082	0.077	0.934
							β_3	0.007	0.098	0.092	0.938
	2-5	100	-6.86	1.69	0.95	8.59	β_1	-0.019	0.102	0.091	0.920
							β_2	-0.022	0.203	0.194	0.944
							β_3	0.034	0.158	0.146	0.932
		200	-3.32	0.84	0.94	4.27	β_1	-0.005	0.077	0.069	0.928
							β_2	-0.011	0.147	0.143	0.950
							β_3	0.013	0.129	0.115	0.922

Correlation	Cluster Size	# of Clusters	Bias (ζ) ($\times 10^{-3}$)	SSD(ζ) ($\times 10^{-2}$)	95% CI(ζ)	Length(ζ) ($\times 10^{-2}$)	Parameters	Bias	SSD	SEE	95% CI
							β_2	-0.023	0.161	0.148	0.922
							β_3	0.021	0.133	0.117	0.914
		200	-1.33	0.47	0.95	2.42	β_1	-0.006	0.054	0.052	0.944
							β_2	-0.010	0.106	0.106	0.948
							β_3	0.010	0.085	0.085	0.944
	20	100	-0.43	0.16	0.94	0.80	β_1	-0.008	0.045	0.042	0.914
							β_2	-0.015	0.078	0.076	0.918
							β_3	0.016	0.081	0.076	0.916
		200	-0.18	0.09	0.92	0.40	β_1	-0.001	0.033	0.030	0.930
							β_2	0.000	0.060	0.055	0.924
							β_3	0.000	0.062	0.055	0.928

NOTE: SSD and SEE stand for sample standard deviation and standard error estimate, respectively.

Table 2 Simulation Results for the Change Point and Regression Parameters ($\zeta = 1$).

Correlation	Cluster Size	# of Clusters	Bias ($\hat{\zeta}$) ($\times 10^{-3}$)	SSD ($\hat{\zeta}$) ($\times 10^{-3}$)	95% CI($\hat{\zeta}$)	Length($\hat{\zeta}$) ($\times 10^{-2}$)	Parameters	Bias	SSD	SEE	95% CI
High	2	100	-8.63	1.70	0.93	8.51	β_1	-0.021	0.102	0.093	0.922
							β_2	-0.028	0.204	0.199	0.940
							β_3	0.040	0.174	0.162	0.928
		200	-3.09	0.87	0.96	4.23	β_1	-0.004	0.071	0.065	0.922
							β_2	-0.008	0.141	0.141	0.952
							β_3	0.012	0.129	0.114	0.926
	2-5	100	-4.24	0.89	0.94	4.82	β_1	-0.011	0.089	0.081	0.912
							β_2	-0.022	0.179	0.161	0.916
							β_3	0.026	0.168	0.148	0.918
		200	-1.29	0.50	0.93	2.39	β_1	-0.010	0.064	0.059	0.932
							β_2	-0.013	0.119	0.116	0.944
							β_3	0.017	0.114	0.107	0.934
Moderate	2	100	-0.49	0.17	0.96	0.89	β_1	-0.012	0.071	0.066	0.936
							β_2	-0.019	0.112	0.108	0.938
							β_3	0.024	0.137	0.129	0.934
		200	-0.13	0.09	0.96	0.43	β_1	-0.003	0.050	0.047	0.942
							β_2	-0.005	0.083	0.077	0.946
							β_3	0.006	0.100	0.091	0.932
	2	100	-8.60	1.84	0.93	8.59	β_1	-0.017	0.094	0.085	0.914
							β_2	-0.016	0.194	0.192	0.958
							β_3	0.031	0.156	0.144	0.924
		200	-3.04	0.82	0.95	4.33	β_1	-0.006	0.064	0.059	0.918
							β_2	-0.006	0.133	0.136	0.946
							β_3	0.011	0.113	0.102	0.920
	2-5	100	-4.36	0.89	0.94	4.86	β_1	-0.010	0.073	0.066	0.904

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Correlation	Cluster Size	# of Clusters	Bias (ζ) ($\times 10^{-3}$)	SSD(ζ) ($\times 10^{-2}$)	95% CI(ζ)	Length(ζ) ($\times 10^{-2}$)	Parameters	Bias	SSD	SEE	95% CI
							β_2	-0.016	0.160	0.147	0.932
							β_3	0.022	0.131	0.116	0.912
		200	-1.27	0.50	0.93	2.41	β_1	-0.006	0.052	0.048	0.940
							β_2	-0.004	0.107	0.105	0.936
							β_3	0.008	0.086	0.084	0.936
	20	100	-0.52	0.18	0.96	0.89	β_1	-0.007	0.045	0.041	0.900
							β_2	-0.010	0.079	0.075	0.924
							β_3	0.014	0.082	0.076	0.900
		200	-0.11	0.09	0.97	0.43	β_1	0.001	0.033	0.029	0.926
							β_2	0.002	0.060	0.054	0.932
							β_3	-0.002	0.062	0.055	0.932

NOTE: SSD and SEE stand for sample standard deviation and standard error estimate, respectively.

Table 3

Simulation Results for the Cumulative Baseline Hazard Function.

Correlation	Cluster Size	Number of Clusters	Failure Time	Bias	SSD	SEE	95% CI
High	2	100	1	0.038	0.195	0.179	0.940
			2	0.120	0.419	0.356	0.902
		200	1	0.024	0.129	0.128	0.944
			2	0.052	0.248	0.251	0.956
	2-5	100	1	0.033	0.170	0.153	0.922
			2	0.077	0.321	0.300	0.932
		200	1	0.014	0.114	0.108	0.928
			2	0.033	0.212	0.213	0.952
	20	100	1	0.007	0.113	0.111	0.940
			2	0.034	0.232	0.218	0.930
		200	1	0.010	0.081	0.079	0.954
			2	0.027	0.163	0.156	0.950
Moderate	2	100	1	0.033	0.180	0.171	0.938
			2	0.103	0.399	0.343	0.910
		200	1	0.017	0.127	0.122	0.932
			2	0.056	0.245	0.242	0.940
	2-5	100	1	0.030	0.157	0.139	0.914
			2	0.084	0.292	0.276	0.928
		200	1	0.013	0.101	0.098	0.940
			2	0.035	0.207	0.196	0.950
	20	100	1	0.007	0.089	0.084	0.930
			2	0.025	0.183	0.172	0.928
		200	1	0.003	0.065	0.060	0.936
			2	0.013	0.134	0.123	0.922

NOTE: SSD and SEE stand for sample standard deviation and standard error estimate, respectively.

Table 4

Type I Error and Power for SUP Tests for the Existence of the Change Point

Correlation	ζ_0	Cluster Size	Number of Clusters	(β_{20}, β_{30})	SUP ₁ (Optimal)	SUP ₃	SUP ₁₁
High	1	2	100	(0, 0)	0.051	0.051	0.052
				(0.2, -0.35)	0.912	0.844	0.840
		2-5	100	(0, 0)	0.052	0.053	0.053
				(0.2, -0.27)	0.952	0.886	0.878
		20	50	(0, 0)	0.048	0.050	0.050
				(0.2, -0.15)	0.934	0.904	0.878
	0.75	2	100	(0, 0)	0.051	0.051	0.052
				(0.2, -0.35)	0.732	0.642	0.756
		2-5	100	(0, 0)	0.052	0.053	0.053
				(0.2, -0.27)	0.754	0.682	0.824
		20	50	(0, 0)	0.048	0.050	0.050
				(0.2, -0.15)	0.754	0.732	0.796
Moderate	1	2	100	(0, 0)	0.050	0.050	0.050
				(0.2, -0.35)	0.922	0.846	0.832
		2-5	100	(0, 0)	0.051	0.052	0.050
				(0.2, -0.27)	0.932	0.888	0.870
		20	50	(0, 0)	0.053	0.051	0.051
				(0.2, -0.15)	0.944	0.908	0.892
	0.75	2	100	(0, 0)	0.050	0.050	0.050
				(0.2, -0.35)	0.712	0.662	0.754
		2-5	100	(0, 0)	0.051	0.052	0.050
				(0.2, -0.27)	0.766	0.722	0.820
		20	50	(0, 0)	0.053	0.051	0.051
				(0.2, -0.15)	0.788	0.764	0.822

Table 5

Analysis Results Based on the Strong Heart Family Study: Model 1 ($\hat{\zeta} = 0.870[0.834, 0.907]$) and Model 2 ($\hat{\zeta}_{\text{ad-hoc}} = 0.872$).

Parameter	Model 1			Model 2		
	Estimate	SE	p-value	Estimate	SE	p-value
TOTAL TRIGLYCERIDES (mg/dL)	-0.001	0.001	0.583	0.001	0.001	0.136
GENDER	-0.333	0.115	0.004	-0.348	0.121	0.004
AGE	-0.002	0.005	0.723	-0.001	0.005	0.838
BMI (25–30)	0.329	0.335	0.326	0.341	0.334	0.308
BMI (> 30)	1.100	0.342	0.001	1.126	0.343	0.001
FASTING GLUCOSE (mg/dL)	0.068	0.006	< 0.001	0.066	0.006	< 0.001
TELOMERE LENGTH(> ζ)	-1.334	0.270	< 0.001	-0.768	0.146	< 0.001
TELOMERE LENGTH(> ζ) $\times$ TOTAL TRIGLYCERIDES	0.003	0.001	0.036			