Statistical Theory and Related Fields

ISSN: 2475-4269 (Print) 2475-4277 (Online) Journal homepage: https://www.tandfonline.com/loi/tstf20

Robust dynamic risk prediction with longitudinal studies

Qian M. Zhou, Wei Dai, Yingye Zheng & Tianxi Cai

To cite this article: Qian M. Zhou, Wei Dai, Yingye Zheng & Tianxi Cai (2017) Robust dynamic risk prediction with longitudinal studies, Statistical Theory and Related Fields, 1:2, 159-170, DOI: 10.1080/24754269.2017.1400418

To link to this article: https://doi.org/10.1080/24754269.2017.1400418

Published online: 27 Nov 2017.

STATISTICAL THEORY AND RELATED FIELDS, 2017
VOL. 1, NO. 2, 159–170
https://doi.org/10.1080/24754269.2017.1400418

Taylor & Francis
Taylor & Francis Group

Robust dynamic risk prediction with longitudinal studies

Qian M. Zhou [a,*], Wei Dai [b,*], Yingye Zheng [c] and Tianxi Cai [a]

[a]Department of Mathematics and Statistics, Mississippi State University, Mississippi, USA; [b]Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA; [c]Department of Biostatistics and Biomathematics, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

ABSTRACT

Providing accurate and dynamic age-specific risk prediction is a crucial step in precision medicine. In this manuscript, we introduce an approach for estimating the τ-year age-specific absolute risk directly via a flexible varying coefficient model. The approach facilitates the utilisation of predictors varying over an individual's lifetime. By using a nonparametric inverse probability weighted kernel estimating equation, the age-specific effects of risk factors are estimated without requiring the specification of the functional form. The approach allows borrowing information across individuals of similar ages, and therefore provides a practical solution for situations where the longitudinal information is only measured sparsely. We evaluate the performance of the proposed estimation and inference procedures with numerical studies, and make comparisons with existing methods in the literature. We illustrate the performance of our proposed approach by developing a dynamic prediction model using data from the Framingham Study.

ARTICLE HISTORY
Received 21 March 2017
Revised 31 October 2017
Accepted 31 October 2017

KEYWORDS
Inverse probability weighting; longitudinal markers; nonparametric smoothing; predictive accuracy; risk prediction; survival analysis

1. Introduction

Accurate and individualised risk prediction is a key component of precision medicine. For example in colorectal cancer, risk calculator can help tailoring individual's screening regimen and making decisions on specific ages for screening initialisation and surveillance. Factors pertaining to specific ages, such as family history and nutrition intake are important to be incorporated in the outcome prediction. For cardiovascular disease (CVD), the Framingham risk score (FRS) (Wolf, D'Agostino, Belanger, & Kannel, 1991) has been developed separately for men and women based on risk factors such as total cholesterol, HDL, systolic blood pressure and smoking status. Patients with 10-year FRS below 10% are considered to be at lower risk for vascular events during the next decade, whereas patients with scores between 10% and 20% are at moderate risk and those larger than 20% are at higher risk. Various intervention strategies can be implemented based on such risk stratification (Mosca et al., 2004). In these clinical settings, the analytical goal is to provide patients with an estimate of the likelihood of developing a disease within the next τ-years given the subject is disease free at age a and his/her risk profile updated by age a, i.e., the age-specific absolute τ-year residual life risk.

Currently available prediction models are often limited in predicting such age-specific absolute residual life risk. For example, the FRS model includes age as a standard risk factor with linear effects. However, it is well recognised that the relationship between age and CVD risk may change in magnitude through complex interactions with other risk factors (Ridker, Buring, Rifai, & Cook, 2007). Naturally, a risk equation should be a more stochastic function of age (Lloyd-Jones, 2010). Such simplistic models, failing to capture the complex age varying effects, may lead to poor risk estimates and prediction models with low discriminatory power. To be clinically useful with sufficiently adequate prediction accuracy, an ideal prediction model should take into account an individual's most up to date health information and reflect the fact that risk factors may have differential effects on the τ-year residual life risk over various stages of an individual's life span.

Constructing a model that accurately captures how risks change dynamically over lifetime, while of great importance, is challenging for several reasons. Often important risk factors change over time. However, collecting such time-varying information for a large prospective cohort can be a major undertaking. Most cohort studies only collect age-specific information intermittently, sometimes irregularly. Characterising changes over time with limited measures at discrete-time points requires much deliberation. Furthermore, the importance of risk factors on disease outcome may change during an individual's lifetime. For example, body mass index may have substantially different effects on future CVD risks depending on the age. Characterising age-varying effects with a powerful yet flexible statistical model can be challenging. Similar challenges arise when assessing the prediction performance of an

CONTACT Tianxi Cai tcai@hsph.harvard.edu
*These authors contributed equally to this work.

age-specific prediction model since the accuracy may also vary with age.

A popular approach in the risk prediction literature is to decompose the absolute risk into two components: the age-dependent disease risk for a baseline risk profile, which can be estimated from a prospective cohort study or external disease incidence data, and the relative risk of developing disease for a particular risk factor profile compared to the baseline (Gail et al., 1989; Liu, Zheng, Prentice, & Hsu, 2014). While such an approach can accommodate differential effects of risk factors at different ages by adding interaction terms of age and the other covariates in the regression model, it does not incorporate time-varying covariates collected over time. To incorporate repeated measurements, joint modelling (JM) of both the covariate process $Z(\cdot)$ and a survival outcome T have been developed in recent years (Tsiatis, DeGruttola, & Wulfsohn, 1995; Wang & Taylor, 2001; Ye, Lin, & Taylor, 2008). Parameter estimation typically involves specifying the covariate process $Z(\cdot)$ and then linking $Z(\cdot)$ to T via a proportional hazard (PH) model with time-varying covariates. In addition to requiring strong modelling assumptions about $Z(\cdot)$ and T, these joint modelling (JM) methods have a limitation of being computationally infeasible when many time-varying covariates are under consideration. Semi-parametric methods have been proposed to directly make prediction of τ-year residual life risk given covariate information at a landmark time t_0 (Parast, Cheng, & Cai, 2012; Zheng & Heagerty, 2005). However, no existing methods allow for incorporating age-specific effects nonparametrically with sparsely measured time-varying covariates. Furthermore, procedures for nonparametrically evaluating such age-specific prediction models with longitudinal markers and censored event times are not yet available.

In this paper, we propose to directly model the age-specific absolute risk function via a flexible varying-coefficient model and estimate the covariate effects as functions of age via inverse probability weighted (IPW) kernel estimating equations. The procedure allows the estimation of age-specific risks flexibly with the longitudinally collected risk factor information on the same patient, while borrowing strength across individuals of similar ages at different study time points. It also handles irregularly measured serial covariates and censoring easily. Our proposed model, by allowing the effects of risk factors to change over age and the target residual life span τ, is more realistic and could potentially lead to improved predictive performance. To quantify the performance of the age-specific models of τ-year residual life risk, it would be desirable to consider measures of prediction performance specific to age and prediction time since the prediction accuracy of such models is likely to vary over both dimensions. A wide range of performance measures has been considered to quantify the time-specific prediction accuracy of τ-year absolute risk models constructed with baseline markers (Gerds, Cai, & Schumacher, 2008; Uno, Cai, Tian, & Wei, 2007; Zheng, Cai, & Feng, 2006). In the longitudinal setting, Zheng and Heagerty (2004) considered a model-based approach for estimating the accuracy in the absence of censoring. To guard against potential model misspecification and incorporate censored outcomes, we propose an IPW kernel estimator to calculate model performance parameters that quantify the accuracy of the proposed prediction models in predicting τ-year residual life at age a. No existing methods provide nonparametric estimates of such prediction performance with longitudinal markers and censored outcomes.

The remainder of this manuscript is organised as follows. In Section 2; we describe the proposed model and estimation framework. Then, we present simulation results comparing our approach to other popular methods in Section 3. We apply the proposed method to the Framingham Heart Study in Section 4, assessing the age-specific effects of routinely used cardiovascular risk factors on the 10-year residual CVD risk and quantifying the performances of the prediction models. We conclude with a brief discussion in Section 5.

2. Methods

Let T_i be the time to event onset since a baseline time such as study entry. Due to censoring, one can only observe $X_i = \min\{T_i, C_i\}$ and $\delta_i = I(T_i \le C_i)$, where C_i is the censoring time and $I(\cdot)$ is the indicator function. To facilitate the calculation of age-specific risks, we also record age at the occurrence of the event and censoring. Let A_{i0} be the age at which subject i enters the study and then $T_i^{\mathbb{A}} = T_i + A_{i0}$ is the age at which the event occurs, and $C_i^{\mathbb{A}} = C_i + A_{i0}$ is the age at which $T_i^{\mathbb{A}}$ might be censored. Let $X_i^{\mathbb{A}} = X_i + A_{i0} = \min\{T_i^{\mathbb{A}}, C_i^{\mathbb{A}}\}$. In addition to the event time information, risk markers are ascertained repeatedly during the follow-up. For the ith subject, let $\mathbf{Z}_i(a) = \left(Z_{i1}(a), Z_{i2}(a), \ldots, Z_{ip}(a)\right)^{\top}$ denote a vector of p risk factors measured at age a, let $\{A_{ik}, k = 0, ..., m_i\}$ be the ages at which these risk factors are collected and $\mathbf{Z}_{ij} = \mathbf{Z}_i(A_{ij})$, where m_i is the total number of measurement times. We assume that $\mathbf{Z}_i(a)$ is potentially observable among those with $T_i^{\mathbb{A}} > a$, the values of $\mathbf{Z}_i(A_{ij})$ are not dependent on the study measurement time $A_{ij} - A_{i0}$ given age A_{ij}. In addition, C_i is assumed independent of T_i , $\mathbf{Z}_i(\cdot)$, entry age A_{i0}, and the underlying study measurement times, with support not shorter than that of $A_{ij} - A_{i0} + \tau$. Figure 1 provides a graphical illustration of the data structure.

2.1. Modelling, estimation and inference

We are interested in estimating the risk of experiencing the event in the next τ-years for subjects who are at age a and event-free, based on the risk factor measured at

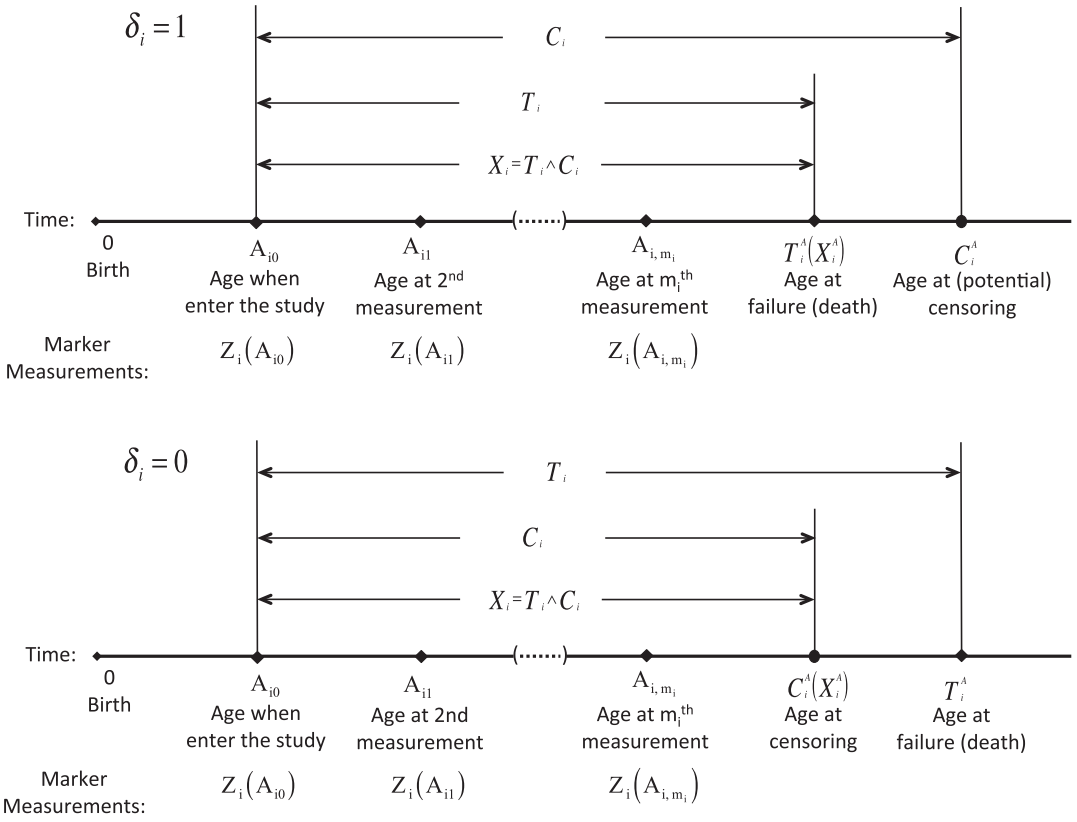


Figure 1. Data structure.

age a, $\mathbf{Z}(a)$. Thus, the goal is to estimate the conditional risk function:

$$\pi_{\tau,a}(\mathbf{z}) = \Pr\left\{T^{\mathbb{A}} - a \le \tau \mid T^{\mathbb{A}} > a, \mathbf{Z}(a) = \mathbf{z}, A = a\right\}.$$

To approximate $\pi_{\tau,a}(\mathbf{z})$, we propose a flexible varying coefficient model:

$$\pi_{\tau,a}\{\mathbf{Z}(a)\} = g\left\{\boldsymbol{\beta}_\tau(a)^{\mathrm{T}}\mathbf{U}(a)\right\}, \tag{1}$$

where $g(\cdot)$ is a known smooth probability distribution function, such as $g(x) = \exp(x)/\{1 + \exp(x)\}$ (which is used in Section 4), $\mathbf{U}(a) = \boldsymbol{\psi}\{\mathbf{Z}(a)\}$ represents transformed risk factors for some known function $\boldsymbol{\psi}$ and $\mathbf{U}(a)$ includes 1 as the first component, such as a log transformed risk factor (used in Section 4), $\boldsymbol{\beta}_\tau(a)$ is an unknown smooth function representing the covariate effects on the τ-year residual life risk at age a. Model (1) allows the effects of risk factors $\mathbf{Z}(a)$ to vary over both age and the residual life span τ. This flexibility is attractive when the risk factors have different effects on long-term versus short-term risks and when certain risk factor profiles have more detrimental effects for younger subjects than for older subjects.

To estimate $\boldsymbol{\beta}_\tau(a)$ for any given age a in the presence of censoring, we propose to obtain $\widehat{\boldsymbol{\beta}}_\tau(a)$ as the solution to the IPW kernel smoothed estimating equation, $\widehat{\boldsymbol{\Phi}}_a(\boldsymbol{\beta}) = 0$, where

$$\widehat{\boldsymbol{\Phi}}_a(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{m_i}\widehat{w}_{\tau ij}\, I(X_i^{\mathbb{A}} \ge A_{ij})\, K_h(A_{ij} - a) \times \mathbf{U}_{ij}\left\{I\left(X_i^{\mathbb{A}} < A_{ij} + \tau\right) - g(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{U}_{ij})\right\}, \tag{2}$$

where $\mathbf{U}_{ij} = \mathbf{U}_i(A_{ij})$, $\widehat{w}_{\tau ij} = \delta_i I(X_i^{\mathbb{A}} \le A_{ij} + \tau)/\widehat{G}(X_i) + I\left(X_i^{\mathbb{A}} > A_{ij} + \tau\right)/\widehat{G}(A_{ij} - A_{i0} + \tau)$ with $\widehat{G}(\cdot)$ being the Kaplan–Meier estimator for the survival function of the censoring time $G(c) = P(C_i > c)$ of C_i, and $K_h(s) = K(s/h)/h$ is a symmetric standard kernel function $K(\cdot)$ with a finite support and with h the smoothing parameter. Note that under the independent censoring assumption, $E(\widehat{w}_{\tau ij} \mid T_i^{\mathbb{A}}, A_{ij}, Z_{ij}, j = 0, , \ldots, m_i) \approx 1$.

Following similar arguments as given in Cai, Tian, Uno, Solomon, and Wei (2010) and Parast et al. (2012), one may show that $\widehat{\boldsymbol{\beta}}_\tau(a)$ converges in probability to a deterministic vector $\bar{\boldsymbol{\beta}}_\tau(a)$ as $n \to \infty$ regardless of whether (1) is correctly specified or not. In addition, one may also show that for $h = O_p(n^{-\upsilon})$ with $\upsilon \in (1/5, 1/2)$, $(nh)^{\frac{1}{2}}\{\widehat{\boldsymbol{\beta}}_\tau(a) - \bar{\boldsymbol{\beta}}_\tau(a)\}$ converges in distribution to a zero-mean normal random vector for any given a. However, it is difficult to directly estimate the asymptotic variance of $(nh)^{\frac{1}{2}}\{\widehat{\boldsymbol{\beta}}_\tau(a) - \bar{\boldsymbol{\beta}}_\tau(a)\}$. To construct confidence interval (CI) for $\boldsymbol{\beta}_\tau(a)$ in practice, we suggest using a perturbation resampling (sometimes referred to as wild bootstrap) method (Park & Wei, 2003; Tian, Zucker, & Wei, 2005; Wu, 1986) to

approximate the distribution of the proposed estimator. Compared to the standard bootstrap, perturbation resampling tends to be more stable especially in the survival setting since all observations take positive weights and contribute to the estimation. Let $\mathcal{V}^{(b)} = \{V_1^{(b)}, \ldots, V_n^{(b)}\}$, $b = 1, \ldots, B$, be B sets of independent positive random variables from a known distribution with mean and variance equal to one. Then, one may obtain perturbed estimates of $\boldsymbol{\beta}_\tau(a)$, $\widehat{\boldsymbol{\beta}}_\tau^{(b)}(a)$, as the solution to the equation:

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{m_i}\widehat{w}_{\tau ij}^{(b)}\, I(X_i^{\mathbb{A}} \geq A_{ij})\, K_h(A_{ij}-a)\, \mathbf{U}_{ij} \times \left\{I\left(X_i^{\mathbb{A}} < A_{ij}+\tau\right) - g(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{U}_{ij})\right\} = 0,$$

where $\widehat{w}_{\tau ij}^{(b)} = V_i^{(b)}\{\delta_i I\left(X_i^{\mathbb{A}} \leq A_{ij}+\tau\right)/\widehat{G}^{(b)}(X_i) + I\left(X_i^{\mathbb{A}} > A_{ij}+\tau\right)/\widehat{G}^{(b)}(A_{ij}-A_{i0}+\tau)\}$ and $\widehat{G}^{(b)}(\cdot)$ is the weighted Kaplan–Meier estimator of $G(\cdot)$ with each subject's contribution to the estimator weighted by $V_i^{(b)}$. The asymptotic variance of $\widehat{\boldsymbol{\beta}}_\tau(a)$, $\hat{\sigma}^2_{\widehat{\boldsymbol{\beta}}_\tau(a)}$, can be estimated by the empirical variance of $\{\widehat{\boldsymbol{\beta}}_\tau^{(b)}(a),\ b = 1, \ldots, B\}$. The $100(1-\alpha)\%$ simultaneous confidence bands for $\{\boldsymbol{\beta}_\tau(a),\ a_l < a < a_u\}$ can be obtained as $\{\widehat{\boldsymbol{\beta}}_\tau(a) \pm \zeta_\alpha \hat{\sigma}_{\widehat{\boldsymbol{\beta}}_\tau(a)}\}$, where ζ_α is the $100(1-\alpha)$th percentile of $\{\sup_{a_l<a<a_u}|\widehat{\boldsymbol{\beta}}_\tau^{(b)}(a) - \widehat{\boldsymbol{\beta}}_\tau(a)|/\hat{\sigma}_{\widehat{\boldsymbol{\beta}}_\tau(a)},\ b = 1, \ldots, B\}$. To justify the resampling method, we may first show that

$$(nh)^{\frac{1}{2}}\{\widehat{\boldsymbol{\beta}}_\tau(a) - \boldsymbol{\beta}_\tau(a)\} = (nh)^{-1/2} \times \sum_{i=1}^{n}\sum_{j=0}^{m_i} K_h(A_{ij}-a)\mathcal{U}(A_{ij}, \mathbf{U}_{ij}, X_i^{\mathbb{A}}) + O_p(h^{\frac{1}{2}})$$

where $\mathcal{U}$ is some deterministic function and $E\{\mathcal{U}(A_{ij}, \mathbf{U}_{ij}, X_i^{\mathbb{A}})\} = 0$. Thus for any fixed a and $h = O_p(n^{-\upsilon})$ with $\upsilon \in (1/5, 1/2)$, $(nh)^{\frac{1}{2}}\{\widehat{\boldsymbol{\beta}}_\tau(a) - \boldsymbol{\beta}_\tau(a)\}$ is asymptotically normal. Furthermore, following similar arguments as given in Tian et al. (2005), we may show that with proper normalisation, $\sup_a |(nh)^{\frac{1}{2}}\{\widehat{\boldsymbol{\beta}}_\tau(a) - \boldsymbol{\beta}_\tau(a)\}|$ converges to an extreme value distribution. In addition,

$$(nh)^{\frac{1}{2}}\{\widehat{\boldsymbol{\beta}}_\tau^{(b)}(a) - \widehat{\boldsymbol{\beta}}_\tau(a)\} = (nh)^{-1/2}\sum_{i=1}^{n}\sum_{j=0}^{m_i} \times K_h(A_{ij}-a)\mathcal{U}(A_{ij}, \mathbf{U}_{ij}, X_i^{\mathbb{A}})(V_i^{(b)}-1) + O_{p^*}(h^{\frac{1}{2}}),$$

where O_{p^*} is with respect to probability space generated by both the observed data and $\mathcal{V}^{(b)}$. Then, we may show that conditional on the data, $(nh)^{-1/2}\sum_{i=1}^{n}\sum_{j=0}^{m_i} K_h(A_{ij}-a)\mathcal{U}(A_{ij}, \mathbf{U}_{ij}, X_i^{\mathbb{A}})(V_i^{(b)}-1)$ converges in distribution to the same limiting unconditional distribution of $(nh)^{-1/2}\sum_{i=1}^{n}\sum_{j=0}^{m_i} K_h(A_{ij}-a)\mathcal{U}(A_{ij}, \mathbf{U}_{ij}, X_i^{\mathbb{A}})$.

2.2. Accuracy measure estimation

With $\boldsymbol{\beta}_\tau(a)$ estimated as $\widehat{\boldsymbol{\beta}}_\tau(a)$, one can then use $\widehat{\pi}_{\tau,a}\{\mathbf{Z}(a)\} = g\{\widehat{\boldsymbol{\beta}}_\tau(a)^{\mathsf{T}}\mathbf{U}(a)\}$ to estimate the τ-year survival probability for event-free subjects with risk factor profile $\mathbf{Z}(a)$ at age a. To assess the accuracy of the limiting risk model $\bar{\pi}_{\tau,a}\{\mathbf{Z}(a)\} = g\{\bar{\boldsymbol{\beta}}_\tau(a)^{\mathsf{T}}\mathbf{U}(a)\}$ in predicting the τ-year residual survival status for different age groups, we extend commonly used time-dependent accuracy parameters, such as true positive rate (TPR), false positive rate (FPR) and area under the receiver operating characteristic curve (AUC), to also incorporate the age dimension. Since the proposed model evaluation method is not limited to a specific prediction model, we next describe these accuracy parameters for a genetic τ-year residual life risk function, $\Pi_{\tau,A}(\mathbf{Z})$, derived based on an age A and the risk factor $\mathbf{Z}(\cdot)$ collected up to age A. For the proposed model, $\Pi_{\tau,A}(\mathbf{Z}) = g\{\boldsymbol{\beta}_\tau(A)^{\mathsf{T}}\mathbf{U}(A)\}$.

2.2.1. Time and age-specific prediction accuracy

When a specific age group a is of interest, we summarise the prediction performance of risk model $\Pi_{\tau,A}(\mathbf{Z})$ using time and age-specific TPR and FPR functions, respectively, defined as

$$\mathrm{TPR}_{\tau,a}(c) = P\left\{\Pi_{\tau,A}(\mathbf{Z}) > c \mid 0 < T^{\mathbb{A}} - a \leq \tau, A = a\right\},$$
$$\mathrm{FPR}_{\tau,a}(c) = P\left\{\Pi_{\tau,A}(\mathbf{Z}) > c \mid T^{\mathbb{A}} - a > \tau, A = a\right\}.$$

We may summarise the overall predictiveness of the model for a given a and τ using

$$\begin{aligned}\mathrm{AUC}_{\tau,a} &= \int \mathrm{TPR}_{\tau,a}(c)d\mathrm{FPR}_{\tau,a}(c)\\ &= P\big[\Pi_{\tau,A_i}(\mathbf{Z}_i) \geq \Pi_{\tau,A_{i'}}(\mathbf{Z}_{i'}) \mid 0 < T_i^{\mathbb{A}} - a \leq \tau, T_{i'}^{\mathbb{A}} - a \geq \tau, A_i = A_{i'} = a\big],\end{aligned}$$

where i and i' index two independent individuals. Similar to the estimation of $\boldsymbol{\beta}_\tau(a)$, these parameters can be estimated using an IPW kernel smoothing approach. For example, $\mathrm{TPR}_{\tau,a}(c)$ can be estimated as

$$\widehat{\mathrm{TPR}}_{\tau,a}(c) = \frac{\sum_{i,j:0<X_i^{\mathbb{A}}-a<\tau}\widehat{w}_{\tau ij}K_h(A_{ij}-a)I\{\widehat{\Pi}_{\tau,A_{ij}}(\mathbf{Z}_i) \geq c\}}{\sum_{i,j:0<X_i^{\mathbb{A}}-a<\tau}\widehat{w}_{\tau ij}K_h(A_{ij}-a)}. \quad (3)$$

and $\mathrm{AUC}_{\tau,a}$ can be estimated as

$$\widehat{\mathrm{AUC}}_{\tau,a} = \frac{\sum_{i,j,i',j':0<X_i^{\mathbb{A}}-a<\tau\leq X_{i'}^{\mathbb{A}}-a}\widehat{w}_{\tau ij}\,\widehat{w}_{\tau i'j'}K_h(A_{ij}-a)\,K_h(A_{i'j'}-a)I\{\widehat{\Pi}_{\tau,A_{ij}}(\mathbf{Z}_i) \geq \widehat{\Pi}_{\tau,A_{i'j'}}(\mathbf{Z}_{i'})\}}{\sum_{i,j,i',j':0<X_i^{\mathbb{A}}-a<\tau\leq X_{i'}^{\mathbb{A}}-a}\widehat{w}_{\tau ij}\,\widehat{w}_{\tau i'j'}K_h(A_{ij}-a)\,K_h(A_{i'j'}-a)}. \quad (4)$$

where $\widehat{\Pi}_{\tau,A_{ij}}(\mathbf{Z}_i)$ is the estimated risk function plugging in estimated model parameters. For the proposed varying coefficient model, $\widehat{\Pi}_{\tau,A_{ij}}(\mathbf{Z}_i) = g\{\widehat{\boldsymbol{\beta}}_\tau(A_{ij})^{\mathrm{T}}\mathbf{U}_{ij}\}$ and

$$n^{-2}\sum_{i,i'}\int \widehat{\mathrm{TPR}}_{\tau,A_{i0}}(c)d\widehat{\mathrm{FPR}}_{\tau,A_{i'0}}(c) \approx \frac{\sum_{i,i':0<X_i^{\mathbb{A}}-A_{i0}<\tau\le X_{i'}^{\mathbb{A}}-A_{i'0}} \hat{w}_{\tau i0}\,\hat{w}_{\tau i'0} I[\widehat{\Pi}_{\tau,A_{i0}}\{\mathbf{Z}_i(A_{i0})\} \ge \widehat{\Pi}_{\tau,A_{i'0}}\{\mathbf{Z}_{i'}(A_{i'0})\}]}{\sum_{i,i':0<X_i^{\mathbb{A}}-A_{i0}<\tau\le X_{i'}^{\mathbb{A}}-A_{i'0}} \hat{w}_{\tau i0}\,\hat{w}_{\tau i'0}}. \quad (6)$$

using similar arguments as those for the consistency of $\widehat{\boldsymbol{\beta}}_\tau(a)$, one may show that $\widehat{\mathrm{AUC}}_{\tau,a}$ is a consistent estimator of $\mathrm{AUC}_{\tau,a}$. Furthermore, $(nh)^{\frac{1}{2}}(\widehat{\mathrm{AUC}}_{\tau,a} - \mathrm{AUC}_{\tau,a})$ converges in distribution to a normal random variable. The CI for $\mathrm{AUC}_{\tau,a}$ can be constructed by perturbed estimates. Specifically, for $b = 1, ..., B$, the bth perturbed estimate of $\widehat{\mathrm{AUC}}_{\tau,a}$ can be obtained as

$$\widehat{\mathrm{AUC}}^{(b)}_{\tau,a} = \frac{\sum_{i,j,i',j':0<X_i^{\mathbb{A}}-a<\tau\le X_{i'}^{\mathbb{A}}-a} \widehat{w}^{(b)}_{\tau ij}\widehat{w}^{(b)}_{\tau i'j'} K_h(A_{ij}-a)\,K_h(A_{i'j'}-a) I\{\widehat{\Pi}^{(b)}_{\tau,A_{ij}}(\mathbf{Z}_i) \ge \widehat{\Pi}^{(b)}_{\tau,A_{i'j'}}(\mathbf{Z}_{i'})\}}{\sum_{i,j,i',j':0<X_i^{\mathbb{A}}-a<\tau\le X_{i'}^{\mathbb{A}}-a} \widehat{w}^{(b)}_{\tau ij}\,\widehat{w}^{(b)}_{\tau i'j'} K_h(A_{ij}-a)\,K_h(A_{i'j'}-a)}. \quad (5)$$

where $\widehat{\Pi}^{(b)}_{\tau,A_{ij}}(\mathbf{Z}_i)$ is the perturbed counterpart of $\widehat{\Pi}_{\tau,A_{ij}}(\mathbf{Z}_i)$ obtained similar to $\widehat{\boldsymbol{\beta}}^{(b)}_\tau(a)$ using weights $\mathcal{V}^{(b)}$.

2.2.2. Time-specific prediction accuracy including age as a predictor

When interest lies in evaluating the accuracy of the prediction model treating age as a risk factor, an overall summary that is not conditional on age would be preferred. That is, incorporating age as a predictor, one may seek to assess the accuracy in predicting τ-year residual life of the risk estimate $\Pi_{\tau,A}(\mathbf{Z})$, constructed using $\mathbf{Z}(\cdot)$ information collected up to a random age A among those with $T^{\mathbb{A}} > A$. For such settings, one may consider

$$\begin{aligned}\mathrm{TPR}_\tau(c) &= P\left[\Pi_{\tau,A}(\mathbf{Z}) > c \mid 0 < T^{\mathbb{A}} - A \le \tau\right] \\ &= \int \mathrm{TPR}_{\tau,a}d\mathcal{F}(a) = E(\mathrm{TPR}_{\tau,A}), \\ \mathrm{FPR}_\tau(c) &= P\left[\Pi_{\tau,A}(\mathbf{Z}) > c \mid T^{\mathbb{A}} - A > \tau\right] \\ &= \int \mathrm{FPR}_{\tau,a}d\mathcal{F}(a) = E(\mathrm{FPR}_{\tau,A}).\end{aligned}$$

where $\mathcal{F}$ is the distribution of the age at measurement. The overall performance of the risk model $\Pi_{\tau,A}(\mathbf{Z})$ for predicting τ-year residual life can be summarised by

$$\begin{aligned}\mathrm{AUC}_\tau &= P\big[\Pi_{\tau,A_i}(\mathbf{Z}_i) > \Pi_{\tau,A_{i'}}(\mathbf{Z}_j) \mid 0 \\ &\quad \le T_i^{\mathbb{A}} - A_i < \tau,\, T_{i'}^{\mathbb{A}} - A_{i'} \ge \tau\big] \\ &= \int\int\int \mathrm{TPR}_{\tau,a}(c)d\mathrm{FPR}_{\tau,a'}(c)d\mathcal{F}(a)d\mathcal{F}(a') \\ &= E\left\{\int \mathrm{TPR}_{\tau,A_i}(c)d\mathrm{FPR}_{\tau,A_{i'}}(c)\right\},\end{aligned}$$

where i and i' index two independent subjects. Plug-in estimates may be constructed for these parameters with an estimated $\mathcal{F}(\cdot)$. For example, in the simple case when only a single measurement is taken at baseline, then AUC_τ may be estimated as

Given the longitudinal data structure, we may also be interested in estimating these accuracy parameters when the age at measurement A follows the marginal distribution of the observed measurement ages in the study. In which case,

$$\begin{aligned}\mathrm{TPR}_\tau(c) &= P\{\Pi_{\tau,A_{ij}}(\mathbf{Z}_{ij}) > c \mid 0 < T_i^{\mathbb{A}} - A_{ij} \le \tau\}, \\ \mathrm{FPR}_\tau(c) &= P\{\Pi_{\tau,A_{i'j'}}(\mathbf{Z}_{i'j'}) > c \mid T_{i'}^{\mathbb{A}} - A_{i'j'} > \tau\},\end{aligned}$$

and $\mathrm{AUC}_\tau = P\{\Pi_{\tau,A_{ij}}(\mathbf{Z}_{ij}) > \Pi_{\tau,A_{i'j'}}(\mathbf{Z}_{i'j'}) \mid 0 < T_i^{\mathbb{A}} - A_{ij} \le \tau,\, T_{i'}^{\mathbb{A}} - A_{i'j'} > \tau\}$.

The accuracy measure AUC_τ in this case can be estimated as

$$\frac{\sum_{i,j,i',j':0<X_i^{\mathbb{A}}-A_{ij}\le\tau<X_{i'}^{\mathbb{A}}-A_{i'j'}} \widehat{w}_{\tau ij}\,\widehat{w}_{\tau i'j'} I\{\widehat{\Pi}_{\tau,A_{ij}}(\mathbf{Z}_i) \ge \widehat{\Pi}_{\tau,A_{i'j'}}(\mathbf{Z}_{i'})\}}{\sum_{i,j,i',j':0<X_i^{\mathbb{A}}-A_{ij}\le\tau<X_{i'}^{\mathbb{A}}-A_{i'j'}} \widehat{w}_{\tau ij}\,\widehat{w}_{\tau i'j'}}.$$

Standard error (SE) and CIs can be constructed similarly to those given above for $\widehat{\mathrm{AUC}}_{\tau,a}$.

2.3. Selection of smoothing parameter

It is known that the choice of the smoothing parameter h is critical as in any nonparametric estimation problem. We employ a K-fold cross validation to select the smoothing parameter. Specifically, the study subjects are divided into K folds of approximately equal sizes. The optimal bandwidth h_{opt} minimises the weighted mean squared prediction error:

$$\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_k}\sum_{j=0}^{m_i} \widehat{w}_{\tau ij}\left[I\left(X_i^{\mathbb{A}} < A_{ij}+\tau\right) - g\left\{\mathbf{U}_{ij}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}^{(-k)}_\tau(A_{ij})\right\}\right]^2,$$

where $\mathcal{S}_k$ is the set of subjects that are in fold k and $\widehat{\boldsymbol{\beta}}^{(-k)}_\tau(A_{ij})$ is the estimate of $\boldsymbol{\beta}_\tau(A_{ij})$ using data excluding those from fold k. To obtain an estimator whose variance dominates bias, we follow the common practice to undersmooth (Cai et al., 2010; Neumann & Polzehl, 1998; Tian et al., 2005) using the final bandwidth $h = h_{\mathrm{opt}}\, n_1^{-0.1}$, where n_1 is the number of observed events by τ years.

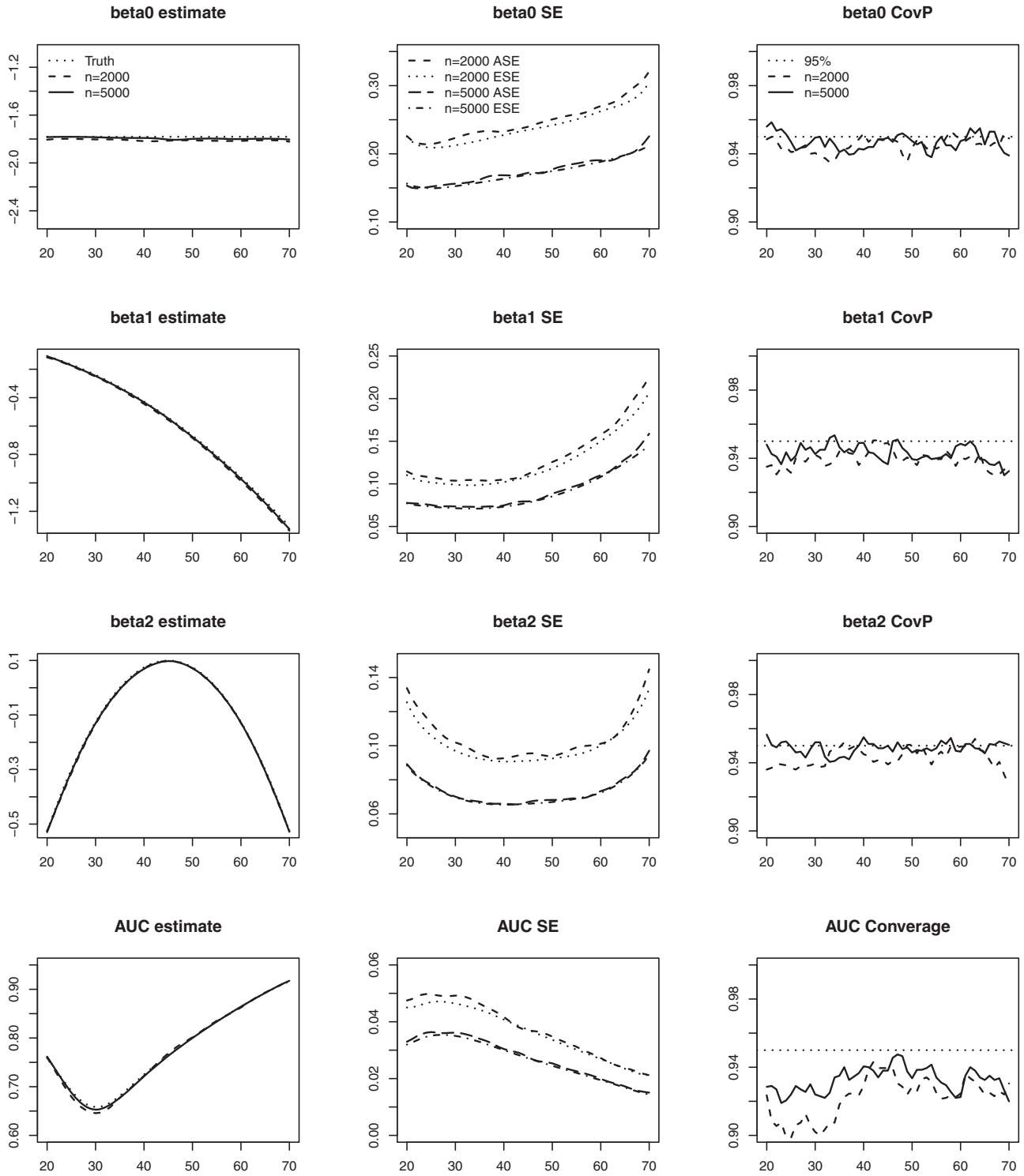


Figure 2. Average of estimates, average of the standard error estimates (ASE), empirical standard errors (ESE) and empirical coverage probabilities (CovP) of the 95% CI. One measurement at baseline. Each entry is based on 1000 simulated samples.The *x*-axis in all the plots is age.

3. Simulations

In this section, we report results from simulation studies that examine the finite-sample performance of the proposed methods and compare our methods with existing methods. Although the proposed methods are catered for survival data with longitudinally measured risk factors, it can be applied to traditional survival data with one single measurement of the risk factor at baseline to flexibly capture the age effect. For the single measurement setting, we compare the proposed procedure to the standard PH model which includes age as a covariate. In the longitudinal setting, we will compare the proposed methods to the commonly used joint modelling (JM) approach. For both settings, we considered $n =$ 2000 and 5000, let $K(\cdot)$ be the Gaussian kernel, and $B =$ 1000 for perturbations. For each configuration, results are summarised based on 1000 simulated data-sets.

3.1. Simulations with a single measurement

For this setting, we simulate A_{i0} from Uniform (15, 75) and two independent baseline covariates $Z_{i1}(A_{i0})$ and $Z_{i2}(A_{i0})$ from $\mathrm{N}N(0, 4)$ and $\mathrm{N}\,N(1, 4)$, respectively. The survival time from entry, T_i, is generated from $\log(T_i) = 2.5 + \beta_1(A_{i0})Z_{i1}(A_{i0}) + \beta_2(A_{i0})Z_{i2}(A_{i0}) + 0.5\epsilon$, where $\beta_1(a) = a^2/7500$, $\beta_2(a) = \{(a-45)^2 - 100\}/2000$ and ϵ follows a standard logistic distribution. The censoring time C_i is generated from $\exp(\tilde{C}_i)$ where $\tilde{C}_i \sim \mathrm{N}\,N(1.6, 0.36)$, resulting in about 80% of censoring. For illustration, we choose τ to be 5 years. Throughout, we choose $g(\cdot)$ to be the logistic link function. Using the proposed bandwidth selection procedure with fivefold cross-validation, h is about 6.4 and 5.4 for $n = 2000$ and 5000, respectively. We obtain the estimates of $\boldsymbol{\beta}_\tau(a)$ and $\mathrm{AUC}_{\tau,a}$ for ages from 20 to 70. Throughout the simulation studies, we let $\mathbf{U}_{ij} = \mathbf{Z}_{ij} = \mathbf{Z}_i(A_{ij})$.

In Figure 2, we present the average of the point estimates, the average of the SE estimates compared with the empirical SEs and the coverage probabilities (CovPs) of the 95% CIs for $\boldsymbol{\beta}_\tau(a)$ and $\mathrm{AUC}_{\tau,a}$ across a range of a. The results suggest that the proposed estimators produce negligible biases, and the estimated SEs are close to the empirical SEs. The empirical CovPs of the 95% CIs are close to their nominal level for $\boldsymbol{\beta}_\tau(a)$ coefficients. For AUC, the CovPs of the CIs are close to the nominal level but slightly below 95% for younger ages when $n = 2000$, possibly due to the fact the curvature of the AUC function is high in that range leading to a slight bias. The results are much improved when n increases to 5000.

For comparison, we obtain an alternative τ-year risk estimate, $\widehat{\Pi}^{\mathrm{cox}}_{\tau,A_{i0}}(\mathbf{Z}_{i0}) = g_{\mathrm{cox}}(\log\widehat{\Lambda}_\tau + \widehat{\gamma}_A A_{i0} + \widehat{\boldsymbol{\gamma}}_Z^{\intercal}\mathbf{Z}_{i0})$, from fitting a cox model including A_{i0} and $\mathbf{Z}_{i0} = (Z_{i1}(A_{i0}), Z_{i2}(A_{i0}))^{\intercal}$ as covariates, where $\widehat{\Lambda}_\tau$ is the estimated baseline cumulative hazard function at τ, $g_{\mathrm{cox}}(x) = 1 - e^{-e^x}$, and $(\widehat{\gamma}_A, \widehat{\boldsymbol{\gamma}}_Z^{\intercal})^{\intercal}$ are the estimated log-hazard ratio for $(A_{i0}, \mathbf{Z}_{i0}^{\intercal})$. For both of the risk estimates from our proposed method and the cox model, we evaluate their age-specific prediction performance as well as the overall prediction performance based on AUC_τ. As shown in Figure 3, the age-specific AUC, $\mathrm{AUC}_{\tau,a}$, of our proposed approach was generally higher than those from the cox model. The overall AUC, AUC_τ, was about 0.817 for the proposed model and 0.77 for the cox model. The average difference between the two overall AUC_τ's was 0.047 (SE = 0.007). These results highlight the improved prediction performance for using the proposed age-specific model.

3.2. Simulations with longitudinal measurements

We also conducted simulation studies to examine the performance of the proposed procedures in longitudinal settings. To simulate the age at the occurrence of event $T^{\mathbb{A}}$ and longitudinal measurements of a risk factor $Z(a)$, we generate two random effects α_0, α_1 from $N(0, 1)$. The age of event $T^{\mathbb{A}}$ is obtained from

$$\log(T^{\mathbb{A}}) = 0.5\log(-\log\epsilon + 2)/\{\Phi(\alpha_1) + 0.5\} + 3$$

where ϵ is generated from an uniform distribution over (0, 1) and $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution. We simulate age of entry to the study A_0 from Uniform (10, 70). Among the subjects who survive by the entry to the study, i.e., $T^{\mathbb{A}} > A_0$, we randomly sample n subjects as our cohort. For the ith subject in this cohort, the survival time since entry is $T_i = T_i^{\mathbb{A}} - A_{i0}$, and C_i is generated from a Uniform (10, 40), which leads to about 25% of censoring. The risk factor is measured at the entry age A_{i0} and ages $A_{ij} = A_{i0} + \Delta_{ij}$ after entering the study until event or censored, whichever comes first, where Δ_{ij} is generated from a $N(4, 1)$ distribution. And the observed marker value at age A_{ij} is $Z_i(A_{ij}) = \alpha_{0i} + \alpha_{1i}\log(A_{ij}) + e_i(A_{ij})$ where $e_i(A_{ij}) \sim N(0, 1.5^2)$. We choose τ to be 10 years. The selected smoothing parameter h is around 5.0 and 1.3 for $n = 2000$ and 5000, respectively, using fivefold cross-validation scheme described in Section 2.3. We obtain $\boldsymbol{\beta}_\tau(a)$ and $\mathrm{AUC}_{\tau,a}$ estimates for ages from 20 to 60.

The average of the point estimates, average of the SE estimators, the empirical SEs and the coverage probabilities of the 95% CIs for the $\boldsymbol{\beta}_\tau(a)$ and $\mathrm{AUC}_{\tau,a}$ at a series of ages are shown in Figure 4. The proposed procedures yield estimators with negligible biases. The estimated SEs obtained through perturbation resampling are close to the empirical SEs.

For comparison, we also fit the data with a JM approach (Rizopoulos, 2010; Tsiatis & Davidian, 2004) for longitudinal and survival data. In particular, we follow the setup in Rizopoulos (2010) to specify a linear

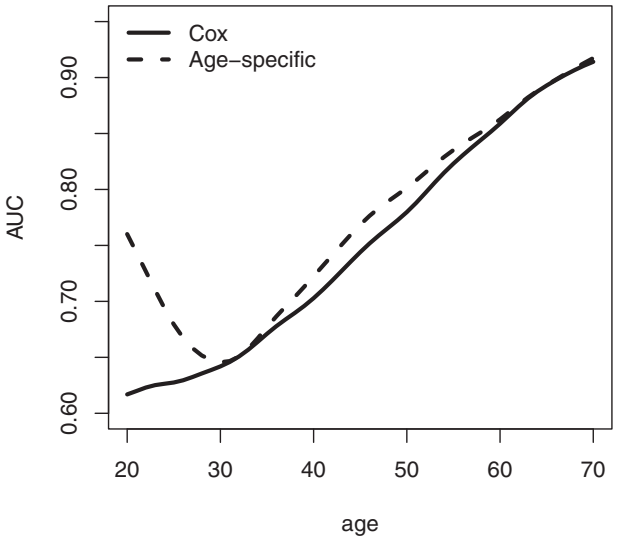


Figure 3. Age-specific AUCs for the proposed method (age-specific) and cox model (Cox). Values are averaged over 1000 repetitions with sample size $n = 5000$.

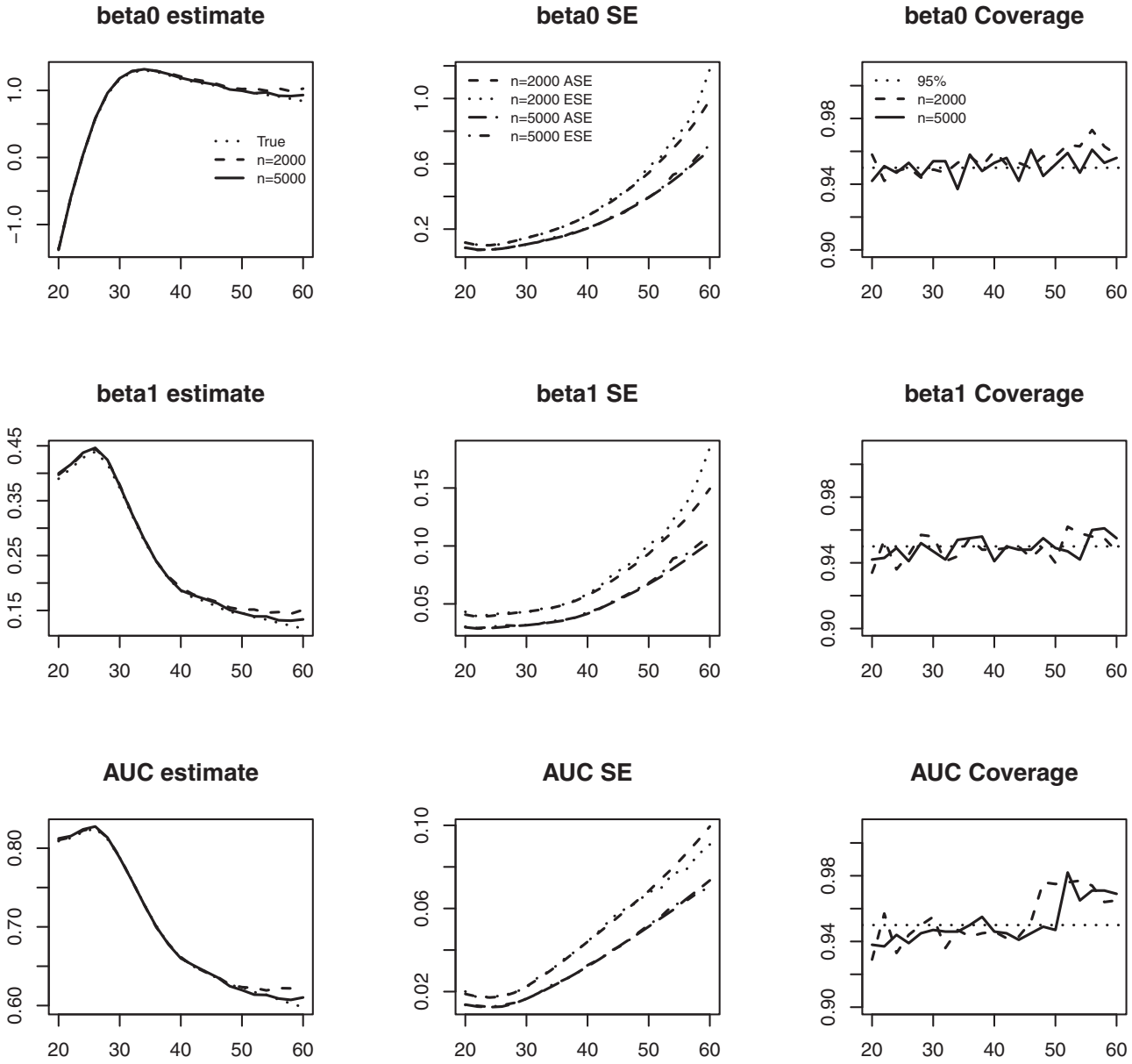


Figure 4. Average of estimates, average of the standard error estimates (ASE), empirical standard errors (ESE) and empirical coverage probabilities of the 95% CI. Longitudinal measurements. Each entry is based on 1000 simulated samples. The *x*-axis in all the plots is age.

mixed effects model with random intercept and slope for the longitudinal measurements Z_{ij} and a PH model relating the hazard function to the random slope, in which the log baseline hazard is approximated using B-splines. With the parameter estimates from the joint model, for subject i with measurement at A_{ij}, one can use a Monte Carlo approach to predict τ-year residual life risk given that the person has survived A_{ij} and we let $\widehat{\Pi}^{\mathrm{JM}}_{\tau,A_{ij}}(\mathbf{Z}_i)$ denote the resulting estimate of the risk function. We can estimate its corresponding $\mathrm{AUC}_{\tau,a}$ as discussed in Section 2.2. We present the average of the estimated $\mathrm{AUC}_{\tau,a}$ for the two modelling approaches at $n = 5000$ in Figure 5. The results suggest that the proposed approach improved prediction accuracy over a wide range of ages compared to the JM approach.

4. Application

In this section, we apply the proposed methods to the Framingham Heart Study to develop and evaluate age-specific CVD or death risk prediction models. The original goal of the study was to identify the common factors that contribute to CVD by following its development over a long period of time in a large group of participants who had not yet developed CVD. Started in 1948 with 5209 adult subjects, the study is now on its third generation of participants. Information on a wide spectrum of risk factors and disease outcomes is collected on each of the many follow-up visits during participants' lifetime. The data-set consists of 3982 subjects (2108 females and 1874 males) with complete information on the risk factors at least one measurement time.

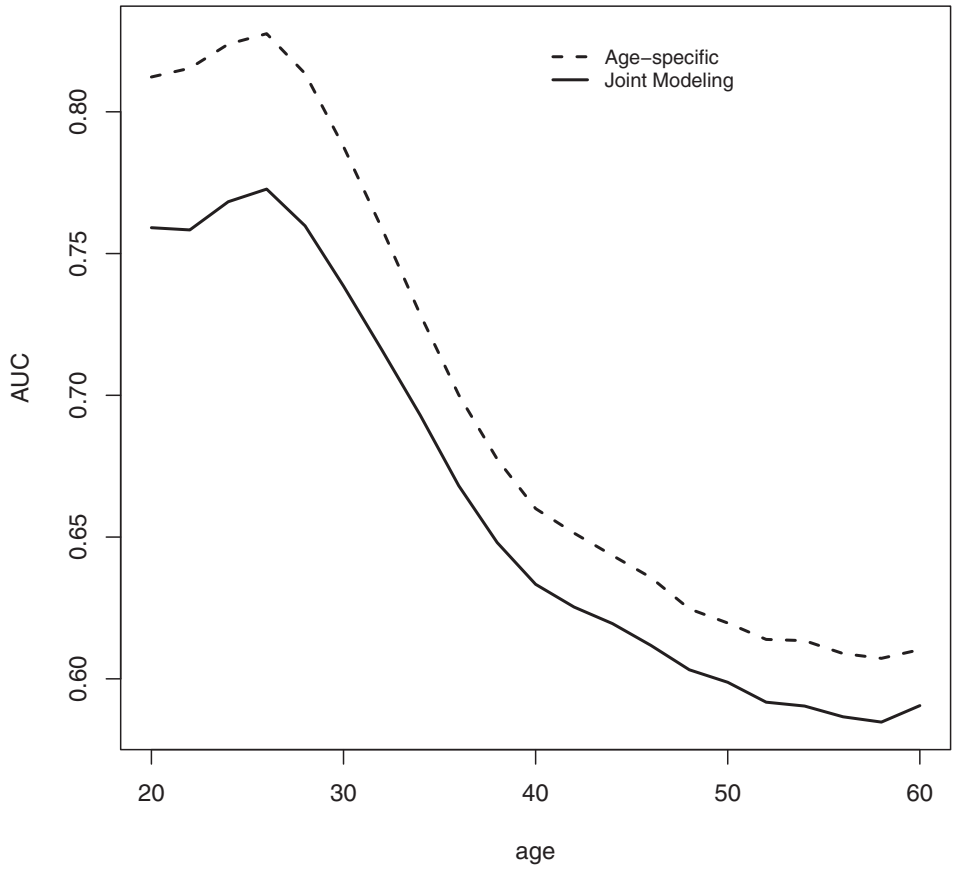
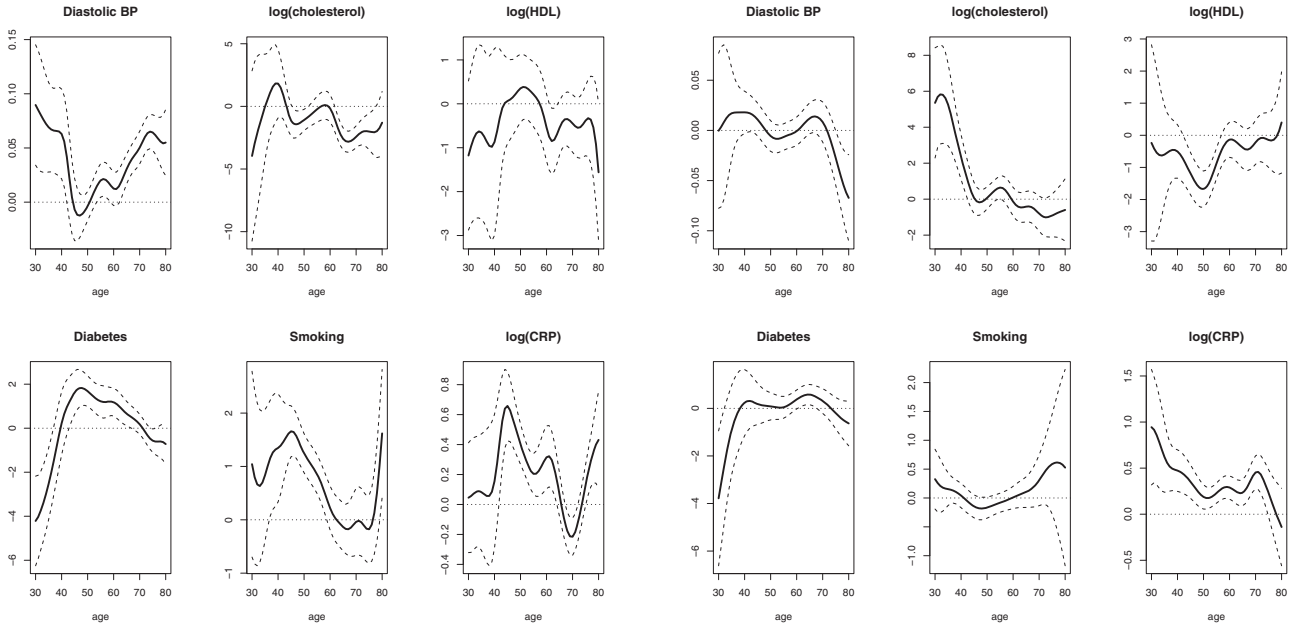


Figure 5. Age-specific AUC for the proposed method (age-specific) and JM with longitudinal measurements. Values are averaged over 1000 repetitions with sample size $n = 5000$.

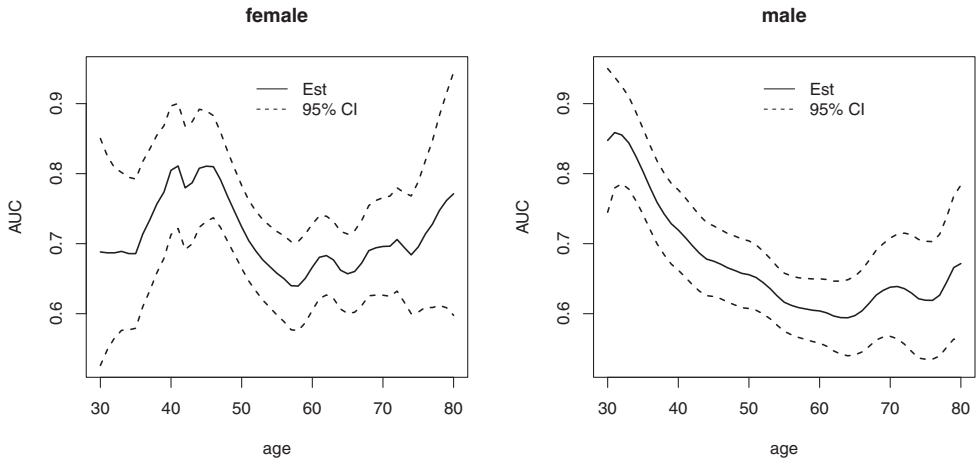
Several traditional Framingham risk factors were collected on these subjects on each of their visits, including age, diastolic blood pressure, cholesterol, high-density lipoprotein (HDL), diabetes and smoking. In addition, an inflammation marker, C-reactive protein (CRP), was also measured at various visits. The median number of measurement times is 3. We use both Framingham risk factors and CRP to estimate age-specific 10-year risk of CVD or death for females and males separately. Thus, we only include visits where all these risk factors are measured based on the study design. The outcome of interest is time to the onset of first major CVD event or death. Such a composite outcome avoids the issues of having to account for the competing risks from other causes of death. Nevertheless, we note that our calculation of the probability of experiencing CVD or death within 10-years in this data-set can be regarded approximately as 10-year CVD risk since majority of the observed events are CVD events, especially at younger age. In the study, there are 54 subjects who had CVD prior to death within 10 years. The cumulative incidence rate for CVD prior to death within 10 year is estimated to be about 1.4%. The median follow-up time was 32 years and the entry ages range from 5 to 70 with a median of 35.

We fit the proposed age-specific 10-year risk model (1) with a logistic link. Predictors include original scale of diastolic blood pressure, diabetes, smoking, the log scale of cholesterol, HDL and CRP. We use a Gaussian kernel for $K(\cdot)$, with the smoothing parameter h selected as 5.2 for males and 3.8 for females using the fivefold cross-validation scheme described in Section 2.3. In Figure 6(a,b), we demonstrate how the effects of major risk factors may vary by age for women and men, respectively. For men, blood pressure does not show any significant association with risk of outcome for all the ages, whereas for women, higher blood pressure significantly increases the risk, especially for younger ages and older ages. For men, as expected, high cholesterol increases risk, and our analysis further reveals that for men, the effects decrease with ages. In other words, having high cholesterol for a younger man exposes them to higher risk compared with older men. For women, diabetes status also shows an age-varying pattern: the effect increases with age initially, reaches a peak at 50 and goes down afterward. Smoking shows significant effect at age 45 for women, but the effect diminishes as age increases. However, for men, neither diabetes status nor smoking shows significant association with the risk at almost all ages. HDL shows no significant association with outcome for women, but is inversely associated with the risk between age 40 and 60 and no significant association at other ages for men. Finally, CRP also exhibits different age-varying effects between women and men.



(a) Effects of major risk factors: Female.

(b) Effects of major risk factors: Male.



(c) Age specific accuracy for 10-year residual life CVD or death risk.

Figure 6. Data analysis: Framingham Heart Study. The panels (a) and (b) show the point and interval estimates of the age-specific effects of the major risk factors for male and female, respectively. The panel (c) shows the point and interval estimates of the age-specific AUC for male and female, respectively.

For men, the effect of CRP is monotonically decreasing as age increases while for women, higher CRP increases risk before 60 years old and peaks at around 45 years old. In Figure 6(c), we show the age-specific AUC of the proposed age-specific risk scores for men and women. For women, the $\mathrm{AUC}_{10,a}$ does not vary substantially over age a with values fluctuating around 0.7. On the contrary, for men, $\mathrm{AUC}_{10,a}$ is substantially higher for younger ages with value as high as 0.9 and decreases to 0.6 at older ages. Thus, the age-specific risk model is highly accurate in predicting 10-year risk of CVD or death for younger males but only moderately accurate for middle aged or older males.

5. Discussion

When subjects are monitored over time for a clinical condition, it is highly desirable to dynamically recalculate risk estimates according to the updated risk factor information. Age is an important risk factor for many diseases such as CVD and the effects of other predictors on the disease risk may vary over age. Current risk prediction models used in clinical practice such as the FRS often incorporate age as an additive risk factor, which may limit the model prediction performance. Our proposed method estimates the age-specific absolute risk directly via a

flexible varying-coefficient model that allows the predictor effects to vary over age and allows for easy incorporation of longitudinally collected risk factor information. We also provide procedures for nonparametrically assessing the prediction performance of such age-specific models, extending existing time-specific accuracy parameters to also incorporate the additional age domain.

Unlike the cox model with time-varying covariates, our proposed model can easily provide age-specific absolute risk estimates without having to specify the full longitudinal marker processes. Compared to the JM approach, our method has the major advantage of allowing for non-linear effects and non-trivial number of time-varying continuous or discrete risk markers. Additionally, our kernel-based procedure allows borrowing information across individuals of similar ages therefore provides a practical solution for situations where the longitudinal information is only measured sparsely and irregularly.

Our method allows for internal covariates in that it aims to make prediction for the residual life among event-free subjects at age a, although it does require the availability of the marker information at age a for those with $T^{\mathbb{A}} > a$. When the outcome is a non-terminal event that is subject to death as a competing risk, then one may easily modify the proposed procedures to instead make prediction of the disease risk for those who are still alive and have not yet developed the disease at age a. The age-specific accuracy parameters can also be modified to accommodate competing risks similar to those considered in Blanche, Dartigues, and Jacqmin-Gadda (2013).

The proposed method employs a working model that requires the specification of g and ψ, both of which could potentially impact the model prediction performance. In general, the prediction performance is less sensitive to the choice of g due to the robustness properties such as the logistic likelihood as noted in Li and Duan (1989) and Eguchi and Copas (2002). One may choose appropriate ψ based on existing literature on the functional form of known risk factors or exploratory analyses. On the other hand, regardless of the choice of ψ or g, the fitted model may be mis-specified yet our proposed method could derive a risk model with good prediction performance. The flexible varying coefficient model is expected to perform well under mis-specification and the proposed inference procedures are always valid regardless of the potential mis-specification in the fitted model. In addition, while the proposed simple IPW method has the advantage of enabling robustness in inference under model mis-specification, it may come at a cost in efficiency loss. If there are auxiliary variables available at baseline, efficiency augmentation methods leveraging such information warrant further research.

Acknowledgments

The Framingham Heart Study and the Framingham SHARe project are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University. The Framingham SHARe data used for the analyses described in this manuscript were obtained through dbGaP (access number: phs000007.v3.p2). This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or the NHLBI.

The work is supported by grants from Natural Sciences and Engineering Research Council of Canada, U01-CA86368, P01-CA053996, R01-GM085047, U54-HG007963, and R01-HL089778 from the National Institutes of Health.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The work is supported by grants from Natural Sciences and Engineering Research Council of Canada [grant number U01-CA86368], [grant number P01-CA053996], [grant number R01-GM085047], [grant number U54-HG007963], [grant number R01-HL089778] from the National Institutes of Health.

Notes on contributors

Qian M. Zhou holds a Ph.D. in statistics from the University of Waterloo. She is now an assistant professor of Statistics in the Department of Mathematics and Statistics at Mississippi State University. Her research interests focus on developing advanced statistical methods in survival analysis, longitudinal data analysis, risk prediction, and model diagnosis.

Wei Dai holds a Ph.D. in biostatistics from Harvard University. She is now working at Citigroup.

Yingye Zheng holds a Ph.D. in biostatistics from the University of Washington. She is now a Full Member of Public Health Science Division in the Department of Biostatistics at the Fred Hutchison Cancer Research Center. Her research interests have been in the developing novel statistical tools for medical decision making in the field of disease screening, diagnosis, prognosis and outcome prediction.

Tianxi Cai holds a Sc.D. in biostatistics from Harvard University. She is now a professor of Biostatistics in the Department of Biostatistics at Harvard School of Public Health. Her research interests are mainly in the area of biomarker evaluation; model selection and validation; prediction methods; personalized medicine in disease diagnosis, prognosis and treatment; statistical inference with high dimensional data; and survival analysis.

ORCID

Qian M. Zhou http://orcid.org/0000-0002-7503-0445

References

Blanche, P., Dartigues, J.-F., & Jacqmin-Gadda, H. (2013). Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Statistics in Medicine, 32*(30), 5381–5397.

Cai, T., Tian, L., Uno, H., Solomon, S. D., & Wei, L. J. (2010). Calibrating parametric subject-specific risk estimation. *Biometrika, 97*(2), 389–404.

Eguchi, S., & Copas, J. (2002). A class of logistic-type discriminant functions. *Biometrika, 89*(1), 1–22.

Gail, M. H., Brinton, L. A., Byar, D. P., Corle, D. K., Green, S. B., Schairer, C., & Mulvihill, J. J. (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of the National Cancer Institute, 81*(24), 1879.

Gerds, T. A., Cai, T., & Schumacher, M. (2008). The performance of risk prediction models. *Biometrical Journal, 50*(4), 457–479.

Li, K.-C., & Duan, N. (1989). Regression analysis under link violation. *The Annals of Statistics, 17*(3), 1009–1052.

Liu, D., Zheng, Y., Prentice, R. L., & Hsu, L. (2014). Estimating risk with time-to-event data: An application to the women's health initiative. *Journal of the American Statistical Association, 109*(506), 514–524.

Lloyd-Jones, D. M. (2010). Cardiovascular risk prediction basic concepts, current status, and future directions. *Circulation, 121*(15), 1768–1777.

Mosca, L., Appel, L. J., Benjamin, E. J., Berra, K., Chandra-Strobos, N., Fabunmi, R. P., … Williams, C.L. (2004). Evidence-based guidelines for cardiovascular disease prevention in women 1. *Journal of the American College of Cardiology, 43*(5), 900–921.

Neumann, M. H., & Polzehl, J. (1998). Simultaneous bootstrap confidence bands in nonparametric regression. *Journal of Nonparametric Statistics, 9*(4), 307–333.

Parast, L., Cheng, S.-C., & Cai, T. (2012). Landmark prediction of long-term survival incorporating short-term event time information. *Journal of the American Statistical Association, 107*(500), 1492–1501.

Park, Y., & Wei, L. J. (2003). Estimating subject-specific survival functions under the accelerated failure time model. *Biometrika, 90*(3), 717–723.

Ridker, P. M., Buring, J. E., Rifai, N., & Cook, N. R. (2007). Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: The Reynolds risk score. *Jama, 297*(6), 611–619.

Rizopoulos, D. (2010). Jm: An r package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software, 35*(9), 1–33.

Tian, L., Zucker, D., & Wei, L. J. (2005). On the cox model with time-varying regression coefficients. *Journal of the American Statistical Association, 100*(469), 172–183.

Tsiatis, A., DeGruttola, V., & Wulfsohn, M. S. (1995). Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and cd4 counts in patients with aids. *Journal of the American Statistical Association, 90*(429), 27.

Tsiatis, A. A., & Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica, 14*(3), 809–834.

Uno, H., Cai, T., Tian, L., & Wei, L. J. (2007). Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association, 102*(478), 527–537.

Wang, Y., & Taylor, J. M. G. (2001). Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of the American Statistical Association, 96*(455), 895–905.

Wolf, P. A., D'Agostino, R. B., Belanger, A. J., and Kannel, W. B. (1991). Probability of stroke: A risk profile from the framingham study. *Stroke, 22*(3), 312–318.

Wu, C.-F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics, 14*(4), 1261–1295.

Ye, W., Lin, X., & Taylor, J. M. G. (2008). Semiparametric modeling of longitudinal measurements and time-to-event data – A two-stage regression calibration approach. *Biometrics, 64*(4), 1238–1246.

Zheng, Y., Cai, T., & Feng, Z. (2006). Application of the time-dependent roc curves for prognostic accuracy with multiple biomarkers. *Biometrics, 62*(1), 279–287.

Zheng, Y., & Heagerty, P. J. (2004). Semiparametric estimation of time-dependent ROC curves for longitudinal marker data. *Biostatistics, 5*(4), 615–632.

Zheng, Y., & Heagerty, P. J. (2005). Partly conditional survival models for longitudinal data. *Biometrics, 61*(2), 379–391.