Published in final edited form as:

Biometrics. 2019 September ; 75(3): 853–863. doi:10.1111/biom.13057.

A Varying-Coefficient Generalized Odds Rate Model with Time-Varying Exposure: An Application to Fitness and CVD Mortality

Jie Zhou¹, Jiajia Zhang^{1,*}, Alexander C. Mcclain¹, Wenbin Lu², Xuemei Sui³, James W. Hardin¹

¹Department of Epidemiology and Biostatistics, University of South Carolina, Columbia, SC, USA

²Department of Statistics, North Carolina State University, Raleigh, NC, USA

³Department of Exercise, University of South Carolina, Columbia, SC, USA

Summary:

Varying-coefficient models have become a common tool to determine whether and how the association between an exposure and an outcome changes over a continuous measure. These models are complicated when the exposure itself is time-varying and subjected to measurement error. For example, it is well known that longitudinal physical fitness has an impact on cardiovascular disease (CVD) mortality. It is not known, however, how the effect of longitudinal physical fitness on CVD mortality varies with age. In this paper, we propose a varying-coefficient generalized odds rate model that allows flexible estimation of age-modified effects of longitudinal physical fitness on CVD mortality. In our model, the longitudinal physical fitness is measured with error and modeled using a mixed effects model, and its associated age-varying coefficient function is represented by cubic B-splines. An expectation-maximization (EM) algorithm is developed to estimate the parameters in the joint models of longitudinal physical fitness and CVD mortality. A modified pseudo-adaptive Gaussian–Hermite quadrature method is adopted to compute the integrals with respect to random effects involved in the E-step. The performance of the proposed method is evaluated through extensive simulation studies and is further illustrated with an application to cohort data from the Aerobic Center Longitudinal Study.

Keywords

B-splines; EM Algorithm; Generalized Odds Rate Model; Joint Modeling; Varying Coefficient

1. Introduction

Physical inactivity, mainly due to a sedentary lifestyle, has been shown to have a positive association with cardiovascular disease (CVD) mortality (Blair et al., 1996; Kohl 3rd, 2001; Mora et al., 2007; Nocon et al., 2008). The Aerobic Center Longitudinal Study (ACLS)

* jzhang@mailbox.sc.edu.

Supporting Information

The conditions for asymptotic properties, details of the EM algorithm, including the Gaussian–Hermite Quadrature and the maximization steps, and more simulation results mentioned in the manuscript may be found online in the Supporting Information section at the end of the article.

involves patients in the Cooper Clinic in Dallas, TX, who attended periodic preventive medical examinations, along with health and lifestyle behaviors consultation. The longitudinal measurements of cardiorespiratory fitness (“fitness”), an objective measure for physical activity (PA, of Sports Medicine, 2013), provides a unique opportunity to advance the current understanding of the association between longitudinal fitness and CVD mortality.

Using a total of 3,980 patients enrolled in the study from 1970 ~ 1980 with follow-up till 2003, we first explore the association between the baseline fitness and CVD mortality using the proportional hazards (PH) (Cox, 1972) model. Adjusting for age, gender, BMI, family history of CVD, and smoking status, the results from the PH model are summarized in Table 1. It shows that baseline fitness has an inverse association with CVD mortality (coefficient= -0.013), but the effect is not significant (p value= 0.572). However, if we consider age as an effect modifier of fitness and include “age×fitness” in the model, we find the interaction term is significant (p value= 0.050). This indicates that the effect of fitness on CVD mortality changes over age, which can be depicted through an age varying coefficient.

Moreover, it is well-known that there are changes in the overall level of fitness with age. Figure 1 displays the fitness profiles for all participants in the ACLS by age. It can be seen that the mean fitness is around 20 and gradually decreases with age. Changes in fitness that occur during follow-up may have important influence on CVD mortality but cannot be detected when the analysis relies on a single baseline assessment. Additionally, even though the standard exercise test is an objective measure of PA which is superior to self-report, the values still appear to be subjected to measurement error. This measurement error could be due to true measurement error in the equipment or small biological fluctuations in the subjects fitness level on the day of measurement (e.g., a bad night of sleep). We seek a model that allows for a time-varying effect of fitness on CVD mortality, where fitness is subject to measurement error.

In practice it is challenging to capture the association between a time-varying covariate and a survival outcome with a varying-coefficient model. Previous studies focused on either estimation of varying coefficients for time-independent variables (Cai and Sun, 2003; Tian et al., 2005), or fixed coefficients for time-dependent variables (Fisher and Lin, 1999; Zeng and Lin, 2006). To the best of our knowledge, there is no literature on survival models that consider both a time-varying covariate and its varying-effect over another variable. What complicates our situation more is that the exposure of interest is an endogeneous covariate, which are time-dependent measures that typically require the subject to survive to be measured and commonly measured with error (Rizopoulos, 2012), so that the previous methods do not apply.

The most popular tools in modeling the association between a survival outcome and an endogeneous covariate with measurement error are joint models. Specifically, a mixed effects model with normal random effects is commonly assumed for the longitudinal observations, and standard survival models are used for the survival outcome. There have been plenty of work on joint models which combine the linear mixed model with PH model (e.g., Wulfsohn and Tsiatis (1997); Bycott and Taylor (1998); Zeng et al. (2005); Zeng and Cai (2005)). Furthermore, the proportional odds (PO) joint model has also been studied in

the literature when the PH assumption is violated (Andrinopoulou et al., 2014). Under the Bayesian framework, Köhler et al. (2017) and Andrinopoulou et al. (2017) studied the nonlinear effect of the longitudinal predictor using P-splines. Various extensions of joint models have been made to account for complex data structures in practice, with considerations of multiple longitudinal outcomes (Song et al., 2002; Brown et al., 2005; Rizopoulos and Ghosh, 2011; Moreno-Betancur et al., 2017), competing risks (Elashoff et al., 2008; Huang et al., 2011), and cure rate models (Yu et al., 2004; Brown and Ibrahim, 2003). More overviews and extensions can be found in Tsiatis and Davidian (2004) and Rizopoulos (2012).

The existing joint models do not allow varying coefficients, so they cannot be used to estimate the age-related association between fitness and CVD mortality. Therefore, we develop a novel joint modeling framework considering the following three features: (1) longitudinal process of fitness, (2) survival process of CVD mortality, and (3) the age-related fitness effects. For the longitudinal process, we assume a flexible pre-specified time function with random coefficients to accommodate subject-specific longitudinal trajectories. For the survival process, we propose to incorporate the generalized odds rate (GOR) model (Dabrowska and Doksum, 1988; Scharfstein et al., 1998; Zhou et al., 2017a), including the PH model and the PO model (Bennett, 1983) as special cases. To investigate the age-related fitness effect on CVD mortality, we include a novel age-dependent varying coefficient for longitudinal fitness in the survival model. The proposed model can improve the understanding of how age-related fitness affects CVD mortality, which can provide direct guidance in behavior consultation.

The rest of the article is organized as follows. We first introduce the notations and model definitions in Section 2. The estimation procedures for the proposed joint model are discussed in Section 3, where the details of the expectation-maximization (EM) algorithm and the corresponding variance estimation are presented in Section 3.3 and Section 3.4, respectively. The results of extensive simulation studies are performed in Section 4. To study the nonlinear age-dependent effect of fitness on CVD mortality, we apply the proposed methods to the ACLS data in Section 5. The final discussion and conclusions are summarized in Section 6.

2. Model and Notation

Consider a longitudinal study of n subjects. Let T_i denote the failure time of interest for subject i , $i = 1, \dots, n$. Let $\mathbf{Z}_i$ denote a vector of baseline covariates of subject i , $A_i(s)$ denote their age at time s , and $W_i(\cdot)$ denote the true underlying time-varying exposure process, such as the longitudinal physical fitness process in our application. Note that $A_i \equiv A_i(0)$ is the baseline age and $A_i(s) = A_i + s$. The filtration of $W_i(\cdot)$ is denoted by $\mathcal{W}_i(t) = \{W_i(s): s \leq t\}$. Let $0\Lambda_i(t)$ denote the conditional cumulative hazard function of T_i given $\mathbf{Z}_i$, A_i and $\mathcal{W}_i(\cdot)$. We propose the following varying-coefficient generalized odds rate model for $\Lambda_i(t)$:

$$\Lambda_i(t) \equiv \Lambda\{t | \mathbf{Z}_i, A_i, \mathcal{W}_i(\cdot)\} = G_r \left(\int_0^t \lambda_0(s) \exp[\mathbf{Z}_i' \boldsymbol{\beta} + \psi\{A_i(s)\} W_i(s)] ds \right), \quad (1)$$

where $\lambda_0(t)$ is an unspecified baseline hazard function, $\boldsymbol{\beta}$ is a vector of coefficients for $\mathbf{Z}_i$, $\psi\{A_i(s)\}$ is an age-varying smooth coefficient function for $W_i(s)$, and $G_r(\cdot)$ is a pre-specified increasing transformation function indexed by a non-negative argument r . In particular, in our implementation, we take the transformation function as $G_r(x) = r^{-1} \log(1 + rx)$ when $r > 0$ and $G_r(x) = x$ when $r = 0$, which reduces to a PH model with $r = 0$ and a PO model with $r = 1$ (Zeng et al., 2016). In addition, we approximate the smooth coefficient function using cubic B-splines, i.e. $\psi(a) = \sum_{l=1}^L \gamma_l B_l(a)$, where $B_l(\cdot)$, $l = 1, \dots, L$, are the B-spline basis functions.

The whole trajectory of the true longitudinal exposure $W_i(\cdot)$ is usually not obtainable in practice. Instead, we can only observe $\mathbf{Y}_i \equiv (Y_{i,1}, \dots, Y_{i,m_i})$, where $Y_{i,j} = Y_i(t_{i,j})$, $j = 1, \dots, m_i$, are m_i contaminated measurements of $W_i(\cdot)$, at a sequence of intermittent observation times $t_i = (t_{i,1}, \dots, t_{i,m_i})$, with $0 \leq t_{i,1} < \dots < t_{i,m_i}$. We assume the following random effects model for $Y_i(t)$:

$$Y_i(t) = W_i(t) + \epsilon_i(t) \equiv g(t)' \mathbf{b}_i + \epsilon_i(t), \quad (2)$$

where $g(t)$ is a d -dimensional vector of known functions of t , for example, $g(t) = (1, t)'$ corresponds to a linear function of t with $d = 2$, $\mathbf{b}_i = (b_{i1}, \dots, b_{id})$ is a d -dimensional vector of random effects assumed to follow a multivariate normal distribution with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\mathbf{D}$, and the error process $\epsilon_i(t)$ is an independent mean-zero normal process with variance σ^2 .

3. Estimation Procedure

3.1 Complete Likelihood Function

In the presence of censoring, we observe $V_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$, $i = 1, \dots, n$, where C_i is the right censoring time of subject i and is assumed to be independent of T_i given $\mathbf{Z}_i$ and A_i . The observed data for subject i can be denoted as $O_i = (V_i, \delta_i, A_i, \mathbf{Z}_i, t_i, \mathbf{Y}_i)$ and the parameters to be estimated include $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \lambda_0, \boldsymbol{\mu}, \mathbf{D}, \sigma^2)$, where $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_L)$ is a vector of coefficients for the B-spline functions. Note that given random effects $\mathbf{b}_i$, $Y_i(\cdot)$ is independent of T_i , C_i , A_i and $\mathbf{Z}_i$. Therefore, the conditional cumulative hazard function of T_i given $\mathbf{Z}_i$, A_i and $\mathcal{W}_i(\cdot)$ is the same as that of T_i given $\mathbf{Z}_i$, A_i and $\mathbf{b}_i$, denoted by $\Lambda(t | \mathbf{Z}_i, A_i, \mathbf{b}_i)$. Let $S(t | \mathbf{Z}_i, A_i, \mathbf{b}_i) = \exp\{-\Lambda(t | \mathbf{Z}_i, A_i, \mathbf{b}_i)\}$ denote the corresponding conditional survival function of T_i given $\mathbf{Z}_i$, A_i and $\mathbf{b}_i$. Under the proposed model (1), the survival function $S(t | \mathbf{Z}_i, A_i, \mathbf{b}_i) = \exp\{-\Lambda(t | \mathbf{Z}_i, A_i, \mathbf{b}_i)\}$ can be written as

$$S(t|Z_i, A_i, \mathbf{b}_i) = \begin{cases} \exp\left[-\int_0^t \lambda_0(s) \exp\{Z_i' \boldsymbol{\beta} + \psi(A_i + s) \times W_i(s)\} ds\right], & r = 0, \\ \left[1 + r \int_0^t \lambda_0(s) \exp\{Z_i' \boldsymbol{\beta} + \psi(A_i + s) \times W_i(s)\} ds\right]^{-1/r}, & r > 0. \end{cases}$$

Let $S(t|Z_i, A_i, \boldsymbol{\phi}_i, \mathbf{b}_i)$ denote the frailty model with

$$S(t|Z_i, A_i, \boldsymbol{\phi}_i, \mathbf{b}_i) = \exp\left(-\phi_i \int_0^t \lambda_0(s) e^{Z_i' \boldsymbol{\beta}} \exp[\psi\{A_i(s)\} W_i(s)] ds\right),$$

and ϕ_i is a frailty variable with gamma distribution. It is easy to show that $S(t|Z_i, A_i, \mathbf{b}_i) = \int S(t|Z_i, A_i, \boldsymbol{\phi}_i, \mathbf{b}_i) f(\phi_i) d\phi_i$ where $f(\cdot)$ is the density function of a gamma distribution with mean 1 and variance r . Thus, the conditional survival function $S(t|Z_i, A_i, \mathbf{b}_i)$ is equivalent to the marginal distribution of gamma frailty model.

The complete likelihood function of $\boldsymbol{\theta}$ given the observed data $\mathbf{O} \equiv (O_1, \dots, O_n)$, the frailty terms $\boldsymbol{\phi} \equiv (\phi_1, \dots, \phi_n)$ and the random effects $\mathbf{b} = (\mathbf{b}_1, \dots, \mathbf{b}_n)$ can be written as:

$$\begin{aligned} \mathcal{L}_c(\boldsymbol{\theta}|\mathbf{O}, \boldsymbol{\phi}, \mathbf{b}) &= \prod_{i=1}^n p(V_i, \delta_i | \phi_i, \mathbf{b}_i; \boldsymbol{\beta}, \boldsymbol{\gamma}, \lambda_0) \times p(Y_i | \mathbf{b}_i; \sigma^2) \times p(\mathbf{b}_i | \boldsymbol{\mu}, \mathbf{D}) \times f(\phi_i) \\ &= \prod_{i=1}^n \left\{ \phi_i \lambda_0(V_i) e^{Z_i' \boldsymbol{\beta}} \eta(V_i | A_i, \mathbf{b}_i; \boldsymbol{\gamma}) \right\}^{\delta_i} \times \exp\left\{ -\phi_i \int_0^{V_i} \lambda_0(s) e^{Z_i' \boldsymbol{\beta}} \eta(s | A_i, \mathbf{b}_i; \boldsymbol{\gamma}) ds \right\} \\ &\quad \times (2\pi\sigma^2)^{-\frac{m_i}{2}} \exp\left\{ -\frac{1}{2\sigma^2} (\mathbf{Y}_i - \mathbf{G}_i \mathbf{b}_i)' (\mathbf{Y}_i - \mathbf{G}_i \mathbf{b}_i) \right\} \\ &\quad \times (2\pi)^{-\frac{d}{2}} |\mathbf{D}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (\mathbf{b}_i - \boldsymbol{\mu})' \mathbf{D}^{-1} (\mathbf{b}_i - \boldsymbol{\mu}) \right\} \times f(\phi_i), \end{aligned} \quad (3)$$

where $\mathbf{G}_i = (g(t_{i,1}), \dots, g(t_{i,m_i}))'$ and $\eta(s|A_i, \mathbf{b}_i; \boldsymbol{\gamma}) = \exp[\psi\{A_i(s)\} g(s)' \mathbf{b}_i]$.

The observed likelihood function $\mathcal{L}(\boldsymbol{\theta}|\mathbf{O})$ can be derived by integrating the frailty terms $\boldsymbol{\phi}$ and the random effects $\mathbf{b}$ out of (3). Direct maximization of the observed likelihood $\mathcal{L}(\boldsymbol{\theta}|\mathbf{O})$ is difficult due to the integrals with respect to the random effects and frailties. Therefore, we develop an EM algorithm to estimate the model parameters, including both finite and infinite dimensional parameters.

3.2 Conditional Expectations

After dropping the terms that do not contain $\boldsymbol{\theta}$, the complete log-likelihood function can be written as the summation of three distinct parts, i.e.

$$l^c(\boldsymbol{\theta}|\boldsymbol{\phi}, \mathbf{b}) = l_1^c(\lambda_0, \boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{\phi}, \mathbf{b}) + l_2^c(\sigma^2|\mathbf{b}) + l_3^c(\boldsymbol{\mu}, \mathbf{D}|\mathbf{b}), \text{ where}$$

$$\begin{aligned}
l_1^c(\lambda_0, \beta, \gamma | \phi, b) &= \sum_{i=1}^n \delta_i [\log\{\lambda_0(V_i)\} + \mathbf{Z}_i' \beta + \log\{\eta(V_i | A_i, \mathbf{b}_i; \gamma)\}] - \phi_i \int_0^{V_i} \lambda_0(s) e^{\mathbf{Z}_i' \beta} \eta(s | A_i, \mathbf{b}_i; \gamma) ds, \\
l_2^c(\sigma^2 | b) &= \sum_{i=1}^n -\frac{m_i}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} (\mathbf{Y}_i - \mathbf{G}_i \mathbf{b}_i)' (\mathbf{Y}_i - \mathbf{G}_i \mathbf{b}_i), \text{ and} \\
l_3^c(\mu, D | b) &= \sum_{i=1}^n -\frac{1}{2} \log(|D|) - \frac{1}{2} (\mathbf{b}_i - \mu)' D^{-1} (\mathbf{b}_i - \mu).
\end{aligned}$$

Let $Q(\theta, \theta^{(k)})$ denote the conditional expectation of the complete log-likelihood function $\ell(\theta, \phi, b)$ given observed data $\mathbf{O} = (O_1, \dots, O_n)$ and current estimates $\theta^{(k)}$. Similar to previous arguments, $Q(\theta, \theta^{(k)})$ can be written as the summation of three distinct parts,

$$\begin{aligned}
Q(\theta; \theta^{(k)}) &= E_b [E_\phi \{l^c(\theta | \phi, b) | \mathbf{O}, b\} | \mathbf{O}, \theta^{(k)}] \\
&= E_b [E_\phi \{l_1^c(\lambda_0, \beta, \gamma | \phi, b) | b, \mathbf{O}, \theta^{(k)}\} | \mathbf{O}, \theta^{(k)}] \\
&\quad + E_b \{l_2^c(\sigma^2 | b) | \mathbf{O}, \theta^{(k)}\} + E_b \{l_3^c(\mu, D | b) | \mathbf{O}, \theta^{(k)}\} \\
&= Q_1(\lambda_0, \beta, \gamma; \theta^{(k)}) + Q_2(\sigma^2; \theta^{(k)}) + Q_3(\mu, D; \theta^{(k)}).
\end{aligned}$$

To evaluate the conditional expectation $Q(\theta, \theta^{(k)})$, we need to calculate both $E(\phi_i | b_i, O_i, \theta^{(k)})$ and the conditional expectations of functions of $\mathbf{b}_i$ given O_i and current estimate $\theta^{(k)}$. The conditional distribution of ϕ_i given $\mathbf{b}_i, O_i$ and $\theta^{(k)}$ is

$$p(\phi_i | \mathbf{b}_i, O_i) \propto \phi_i^{\delta_i} \times \exp \left\{ -\phi_i \int_0^{V_i} \lambda_0(s) e^{\mathbf{Z}_i' \beta} \eta(s | A_i, \mathbf{b}_i; \gamma) ds \right\} \times f(\phi_i).$$

Plugging in the density for Gamma($1/r, r$) and doing some algebra, it can be shown that the resulting conditional distribution is a gamma distribution with shape parameter $\delta_i + 1/r$ and scale parameter $\left\{ 1/r + \int_0^{V_i} \lambda_0(s) e^{\mathbf{Z}_i' \beta} \eta(s | A_i, \mathbf{b}_i; \gamma) ds \right\}^{-1}$.

Expectations with respect to the conditional distribution of $\mathbf{b}_i$ given O_i and $\theta^{(k)}$ can be approximated using a modified version of adaptive Gaussian–Hermite (GH) quadrature. Details can be found in Appendix A.

3.3 EM Algorithm

An EM algorithm is derived to obtain the maximum likelihood estimator (MLE) of θ . The algorithm is described as follows.

Initialization: obtain initial values $\theta^{(0)}$ based on the following two-step approach:

Step 1: Fit the mixed effect model $Y(t) = g(t)' \mathbf{b} + \epsilon(t)$ based on the data (t_i, Y_i) , $i = 1, \dots, n$, and set the estimated mean and covariance matrix of $\mathbf{b}$ as $\mu^{(0)}$ and $D^{(0)}$, and the estimated

variance of error term as $(\sigma^2)^{(0)}$. This can be realized by using the “nlme” package (Pinheiro et al., 2016) (<https://cran.r-project.org/web/packages/nlme/>) in R (R Core Team, 2013).

Step 2: For a pre-specified r , fit the GOR model only with $\mathbf{Z}$ as the covariates, and set the estimated regression coefficients as $\boldsymbol{\beta}^{(0)}$ and estimated baseline cumulative hazard function as $\Lambda_0^{(0)}(t)$. This can be done using the R package “TransModel” (Zhou et al., 2017b) (<https://cran.r-project.org/web/packages/TransModel/index.html>). The initial values for $\boldsymbol{\gamma}^{(0)}$ are set to be 0.

In the k th iteration,

E-step: Compute the conditional expectations described in Section 3.2 based on $\mathbf{O}$ and current estimate $\boldsymbol{\theta}^{(k)}$ using adaptive GH quadrature.

M-step: Maximize the expectation of the log-likelihood functions, $Q_1(\lambda_0(\cdot), \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}^{(k)})$, $Q_2(\sigma^2; \boldsymbol{\theta}^{(k)})$ and $Q_3(\boldsymbol{\mu}, \mathbf{D}; \boldsymbol{\theta}^{(k)})$, respectively, and update the parameters as $\boldsymbol{\theta}^{(k+1)}$. The details are below.

- i. By solving the partial derivative of $Q_1(\lambda_0(\cdot), \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}^{(k)})$ with respect to $\lambda_0(\cdot)$, we obtain

$$\tilde{\lambda}_0(t; \boldsymbol{\beta}, \boldsymbol{\gamma}) = \frac{\sum_{i=1}^n \delta_i I(V_i = t)}{\sum_{i=1}^n I(V_i \geq t) e^{\mathbf{Z}_i' \boldsymbol{\beta}} E_{\mathbf{b}_i} \left\{ E(\phi_i | \mathbf{b}_i, \mathbf{O}_i, \boldsymbol{\theta}^{(k)}) \eta(t | A_i, \mathbf{b}_i; \boldsymbol{\gamma}) | \boldsymbol{\theta}^{(k)}, \mathbf{O}_i \right\}}$$

as a function of $(\boldsymbol{\beta}, \boldsymbol{\gamma})$. Apply the Newton–Raphson algorithm to maximize the expectation of the profile log-likelihood function, $Q_1(\tilde{\lambda}_0(t; \boldsymbol{\beta}, \boldsymbol{\gamma}), \boldsymbol{\beta}, \boldsymbol{\gamma}; \boldsymbol{\theta}^{(k)})$ to obtain the updates $\boldsymbol{\beta}^{(k+1)}$ and $\boldsymbol{\gamma}^{(k+1)}$. Then, the baseline hazard function can be updated by $\lambda_0^{(k+1)}(t) = \tilde{\lambda}_0(t; \boldsymbol{\beta}^{(k+1)}, \boldsymbol{\gamma}^{(k+1)})$. More details can be found in Appendix B.

- ii. From $Q_2(\sigma^2; \boldsymbol{\theta}^{(k)})$ and $Q_3(\boldsymbol{\mu}, \mathbf{D}; \boldsymbol{\theta}^{(k)})$, we update with the following formula

$$\begin{aligned} \boldsymbol{\mu}^{(k+1)} &= \sum_{i=1}^n E\{\mathbf{b}_i | \boldsymbol{\theta}^{(k)}, \mathbf{O}_i\} / n, \quad \mathbf{D}^{(k+1)} = \sum_{i=1}^n E\{\mathbf{b}_i \mathbf{b}_i' | \boldsymbol{\theta}^{(k)}, \mathbf{O}_i\} / n, \quad \text{and} \\ (\sigma^2)^{(k+1)} &= \frac{\sum_{i=1}^n \sum_{j=1}^{m_i} E\{(Y_i - G_i \mathbf{b}_i)'(Y_i - G_i \mathbf{b}_i) | \boldsymbol{\theta}^{(k)}, \mathbf{O}_i\}}{\sum_{i=1}^n m_i}. \end{aligned}$$

Iterate the E-step and M-step until $(\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)})^2 < 0.001$. From our numerical experience, the above algorithm usually converges within 100 iterations.

Remarks: Following the proof in Kim et al. (2017), the consistency of the MLEs, $\hat{\boldsymbol{\theta}}$, $\hat{\psi}$, $\hat{\Lambda}$, and the asymptotic normality and the semiparametric efficiency of $\hat{\boldsymbol{\theta}}$ can be established given the conditions listed in the supporting information.

3.4 Variance Estimation

After the EM algorithm converges, we have the maximum likelihood estimate $\hat{\theta}$. Let $\theta^* = \theta \setminus \lambda_0$ denote the vector of finite dimensional parameters. Suppose the length of the vector θ^* is m . The $m \times m$ variance-covariance matrix of $\hat{\theta}^*$ can be estimated by inverting the observed information matrix based on the profile likelihood.

To be specific, we define $pl(\theta^*) = \max_{\lambda_0} n^{-1} \sum_{i=1}^n pl_i(\theta^*, \lambda_0)$ as the logarithm of the profile likelihood for θ^* , where $pl_i(\theta^*, \lambda_0)$ denote the logarithm of the observed likelihood for subject i , $i = 1, \dots, n$. Let $I(\theta^*) = \{v_{ll'}\}$, $l, l' = 1, \dots, m$, denote the observed information matrix for $\hat{\theta}^*$. The element $v_{ll'}$ can be approximated by the second-order numerical difference of $pl(\theta^*)$ (Murphy and Van der Vaart, 2000; Zeng and Cai, 2005; Zeng et al., 2005). Specifically,

$$v_{ll'} = \frac{\{q(\hat{\theta}^* + h_n e_l) - q(\hat{\theta}^*)\}' \{q(\hat{\theta}^* + h_n e_{l'}) - q(\hat{\theta}^*)\}}{h_n^2},$$

where $q(\hat{\theta}^*) = (pl_1(\hat{\theta}^*), \dots, pl_n(\hat{\theta}^*))$ is the vector of profile likelihood functions being evaluated at $\hat{\theta}^*$, e_l is the unit vector of length m that has the l th element being 1 and other elements being 0, and $h_n = O(1/\sqrt{n})$ is a pre-specified constant that is bounded by $1/\sqrt{n}$.

4. Simulation Study

To study the properties of the proposed methods simulation studies were ran. We generated data for the proposed joint models from the following varying-coefficient GOR model

$$s(t|Z_i) = \begin{cases} \exp\left[-\int_0^t \lambda_0(s) \exp\{Z_i \beta + \psi(A_i + s) \times W_i(s)\} ds\right], & r = 0, \\ \left[1 + r \int_0^t \lambda_0(s) \exp\{Z_i \beta + \psi(A_i + s) \times W_i(s)\} ds\right]^{-1/r}, & r > 0. \end{cases}$$

The baseline hazard function $\lambda_0(\cdot)$ is chosen to be Weibull with the shape and scale parameters set at 2. The varying coefficient function is chosen as $\psi(a) = -0.2 \sin(a)$. Baseline standardized age A_i is generated from a standard normal distribution and two baseline covariates are included: Z_1 follows Uniform (0,2) and Z_2 follows Bernoulli (0.5). Coefficients for $Z = (Z_1, Z_2)$ are $\beta = (1, -1)$. Different GOR models with the transformation parameter $r = 0, 0.5, 1$ and 2 are considered.

A linear function for the fitness over time is assumed, i.e., $W_i(t) = b_{j0} + b_{j1}t$. The random effects $b_j = (b_{j0}, b_{j1}) \sim N(\mu, D)$, where $\mu = (2, 1)$ and the covariance matrix $D = \{v_{ij}\}$ is assumed to be $v_{ij} = I(i=j) + 0.5I(i \neq j)$, that is, the variances are 1 and the covariance is 0.5. The error terms is assumed to follow the normal distribution with mean zero and variance $\sigma^2 = 0.5$.

The censoring time C is generated from the uniform distribution, $U(0, a)$, where a is adjusted to have 50% censoring rate. Subject i is assumed to have visits $0 \leq t_{i1} < \dots < t_{im_i} < \min\{T_i, C_i\}$, and the length between two consecutive visits are set to be 0.1. Sample size of $n = 500$ is used and 1000 replications are conducted for each setting. We use 5 nodes in our adaptive GH quadrature method and $L = 3$ knots at the percentiles of observed age for the B-splines to estimate the varying coefficient function $\psi(\cdot)$.

The simulation results are summarized in Table 2, where we report the bias, empirical standard deviation (StDev), mean of the estimated standard error (StdErr) and the coverage probability (CP) of 95% Wald confidence intervals. The bias of all the parameters is small, the estimated standard errors based on the profile likelihood are close to the empirical estimates and the CP is close to the nominal level 0.95. The estimated baseline cumulative hazard functions and the varying coefficient functions $\psi(\cdot)$ are compared with the true curves in Figures 2 and 3, respectively, overlaid with 2.5th and 97.5th quantiles of the estimates. All the curves are found to be close to the true values.

The Akaike information criterion (AIC) can be used to select appropriate transformation parameter r and knots L in practice. To evaluate its performance, we do a small cross validation for the setting of $r = 1$. We search among the grids with $r = 0, 1, 2$ and $L = 3, 5, 7$ based on AIC. Proportions of selecting each grid are reported in Table 3 for sample sizes $n = 200$ and 500. As we can see from the result, when we have small sample such as 200, the PH model is more likely to be selected. The proportion of selecting the true r reaches to 72.9% as sample size increases to 500.

More simulation settings are conducted and results are summarized in the supporting information. First, we investigate more settings with regard to different functions for the baseline distribution, different sample sizes and censoring, which give similar findings. Second, we evaluate the impact of the initial value. We conduct simulation studies using the initial values from the linear mixed model and PH model separately, and simulation results are similar. Third, a setting with conditionally independent censoring, where C is generated based on A , is also considered in the supporting information, and the results are found to be similar.

5. Application to the ACLS Data

In order to assess the longitudinal effect of fitness on the CVD mortality, we apply the proposed model to the subset of ACLS Data set. We include patients who were enrolled between 1970 and 1980, and were followed till the end of 2003. Fatal outcomes (e.g. CVD mortality) were extracted from mortality surveillance, principally through the National Death Index. The main exposure variable is the cardiorespiratory fitness (fitness), which is quantified as the maximal treadmill time in minutes during a symptom limited exercise test. As an objective measure of physical activity, fitness is a more reliable measure of recent activity levels than self-reported values. Other potential confounders we adjust in the model include gender, BMI, smoking status and family history of CVD. There are a total of 3,980 patients with 437 females and 3,543 males. The number of follow-up visits for all

participants ranges from 3 to 30 with median number of follow up being equal to 5. Among all participants, 145 participants died from CVD by the end of year 2003.

We assume a linear form for the fitness trajectory over time. Similar to the simulation, we use GH quadrature with 5 nodes for the approximation in the E-step, which gives similar results to those using a larger number of nodes. We apply cubic B-splines with L knots being placed at percentiles of observed age to estimate the varying coefficient, where the number L is selected based on the AIC. For illustration, in Figure 4 we plot the AIC versus number of knots for three different models: a PH model ($r=0$), a PO model ($r=1$) and a variant of PO model ($r=2$). Based on the curves, the PH model with 4 knots results in the smallest AIC.

We summarize the estimated coefficients in the PH model with 4 knots in Table 4. Based on the results, higher BMI will increase the risk of CVD mortality and females generally have lower risk of dying from CVD. Smoking and family history are positively associated with a risk of dying from CVD. All the terms in the longitudinal process are found to be highly significant, indicating a significant decreasing linear trend of fitness with time. The baseline cumulative hazard curve is plotted in Figure 5(a), which is a step function with jumps at the event times.

Based on the estimated γ coefficients in B-splines, we also can test the hypothesis “ H_0 : the varying coefficient is constant with age”, which is equivalent to $H_{01} : M_1 \gamma = 0$, where

$$M_1 = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix}$$

The test statistic $(M_1 \hat{\gamma}) (M_1 \widehat{V}_{\gamma} M_1')^{-1} (M_1 \hat{\gamma})'$ follows a chi-squared distribution with 4 degree of freedom under H_{01} , where $\widehat{V}_{\gamma}$ is the estimated covariance matrix of $\hat{\gamma}$. The calculated test statistic value is 138.12, which yields a highly significant p-value. Therefore, we can conclude that the fitness effect on CVD mortality is significantly non-constant with age.

The age-dependent varying coefficient is illustrated in Figure 5(b) along with its 95% pointwise confidence intervals, which shows a clear pattern of the effect of fitness on CVD mortality with age.

In addition, based on the estimated point-wise confidence intervals for the varying coefficient, there is a negative association between fitness and CVD mortality for the age period of study. Based on the curve, physical activity has significant protective effects on CVD mortality till age 70, and no significant impact from 70 to 80. An explanation for this finding could be that after 70, age genetic factors take over as the dominate reason for CVD related mortality, and that the individuals' physical activity is not a significant factor. The protective effect of physical activity is the strongest around age 40, suggesting that more exercise during middle-aged population is the most effective in reducing CVD associated mortality. Note that the standard errors reported are from the optimal selected model and no post-model selection effects have been adjusted here.

6. Discussion

We proposed a joint model with an age-dependent varying coefficient for GOR model with a longitudinal endogenous covariate measured with error. The age-related varying coefficient was flexibly modeled with cubic B-splines. The function $g(t)$ presents how individual longitudinal observations change over time. In practice, the individual longitudinal profile can be plotted along with the smoothed curve. The plot can be used to provide evidence for a linear, quadratic or other forms of change. Then, the linear, quadratic or cubic spline function of t will be suggested for $g(t)$. The EM algorithm is applied in estimating the proposed joint model, while the variance of the estimates are approximated based on a profile likelihood function. The estimation methods are discussed and evaluated by simulation studies.

The ACLS dataset is used to illustrate the usage of the model, where we study the longitudinal effect of fitness on the CVD mortality. The effect of fitness on CVD mortality is found to change over age, and the trajectory can be clearly described by the estimated varying coefficient curve as illustrated in Section 5.

Aging is the most important factor in many chronic diseases. The change in age-related behavior plays an important role in disease development and corresponding disease-related mortality. The proposed model can be broadly used in modeling survival outcomes with time-varying effects of longitudinal predictors, and helps improving the understanding of the real impact of some age-related chronic behaviors on survival outcomes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We greatly appreciate Dr. Steven N. Blair in the University of South Carolina for providing the ACLS study data.

References

- Andrinopoulou E-R, Eilers PH, Takkenberg JJ, and Rizopoulos D (2017). Improved dynamic predictions from joint models of longitudinal and survival data with time-varying effects using p-splines. *Biometrics*.
- Andrinopoulou E-R, Rizopoulos D, Takkenberg JJ, and Lesaffre E (2014). Joint modeling of two longitudinal outcomes and competing risk data. *Statistics in medicine* 33, 3167–3178. [PubMed: 24676841]
- Bennett S (1983). Analysis of survival data by the proportional odds model. *Statistics in medicine* 2, 273–277. [PubMed: 6648142]
- Blair SN, Kampert JB, Kohl HW, Barlow CE, Macera CA, Paffenbarger RS, and Gibbons LW (1996). Influences of cardiorespiratory fitness and other precursors on cardiovascular disease and all-cause mortality in men and women. *Jama* 276, 205–210. [PubMed: 8667564]
- Brown ER and Ibrahim JG (2003). Bayesian approaches to joint cure-rate and longitudinal models with applications to cancer vaccine trials. *Biometrics* 59, 686–693. [PubMed: 14601770]
- Brown ER, Ibrahim JG, and DeGruttola V (2005). A flexible b-spline model for multiple longitudinal biomarkers and survival. *Biometrics* 61, 64–73. [PubMed: 15737079]

- Bycott P and Taylor J (1998). A comparison of smoothing techniques for cd4 data measured with error in a time-dependent cox proportional hazards model. *Statistics in medicine* 17, 2061–2077. [PubMed: 9789914]
- Cai Z and Sun Y (2003). Local linear estimation for time-dependent coefficients in cox's regression models. *Scandinavian Journal of Statistics* 30, 93–111.
- Cox DR (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society* 34, 187–220.
- Dabrowska DM and Doksum KA (1988). Estimation and testing in a two-sample generalized odds-rate model. *Journal of the American Statistical Association* 83, 744–749.
- Elashoff RM, Li G, and Li N (2008). A joint model for longitudinal measurements and survival data in the presence of multiple failure types. *Biometrics* 64, 762–771. [PubMed: 18162112]
- Fisher LD and Lin DY (1999). Time-dependent covariates in the cox proportional-hazards regression model. *Annual review of public health* 20, 145–157.
- Huang X, Li G, Elashoff RM, and Pan J (2011). A general joint model for longitudinal measurements and competing risks survival data with heterogeneous random effects. *Lifetime data analysis* 17, 80–100. [PubMed: 20549344]
- Kim S, Zeng D, and Taylor JM (2017). Joint partially linear model for longitudinal data with informative drop-outs. *Biometrics* 73, 72–82. [PubMed: 27479944]
- Kohl H 3rd (2001). Physical activity and cardiovascular disease: evidence for a dose response. *Medicine and science in sports and exercise* 33, S472–83. [PubMed: 11427773]
- Köhler M, Umlauf N, and Greven S (2017). Nonlinear association structures in flexible bayesian additive joint models. *arXiv preprint arXiv:1708.06337*.
- Mora S, Cook N, Buring JE, Ridker PM, and Lee I-M (2007). Physical activity and reduced risk of cardiovascular events. *Circulation* 116, 2110–2118. [PubMed: 17967770]
- Moreno-Betancur M, Carlin JB, Brilleman SL, Tanamas SK, Peeters A, and Wolfe R (2017). Survival analysis with time-dependent covariates subject to missing data or measurement error: Multiple imputation for joint modeling (mijm). *Biostatistics* page kxx046.
- Murphy SA and Van der Vaart AW (2000). On profile likelihood. *Journal of the American Statistical Association* 95, 449–465.
- Nocon M, Hiemann T, Müller-Riemenschneider F, Thalau F, Roll S, and Willich SN (2008). Association of physical activity with all-cause and cardiovascular mortality: a systematic review and meta-analysis. *European Journal of Cardiovascular Prevention & Rehabilitation* 15, 239–246. [PubMed: 18525377]
- of Sports Medicine, A. C (2013). ACSM's guidelines for exercise testing and prescription. Lippincott Williams & Wilkins.
- Pinheiro J, Bates D, DebRoy S, Sarkar D, and R Core Team (2016). nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1–128.
- R Core Team (2013). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
- Rizopoulos D (2012). Joint models for longitudinal and time-to-event data: With applications in R. Chapman and Hall/CRC.
- Rizopoulos D and Ghosh P (2011). A bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Statistics in medicine* 30, 1366–1380. [PubMed: 21337596]
- Scharfstein DO, Tsiatis AA, and Gilbert PB (1998). Semiparametric efficient estimation in the generalized odds-rate class of regression models for right-censored time-to-event data. *Lifetime data analysis* 4, 355–391. [PubMed: 9880995]
- Song X, Davidian M, and Tsiatis AA (2002). An estimator for the proportional hazards model with multiple longitudinal covariates measured with error. *Biostatistics* 3, 511–528. [PubMed: 12933595]
- Tian L, Zucker D, and Wei L (2005). On the cox model with time-varying regression coefficients. *Journal of the American statistical Association* 100, 172–183.

- Tsiatis AA and Davidian M (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica* 14, 809–834.
- Wulfsohn MS and Tsiatis AA (1997). A joint model for survival and longitudinal data measured with error. *Biometrics* 53, 330–339. [PubMed: 9147598]
- Yu M, Law NJ, Taylor JM, and Sandler HM (2004). Joint longitudinal-survival-cure models and their application to prostate cancer. *Statistica Sinica* 14, 835–862.
- Zeng D and Cai J (2005). Simultaneous modelling of survival and longitudinal data with an application to repeated quality of life measures. *Lifetime Data Analysis* 11, 151–174. [PubMed: 15940822]
- Zeng D, Cai J, et al. (2005). Asymptotic results for maximum likelihood estimators in joint analysis of repeated measurements and survival time. *The Annals of Statistics* 33, 2132–2163.
- Zeng D and Lin D (2006). Efficient estimation of semiparametric transformation models for counting processes. *Biometrika* 93, 627–640.
- Zeng D, Mao L, and Lin D (2016). Maximum likelihood estimation for semiparametric transformation models with interval-censored data. *Biometrika* 103, 253–271. [PubMed: 27279656]
- Zhou J, Zhang J, and Lu W (2017a). An expectation maximization algorithm for fitting the generalized odds-rate model to interval censored data. *Statistics in medicine* 36, 1157–1171. [PubMed: 28004414]
- Zhou J, Zhang J, and Lu W (2017b). TransModel: Fit Linear Transformation Models for Right Censored Data. R package version 2.1.

Figure 1.
Profile Plots of Longitudinal Fitness. (This figure appears in color in the electronic version of this article.)

Figure 2.

Estimated Baseline Cumulative Hazard Curves (solid lines are the mean of estimates, dashed lines are the true curve and the dotted lines are the 2.5 and 97.5 quantiles of the estimates).

(This figure appears in color in the electronic version of this article.)

Figure 3.

Estimated Varying Coefficient Curves $\psi(A(t))$ (solid lines are the mean of estimates, dashed lines are the true curve and the dotted lines are the 2.5 and 97.5 quantiles of the estimates). (This figure appears in color in the electronic version of this article.)

Figure 4.

ACLS Data: Choose Knots and r Based On AIC. (This figure appears in color in the electronic version of this article.)

Figure 5.

ACLS Data: Estimated Baseline Cumulative Hazard and Age-dependent Varying Coefficient for Fitness.

Table 1

Fit PH Models for ACLS Baseline Data

<i>Variable</i>	Without Interaction			With Interaction		
	<i>Estimate</i>	<i>StDev</i>	<i>P value</i>	<i>Estimate</i>	<i>StDev</i>	<i>P value</i>
BMI	0.103	0.030	0.001	0.110	0.030	0.000
FamilyCVD	0.124	0.168	0.462	0.133	0.168	0.427
Smoke	0.218	0.233	0.349	0.230	0.233	0.323
Female	−0.342	0.371	0.356	−0.348	0.371	0.347
AGE	0.117	0.011	0.000	0.181	0.035	0.000
Fitness	−0.013	0.023	0.572	0.192	0.107	0.072
AGE×Fitness	-	-	-	−0.004	0.002	0.050

Table 2

Simulation Results

Variable	Bias	StDev	StdErr	CP	Bias	StDev	StdErr	CP
r=0					r=0.5			
β_1	0.005	0.125	0.124	0.944	-0.003	0.139	0.145	0.954
β_2	-0.013	0.138	0.141	0.948	-0.005	0.169	0.165	0.946
μ_0	-0.001	0.047	0.048	0.952	0.003	0.046	0.048	0.964
μ_1	-0.006	0.059	0.060	0.952	-0.008	0.058	0.057	0.944
σ^2	-0.004	0.010	0.010	0.934	-0.005	0.009	0.010	0.912
v_{11}	-0.007	0.069	0.075	0.968	-0.011	0.074	0.073	0.940
v_{12}	0.012	0.066	0.065	0.940	0.014	0.062	0.062	0.952
v_{22}	-0.008	0.099	0.099	0.938	-0.005	0.098	0.093	0.944
r=1					r=2			
β_1	0.009	0.170	0.164	0.942	0.000	0.203	0.194	0.946
β_2	-0.007	0.188	0.187	0.946	0.025	0.228	0.223	0.940
μ_0	0.000	0.049	0.048	0.946	0.002	0.047	0.047	0.944
μ_1	-0.003	0.060	0.056	0.926	-0.001	0.058	0.054	0.926
σ^2	-0.006	0.009	0.009	0.886	-0.006	0.008	0.008	0.884
v_{11}	-0.008	0.071	0.072	0.942	-0.017	0.069	0.071	0.932
v_{12}	0.021	0.062	0.062	0.938	0.013	0.058	0.059	0.964
v_{22}	0.007	0.092	0.091	0.948	-0.007	0.087	0.084	0.924

Table 3

Cross validation based on AIC

	<i>n</i> = 200				<i>n</i> = 500			
	<i>L</i> = 3	<i>L</i> = 5	<i>L</i> = 7	<i>Row sums</i>	<i>L</i> = 3	<i>L</i> = 5	<i>L</i> = 7	<i>Row sums</i>
<i>r</i> = 0	27.2	25.8	4.0	57.0	16.3	3.1	1.6	21.0
<i>r</i> = 1	20.2	16.0	2.7	38.9	57.7	11.0	4.2	72.9
<i>r</i> = 2	2.1	1.6	0.4	4.1	4.7	1.0	0.3	6.0

Table 4

ACLS Data: Parameter Estimates

PH Model			
<i>Parameter</i>	<i>Estimate</i>	<i>StDev</i>	<i>P value</i>
BMI	0.092	0.029	0.002
FamilyCVD	0.198	0.179	0.269
Smoke	0.167	0.234	0.476
Female	−0.442	0.386	0.253
μ_0	18.337	0.073	< 0.001
μ_1	0.063	0.005	< 0.001
σ^2	3.984	0.022	< 0.001
v_{11}	19.056	0.508	< 0.001
v_{12}	−0.280	0.026	< 0.001
v_{22}	0.052	0.002	< 0.001