Journal of Computational and Graphical Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ucgs20

Joint Modeling of Longitudinal Imaging and Survival Data

Kai Kang & Xinyuan Song

To cite this article: Kai Kang & Xinyuan Song (2022): Joint Modeling of Longitudinal Imaging and Survival Data, Journal of Computational and Graphical Statistics, DOI: 10.1080/10618600.2022.2102027

To link to this article: https://doi.org/10.1080/10618600.2022.2102027

Joint Modeling of Longitudinal Imaging and Survival Data

Kai Kang and Xinyuan Song

Department of Statistics, The Chinese University of Hong Kong, Hong Kong, China

xysong@sta.cuhk.edu.hk

Abstract

This paper considers a joint modeling framework for simultaneously examining the dynamic pattern of longitudinal and ultrahigh-dimensional images and their effects on the survival of interest. A functional mixed effects model is considered to describe the trajectories of longitudinal images. Then, a high-dimensional functional principal component analysis (HD-FPCA) is adopted to extract the principal eigenimages to reduce the ultrahigh dimensionality of imaging data. Finally, a Cox regression model is used to examine the effects of the longitudinal images and other risk factors on the hazard. A theoretical justification shows that a naive two-stage procedure that separately analyzes each part of the joint model produces biased estimation even if the longitudinal images have no measurement error. We develop a Bayesian joint estimation method coupled with efficient Markov chain Monte Carlo sampling schemes to perform statistical inference for the proposed joint model. A Monte Carlo dynamic prediction procedure is proposed to predict the future survival probabilities of subjects given their historical longitudinal images. The proposed model is assessed through extensive simulation studies and an application to Alzheimer's Disease Neuroimaging Initiative, which turns out to hold the promise of accuracy and possess higher predictive capacity for survival outcome compared with existing methods.

Keywords: HD-FPCA; Imaging data; longitudinal response; MCMC methods; time-to-event outcome

1 Introduction

The current research is motivated by the Alzheimer's Disease Neuroimaging Initiative (ADNI) study, which was first launched in 2004 with overarching objective of detecting Alzheimer's disease (AD) at the earliest stage and identifying ways to track the disease progression with biomarkers. ADNI recruited participants with age from 55 to 90 and collected their imaging, genetic, clinical, cognitive, and biochemical characteristics across the long-term study. The preprocessed magnetic resonance imaging (MRI) data (Appendix 1 of Supplementary Material), which can be represented by a $133 \times 170 \times 129$ array, were scanned for each participants at several visits. During the cohort study, participants who met the specific inclusion criteria shown in the general procedures manual would be diagnosed as suffering AD, and thus, a possibly censored time-to-AD outcome was also recorded. A portion of participants with mild cognitive impairment (MCI) at baseline quickly progressed to AD, whereas some remained stable or even revert to normal cognition. For instance, the ventricle (butterfly-shaped cavity) of the subject with MCI to AD conversion was increasingly enlarged across time, whereas that of subject staying at MCI stage remained nearly unchanged (See Figure S1 in Supplementary Material). Such difference, in turn, indicates that the dynamic changes of specific brain regions are likely to be associated with AD progression. Hence, investigating the association between the longitudinal feature of MRI images and time to AD and obtaining a prognostic model for doctors to plan intervention based on the history of patients' MRI images are important in AD prevention and targeted treatment.

Joint modeling of longitudinal and survival data is a useful tool in simultaneously analyzing repeatedly measured outcomes and time-to-event responses (e.g., DeGruttola and Tu, 1994; Wang and Taylor, 2001; Lin et al., 2002; Brown and Ibrahim, 2003). A comprehensive overview can be found in Tsiatis and Davidian

(2004) and the references therein. The basic idea behind existing joint models is that certain common factors jointly influence the time-to-event response and longitudinal variables, thereby inducing correlation between the two observation processes. Thus, existing joint models often consist of a mixed effects model for describing the dynamic variations or changing patterns of the longitudinal outcomes and a survival model, such as a proportional hazards (Cox) model (Cox, 1972), for delineating the association between the longitudinal outcomes and the survival of interest.

Despite the rapid development of the joint modeling approach for longitudinal and survival data, little is known when the longitudinal observations are ultrahigh-dimensional images, such as the MRI data collected in the ADNI study. Incorporating longitudinal images into the existing joint modeling framework is confronted with several challenges. The first challenge is the ultrahigh dimensionality of images. Basically, images are voxelwise stored and the number of voxels is often huge compared with the sample size. Therefore, applying the preceding joint models at each voxel is computationally infeasible. The second challenge lies in performing efficient estimation for a joint analysis of the mixed effects and survival models. A naive two-stage estimation procedure that separately analyzes each part of the joint model leads to biased estimation in the analysis of a survival model with covariate measurement errors (Prentice, 1982). Such kind of bias is also expected when applying the two-stage method to the present study with imaging predictors. The third challenge is conducting dynamic prediction. Given the history of longitudinal images and other relevant predictors, dynamically predicting the survival outcome and thus making highly informed decision is of great importance. However, the existing prediction procedures (e.g., Rizopoulos, 2011; Andrinopoulou et al., 2018) only accommodate longitudinal low-dimensional observations. How to make use of longitudinal and ultrahigh dimensional imaging information to facilitate a dynamic prediction for survival probabilities has not ever been investigated.

This study aims to address the aforementioned challenges. First, we develop a novel joint modeling approach to simultaneously analyze longitudinal imaging and survival data. We consider a functional mixed effects model to describe the dynamic pattern of longitudinal images and address the between-image correlation through subject-specific random imaging coefficients. The subject-specific imaging covariates are then incorporated into a Cox model to examine the effects of time-variant images and time-invariant predictors on time-to-event. Moreover, a high-dimensional FPCA (HD-FPCA) technique (Zipunnikov et al., 2014) is adopted to extract the principal eigenimages of subject- and subject-visit-specific variations in ultrahigh longitudinal imaging data. Second, we establish valid estimation procedure for the proposed model. We propose to use a Bayesian approach coupled with efficient Markov chain Monte Carlo (MCMC) sampling schemes to conduct statistical inference. A theoretical justification shows that a naive two-stage method produces biased estimation even if the longitudinal imaging data has no measurement error, and empirical evidences demonstrate that the proposed Bayesian joint estimation procedure is consistently superior to the two-stage procedure in all scenarios under consideration. Third, we develop a mechanism for dynamic prediction of the time-to-event outcome. A Monte Carlo estimate of the predictive survival probability given the history of longitudinal images and baseline covariates is derived. Numerical studies show that the proposed joint model outperforms several existing models in terms of predictive capacity.

Relevant researches in the existing literature have worked either on joint models for low-dimensional functional observations or on survival models with time-invariant imaging predictors. Lee et al. (2015) utilized the FPCA technique to analyze Cox models with baseline functional predictors. Such time-invariant assumption ignores the possibly time-dependent association between functional/imaging observations and the hazard of interest, which will be demonstrated in Sections 4 and 5 to result in lower predictive accuracy than the proposed model. Li and Luo (2019) developed a joint model for longitudinal low-

dimensional functional and survival data. However, they only used a functional random intercept to capture the subject-specific effects and restricted the functional parameter in the survival model to be a constant. All of these methods are not applicable to the current model setting where the imaging data are ultra-high dimensional and longitudinal with dynamic structure.

The rest of this paper is organized as follows. Section 2 introduces the proposed joint model for longitudinal imaging and survival data. Section 3 presents the Bayesian joint estimation procedure. Meanwhile, a theoretical justification is provided to show that a naive two-stage method produces biased estimation. Section 4 presents simulation studies to evaluate the empirical performance of the proposed method and demonstrates that the proposed method consistently outperforms the two-stage method. Section 5 applies the proposed methodology to the ADNI dataset. Section 6 concludes the study with a brief discussion. Technical details are provided in the Supplementary Material.

2 Joint modeling of longitudinal imaging and survival data

Let $\mathbf{Y}_{ij}$ denote the sample of image for subject i $(i = 1, \cdots, I)$ at time T_{ij} ($j = 1, \cdots, J_i$), where I is the total number of subjects and J_i is the number of visits for subject i. All the images are stored in 3-dimensional (3D) array structures of dimension $p = p_1 \times p_2 \times p_3$ with voxels being the same across subjects and visits. The images can be unfolded into a $p \times 1$ vector containing the voxels in a particular order, where the order is preserved across all subjects and visits, that is, $\mathbf{Y}_{ij} = (Y_{ij}(1), \cdots, Y_{ij}(p))'$. Following Greven et al. (2010) and Zipunnikov et al. (2014), we consider a functional mixed effects model as follows:

$$Y_{ij}(v) = \mu(v) + X_{i0}(v) + X_{i1}(v)T_{ij} + W_{ij}(v), \qquad (1)$$

where v denotes the voxel in a compact space $\mathcal{V}$, $\mu(v)$ is a fixed image, $X_{i0}(v)$ and $X_{i1}(v)$ denote the random imaging intercept and slope for subject i, and

$W_{ij}(v)$ is the random subject-visit-specific imaging deviation from subject-specific mean image. Similar to Zipunnikov et al. (2014), we assume that $X_{i0}(v), X_{i1}(v)$, and $W_{ij}(v)$ are zero-mean, square-integrable, and mutually uncorrelated random processes on a compact space $\mathcal{V}$. Model (1) excludes a measurement error term because the images are usually presmoothed. The linear form of T_{ij} in (1) is only for presentation simplicity. An extension of incorporating polynomial terms of T_{ij}, spline growth curves, and/or other covariates of interest is straightforward.

For each subject *i*, let T_i^* and C_i denote time to event and the censoring time, respectively. The observed time $R_i = \min(T_i^*, C_i)$ and failure indicator $\Delta_i = I(T_i^* \le C_i)$ are recorded. Let $\mathbf{Z}_i = (Z_{i1}, \cdots, Z_{iq})'$ denote time-invariant covariates. A Cox model that relates the hazard of interest to imaging trajectories and baseline covariates $\mathbf{Z}_i$ is defined as follows:

$$h(t \mid X_{i0}(v), X_{i1}(v), \mathbf{Z}_i) = h_0(t) \exp\Big[\int_{\mathcal{V}} \{\beta_0(v) X_{i0}(v) + t\beta_1(v) X_{i1}(v)\} dv + \boldsymbol{\gamma}' \mathbf{Z}_i\Big], \qquad (2)$$

where $h_0(t)$ denotes an unspecified baseline hazard function, and $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_q)'$ is a $q \times 1$ vector of regression coefficients. Similar to conventional joint models for longitudinal and survival data (e.g., Tsiatis and Davidian, 2004), the effect of longitudinal images on the hazard of interest is assessed through $\beta_0(v) X_{i0}(v) + t\beta_1(v) X_{i1}(v)$ in model (2). Here, $\beta_0(v)$ and $\beta_1(v)$ are imaging parameters that link $X_{i0}(v)$ and $tX_{i1}(v)$, respectively, to $h(t \mid \cdot)$, thereby characterizing the effects of longitudinal imaging data on the hazard function. In the ADNI study, for example, $\beta_0(v)$ is significantly positive near the brain regions of the "lateral ventricle" (see the upper panel of Figure 2). That is to say, subjects with larger lateral ventricles are at higher risk of suffering AD. Notably, $W_{ij}(v)$ in model (1) plays a role similar to measurement error and is thus not included in model (2).

Let $\mathbf{K}^{X0}(v_1,v_2), \mathbf{K}^{X1}(v_1,v_2)$, and $\mathbf{K}^{W}(v_1,v_2)$ denote the covariance operators of $X_{i0}(v)$, $X_{i1}(v)$, and $W_{ij}(v)$, respectively. Assume that $\mathbf{K}^{X0}(v_1,v_2), \mathbf{K}^{X1}(v_1,v_2)$, and $\mathbf{K}^{W}(v_1,v_2)$ are continuous. Based on the Karhunen-Loève (KL) decomposition (Karhunen, 1947; Loève, 1978), $X_{i0}(v)=\sum_{k=1}^{\infty}\xi_{ik}\psi_k^{X0}(v)$, $X_{i1}(v)=\sum_{l=1}^{\infty}\zeta_{il}\psi_l^{X1}(v)$, and $W_{ij}(v)=\sum_{m=1}^{\infty}\eta_{ijm}\psi_m^{W}(v)$, where $\psi_k^{X0}(\cdot), \psi_l^{X1}(\cdot)$, and $\psi_m^{W}(\cdot)$ are the eigenfunctions of $\mathbf{K}^{X0}, \mathbf{K}^{X1}$, and $\mathbf{K}^{W}$ operators, respectively, and $\xi_{ik}=\int_{\mathcal{V}}X_{i0}(v)\psi_k^{X0}(v)dv, \zeta_{il}=\int_{\mathcal{V}}X_{i1}(v)\psi_l^{X1}(v)dv$, and $\eta_{ijm}=\int_{\mathcal{V}}W_{ij}(v)\psi_m^{W}(v)dv$ are their corresponding eigenscores with $E(\xi_{ik})=E(\zeta_{il})=E(\eta_{ijm})=0$ and $\mathrm{Var}(\xi_{ik})=\lambda_k^{X0}$, $\mathrm{Var}(\zeta_{il})=\lambda_l^{X1}$, and $\mathrm{Var}(\eta_{ijm})=\lambda_m^{W}$. Then, model (1) can be rewritten as

$$Y_{ij}(v)=\mu(v)+\sum_{k=1}^{\infty}\xi_{ik}\psi_k^{X0}(v)+T_{ij}\sum_{l=1}^{\infty}\zeta_{il}\psi_l^{X1}(v)+\sum_{m=1}^{\infty}\eta_{ijm}\psi_m^{W}(v). \qquad (3)$$

The assumption that $X_{i0}(\cdot), X_{i1}(\cdot)$, and $W_{ij}(\cdot)$ are mutually uncorrelated is ensured by the assumption that $\{\xi_{ik}\}_{k=1}^{\infty}, \{\zeta_{il}\}_{l=1}^{\infty}$, and $\{\eta_{ijm}\}_{m=1}^{\infty}$ are mutually uncorrelated.

Model (3) can be approximated by truncating the summations to the first N_0, N_1, and N_W coordinates, respectively, as follows:

$$Y_{ij}(v)=\mu(v)+\sum_{k=1}^{N_0}\xi_{ik}\psi_k^{X0}(v)+T_{ij}\sum_{l=1}^{N_1}\zeta_{il}\psi_l^{X1}(v)+\sum_{m=1}^{N_W}\eta_{ijm}\psi_m^{W}(v), \qquad (4)$$

where the eigenscores ξ_{ik}, ζ_{il}, and η_{ijm} are assumed to be normally distributed as $\xi_{ik}\sim N(0,\lambda_k^{X0}), \zeta_{il}\sim N(0,\lambda_l^{X1})$, and $\eta_{ijm}\sim N(0,\lambda_m^{W})$. The normality assumption on the prior distributions of ξ_{ik}, ζ_{il}, and η_{ijm} can be relaxed without much difficulty. Other distributions with the existence of second order moments can also be considered. The numbers of principal components, N_0, N_1, and N_W, are typically

small and determined using criterion-based methods (Müller and Stadtmüller, 2005).

Suppose $\beta_0(v)$ and $\beta_1(v)$ in (2) can be expanded on $\psi_k^{X0}(\cdot)$ and $\psi_l^{X1}(\cdot)$, respectively. Then, $\beta_0(\cdot)$ and $\beta_1(\cdot)$ can likewise be approximated by $\beta_0(v)=\sum_{k=1}^{N_0}\beta_{0k}\psi_k^{X0}(v)$ and $\beta_1(v)=\sum_{l=1}^{N_1}\beta_{1l}\psi_l^{X1}(v)$, respectively. Based on the orthonormality of eigenimages, model (2) can be rewritten as

$$\begin{aligned} h(t \mid X_{i0}(v), X_{i1}(v), \mathbf{Z}_i) &= h_0(t)\exp\Big(\sum_{k=1}^{N_0}\beta_{0k}\xi_{ik} + t\sum_{l=1}^{N_1}\beta_{1l}\zeta_{il} + \boldsymbol{\gamma}'\mathbf{Z}_i\Big) \\ &= h_0(t)\exp\big(\boldsymbol{\beta}_0'\boldsymbol{\xi}_i + t\boldsymbol{\beta}_1'\boldsymbol{\zeta}_i + \boldsymbol{\gamma}'\mathbf{Z}_i\big), \end{aligned} \quad (5)$$

where $\boldsymbol{\beta}_0=(\beta_{01},\cdots,\beta_{0N_0})'$, $\boldsymbol{\xi}_i=(\xi_{i1},\cdots,\xi_{iN_0})'$, $\boldsymbol{\beta}_1=(\beta_{11},\cdots,\beta_{1N_1})'$, and $\boldsymbol{\zeta}_i=(\zeta_{i1},\cdots,\zeta_{iN_1})'$. Figure 1 presents a path diagram to illustrate the modeling procedure described by (1)–(5).

To derive the eigenimages in (4), the spectral decompositions of $\mathbf{K}^{X0}, \mathbf{K}^{X1}$, and $\mathbf{K}^{W}$ should be constructed. We start with centering the longitudinal imaging data as $\tilde{Y}_{ij}(v)=Y_{ij}(v)-\hat{\mu}(v)$, where the estimated overall mean image is $\hat{\mu}(v)=n^{-1}\sum_{i=1}^{I}\sum_{j=1}^{J_i}Y_{ij}(v)$, and $n=\sum_{i=1}^{I}J_i$. Under the assumption of model (1), we have

$$\mathrm{E}(\mathbf{Y}_{ij_1}\mathbf{Y}_{ij_2}^{'}) = \mathrm{Cov}(\mathbf{Y}_{ij_1},\mathbf{Y}_{ij_2}^{'}) = \mathbf{K}^{X0} + T_{ij_1}T_{ij_2}\mathbf{K}^{X1} + \delta_{j_1j_2}\mathbf{K}^{W}, \quad (6)$$

where $j_1, j_2 = 1,\cdots,J_i$, $\delta_{ij}=1$ if $i = j$ and 0 otherwise. Equation (6) suggests a straightforward solution for the covariance operators estimation by regressing linearly the "outcome" $\mathbf{Y}_{ij_1}\mathbf{Y}_{ij_2}^{'}$ on the "covariates" $(1, T_{ij_1}T_{ij_2}, \delta_{j_1j_2})$. Then, Equation (6) can be formulated as $\mathrm{E}(\mathbf{Y}_{ij_1j_2})=\mathbf{K}^{v}\mathbf{f}_{ij_1j_2}$, where $\mathbf{Y}_{ij_1j_2}=\mathbf{Y}_{ij_2}\otimes\mathbf{Y}_{ij_1}$ is a $p^2\times 1$ vector, $\otimes$ denotes the Kronecker product, $\mathbf{K}^{v}=[\mathrm{vec}(\mathbf{K}^{X0}),\mathrm{vec}(\mathbf{K}^{X1}),\mathrm{vec}(\mathbf{K}^{W})]$ is a $p^2\times 3$ matrix with the parameters of interest, $\mathrm{vec}(\cdot)$ stands for the vectorization

that stacks the columns of a matrix on top of each other, and $\mathbf{f}_{ij_1j_2} = (1, T_{ij_1} T_{ij_2}, \delta_{j_1j_2})'$ is a 3 × 1 vector of covariates. Let $\mathbf{Y}^v$ be a $p^2 \times m$ matrix with $m = \sum_{i=1}^{I} J_i^2$, and $\mathbf{F}$ be a $3 \times m$ matrix with columns of $\mathbf{f}_{ij_1j_2}, i = 1, \cdots, I,\ j_1, j_2 = 1, \cdots, J_i$. With these notations, We have $\mathrm{E}(\mathbf{Y}^v) = \mathbf{K}^v \mathbf{F}$. Thus, the moment estimator of $\mathbf{K}^v$ is $\mathbf{K}^v = \mathbf{Y}^v \mathbf{F}'(\mathbf{F}\mathbf{F}')^{-1}$, which provides the unbiased estimators of covariance operator $\mathbf{K}^{X0}, \mathbf{K}^{X1}$, and $\mathbf{K}^W$ as follows:

$$\mathbf{K}^{X0} = \sum_{i=1}^{I} \sum_{j_1=1}^{J_i} \sum_{j_2=1}^{J_i} \mathbf{Y}_{ij_1} \mathbf{Y}_{ij_2}' h_{ij_1j_2}^1, \quad \mathbf{K}^{X1} = \sum_{i=1}^{I} \sum_{j_1=1}^{J_i} \sum_{j_2=1}^{J_i} \mathbf{Y}_{ij_1} \mathbf{Y}_{ij_2}' h_{ij_1j_2}^2,$$
$$\mathbf{K}^{W} = \sum_{i=1}^{I} \sum_{j_1=1}^{J_i} \sum_{j_2=1}^{J_i} \mathbf{Y}_{ij_1} \mathbf{Y}_{ij_2}' h_{ij_1j_2}^3, \tag{7}$$

where $h_{ij_1j_2}^q, q = 1, 2, 3$ are the elements of the *q*th colomn of $\mathbf{H} = \mathbf{F}'(\mathbf{F}\mathbf{F}')^{-1}$. Despite an explicit representation of the covariance operators $\mathbf{K}^{X0}, \mathbf{K}^{X1}$, and $\mathbf{K}^W$ in (7), constructing and diagonalizing $\mathbf{K}^{X0}, \mathbf{K}^{X1}$, and $\mathbf{K}^W$ with a brute-force implementation fails when $\mathbf{Y}_{ij}$ is ultrahigh dimensional. For example, in the analysis of ADNI dataset, the MRI images of dimension $133 \times 170 \times 129$ ($p = 2,916,690$) result in a covariance operator $\mathbf{K}^{X0} = \{\mathbf{K}^{X0}(v_1, v_2)\}$ of dimension $2,916,690 \times 2,916,690$. A standard eigenanlysis on $\mathbf{K}^{X0}$ requires $O(p^3)$ operations and is essentially of infeasible level.

To overcome this challenge, we follow the idea of Zipunnikov et al. (2014) to manage longitudinal images using the HD-FPCA technique. Let $\mathbf{Y} = (\mathbf{Y}_1, \cdots, \mathbf{Y}_I)$, where $\mathbf{Y}_i = (\mathbf{Y}_{i1}, \cdots, \mathbf{Y}_{iJ_i})$ is a centralized $p \times J_i$ matrix and the colomn *j*, $j = 1, \cdots, J_i$, contains the unfolded image for subject *i* at visit *j*. We construct the singular value decomposition (SVD) of the matrix $\mathbf{Y}$ as $\mathbf{Y} = \mathbf{V}\mathbf{S}^{1/2}\mathbf{U}'$, where $\mathbf{V}$ is a *p* × *n* matrix with *n* orthonormal columns, $\mathbf{S}$ is a *n* × *n* diagonal matrix, and $\mathbf{U}$ is a *n* × *n* orthogonal matrix. The calculation of the SVD of $\mathbf{Y}$ requires only

$O(pn^2+n)$ operators. Meanwhile, we consider the $n \times n$ symmetric matrix $\mathbf{Y}'\mathbf{Y}$ with its spectral decomposition $\mathbf{Y}'\mathbf{Y}=\mathbf{USU}'$. Then, we obtain that $\mathbf{V}=\mathbf{YUS}^{-1/2}$. In an extremely ultrahigh-dimensional case where $\mathbf{Y}$ cannot be directly loaded into the memory, we partition it into D slices as $\mathbf{Y}'=[(\mathbf{Y}^1)'|(\mathbf{Y}^2)'|\cdots,(\mathbf{Y}^D)']$, where D is chosen to ensure that the dth slice $\mathbf{Y}^d$ of dimension $p/D\times n$ can be adapted to the available computer memory. $\mathbf{Y}'\mathbf{Y}$ can then be sequentially calculated as $\sum_{d=1}^{D}(\mathbf{Y}^d)'\mathbf{Y}^d$. Accordingly, $\mathbf{V}$ is partitioned into D slices as $\mathbf{V}'=[(\mathbf{V}^1)'|(\mathbf{V}^2)'|\cdots,(\mathbf{V}^D)']$, and $\mathbf{V}^d=\mathbf{Y}^d\mathbf{US}^{-1/2}$.

After obtaining the SVD of $\mathbf{Y}$, each image can be represented as $\mathbf{Y}_{ij}=\mathbf{VS}^{1/2}\mathbf{U}_{ij}$, where $\mathbf{U}_{ij}$ is the corresponding column of matrix $\mathbf{U}'$. Therefore, vectors $\mathbf{Y}_{ij}, i=1,\cdots,I, j=1,\cdots,J_i$ differ only through factors $\mathbf{U}_{ij}$ of dimension $n\times 1$. By substituting $\mathbf{Y}_{ij}=\mathbf{VS}^{1/2}\mathbf{U}_{ij}$ into Equation (7), the estimated covariance operators can be rewritten as

$$\mathbf{K}^{X0}=\mathbf{VS}^{1/2}\mathbf{X}_0^{\mathbf{U}}\mathbf{S}^{1/2}\mathbf{V}',\quad \mathbf{K}^{X1}=\mathbf{VS}^{1/2}\mathbf{X}_1^{\mathbf{U}}\mathbf{S}^{1/2}\mathbf{V}',\quad \mathbf{K}^{W}=\mathbf{VS}^{1/2}\mathbf{X}_W^{\mathbf{U}}\mathbf{S}^{1/2}\mathbf{V}', \quad (8)$$

where the inner matrices are $\mathbf{X}_0^{\mathbf{U}}=\sum_{i=1}^{I}\sum_{j_1=1}^{J_i}\sum_{j_2=1}^{J_i}\mathbf{U}_{ij_1}\mathbf{U}_{ij_2}'h_{ij_1j_2}^1$, $\mathbf{X}_1^{\mathbf{U}}=\sum_{i=1}^{I}\sum_{j_1=1}^{J_i}\sum_{j_2=1}^{J_i}\mathbf{U}_{ij_1}\mathbf{U}_{ij_2}'h_{ij_1j_2}^2$, and $\mathbf{X}_W^{\mathbf{U}}=\sum_{i=1}^{I}\sum_{j_1=1}^{J_i}\sum_{j_2=1}^{J_i}\mathbf{U}_{ij_1}\mathbf{U}_{ij_2}'h_{ij_1j_2}^3$. Equation (8) corresponds high-dimensional operators $\{\mathbf{K}^{X0},\mathbf{K}^{X1},\mathbf{K}^{W}\}$ with their low-dimensional counterparts $\{\mathbf{X}_0^{\mathbf{U}},\mathbf{X}_1^{\mathbf{U}},\mathbf{X}_W^{\mathbf{U}}\}$. Based on the spectral decomposition $\mathbf{S}^{1/2}\mathbf{X}_0^{\mathbf{U}}\mathbf{S}^{1/2}=\mathbf{A}_0\mathbf{\Sigma}_0\mathbf{A}_0'$, $\mathbf{S}^{1/2}\mathbf{X}_1^{\mathbf{U}}\mathbf{S}^{1/2}=\mathbf{A}_1\mathbf{\Sigma}_1\mathbf{A}_1'$, and $\mathbf{S}^{1/2}\mathbf{X}_W^{\mathbf{U}}\mathbf{S}^{1/2}=\mathbf{A}_W\mathbf{\Sigma}_W\mathbf{A}_W'$, the covariance operators in (8) can be written as

$$\mathbf{K}^{X0}=(\mathbf{VA}_0)\mathbf{\Sigma}_0(\mathbf{VA}_0)',\quad \mathbf{K}^{X1}=(\mathbf{VA}_1)\mathbf{\Sigma}_1(\mathbf{VA}_1)',\quad \mathbf{K}^{W}=(\mathbf{VA}_W)\mathbf{\Sigma}_W(\mathbf{VA}_W)', \quad (9)$$

where $\mathbf{A}_0, \mathbf{A}_1$, and $\mathbf{A}_W$ are $n \times n$ orthogonal matrices, and $\mathbf{\Sigma}_0, \mathbf{\Sigma}_1$, and $\mathbf{\Sigma}_W$ are diagonal matrices. The orthogonality of $\mathbf{A}_0, \mathbf{A}_1$, and $\mathbf{A}_W$ ensures that the $p \times n$ matrices $\mathbf{VA}_0, \mathbf{VA}_1$ and $\mathbf{VA}_W$ have orthonormal columns. Given the uniqueness of the spectral decomposition of a symmetric matrix, the expressions in (9) are the spectral decompositions of $\mathbf{K}^{X0}, \mathbf{K}^{X1}$, and $\mathbf{K}^{W}$, where the columns of $\mathbf{VA}_0, \mathbf{VA}_1$, and $\mathbf{VA}_W$ are the eigenimages, and the diagonal elements of $\mathbf{\Sigma}_0, \mathbf{\Sigma}_1$, and $\mathbf{\Sigma}_W$ are the eigenvalues.

3 Estimation

Denote $\boldsymbol{\eta}_{ij} = (\eta_{ij1}, \cdots, \eta_{ijN_W})'$, $\boldsymbol{\psi}_k^{X0} = (\psi_k^{X0}(1), \cdots, \psi_k^{X0}(p))'$, $\boldsymbol{\psi}_l^{X1} = (\psi_l^{X1}(1), \cdots, \psi_l^{X1}(p))'$, and $\boldsymbol{\psi}_m^{W} = (\psi_m^{W}(1), \cdots, \psi_m^{W}(p))'$. Let $\mathbf{\Psi}^{X0} = (\boldsymbol{\psi}_1^{X0}, \cdots, \boldsymbol{\psi}_{N_0}^{X0})$ be the $p \times N_0$ matrix of eigenfunctions obtained by binding the column vectors $\boldsymbol{\psi}_k^{X0}$, $\mathbf{\Psi}^{X1} = (\boldsymbol{\psi}_1^{X1}, \cdots, \boldsymbol{\psi}_{N_1}^{X1})$, and $\mathbf{\Psi}^{W} = (\boldsymbol{\psi}_1^{W}, \cdots, \boldsymbol{\psi}_{N_W}^{W})$. Let $\mathbf{\Lambda}^{X0} = \mathrm{diag}\{\lambda_1^{X0}, \cdots, \lambda_{N_0}^{X0}\}$, $\mathbf{\Lambda}^{X1} = \mathrm{diag}\{\lambda_1^{X1}, \cdots, \lambda_{N_1}^{X1}\}$, and $\mathbf{\Lambda}^{W} = \mathrm{diag}\{\lambda_1^{W}, \cdots, \lambda_{N_W}^{W}\}$ be the $N_0 \times N_0$, $N_1 \times N_1$, and $N_W \times N_W$ diagonal matrices of eigenvalues, respectively.

The most common method for modeling the baseline hazard function is to assume that it is piecewise constant as follows: $h_0(t) = h_g$, for $s_{g-1} < t \le s_g$, $g = 1, \ldots, G$, where $0 = s_0 < s_1 < \cdots < s_G$ define the intervals for $h_0(t)$ and are frequently selected according to the quantiles of the observed time R_i with $s_G > \max_i R_i$ for $i = 1, \cdots, I$. We define r_{ig} = 1 if $R_i \in (s_{g-1}, s_g]$ (i.e., subject i fails or is censored in the gth interval) and 0 otherwise.

Let θ be the vector that contains all the unknown parameters, $\mathbf{Y}_i = (\mathbf{Y}_{i1}, \cdots, \mathbf{Y}_{iJ_i})$, $\mathbf{Z} = (\mathbf{Z}_1^T, \cdots, \mathbf{Z}_I^T)^T$, $\mathbf{\Delta} = (\Delta_1, \cdots, \Delta_I)^T$, and $\mathbf{R} = (R_1, \cdots, R_I)^T$. Based on the proposed model, the observed-data likelihood function can be written as follows:

$$p(\mathbf{R},\boldsymbol{\Delta},\mathbf{Y},\mathbf{Z}\,|\,\boldsymbol{\theta})=\prod_{i=1}^{I}\int\int\int p(R_i,\Delta_i\,|\,\boldsymbol{\xi}_i,\boldsymbol{\zeta}_i,\mathbf{Z}_i,\boldsymbol{\theta})\,p(\mathbf{Y}_i\,|\,\boldsymbol{\xi}_i,\boldsymbol{\zeta}_i,\boldsymbol{\theta})\,p(\boldsymbol{\xi}_i\,|\,\boldsymbol{\theta})\,p(\boldsymbol{\zeta}_i\,|\,\boldsymbol{\theta})d\boldsymbol{\xi}_i d\boldsymbol{\zeta}_i, \quad (10)$$

which does not have an explicit form due to model complexity. The detailed form of (10) is provided in Appendix 2 of Supplementary Material.

3.1 Estimation procedures

3.1.1 Two-stage procedure

A naive method for avoiding the intractable likelihood (10) is the use of a two-stage procedure, which is implemented in two stages as follows: (i) obtain the best linear unbiased estimator of eigenscores based on model (4): $(\boldsymbol{\xi}_i',\boldsymbol{\zeta}_i')'=\left[\mathbf{B}_i'\{\boldsymbol{\Lambda}_i^W\}^{-1}\mathbf{B}_i+\{\boldsymbol{\Lambda}^X\}^{-1}\right]^{-1}\mathbf{B}_i'\{\boldsymbol{\Lambda}_i^W\}^{-1}vec(\mathbf{V}_i)$, where $\mathbf{B}_i=(\mathbf{1}_{J_i}\otimes\mathbf{B}_0,\mathbf{T}_i\otimes\mathbf{B}_1),\mathbf{B}_0=(\boldsymbol{\Psi}^{W'}\boldsymbol{\Psi}^W)^{-1}\boldsymbol{\Psi}^{W'}\boldsymbol{\Psi}^{X0},\mathbf{B}_1=(\boldsymbol{\Psi}^{W'}\boldsymbol{\Psi}^W)^{-1}\boldsymbol{\Psi}^{W'}\boldsymbol{\Psi}^{X1},\mathbf{T}_i=(T_{i1},\cdots,T_{iJ_i})',\boldsymbol{\Lambda}_i^W=\mathbf{I}_{J_i}\otimes\boldsymbol{\Lambda}^W,\boldsymbol{\Lambda}^X=\mathrm{diag}(\boldsymbol{\Lambda}^{X0},\boldsymbol{\Lambda}^{X1}),\mathbf{V}_i=(\mathbf{V}_{i1},\cdots,\mathbf{V}_{iJ_i})$, and $\mathbf{V}_{ij}=(\boldsymbol{\Psi}^{W'}\boldsymbol{\Psi}^W)^{-1}\boldsymbol{\Psi}^{W'}\mathbf{Y}_{ij}$ and (ii) estimate model (5) by replacing $\boldsymbol{\xi}_i$ and $\boldsymbol{\zeta}_i$ by $\boldsymbol{\xi}_i$ and $\boldsymbol{\zeta}_i$. However, as shown by Prentice (1982), such a two-stage procedure leads to considerable bias in the joint analysis of longitudinal and survival data when the measurement error of the longitudinal observation is large. In the proposed joint model with imaging data, the longitudinal images are modeled through equation (4), which includes a term $\sum_{m=1}^{N_W}\eta_{ijm}\psi_m^W(v)$ to accommodate subject-visit-specific deviations from the subject-specific mean image. When the repeatedly collected images vary markedly across visits, this term tends to be large and plays a role similar to measurement error. Thus, applying a two-stage procedure to the proposed model is problematic even though the observed images are perfectly scanned. We provide a theoretical justification below from a frequentist perspective to show that the parameter estimators obtained using the two-stage approach are biased. Given that the main issue in a two-stage approach originates from the subject-visit-specific term, for notation simplicity, we keep $\boldsymbol{\eta}_{ij}$ and $\boldsymbol{\xi}_i$ but exclude $\boldsymbol{\zeta}_i$ (and thus $\boldsymbol{\beta}_1$) in the model and assume a

balanced design with $J_i = J$ in Theorem 1 and its proof (see Appendix 3 of Supplementary Material).

Theorem 1. *For a joint model defined in (4) and (5), let* $\boldsymbol{\alpha}^* = (\boldsymbol{\beta}_0^*, \boldsymbol{\gamma}^*)$ *and* $\boldsymbol{\alpha} = (\boldsymbol{\beta}_0, \hat{\boldsymbol{\gamma}})$ *be the true parameters and the maximum likelihood estimator obtained from a two-stage approach, respectively. Under the regularity conditions (C1)-(C2) as stated in the Appendix 2 of Supplementary Material, we have* $E(\boldsymbol{\alpha}) \neq \boldsymbol{\alpha}^*$.

3.1.2 Bayesian joint estimation procedure

Given the inferiority of the two-stage estimation procedure for the proposed joint model, we propose a Bayesian approach to jointly estimate all unknowns. To conduct a Bayesian inference of the joint model defined by (4)–(5), we must specify prior distributions for all unknown parameters in $\boldsymbol{\theta}$. For $\boldsymbol{\beta}_0, \boldsymbol{\beta}_1$, and $\boldsymbol{\gamma}$, we assign normal priors as follows: $\boldsymbol{\beta}_0 \sim N(\mathbf{0}, \sigma_{\beta 0}^2 \mathbf{I}_{N_0}), \boldsymbol{\beta}_1 \sim N(\mathbf{0}, \sigma_{\beta 1}^2 \mathbf{I}_{N_1})$, and $\boldsymbol{\gamma} \sim N(\mathbf{0}, \sigma_\gamma^2 \mathbf{I}_q),$ where $\mathbf{I}_a$ denotes the *a*-dimensional identity matrix, $\sigma_{\beta 0}^2, \sigma_{\beta 1}^2$, and σ_γ^2 are chosen to be very large when prior knowledge about $\boldsymbol{\beta}_0, \boldsymbol{\beta}_1$, and $\boldsymbol{\gamma}$ is unavailable. For variance components $\lambda_k^{X_0}, \lambda_l^{X_1}$, and λ_m^W, we assume independent inverse-gamma priors as follows: For $k = 1, \cdots, N_0, l = 1, \cdots, N_1, m = 1, \cdots, N_W$, $\lambda_k^{X0} \sim \mathrm{IG}(a_k^{X0}, b_k^{X0}), \lambda_l^{X1} \sim \mathrm{IG}(a_l^{X1}, b_l^{X1})$, and $\lambda_m^W \sim \mathrm{IG}(a_m^W, b_m^W),$ where $\mathrm{IG}(a, b)$ denotes the inverse-gamma distribution with hyperparameters *a* and *b*. Here, we regard $\lambda_k^{X0}, \lambda_l^{X1}$, and λ_m^W as parameters to be estimated. However, a simple empirical Bayes (EB) method, which estimate $\lambda_k^{X0}, \lambda_l^{X1}$, and λ_m^W by diagonalizing the covariance operators as discussed in Section 2, has been proved to work as well in practice. The EB estimators provide some information in choosing proper hyperparameters in Bayesian inference. Specifically, one may choose $a_k^{X0} \leq 0.01$ and $b_k^{X0} \leq 0.01 \lambda_k^{X0}$, where λ_k^{X0} can be taken as the EB estimator $\tilde{\lambda}_k^{X0}$. Similar specifications hold for $a_l^{X1}, b_l^{X1}, a_m^W$,

and b_m^W . For the hazard values h_gs that are involved in the baseline hazard function, we assume independent gamma distribution as follows: For $g = 1, \cdots, G, h_g \sim \text{Gamma}(\alpha_{1g}, \alpha_{2g})$, where α_{1g} and α_{2g} are hyperparameters. Setting $(\alpha_{1g}, \alpha_{2g}) = (0.2, 0.4)$ or (0.5,1) results in a flat prior for the h_gs (Lee et al., 2015).

The Bayesian estimate of θ can be obtained using the mean or mode of the posterior samples drawn from $p(\theta \mid \mathbf{D})$, where $\mathbf{D} = (\mathbf{R}, \mathbf{\Delta}, \mathbf{Y}, \mathbf{Z})$ denotes the observed data. However, direct sampling from $p(\theta \mid \mathbf{D})$ is intractable because of the existence of latent variables $\xi = \{\xi_i, i = 1, \cdots, I\}$ and $\zeta = \{\zeta_i, i = 1, \cdots, I\}$. Instead, we work on $p(\theta, \xi, \zeta \mid \mathbf{D})$ and employ the Gibbs sampler to iteratively simulate each unknown from its full conditional distribution. Here, we denote $(\mathbf{\Lambda}^A)^{-1} = \mathbf{\Lambda}^{-A}$ for notation simplicity. Notably, although η_{ij} s are also unknown, we do not incorporate them into the set of parameters because they can be automatically updated by using $\eta_{ij} = (\mathbf{\Psi}^{W\prime}\mathbf{\Psi}^W)^{-1}\mathbf{\Psi}^{W\prime}(\mathbf{Y}_{ij} - \mathbf{\Psi}^{X0}\xi_i - T_{ij}\mathbf{\Psi}^{X1}\zeta_i)$. Then, the full conditional distributions involved in the Gibbs sampler are provided as follows:

$$
\begin{aligned}
[\xi_i, \zeta_i \mid \text{others}] \propto \quad & \exp\Big[-\frac{1}{2}\sum_{j=1}^{J_i}(\mathbf{Y}_{ij} - \mathbf{\Psi}^{X0}\xi_i - T_{ij}\mathbf{\Psi}^{X1}\zeta_i)'\mathbf{\Psi}^W(\mathbf{\Psi}^{W\prime}\mathbf{\Psi}^W)^{-1}\mathbf{\Lambda}^{-W} \\
& \times(\mathbf{\Psi}^{W\prime}\mathbf{\Psi}^W)^{-1}\mathbf{\Psi}^{W\prime}(\mathbf{Y}_{ij} - \mathbf{\Psi}^{X0}\xi_i - T_{ij}\mathbf{\Psi}^{X1}\zeta_i) + \Delta_i(\boldsymbol{\beta}_0'\xi_i + R_i\boldsymbol{\beta}_1'\zeta_i) \\
& - \sum_{g=1}^{G} r_{ig} \sum_{k=1}^{g} \int_{s_{k-1}}^{\min(s_k, R_i)} h_k \exp(\boldsymbol{\beta}_0'\xi_i + t\boldsymbol{\beta}_1'\zeta_i + \boldsymbol{\gamma}'\mathbf{Z}_i)dt \\
& - \frac{1}{2}\xi_i'\mathbf{\Lambda}^{-X0}\xi_i - \frac{1}{2}\zeta_i'\mathbf{\Lambda}^{-X1}\zeta_i\Big],\ i = 1, \cdots, I, \\
[\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\gamma} \mid \text{others}] \propto \quad & \exp\Big[-\sum_{g=1}^{G} r_{ig} \sum_{k=1}^{g} \int_{s_{k-1}}^{\min(s_k, R_i)} h_k \exp(\boldsymbol{\beta}_0'\xi_i + t\boldsymbol{\beta}_1'\zeta_i + \boldsymbol{\gamma}'\mathbf{Z}_i)dt \\
& + \Delta_i(\boldsymbol{\beta}_0'\xi_i + R_i\boldsymbol{\beta}_1'\zeta_i + \boldsymbol{\gamma}'\mathbf{Z}_i) - \frac{1}{2\sigma_{\beta 0}^2}\boldsymbol{\beta}_0'\boldsymbol{\beta}_0 - \frac{1}{2\sigma_{\beta 1}^2}\boldsymbol{\beta}_1'\boldsymbol{\beta}_1 - \frac{1}{2\sigma_\gamma^2}\boldsymbol{\gamma}'\boldsymbol{\gamma}\Big], \\
[\lambda_k^{X0} \mid \text{others}] = \quad & \text{IG}\Big[I/2 + a_k^{X0}, \sum_{i=1}^{I}\xi_{ik}^2/2 + b_k^{X0}\Big],\ k = 1, \cdots, N_0, \\
[\lambda_l^{X1} \mid \text{others}] = \quad & \text{IG}\Big[I/2 + a_l^{X1}, \sum_{i=1}^{I}\zeta_{il}^2/2 + b_l^{X1}\Big],\ l = 1, \cdots, N_1,
\end{aligned}
$$

$$[\lambda_m^W \mid \text{others}] = \quad \text{IG}\Big[\sum_{i=1}^{I} J_i / 2 + a_m^W, \sum_{i=1}^{I}\sum_{j=1}^{J_i} \eta_{ijm}^2 / 2 + b_m^W\Big],\ m = 1, \cdots, N_W,$$

$$[h_g \mid \text{others}] = \quad \text{Gamma}\Big[\alpha_{1g} + \sum_{i=1}^{n} r_{ig}\Delta_i, \alpha_{2g} + \sum_{i=1}^{n} \exp(\boldsymbol{\beta}_0'\boldsymbol{\xi}_i + \boldsymbol{\gamma}'\mathbf{Z}_i)\Big\{r_{ig}\int_{s_{g-1}}^{R_i} \exp\big(t\boldsymbol{\beta}_1'\boldsymbol{\zeta}_i\big)dt$$

$$+ I(g < G)\sum_{l=g+1}^{G} r_{il}\int_{s_{g-1}}^{s_g} \exp\big(t\boldsymbol{\beta}_1'\boldsymbol{\zeta}_i\big)dt\Big\}\Big],\ g = 1, \cdots, G.$$

Among the above full conditional distributions, the last four are familiar distributions and can be easily sampled from, whereas the first two do not have explicit forms and require the use of Metropolis-Hastings (MH) algorithm (Metropolis et al., 1953; Hastings, 1970). The MH algorithm can be implemented by choosing a multivariate normal distribution with the mean centered at the current value and a covariance matrix proportional to the inverse of the Fisher information matrix of the posterior distribution as the proposal distribution. A small variance parameter is often used to tune the acceptance rate at approximately 35%.

3.2 Dynamic prediction

Prediction of survival probabilities for a new subject *i* given a set of longitudinal images $\mathcal{Y}_i^{\{t\}} = \{\mathbf{Y}_{ij}; j = 1, \cdots, J_i, T_{ij} \le t\}$ and baseline covariates $\mathbf{Z}_i$ is also of scientific interest. We focus on a time frame $(t, t+\delta]$, within which an intervention for improving patients' survival is available. The conditional probability of survival time $t+\delta$ given survival up to t can be calculated as

$$\pi_i(t+\delta \mid t) = \Pr(T_i^* \ge t+\delta \mid T_i^* > t, \mathcal{Y}_i^{\{t\}}, \mathbf{Z}_i, \mathbf{D}) = \int \Pr(T_i^* \ge t+\delta \mid T_i^* > t, \mathcal{Y}_i^{\{t\}}, \mathbf{Z}_i; \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathbf{D})d\boldsymbol{\theta}$$

, where the second part of the integrand is the posterior distribution of the parameters given the observed data $\mathbf{D}$. Based on the proposed joint model, the first part of the integrand is

$$\Pr(T_i^* \ge t+\delta \mid T_i^* > t, \mathcal{Y}_i^{\{t\}}, \mathbf{Z}_i, \mathbf{D}; \boldsymbol{\theta}) = \int\int \frac{\Pr(T_i^* \ge t+\delta \mid \mathbf{Z}_i, \boldsymbol{\xi}_i, \boldsymbol{\zeta}_i, \mathbf{D}; \boldsymbol{\theta})}{\Pr(T_i^* \ge t \mid \mathbf{Z}_i, \boldsymbol{\xi}_i, \boldsymbol{\zeta}_i, \mathbf{D}; \boldsymbol{\theta})} p(\boldsymbol{\xi}_i, \boldsymbol{\zeta}_i \mid T_i^* > t, \mathcal{Y}_i^{\{t\}}, \mathbf{Z}_i, \mathbf{D}; \boldsymbol{\theta})d\boldsymbol{\xi}_i d\boldsymbol{\zeta}_i$$

$$\Pr(T_i^* \ge t \mid \mathbf{Z}_i, \boldsymbol{\xi}_i, \boldsymbol{\zeta}_i, \mathbf{D}; \boldsymbol{\theta}) = \exp\Big\{-\sum_{g=1}^{G} r_{ig}\sum_{k=1}^{g}\int_{s_{k-1}}^{\min(s_k, t)} h_k$$

, where

$\exp\left[\boldsymbol{\beta}_0'\boldsymbol{\xi}_i + s\boldsymbol{\beta}_1'\boldsymbol{\zeta}_i + \boldsymbol{\gamma}'\mathbf{Z}_i\right]ds\}$. Thus, a Monte Carlo estimates of $\pi_i(t+\delta \mid t)$ can be obtained as

$$\hat{\pi}_i(t+\delta \mid t) = \frac{1}{B}\sum_{b=1}^{B} \Pr(T_i^* \geq t+\delta \mid \mathbf{Z}_i, \boldsymbol{\xi}_i^{(b)}, \boldsymbol{\zeta}_i^{(b)}, \mathbf{D}; \boldsymbol{\theta}^{(b)}) / \Pr(T_i^* \geq t \mid \mathbf{Z}_i, \boldsymbol{\xi}_i^{(b)}, \boldsymbol{\zeta}_i^{(b)}, \mathbf{D}; \boldsymbol{\theta}^{(b)}),$$

where $\boldsymbol{\theta}^{(b)} \sim N(\boldsymbol{\theta}, \mathcal{H})$, $(\boldsymbol{\xi}_i^{(b)}, \boldsymbol{\zeta}_i^{(b)}) \sim p(\boldsymbol{\xi}_i, \boldsymbol{\zeta}_i \mid T_i^* > t, \mathcal{Y}_i^{\{t\}}, \mathbf{Z}_i, \mathbf{D}; \boldsymbol{\theta}^{(b)})$, and B is the number of Monte Carlo samples. We use the time-dependent integrated area under the receiver operating characteristic curve (AUC) to assess the performance of the proposed dynamic prediction procedure. The AUC measures how well the proposed model can discriminate patients who will experience the event from patients who will not (Heagerty and Zheng, 2005; Andrinopoulou et al., 2018). Given a randomly selected pair of subjects (i_1, i_2), the AUC is a function of (t, δ) and can be defined as

$$\mathrm{AUC}(t,\delta) = P\left(\pi_{i_1}(t,\delta) < \pi_{i_2}(t,\delta) \mid \{T_{i_1}^* \in (t, t+\delta]\} \cap \{T_{i_2}^* > t+\delta\}\right).$$

4 Simulation study

4.1 Simulation 1

We simulate 100 datasets based on model (4)–(5) with $\mathcal{V} = [1,50]\times[1,50]\times[1,50]$ and $\mu(v) = 0$. We consider two sample sizes: I = 300 and I = 500. Under each sample size, we assume $J_i \in \{7,8,9\}$, $N_0 = N_1 = 2$, and N_W = 4. The eigenimages $\psi_k^{X0}(v)$, $\psi_l^{X1}(v)$, and $\psi_m^{W}(v)$ are presented in Figure S2 of Supplementary Material. These eigenimages can be regarded as 3D images with voxel intensities on the scale of $[0,1]$, where the voxels within each sub-block are set to 1, and the outside white voxels are set to 0. The two grey, two green, and four red sub-blocks represent $\boldsymbol{\psi}_k^{X0}, \boldsymbol{\psi}_l^{X1}$, and $\boldsymbol{\psi}_m^{W}$, respectively. Each of the eigenimages is normalized to have norm one. The true eigenvalues are set as $(\lambda_1^{X0}, \lambda_2^{X0}) = (\lambda_1^{X1}, \lambda_2^{X1}) = (1.0, 0.5)$ and $(\lambda_1^{W}, \lambda_2^{W}, \lambda_3^{W}, \lambda_4^{W}) = (0.25, 0.2, 0.15, 0.1)$. Two baseline covariates $\mathbf{Z}_i = (Z_{i1}, Z_{i2})'$ are considered and independently generated from N(0, 1) and t(5), respectively, where t(5) denotes the t distribution with a

degree of freedom 5. The true population values of the regression coefficient are set as $\boldsymbol{\beta}_0 = (1,-1)'$, $\boldsymbol{\beta}_1 = (-1,1)'$, and $\boldsymbol{\gamma} = (1,-1)'$. The failure time T_i^* is generated based on model (5) with three types of baseline hazard functions: (i) $h_0(t) = 1$ (constant); (ii) $h_0(t) = t + 0.5$ (linear), and (iii) $h_0(t) = t^2 + 0.3$ (nonlinear). The censoring time C_i is independently generated from a uniform distribution $U[c_1, c_2]$, which is a common setting in survival analysis (e.g., Tsiatis and Davidian, 2004; Lee et al., 2015; Andrinopoulou et al., 2018). In particular, if $T^* < c1$, then $T^* < C$, and thus censoring does not occur. Instead, if $T^* > c2$, then $T^* > C$, leading to censoring. Here, c_1 and c_2 are chosen to achieve censoring rates of 30% and 50%, respectively.

We unfold each images $\mathbf{Y}_{ij}$ and obtain an ultrahigh dimensional vector of size $p = 50 \times 50 \times 50 = 125,000$. Then, we utilize the HD-FPCA technique described in Section 2 to obtain the estimated eigenimages $\hat{\psi}_k^{X0}$, $\hat{\psi}_l^{X1}$, and $\hat{\psi}_m^W$. In modeling the baseline hazard function $h_0(t)$, we set G = 5 and the cut points $s_0, \cdots, s_G$ that define the sub-intervals as the quantiles of the observed survival time. The hyperparameters of the prior distributions discussed in Section 3.1.2 are assigned as follows: for $k, l = 1,2, m = 1, \cdots, 4$, and $g = 1, \cdots, 5$, $\sigma_{\beta 0}^2 = \sigma_{\beta 1}^2 = \sigma_\gamma^2 = 10^4$, $a_k^{X0} = a_l^{X1} = a_m^W = 0.01$, $(a_{1g}, a_{2g}) = (0.2, 0.4)$, $b_k^{X0} = 0.01\tilde{\lambda}_k^{X0}$, $b_l^{X1} = 0.01\tilde{\lambda}_l^{X1}$, and $b_m^W = 0.01\tilde{\lambda}_m^W$, where $\tilde{\lambda}_k^{X0}$, $\tilde{\lambda}_l^{X1}$, and $\tilde{\lambda}_m^W$ are their EB estimators obtained as a by-product of the HD-FPCA procedure. We conduct a few test runs to decide the number of burn-in iterations required for convergence and find that the MCMC algorithm converges within 5,000 iterations. To be conservative, we collect 10,000 MCMC samples after 10,000 burn-in iterations. We calculate the bias (BIAS), root mean square error (RMSE), and standard error (SE) between the posterior means of the unknown parameters and their true population values to assess the empirical performance of Bayesian estimation. Table 1 summarizes the estimation results of regression parameters $\boldsymbol{\beta}_0, \boldsymbol{\beta}_1$, and $\boldsymbol{\gamma}$ on the basis of the 100 replicated datasets. The BIAS, RMSE, and SE for most of the parameters

are fairly small, indicating the satisfactory performance of parameter estimation in all the settings under consideration. As expected, the performance is improved as either the sample size increases from $I = 300$ to $I = 500$ or the censoring rate (CR) decrease from 50% to 30%. The Bayesian estimates of the eigenscores λ_k^{X0}, λ_l^{X1}, and λ_m^W are close to the true population values in all simulation settings. We report the results obtained in the setting of $I = 500$, CR $= 30\%$, and $h(t) = 1$ in the column "Normal" of Table S1 of Supplementary Material. Moreover, we follow the existing literature (Li et al., 2013; Feng et al., 2020) to adopt BIC to select N_0, N_1 and N_W ($N_0 = N_1$ is assumed for simplicity). Given that the main interest is the selection of N_0 and N_1 that determine the number of imaging parameters in model (5), we let $N_0 = N_1$ vary from 1 to 5 and fix N_W at 4 to avoid a tedious model comparison. Additional simulation results (not reported) also confirm that the choice of N_W hardly affects the estimation of the survival model. As shown in Table S2 of Supplementary Material, BIC correctly selects the true number of principle components in 99 of 100 replications. The computation time for implementing the MCMC algorithm in each replication of Simulation 1 is 4.40 minutes using a PC Intel Core i7-9700 3.00GHz CPU and 16.0 GB RAM.

4.2 Simulation 2

In this section, we compare the empirical performances of the proposed joint estimation procedure and the naive two-stage procedure that is discussed in Section 3.1.1.

We consider the 100 replicated datasets generated in the setting of $\mathrm{CR} = 30\%$ and $h_0(t) = 1$ in Simulation 1 and an additional case of $(\lambda_1^W, \lambda_2^W, \lambda_3^W, \lambda_4^W) = (1, 0.75, 0.5, 0.25)$ to examine how the performances of the proposed and naive procedures vary with the increased magnitude of the subject-visit-specific variation of the imaging part. The proposed method is performed in a similar manner as in Simulation 1. The two-stage procedure is implemented using the steps described in Section 3.1.1. Table S3 of

Supplementary Material presents the estimates of $\boldsymbol{\beta}_0, \boldsymbol{\beta}_1$, and $\boldsymbol{\gamma}$ obtained from the proposed and two-stage procedures. The two-stage approach tends to produce biased estimation results, and its performance does not improve when the sample size increases. Moreover, the bias becomes large when the subject-visit-specific variation of images increases. This finding agrees with the theoretical result of Theorem 1, which indicates that the parameter estimates obtained from the two-stage approach are biased even if the longitudinal images are perfectly observed. On the contrary, the performance of the proposed method significantly improves with the sample size.

We also conduct simulations under other settings and examine the sensitivity of Bayesian results to the model and prior assumptions. The details are provided in Appendix 4 of Supplementary Material.

5 Real Data Analysis

The proposed method is applied to the ADNI dataset (adni.loni.usc.edu). We considered subjects who suffered from MCI at baseline and included only those who had at least three longitudinal MRI scans because too few information is insufficient to describe individuals' time-dependent features through model (1). The final dataset contained $I = 339$ subjects who participated 3 to 6 follow-up visits, including baseline, 6 months, 12 months, 18 months, 24 months, and 36 months. The exact visiting time $\tilde{T}_{ij}$ may be slightly earlier or later than these preassigned date and the maximum visiting time is around 37.14 months. The time variable T_{ij} was standardized according to the procedure in Greven et al. (2010). Among the 339 MCI subjects included in the current study, 127 of them met the inclusion criteria shown in the general procedure manual of the ADNI study and were detected as AD during the cohort study, whereas 212 still remained at MCI stage. The observed time was calculated as the period from the baseline to the date of the first diagnosis of AD if $\Delta_i = 1$ or the date of last visit if $\Delta_i = 0$. The censoring rate was approximately 62%.

For each subject, the MRI data, which were preprocessed to $133 \times 170 \times 129$ dimension, were collected over time. Besides the longitudinal MRI data and time-to-AD observations, we also included subjects' demographic variables and biomarkers at baseline. The demographic variables include "gender (Z_{i1}, 1 = male)" and "whether or not the subject was married (Z_{i2}, 1 = married)." A biomarker variable, APOE-$\epsilon 4$ gene, which was identified as important in the existing medical research, was included as well. The APOE-$\epsilon 4$ was coded using two dummy variables: "one APOE-$\epsilon 4$ allele carrier (Z_{i3})" and "two APOE-$\epsilon 4$ alleles carrier (Z_{i4})".

The main purpose of this study is to (1) characterize the dynamic patterns of individuals' MRI images across time, (2) investigate the possibly time-variant effects of MRI images on the hazards of AD, and (3) examine the effectiveness of the use of longitudinal MRI images for dynamic prediction of AD onset. We considered the proposed joint model defined by (4) and (5). For statistical inference, we first standardized the longitudinal MRI data and then adopted HD-FPCA to obtain the eigenimages accordingly. We utilized BIC to determine N_0, N_1, and N_W. Although separately fine-tuning N_0, N_1, and N_W is not technically involved, it is rather cumbersome. Thus, we simply assumed that $N_0 = N_1 = N_W = N$. The BIC values corresponding to N from 1 to 10 are reported in Figure S4 of Supplementary Material, and those corresponding to $N > 10$ are considerably large and not reported. The BIC value is the smallest when $N = 7$. Thus, the joint model with the first seven eigenimages was chosen for the subsequent analysis. In modeling the baseline hazard function $h_0(t)$, we chose $G = 5$ and selected cut points $s_0, \cdots, s_G$ as the quantiles of the observed survival time. Owing to the lack of prior information, the hyperparameters of the prior distributions were assigned to reflect flat priors as in Simulation 1. We ran several parallel chains with different initial values and found that the MCMC algorithm converged within 10,000 iterations. Thus, we collected 10,000 posterior samples after 10,000 burn-in iterations for posterior inference. The computation

time for implementing the MCMC algorithm in this study is 5.03 minutes using the same PC provided in Simulation 1. The Bayesian estimates of parameters together with their standard error estimates are presented in Table S6 of Supplementary Material. The following findings are obtained.

For the imaging predictors, we observe several significant effects, such as $\hat{\beta}_{01} = 0.138(0.029), \hat{\beta}_{05} = 0.138(0.049)$, and $\hat{\beta}_{11} = -0.769(0.246)$. That is, the 1st and 5th eigenimages of random imaging intercept and the 1st eigenimage of random imaging slope exhibit significant effects on AD hazards. Figure S5 of Supplementary Material depicts a full picture of the estimated imaging parameters $\beta_0(\cdot)$ and $\beta_1(\cdot)$, which are calculated using the estimated coefficients $\hat{\beta}_{01}, \hat{\beta}_{05}$, and $\hat{\beta}_{11}$ along with the corresponding eigenimages based on the KL decomposition. Several brain regions are detected to be highly associated with AD hazards. For example, the positive effect of "lateral ventricles" (Figure 2, the top row) is evident, implying that the enlargement of the lateral ventricles is positively associated with the development of AD. This finding is consistent with the previous medical studies (Nestor et al., 2008). In the brain regions that depict negative effects, the magnitudes of the effects of "inferior parietal lobe" (Figure 2, the second row) and "inferior temporal gyrus and fusiform gyrus" (Figure 2, the third row) are relatively large. This result is in line with the existing research, which indicated that AD is negatively correlated to the volume or cortical thickness of inferior parietal lobe, inferior temporal gyrus, and fusiform gyrus. Moreover, for the brain regions detected as significant in $\beta_1(\cdot)$, such as the " lateral ventricles" (Figure S6, the first row), "precentral gyrus, posterior cingulate and pars triangularis" (Figure S6, the second row), and "cerebellum exterior" (Figure S6, the third row), they have increasingly positive effects on AD hazards over time because $\beta_1(\cdot)$ is the parameter of the random imaging slope and thus represents a time-dependent effect. Although previous medical researches (Canu et al., 2011) suggested that these brain regions are strongly associated with the progression of AD, our study further reveals that the volume loss of

these regions increases the risk of developing AD in an incremental manner with aging.

For the baseline covariates, gender had a significant negative effect on AD hazards, indicating that females have a higher risk of suffering from AD than males. This finding agrees with the current medical literature. A surprising result was that the APOE-$\epsilon 4$ gene, which was found to be an important risk factor of AD in previous research (Okuizumi et al., 1994), had an insignificant effect on the conversion to AD in our study. One possible explanation is that APOE-$\epsilon 4$ gene is somehow a confounding variable between certain brain regions and the hazards of AD. To verify this guess, we regressed the Bayesian estimates of eigenscores, $\hat{\xi}_{ik}$ and $\hat{\zeta}_{il}$, on the APOE-$\epsilon 4$ gene (Z_{i3} and Z_{i4}) for $k = 1, \cdots, 7$ and $l = 1, \cdots, 7$. The results (not reported) show that APOE-$\epsilon 4$ alleles have significant effects on $\zeta_{i2}, \zeta_{i3}, \zeta_{i5}, \zeta_{i6}$, and ζ_{i7}, thereby implying that APOE-$\epsilon 4$ alleles are associated with AD through certain brain regions. Hence, the effects of APOE-$\epsilon 4$ alleles on AD hazards become insignificant when the imaging predictors are controlled.

In addition to the proposed model M_p, we further considered three competing models, M_{r1}, M_{r2}, and M_{r3}. M_{r1} replaces the longitudinal MRI data incorporated in M_p by patients' longitudinal ventricular volume, which measures four brain cavities containing cerebral spinal fluid. Such a scalar measure was used in previous medical studies (Nestor et al., 2008) as a marker of AD progression, and we included it in M_{r1} for comparison of the prediction capacity between the dynamic imaging information and longitudinal scalar measure in AD diagnosis. M_{r2} and M_{r3} are two reduced models of M_p. M_{r2} is a Cox regression model considered by Lee et al. (2015) that only includes the baseline MRI images. M_{r3} does not include any imaging predictor in the Cox model. We utilized BIC to determine the number of eigenimages in M_{r2}, and $N = 3$ was chosen. The estimation results of M_{r1}, M_{r2}, and M_{r3} are presented in Table S6 of Supplementary Material, together with those of M_p. Similar to the results of M_p,

the effects of APOE-$\epsilon 4$ alleles are insignificant in M_{r1} and M_{r2} wherein the imaging predictors/ventricular volume are controlled, but become significant in M_{r3} wherein the imaging predictors/ventricular volume are excluded. Although a significant positive effect of "pateral ventricles" on AD progression (see Figure S7 of Supplementary Material) was revealed by M_{r2}, other brain regions that were identified as important indicators of AD progression by M_p were not detected by M_{r2}.

Furthermore, we used a Monte Carlo cross-validation method to calculate the integrated AUC for the four models at t = 19 and 25 months and δ = 6 and 12 months. The full dataset was randomly split into a training set $\mathbf{A}_{train}$ with 239 subjects and a test set $\mathbf{A}_{test}$ with 100 subjects. Such random split was repeated 100 times. For each split, we fitted all the competing models to the training set, estimated the parameters, and then calculated AUC using the test set. The AUC value of competing model M_c is calculated by

$$\mathrm{AUC}_{M_c}(t,\delta)=\frac{1}{\sum_{i_1,i_2\in\mathbf{A}_{test}} I(\{T_{i_1}^*\in(t,t+\delta]\}\cap\{T_{i_2}^*>t+\delta\})}\sum_{i_1,i_2\in\mathbf{A}_{test}} P\left(\pi_{i_1}(t,\delta)<\pi_{i_2}(t,\delta)\,|\,\{T_{i_1}^*\in(t,t+\delta]\}\cap\{T_{i_2}^*>t+\delta\}\right)$$

for $t\in\{19,25\}$ and $\delta\in\{6,12\}$. Here $\pi_i(t,\delta)$ corresponds to $\Pr(T_i^*\geq t+\delta\,|\,\boldsymbol{\xi}_i^{(train)},\boldsymbol{\zeta}_i^{(train)},\mathbf{A}_{train};\boldsymbol{\theta}^{(train)})/\Pr(T_i^*\geq t\,|\,\boldsymbol{\xi}_i^{(train)},\boldsymbol{\zeta}_i^{(train)},\mathbf{A}_{train};\boldsymbol{\theta}^{(train)})$, where $\boldsymbol{\theta}^{(train)}$ is the posterior estimate of θ based on $\mathbf{A}_{train}$ and $(\boldsymbol{\xi}_i^{(train)},\boldsymbol{\zeta}_i^{(train)})\sim p(\boldsymbol{\xi}_i,\boldsymbol{\zeta}_i\,|\,T_i^*>t,\mathbf{A}_{train};\boldsymbol{\theta}^{(train)})$. Figure 3 displays the boxplots of AUC values based on the 100 random splits for the four models. Overall, M_p consistently outperform all the three competing models for all the combinations of t and δ. Specifically, M_p, M_{r1}, and M_{r2} that incorporate imaging covariates or ventricular volume are much superior to M_{r3} that considers baseline covariates only. M_p and M_{r2} has noticeably higher AUC values than M_{r1}, thereby indicating that MRI data, in comparison with ventricular volume, are a more sensitive measure to discriminate between subjects who remained at MCI and those who progressed to AD. Furthermore, M_p that incorporates the longitudinal images is

superior to M_{r2} that includes the baseline images only. This result confirms that dynamic MRI images can provide more accurate prediction of MCI-to-AD progression compared with cross-sectional MRI images. Such an enhanced diagnostic model is crucial in evaluating drug therapy in AD clinical trials and the prodromal stages of dementia in subjects with MCI.

Finally, we conducted additional analyses to check the performance of the tail regions of the imaging parameters, the comparison of alternative models with other longitudinal or baseline variables, and the plausibility of model assumptions. The details are provided in Appendix 5 of Supplementary Material.

6 Discussion

The present study can be extended in several directions. First, we use voxel-based morphometry to extract useful information from the MRI data. Recent studies (e.g., Wang et al. 2011) argued that surface-based morphometry may exhibit some advantages over voxel measures. However, the surface morphometry statistics obtained by Wang at al. (2011) mainly provide the morphometric information and cannot identify brain regions of interest. Developing new interpretable surface morphometry statistics and investigating its utility in detecting AD progression are of great interest in the future research. Second, the proposed model assumes that the imaging parameters $\beta_0(\cdot)$ and $\beta_1(\cdot)$ can be fully represented by the eigenfunctions of $X_{i0}(\cdot)$ and $X_{i1}(\cdot)$, respectively. While this assumption may restrict the flexibility of $\beta(v)$s and may not hold in practice, it can dramatically simplify the analysis and has thus been used in a various imaging data analyses (e.g., Kong et al., 2018; Li et al., 2018; Feng et al., 2020). Moreover, Wang et al. (2021) demonstrated, based on simulation studies, that the estimation obtained under this assumption could give a good recovery of the reality even though the assumption is violated. Recent works (Wang and Zhu, 2017; Kang et al., 2018) developed a functional penalization approach to estimate $\beta_0(v)$ and $\beta_1(v)$ directly. However, the

feasibility of adapting their approach to the current complex model framework requires further investigation. Third, the proposed joint modeling framework can be modified to handle informative dropout without difficulty. Similar to the existing literature (Wu et al. 2012), we can regard dropout as another event and jointly model time to AD and time to dropout. Shared or correlated random effects can be introduced to address possible correlation between the two time-to-event outcomes. Another direction is to assume a joint distribution for the time-to-AD and dropout processes (Han et al. 2014). Lastly, we unfold the 3D images and treat them as one-dimensional functions. This may destroy the underlying spatial structure of the image. Extending the existing model to treat MRI data as 3D vectors is a promising way to keep complete spatial structure of images. For example, tensor regression/partition techniques (Zhu et al., 2011; Zhou et al., 2013; Miranda et al., 2018) can be considered to manage MRI data in the form of multidimensional array. Incorporating these methods into the framework of longitudinal imaging and survival data remains to be addressed in the future research.

Supplementary Material

In the Supplementary Material, Appendix 1 describes the preprocessing of MRI data. Appendix 2 provides the likelihood function in (10). Appendix 3 provides the proof of Theorem 1 in Section 3. Appendices 4 and 5 provides additional numerical results in the simulation and ADNI study, respectively.

Acknowledgment

This research was supported by GRF Grants (14301918, 14302220) from the Research Grant Council of the Hong Kong Special Administration Region. The authors are thankful to the editor, the associate editor, and two anonymous reviewers for their valuable comments and suggestions, which have helped improve the article substantially.

References

Anderson, K., Anderson, B. B., and Pakkenberg, B. (2012). Stereological Quantification of the Cerebellum in Patients with Alzheimer's Disease. *Neurobiology of Aging* **33,** 197.e11–197.e20.

Andrinopoulou, E., Eilers, P. H., Takkenberg, J. J., and Rizopoulos, D. (2018). Improved Dynamic Prediction from Joint Models of Longitudinal and Survival Data with Time-Varying Effects Using P-Splines. *Biometrics* **74,** 685–693.

Brown, E. R. and Ibrahim, J. G. (2003). A Bayesian Semiparametric Joint Hierarchical Model for Longitudinal and Survival data. *Biometrics* **59,** 221-228.

Canu, E., McLaren D. G., Fitzgerald, M. E., Bendlin, B. B., Zoccatelli, G., alessandrini, F., Pizzini, F. B., Ricciardi, G. K., Beltramello, A., Johnson, S. C., and Frisoni, G. B. (2001). Mapping the Structural Brain Changes in Alzheimer's Disease: the Independent Contribution of Two Imaging Modalities. *Journal of Alzheimer's Disease* **26,** 263–274.

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

DeGruttola, V. and Tu, X. (1994). Modelling Progression of CD4-Lymphocyte Count and Its Relationship to Survival Time. *Biometrics* **50,** 1003–1014.

Feng, X. N., Li, T. F., Song, X. Y., and Zhu, H. T. (2020). Bayesian Scalar on Image Regression With Nonignorable Nonresponse. *Journal of American Statistical Association* **115,** 1574–1597.

Greven, S., Crainiceanu, C., Caffo, B., and Reich, D. (2010). Longitudinal Functional Principal Component Analysis. *Electronic Journal of Statistics* **4,** 1022–1054.

Han, M., Song, X. Y., Sun, L. Q., and Liu, L. (2014). Joint Modeling of Longitudinal Data with Informative Observation Times and Dropouts. *Statistica Sinica* **24,** 1487–1504.

Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrics* **57**, 97–109.

Henderson, R., Diggle, P., and Dobson, A. (2000). Joint Modeling of Longitudinal Measurements and Event Time Data. *Biostatistics* **1,** 465–480.

Heagerty, P. J. and Zheng, Y. Y. (2005). Survival Model Predictive Accuracy and ROC Curves. *Biometrics* **61,** 92–105.

Kang, J., Reich, B. J., and Staicu, A. M. (2018). Scalar-on-Image Regression via the Soft-thresholded Gaussian Process. *Biometrika* **105,** 165–184.

Kong, D., Giovanello, K. S., Wang, Y., Lin, W., Lee, E., Fan, Y., Murali Doraiswamy, P., and Zhu, H. (2015). Predicting Alzheimer's Disease Using Combined Imaging-whole Genome SNP Data. *Journal of Alzheimer's Disease* **46,** 695–702.

Kong, D., Ibrahim, J. G., Lee, E., and Zhu, H. (2018). FLCRM: Functional Linear Cox Regression Model. *Biometrics* **74,** 109–117.

Karhunen, K. (2017). Über lineare Methoden in der Wahrscheinlichkeitsrechnung. *Annales Academie Scientiarum Fennicae* **37,** 1–79.

Lee, E. J., Zhu, H. T., Kong, D. H., Wang, Y. L., Giovanello, K. S., and Ibrahim, J. G. (2015). A Bayesian Functional Linear Cox Regression Model for Predicting Time to Conversion to Alzheimer's Disease. *The Annals of Applied Statistics* **9,** 2153–2178.

Li, K. and Luo, S. (2019). Bayesian Functional Joint Models for Multivariate Longitudinal and Time-to-event Data. *Computational Statistics & Data Analysis* **129,** 14–29.

Li, T., Xie, F., Feng, X., Ibrahim, J. G., and Zhu, H. (2018). Functional Linear Regression Models for Nonignorable Missing Scalar Responses. *Statistica Sinica* **28,** 1867–1886.

Li, Y. H., Wang, N. S., and Caroll, R. J. (2013). Selecting the Number of Principle Components in Functional Data. *Journal of American Statistical Association* **108,** 1284–1294

Lin, H., Turnbull, B. W., McCulloch, C. E., and Slate, E. H. (2002). Latent Class Models for Joint Analysis of Longitudinal Biomarker and Event Process Data: Application to Longitudinal Prostate-specific Antigen Readings and Prostate Cancer. *Journal of American Statistical Association* **97,** 53–65.

Loève, M. (1978). *Probability Theory II*, 4th ed. Springer, New York.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equations of State Calculations by Fast Computing Machine. *The Journal of Chemical Physics* **21,** 1087-1091.

Miranda, M. F., Zhu, H. T., and Ibrahim, J. G. (2018). Tensor Partition Regression Models with Application in Imaging Biomarker Detection. *The Annals of Applied Statistics* **12,** 1422–1450.

Müller, H. G. and Stadtmüller, U. (2005). Generalized Functional Linear Models. *The Annals of Statistics* **33,** 774–805.

Nestor, S. M., Rupsingh, R., Borrie, M., et al. (2008). Ventricular Enlargement as a Possible Measure of Alzheimer's Disease Progression Validated Using the Alzheimer's Disease Neuroimaging Initiative Database. *Brain* **131,** 2443–2454.

Okuizumi, K., Onodera, O., Tanaka, H., et al. (1994). ApoE-ϵ4 and Early-onset Alzheimer's. *Nature Genetics* **7,** 10–11.

Prentice, R. (1982). Covariate Measurement Errors and Parameter Estimation in a Failure Time Regression Model. *Biometrika* **69,** 331–342.

Rizopoulos, D. (2011). Dynamic Prediction and Prospective Accuracy in Joint Models for Longitudinal and Time-to-Event Data. *Biometrics* **67,** 819–829.

Tsiatis, A. A. and Davidian, M. (2004). Joint Modeling of Longitudinal and Time-to-event Data: an Overview. *Statistica Sinica* **14,** 809–834.

Wang, X. and Zhu, H. T. (2017). Generalized Scalar-on-Image Regression Models via Total Variation. *Journal of the American Statistical Association* **112,** 1156–1168.

Wang, X., Song, X., and Zhu, H.. (2021). Bayesian Latent Factor on Image Regression with Nonignorable Missing Data. *Statistics in Medicine*, 40:920–932.

Wang, Y., Ibrahim, J. G., and Zhu, H. T. (2020). Partial Least Squares for Functional Joint Models with Applications to the Alzheimer's Disease Neuroimaging Initiative Study. *Biometrics* **76,** 1109–1119.

Wang, Y. and 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,** 895–905.

Wu, L., Liu, W., Yi, G. Y., and Huang, Y. (2012) Analysis of Longitudinal and Survival Data: Joint Modeling, Inference Methods, and Issues. *Journal of Probability and Statistics*, 2012.

Zhou, H., Li, L. X., and Zhu, H. T.(2013). Tensor Regression with Applications in Neuroimaging Data Analysis. *Journal of the American Statistical Association* **108,** 540–552.

Zipunnikov, V., Greven, S., Shou, H., Caffo, B. S., Reich, D. S., and Crainiceanu, C. M. (2014). Longitudinal High-dimensional Principal Component Analysis with Application to Diffusion Tensor Imaging of Multiple Sclerosis. *The Annals of Applied Statistics* **8,** 2175–2202.

Zhu, H., Kong, L., Li, R., Styner, M., Grieg, G., Lin, W. and Gilmore, J. (2014). FADTTS: Functional Analysis of Diffusion Tensor Tract Statistics. *Neuroimage* **56,** 1412–1425.

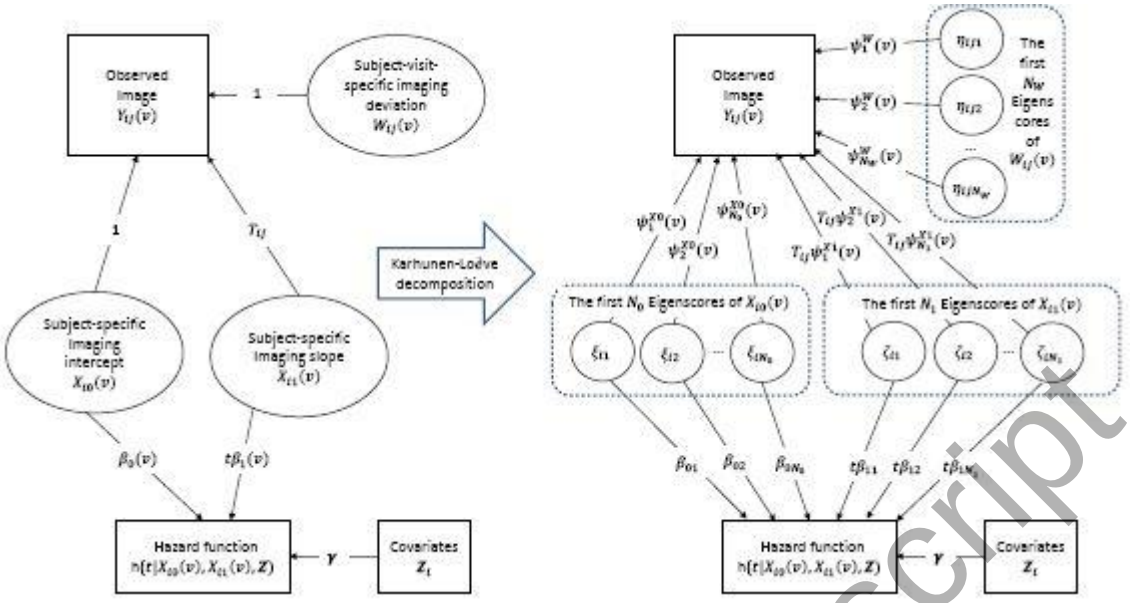

Fig. 1 Path diagram of the proposed model, where the left part is the original model defined in (1)–(2), and the right part is the model approximated by KL decomposition defined in (4)–(5). The fixed image $\mu(v)$ is omitted in path diagram for presentation simplicity.

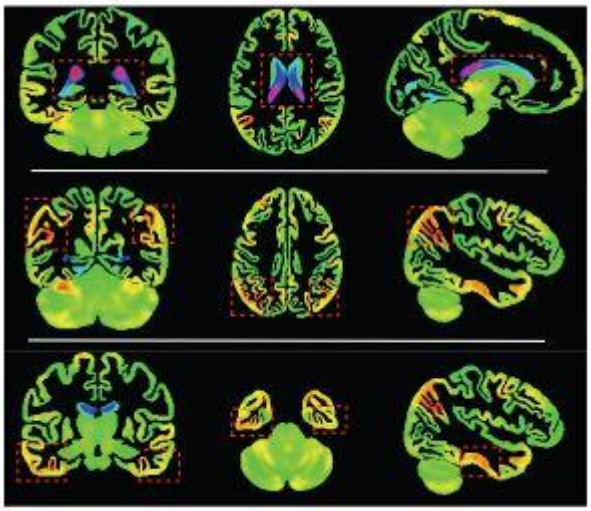

Fig. 2 The coronal, axial, and sagittal planes of the estimated imaging coefficient $\beta_0(\cdot)$, where the upper row is the positive part that is in the brain regions of “lateral ventricle”, and the lower two rows are the negative parts that are in the brain regions of “inferior parietal lobe” and “inferior temporal gyrus and fusiform gyrus”, respectively. All the related regions are marked using red dashed squares. Purple and dark blue represent positive effects; orange and red represent negative effects; and green represents effects near to zero.

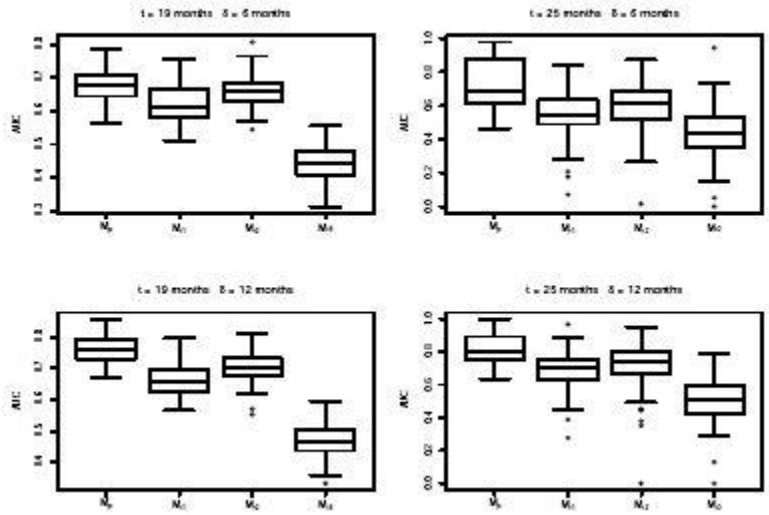

Fig. 3 Boxplots of AUC under M_p, M_{r1}, M_{r2}, and M_{r3} at $t=\{19,25\}$ months and $\delta=\{6,12\}$ months in the ADNI study.

Table 1 Estimation of parameters in the Cox model in Simulation 1

			$h_0(t)=1$			$h_0(t)=t+0.5$			$h_0(t)=t^2+0.3$		
/	CR	Para	BIAS	RMS	SE	BIAS	RMS	SE	BIAS	RMS	SE
300	30%	β_{01}	–0.009	0.124	0.102	0.003	0.137	0.099	0.012	0.140	0.102
		β_{02}	0.020	0.159	0.133	–0.020	0.154	0.131	0.019	0.161	0.136
		β_{11}	–0.030	0.180	0.137	–0.019	0.179	0.146	0.003	0.151	0.129
		β_{12}	0.003	0.190	0.183	–0.020	0.167	0.198	–0.008	0.173	0.169
		γ_1	–0.042	0.111	0.094	–0.052	0.101	0.092	–0.050	0.101	0.094

			$h_0(t)=1$			$h_0(t)=t+0.5$			$h_0(t)=t^2+0.3$		
		γ_2	0.031	0.110	0.078	0.059	0.103	0.076	0.051	0.096	0.080
	50%	β_{01}	0.015	0.134	0.111	0.017	0.144	0.113	0.004	0.129	0.114
		β_{02}	0.022	0.175	0.148	–0.011	0.175	0.150	0.035	0.164	0.156
		β_{11}	–0.027	0.342	0.316	0.030	0.255	0.288	0.021	0.217	0.198
		β_{12}	0.064	0.338	0.431	–0.005	0.278	0.382	–0.024	0.226	0.268
		γ_1	–0.041	0.112	0.102	–0.055	0.113	0.103	–0.079	0.135	0.106
		γ_2	0.049	0.119	0.083	0.081	0.123	0.084	0.063	0.119	0.087
500	30%	β_{01}	0.005	0.112	0.077	–0.001	0.119	0.077	0.008	0.120	0.078
		β_{02}	–0.007	0.140	0.103	–0.002	0.125	0.100	–0.011	0.138	0.101
		β_{11}	–0.004	0.130	0.108	0.009	0.118	0.115	0.007	0.116	0.098
		β_{12}	–0.012	0.162	0.146	–0.024	0.137	0.154	–0.008	0.125	0.131
		γ_1	–0.031	0.074	0.072	–0.041	0.086	0.072	–0.039	0.087	0.072
		γ_2	0.027	0.100	0.059	0.050	0.085	0.060	0.042	0.089	0.060

			$h_0(t)=1$			$h_0(t)=t+0.5$			$h_0(t)=t^2+0.3$		
	50 %	β_{01}	–0.011	0.122	0.085	0.014	0.113	0.087	0.011	0.114	0.089
		β_{02}	0.030	0.121	0.113	0.011	0.153	0.114	–0.007	0.131	0.119
		β_{11}	0.024	0.272	0.257	–0.004	0.252	0.227	0.003	0.173	0.155
		β_{12}	–0.005	0.297	0.348	0.009	0.254	0.306	–0.019	0.171	0.208
		γ_1	–0.053	0.104	0.078	–0.054	0.092	0.080	–0.057	0.097	0.083
		γ_2	0.045	0.088	0.065	–0.071	0.114	0.065	0.056	0.080	0.068