


RESEARCH ARTICLE OPEN ACCESS

Supervised Functional Principal Component Analysis Under the Mixture Cure Rate Model: An Application to Alzheimer's Disease

Jiahui Feng¹ | Haolun Shi¹ | Da Ma² | Mirza Faisal Beg³ | Jiguo Cao¹ ¹Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, British Columbia, Canada | ²School of Medicine, Wake Forest University, Winston-Salem, North Carolina, USA | ³School of Engineering, Simon Fraser University, Burnaby, British Columbia, Canada**Correspondence:** Jiguo Cao (jiguo_cao@sfu.ca)**Received:** 11 January 2024 | **Revised:** 8 October 2024 | **Accepted:** 13 December 2024**Funding:** This work was supported by the Natural Sciences and Engineering Research Council of Canada under Grant RGPIN-2023-04057, RGPIN-2021-02963.**Keywords:** functional principal component analysis | image analysis | survival analysis | triangulation

ABSTRACT

Brain imaging data is one of the primary predictors for assessing the risk of Alzheimer's disease (AD). This study aims to extract image-based features associated with the possibly right-censored time-to-event outcomes and to improve predictive performance. While the functional proportional hazards model is well-studied in the literature, these studies often do not consider the existence of patients who have a very low risk and are approximately insusceptible to AD. We introduce a functional mixture cure rate model that extends the proportional hazards model by allowing a proportion of event-free patients. We propose a novel supervised functional principal component analysis (sFPCA) method to extract image features associated with AD risk while accounting for the complexity arising from right censoring. The proposed method accommodates the irregular boundary issue inherent in brain images with bivariate splines over triangulations. We demonstrate the advantages of the proposed method through extensive simulation studies and provide an application to the Alzheimer's Disease Neuroimaging Initiative (ADNI) study.

1 | Introduction

Alzheimer's disease (AD) is one of the most prevalent age-related neurodegenerative disorders worldwide. In 2018, Alzheimer's Disease International reported that over 50 million individuals globally were living with dementia, a figure anticipated to triple by 2050 [1]. In medical research, structural magnetic resonance imaging (MRI) has become a critical tool for early detection of presymptomatic AD in numerous studies [2–4]. An intriguing research focus in recent years pertains to deriving features from the imaging data that are significantly associated with AD progression and thus enhance the identification of individuals at varying risk levels of AD.

When the outcome of interest is time-to-event, such as the time to disease progression, the proportional hazards model is widely used to model the relationship between a survival outcome and some explanatory variables. These models are constructed under the assumption that all subjects under study would eventually experience the event of interest. However, with the advancements in disease diagnosis and treatment, there may exist a subset of patients who can maintain long-term stability in terms of health status. The cure model is applied with the aim of capturing such a subgroup of low-risk subjects in the study of Alzheimer's disease. Our rationale for accommodating a cured fraction in the model is based upon scientific plausibility. Contrary to the proportional hazards model assumption that all subjects would eventually have AD, recent research suggests that

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2025 The Author(s). *Statistics in Medicine* published by John Wiley & Sons Ltd.

there may exist a subgroup of approximately event-free subjects that may remain stable with a very low risk of disease progression. For instance, certain genetic variations may offer protection against Alzheimer's, suggesting that disease immunity is possible for some individuals [5]. In addition, empirically, epidemiological research (e.g., studies on centenarians) shows that some individuals (and their offspring) oftentimes have shown good resilience to AD pathology, which might be attributed to genetic, lifestyle, and environmental factors that significantly reduce the risk of Alzheimer's disease to the extent of being approximately insusceptible to it [6, 7]. These all point to the existence of a very low-risk subgroup within the population, which motivates the use of a cure model. It is worth noting that the term "cured" in this context is used to describe a subgroup of individuals who are potentially resistant to AD progression and thus have a significantly lower risk of developing AD to the extent of being approximately insusceptible, rather than a complete reversal of brain degeneration. More precisely, such a group of patients shall be referred to as a "low-risk" group, and our model targets to identify the individuals who potentially belong to such a group. The cure rate model, adapted for our purposes, can be used to highlight the different risks of progression into AD within the cohort and to explore the potential impacts of various predictors on these progression trajectories. This approach allows us to identify and model the heterogeneity in disease progression among the ADNI participants, offering insights that otherwise might be obscured in the proportional hazards model.

We consider the mixture cure rate model [8, 9] which simultaneously estimates the cure probability and the survival time for uncured patients. One main challenge in incorporating imaging data into a mixture cure rate model is the high dimensionality of the imaging data. To tackle this, functional principal component analysis (FPCA) is often used to reduce the dimension of functional data by identifying major sources of variability contained in the data. The FPCA method in this context operates in two steps: first, it estimates the FPCs for the functional imaging data, and second, it incorporates several leading FPCs as predictors in the mixture cure rate model.

Multiple studies have been conducted on estimating FPCs for imaging data. For example, Huang et al. [10] introduced a bi-regularization approach that simultaneously penalizes rows and columns of the data matrix. Zipunnikov et al. [11] used singular value decomposition to extract functional features from high-dimensional data and applied the method to brain imaging data to identify the primary variation directions in brain volumes. However, these methods derive FPCs from the functional data alone, without considering the relationship between functional features and other available variables such as the outcome variable. Recently, some researchers considered supervised FPCA methods to enhance prediction accuracy. For instance, Li et al. [12] proposed a supervised sparse and functional principal component analysis method where additional information across the dataset is incorporated into the analysis to yield more interpretable FPCs. Zhang et al. [13] developed a supervised principal component regression method, leveraging the relationship between responses and functional predictors based on the integrated residual sum of squares.

A specific challenge in brain imaging data analysis is the irregularly shaped boundaries of brain images. Conventional smoothing methods suffer from the problem of boundary leakage and hence have poor performance in terms of prediction when used to smooth data over complex domains [14]. One method to handle images with complex boundaries and/or interior holes is partitioning the domain into triangular sections and applying bivariate splines over these triangulations [15, 16]. Yu et al. [17] and Wang et al. [18] implemented such bivariate splines in spatial statistical models. Additionally, Jiang et al. [19] applied sFPCA over triangulation to mammogram imaging data with complex boundaries, which addressed the complexities of irregular boundaries effectively.

As for the second step of using the estimated FPCs/sFPCs in survival models, some statistical methods have been studied in the literature. For example, Kong et al. [20] considered a functional linear proportional hazards regression model to investigate the relationship between the survival outcome and functional predictors. Lee et al. [21] developed a Bayesian functional proportional hazards model with application to the AD brain imaging data. More recent studies by Jiang et al. [19, 22] have advanced this field significantly. In 2023, they introduced an inverse-weighted sFPCA method within a proportional hazards outcome model, specifically designed for right-censored survival outcomes in rectangular-shaped mammogram imaging data [22]. Jiang et al. [19] further extended this approach to mammogram data with irregular shapes. However, a notable limitation in these studies is the assumption that all patients are susceptible to the events of interest. This assumption restricts the applicability of their methods to mixture cure rate models, where a subset of patients are considered insusceptible to the event (e.g., disease recurrence). Addressing this gap could significantly enhance the utility of FPCs and sFPCs in more complex survival models.

This paper makes several significant contributions to the field of functional cure rate modeling and its application in brain imaging data analysis. Firstly, we introduce an advanced functional cure rate model that expands upon the proportional hazards model. This new model uniquely accounts for a subset of cured patients within the study population and investigates the covariate effects on the cured proportion. Secondly, we address the irregular boundary issue that is often encountered in brain imaging data by employing bivariate splines over triangulation. Thirdly, we propose an innovative supervised FPCA over triangulation method to extract image-based features associated with the failure time. Lastly, we apply the proposed method to brain imaging data from the ADNI study, deriving new insights that could have significant implications in the field. This practical application demonstrates the utility and effectiveness of our proposed method in a real-world context. The R code for the implementation of the method, together with a sample input dataset and documentation, is available on GitHub <https://github.com/jiahfeng/sFPCAure>.

The rest of this paper is structured as follows. In Section 2, we introduce the mixture cure rate model, discuss the bivariate splines over triangulation, and propose a novel sFPCA method to estimate the supervised FPCs with time-to-event outcomes. In Section 3, we report the results from simulation studies. In

Section 4, we provide an application to the ADNI dataset. Finally, we conclude the paper with a few remarks.

2 | Methodology

2.1 | Functional Mixture Cure Rate Model

Let T denote a nonnegative time-to-event outcome. Let \mathbf{X} denote a set of demographic variables. Also, let Ω be a bounded two-dimensional domain of arbitrary shape, and $\mathbf{s} = (s_1, s_2)$ represent a particular point $\in \Omega \subset \mathbb{R}^2$. The imaging data can be represented as $\{Z(\mathbf{s}), \mathbf{s} \in \Omega\}$. The mixture cure rate model assumes that a fraction of subjects in the study are potentially cured. Under the mixture cure rate model, the survival function of T given the covariates \mathbf{X} and imaging data \mathbf{Z} is defined as

$$S(t|\mathbf{X}, \mathbf{Z}) = 1 - \pi(\mathbf{X}, \mathbf{Z}) + \pi(\mathbf{X}, \mathbf{Z})S_u(t|\mathbf{X}, \mathbf{Z}) \quad (1)$$

where $S_u(t|\mathbf{X}, \mathbf{Z})$ is the survival function of the uncured subjects, and $\pi(\mathbf{X}, \mathbf{Z})$ is the proportion of uncured subjects, which may depend on (\mathbf{X}, \mathbf{Z}) .

Let R denote the latent uncured status, that is, $R = 1$ if the subject is uncured and $R = 0$ if otherwise. A logistic regression model is assumed for the probability of being uncured $\pi(\mathbf{X}, \mathbf{Z}) = P(R = 1|\mathbf{X}, \mathbf{Z})$ so that

$$\pi(\mathbf{X}, \mathbf{Z}) = \frac{\exp\{\boldsymbol{\alpha}^T \mathbf{X} + \int_{\Omega} \eta(\mathbf{s})Z(\mathbf{s}) d\mathbf{s}\}}{1 + \exp\{\boldsymbol{\alpha}^T \mathbf{X} + \int_{\Omega} \eta(\mathbf{s})Z(\mathbf{s}) d\mathbf{s}\}} \quad (2)$$

where $\boldsymbol{\alpha}$ is a vector of regression parameters, and $\eta(\cdot)$ is a coefficient function for the imaging data. We propose to model the survival function of the uncured group by the proportional hazards assumption, which is in the form of

$$S_u(t|\mathbf{X}, \mathbf{Z}) = S_0(t) \exp\{\boldsymbol{\beta}^T \mathbf{X} + \int_{\Omega} \zeta(\mathbf{s})Z(\mathbf{s}) d\mathbf{s}\} \quad (3)$$

where $S_0(t) = \exp\{-\int_0^t h_0(u)du\}$ with $h_0(t)$ being some unknown baseline hazard function, $\boldsymbol{\beta}$ is a vector of regression parameters, and $\zeta(\cdot)$ is a coefficient function. Note that the demographic variables in the logistic regression model and the survival model are not necessarily the same; here we use \mathbf{X} for both models to simplify the notation.

Suppose that T is possibly right-censored at C , which is assumed to be independent with T given \mathbf{X} and \mathbf{Z} . Let $Y = \min(T, C)$ and $\delta = I(T \leq C)$, where $I(\cdot)$ is the indicator function. The observed data from a random sample of n subjects consist of $D = (Y_i, \delta_i, \mathbf{X}_i, \mathbf{Z}_i)_{i=1}^n$.

Note that the observed imaging data $Z_i(\mathbf{s})$ is actually a realization of the stochastic process $\{Z(\mathbf{s}), \mathbf{s} \in \Omega\}$. By the Karhunen-Loève expansion [23], provided that \mathbf{Z} is a square-integrable process in space with a continuous covariance function, this stochastic process can be written as

$$Z(\mathbf{s}) = \mu(\mathbf{s}) + \sum_{k=1}^{\infty} \xi_k \phi_k(\mathbf{s})$$

where $\mu(\mathbf{s})$ is the mean function, $\phi_k(\mathbf{s})$ is the k th basis function, and $\xi_k = \langle \mathbf{Z} - \mu, \phi_k \rangle = \int_{\Omega} \{Z(\mathbf{s}) - \mu(\mathbf{s})\} \phi_k(\mathbf{s}) d\mathbf{s}$ is the

functional score between \mathbf{Z} and ϕ_k . The scores ξ_k are assumed to be uncorrelated with mean zero and variance σ_k^2 . Without loss of generality, we assume that $\mu(\mathbf{s}) = 0$. Typically, we can approximate $Z(\mathbf{s})$ using a relatively small number, say K , of the leading basis functions that correspond to the largest variances of their respective scores. However, the selected basis functions, while reflecting the variation in the functional data, may not be associated with the time-to-event outcome Y .

To address this concern, we will first consider the estimation of $Z(\mathbf{s})$ using bivariate splines in Section 2.2. These splines are constructed as piecewise polynomial functions over a two-dimensional triangulated domain, which accommodates the semicircular-shaped domain of brain images. Then in Section 2.3, we develop the sFPCA method for the imaging data approximations over triangulations. The extracted functional features will be ordered by the degree of association with both the time-to-event outcome and the cure status. We illustrate how the sFPCA method can be effectively integrated into survival analysis frameworks.

2.2 | Bivariate Splines Over Triangulation

We consider using bivariate splines [16, 18] that are piecewise polynomial functions over a two-dimensional triangulated domain to approximate the functional data \mathbf{Z} . This approach can effectively handle imaging data with complex boundaries. Specifically, we define a triangle ν as a convex hull of three non-collinear points. A triangulation of a domain Ω is then formed by a collection of triangles $\Delta = \{\nu_1, \dots, \nu_N\}$. This triangulation $\Omega = \bigcup_{i=1}^N \nu_i$ must satisfy the condition that the intersection of any two triangles in Δ is restricted to either a shared vertex or a common edge. For a triangle $\nu \in \Delta$, let $\mathbf{x}_1, \mathbf{x}_2$, and \mathbf{x}_3 be the vertices. Then, for any points $\mathbf{x} \in \mathbb{R}^2$, we can write it in the form of $\mathbf{x} = b_1 \mathbf{x}_1 + b_2 \mathbf{x}_2 + b_3 \mathbf{x}_3$, where b_1, b_2 , and b_3 are the barycentric coordinates of the point \mathbf{x} relative to ν . Bernstein polynomials of degree d associated with the triangle ν are then defined as $B_{ijk}^{v,d} = (d!/i!j!k!)b_1^i b_2^j b_3^k$, with $i + j + k = d$. For a non-negative integer r , let $C^r(\Omega)$ be the set of all functions that are r -times continuously differentiable over Ω . Within a given triangulation Δ , we define the spline space of degree d and smoothness r as $\mathbb{H}_d^r(\Delta) = \{h \in C^r(\Omega) : h|_{\nu} \in \mathbb{P}_d, \nu \in \Delta\}$, where $h|_{\nu}$ is the polynomial piece of spline h on triangle ν , and \mathbb{P}_d is the space of all polynomials with a degree less than or equal to d . Subsequently, for any triangle $\nu \in \Delta$, the polynomial piece of a spline h restricted on ν can be expressed as $h|_{\nu} = \sum_{i+j+k=d} \gamma_{ijk}^{\nu} B_{ijk}^{v,d}$, where $\gamma_{\nu} = \{\gamma_{ijk}^{\nu} : i + j + k = d\}$ is the set of B-coefficients of h on ν .

Let $\mathbf{B}(\mathbf{s}) = (B_1(\mathbf{s}), \dots, B_K(\mathbf{s}))^T, \mathbf{s} \in \Omega$ denote a set of degree d bivariate Bernstein basis polynomials for $\mathbb{H}_d^r(\Delta)$, where K is the number of Bernstein basis polynomials. Then, for any function $h \in \mathbb{H}_d^r(\Delta)$, we can rewrite it in the form of basis expansion:

$$h(\mathbf{s}) = \boldsymbol{\gamma}^T \mathbf{B}(\mathbf{s})$$

where $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_K)^T$ is the vector of spline coefficients. Note that $\boldsymbol{\gamma}$ should satisfy the constraint $\mathbf{H}\boldsymbol{\gamma} = 0$ to ensure that the smoothness conditions are met. In practice, we often define a finite grid of pixels for the measurement of imaging data. Let $\{s_j \in \Omega, j = 1, \dots, J\}$ denote the set of pixel points. Given the

dataset $\{s_j, Z_i(s_j), i = 1, \dots, n, j = 1, \dots, J\}$, the values of γ can be estimated by solving the following minimization problem:

$$\min_{h_1, \dots, h_n} \sum_{i=1}^n \sum_{j=1}^J \{Z_i(s_j) - h_i(s_j)\}^2 + \lambda_i \mathcal{E}(h_i) \quad (4)$$

where $\lambda_1, \dots, \lambda_n$ are nonnegative roughness penalty parameters, and

$$\mathcal{E}(h_i) = \sum_{v \in \Delta} \int_v \sum_{p+q=2} \binom{2}{p} (D_{s_1}^p D_{s_2}^q h_i)^2 ds_1 ds_2$$

where $D_{s_1}^q h(s)$ is the q th order of derivative of h along the direction of s_1 with $s = (s_1, s_2)$.

2.3 | Supervised FPCA With Mixture Cure Rate Model

Using bivariate splines over triangulation, we can approximate $Z_i(s)$ with $\gamma_i^T \mathbf{B}(s)$ by solving (4). However, this spline-based method does not take the relationship between the basis functions and the time-to-event outcome into consideration. To address this problem, we consider the method of sFPCA, which aims to identify a set of orthonormal basis functions $\phi = (\phi_1, \phi_2, \dots)$, with $\|\phi\| = 1$ and $\langle \phi_k, \phi_{k'} \rangle = 0$ for $k < k'$, to maximize the following objective function

$$Q(\phi) = \frac{\theta \text{var}(\langle \mathbf{Z}, \phi \rangle) + (1 - \theta) \text{cov}^2\{\log(Y), \langle \mathbf{Z}, \phi \rangle\}}{\|\phi\|^2} \quad (5)$$

for $0 < \theta \leq 1$. The second term in the numerator represents the association between the projection of \mathbf{Z} onto the basis functions and the outcome Y . When $\theta = 1$, the above problem reduces to the standard FPCA approach. The optimal value of θ can be determined through cross-validation, which is conducted over a finite grid of candidate values. The selection criterion for θ is to maximize the prediction accuracy of the model.

When the outcome variable is possibly right-censored, Jiang et al. [22] proposed a method that adjusts the covariance between observed survival times and functional scores by applying inverse probability of censoring weights (IPCW) to account for right censoring. Specifically, for the i th subject, the IPCW can be written as

$$\omega_i = \frac{\delta_i}{\hat{G}\{\log(Y_i)\}}$$

where $\hat{G}(t) = P\{(\log(C) > t)\}$ is the Kaplan–Meier estimate [24] of the survival function for censoring times. Let $\omega = (\omega_1, \dots, \omega_n)^T$ denote the vector of IPCW. To estimate the covariance term, we first compute the weighted average of log-transformed survival time, denoted by $\bar{Y} = \frac{1}{n} \sum_{i=1}^n \omega_i \log(Y_i)$. The covariance between the log-transformed survival times and the projection of the imaging data onto the basis functions can then be estimated by

$$\text{cov}\{\log(Y), \langle \mathbf{Z}, \phi \rangle\} = \frac{1}{n} \sum_{i=1}^n \omega_i \langle Z_i, \phi \rangle \{\log(Y_i) - \bar{Y}\}$$

Next, we introduce an eigenvalue decomposition method to optimize the objective function (5) in solving for ϕ . Suppose that

each basis function $\phi_k(s)$ can be expressed as a linear combination of the bivariate Bernstein polynomials, specifically $b_k^T \mathbf{B}(s)$, where $b_k = (b_{k,1}, \dots, b_{k,K})^T$. The empirical score for the i th subject can be written as $\langle \mathbf{Z}_i, \phi \rangle = \mathbf{b}^T \mathbf{M} \gamma_i$, where $\mathbf{b} = (b_1, \dots, b_K)^T$ and \mathbf{M} is a $K \times K$ -matrix with $\mathbf{M}(k, k') = \langle \mathbf{B}_k, \mathbf{B}_{k'} \rangle$. Then we can estimate the variance of score by $\frac{1}{n} \mathbf{b}^T \mathbf{M} \gamma^T \gamma \mathbf{M} \mathbf{b}$ with $\gamma = (\gamma_1, \dots, \gamma_n)^T$. Similarly, the covariance term can be estimated by $\frac{1}{n} \mathbf{b}^T \mathbf{M} \gamma (\tilde{\mathbf{Y}} \circ \omega)$, where $\tilde{\mathbf{Y}} = (\log(Y_1) - \bar{Y}, \dots, \log(Y_n) - \bar{Y})^T$ and $\mathbf{x} \circ \mathbf{y}$ is the element-wise multiplication between vectors \mathbf{x} and \mathbf{y} .

With the above estimates, the objective function (5) becomes

$$Q(\phi) = \frac{\mathbf{b}^T \mathbf{U} \mathbf{b}}{\mathbf{b}^T \mathbf{M} \mathbf{b}} \quad (6)$$

where $\mathbf{U} = \frac{\theta}{n} \mathbf{M} \gamma^T \gamma \mathbf{M} + \frac{1-\theta}{n^2} \mathbf{M} \gamma^T (\tilde{\mathbf{Y}} \circ \omega) (\tilde{\mathbf{Y}} \circ \omega)^T \gamma \mathbf{M}^T$. The maximization of the function (6) is conducted through eigenvalue decomposition. Let $\mathbf{a} = \mathbf{M}^{1/2} \mathbf{b}$, and we maximize $\mathbf{a}^T (\mathbf{M}^{-1/2})^T \mathbf{U} \mathbf{M}^{-1/2} \mathbf{a}$ subject to the constraint $\mathbf{a}^T \mathbf{a} = \mathbf{I}$. We can estimate $\mathbf{a}_1, \dots, \mathbf{a}_{K^*}$ by finding the leading K^* eigenvectors of the matrix $(\mathbf{M}^{-1/2})^T \mathbf{U} \mathbf{M}^{-1/2}$. Then we can estimate $\hat{\mathbf{b}}_k = \mathbf{M}^{-1/2} \mathbf{a}_k$, and consequently $\hat{\phi}_k(s) = \hat{\mathbf{b}}_k^T \mathbf{B}(s)$ for $s \in \Omega$ and $k = 1, \dots, K^*$. To determine an appropriate value of K^* , we compare the predictive performance of models utilizing different numbers of basis functions. The objective is to identify an optimal balance between model complexity and predictive performance.

Under the framework of sFPCA, the functional score for the i th imaging data can be obtained by the inner product $\hat{\xi}_{i,k} = \langle \hat{\mathbf{Z}}_i, \hat{\phi}_k \rangle$, $k = 1, \dots, K^*$. Since $Z_i(s)$ can be approximated by $\hat{\gamma}_i^T \mathbf{B}(s)$ for $s \in \Omega$, the inner product can be estimated by $\sum_{j=1}^J \hat{\gamma}_i^T \mathbf{B}(s_j) \hat{\phi}_k(s_j)$ for large enough J . With the estimated scores $\hat{\xi}_i = (\hat{\xi}_{i,1}, \dots, \hat{\xi}_{i,K^*})^T$, the probability of being uncured and the survival function of the uncured group for subject i , $i = 1, \dots, n$ at time t can be written as

$$\pi(\mathbf{X}_i, \mathbf{Z}_i) = \frac{\exp(\alpha^T \mathbf{X}_i + \boldsymbol{\eta}^T \hat{\xi}_i)}{1 + \exp(\alpha^T \mathbf{X}_i + \boldsymbol{\eta}^T \hat{\xi}_i)},$$

$$S_u(t | \mathbf{X}_i, \mathbf{Z}_i) = S_0(t)^{\exp(\beta^T \mathbf{X}_i + \boldsymbol{\zeta}^T \hat{\xi}_i)}$$

where $\boldsymbol{\eta} = (\eta_1, \dots, \eta_{K^*})^T$ and $\boldsymbol{\zeta} = (\zeta_1, \dots, \zeta_{K^*})^T$ are the vectors of regression coefficients for the logistic regression model and survival model, respectively. The parameters of interest $(\alpha, \boldsymbol{\eta}, \beta, \boldsymbol{\zeta})$ can be estimated by the expectation-maximization algorithm [25] based on the observed data and estimated scores. Under such a cure rate model, standard errors for the estimated parameters can be obtained through random bootstrap sampling. The interval estimate of the cure probability of a certain subject can be constructed by plugging in the bootstrapped samples of the estimated parameters into the logistic function in (2). It is worth noting, however, due to the fact that the sFPC scores are intrinsically correlated with the outcomes, we do not have a “true” coefficient value for their estimated regression coefficient to compare against, except for cases where $\theta = 0$. In the [Supporting Information](#), we discuss the details of constructing the confidence interval and conduct a numerical study to verify the coverage probabilities of the interval estimate obtained via bootstrap for the case where $\theta = 0$.

Under the proposed method, the same set of survival data is used to estimate the sFPC and to subsequently fit the mixture cure rate

model with the estimated sFPC scores. This is a standard way of performing sFPCA and model estimation in the literature; see, for example, the methodologies described in Jiang et al. [19, 22] which involve using sFPC scores under a proportional hazards model. Such a double use of data is in a similar vein to the idea of partial least square regression, where the latent components that capture the covariance in the response and the dependent variables are extracted first and then a regression model is built using these components as predictors to predict the response variable. Furthermore, to avoid over-optimism, we utilize a 5-fold nested cross-validation procedure. Within each of the primary folds, a secondary level of 5-fold cross-validation is performed specifically for selecting the tuning parameter θ with fine increments on the grid $(0, 1]$, and the primary cross-validation is used to objectively evaluate the model performance in an independent held-out sample.

3 | Simulation Studies

We conducted a large-scale simulation study to investigate the finite-sample performance of the proposed method. We simulated $K = 3$ two-dimensional basis functions, denoted as $\boldsymbol{\psi} = (\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \boldsymbol{\psi}_3)^T$, by tensor products of orthogonal Legendre polynomials basis functions over the domain $[0, 1] \times [0, 1]$ under the resolution of 40×40 . Approximately 60% of the pixels were located within the semicircular-shaped domain Ω . The plots of all three basis functions are presented in Figure 1. We further simulated the individual-specific scores $\boldsymbol{\lambda}_i = (\lambda_{i,1}, \lambda_{i,2}, \lambda_{i,3})^T$ for each subject i , with a mean vector of $\mathbf{0}$ and a covariance matrix given by $\text{diag}(10, 8, 4)$. Given the orthogonal basis functions and scores, the individual-specific images were generated from the following model:

$$Z_i(s) = \mu(s) + \sum_{k=1}^3 \lambda_{i,k} \psi_k(s), s \in \Omega$$

where $\mu(s) = 0$ without loss of generality.

We modeled the probability of being uncured with the following logistic regression model:

$$\pi(Z_i) = \frac{\exp \left[Z_0 + \int_{s \in \Omega} c(s) \{ Z_i(s) - \mu(s) \} ds \right]}{1 + \exp \left[Z_0 + \int_{s \in \Omega} c(s) \{ Z_i(s) - \mu(s) \} ds \right]}$$

We further considered a proportional hazards model for the uncured group such that the hazard function is

$$h_i(t) = h_0(t) \exp \left[\int_{s \in \Omega} c(s) \{ Z_i(s) - \mu(s) \} ds \right]$$

We considered two settings for the coefficient function $c(s)$. In Setting 1, we set $c(s) = 10\psi_3(s)$, such that the probability of being uncured and the hazard over time for the uncured group only depend on the third basis function. In Setting 2, we set $c(s) = \psi_1(s) + 2\psi_2(s) + 8\psi_3(s)$. This made the uncured probability and hazard rate associated with a linear combination of all three basis functions. For the baseline hazard function, we assumed a Weibull distribution $h_0(t) = \kappa \rho (\rho t)^{\kappa-1}$ with $\kappa = 2$ and $\rho = 0.16$. The censoring times were generated by a uniform distribution between 0 and C . We considered two sets of censoring and cure rates. The first scenario involves a censoring rate of approximately 45% coupled with a cure rate of 20%, which implies that around 25% of uncured patients were censored. The second scenario is a censoring rate of 25% with a cure rate of 10%, indicating that about 15% of uncured patients were censored. The above settings can be achieved by adjusting the values of C and Z_0 .

For each setting, we set the sample size as 400. Among the dataset, a random split was made wherein 300 individuals were allocated for training the model and the remaining 100 were used for validation to avoid over-optimism. Within the training dataset, the tuning parameter θ in (5) was chosen by conducting a 5-fold cross-validation over a grid ranging from 0.001 to 1, with incremental steps of 0.111. Specifically, the training dataset was divided into $F = 5$ folds, denoted by d_1, \dots, d_F , each containing an approximately equal number of subjects. For each iteration of the cross-validation, we take $F - 1$ folds as the training set and the remaining one as the validation set. The predictive performance is assessed by the integrated area under the receiver operating characteristic (ROC) curve (AUC) [26].

In addition, we include the simulation results with conventional FPCA to illustrate the superiority of our proposed method. Figures 2 and 3 display the results of both the sFPCA and FPCA methods under Setting 1 with censoring rates of 45% and 25%, respectively, considering scenarios with one and two FPCs/sFPCs. We observe that sFPCA consistently achieves a

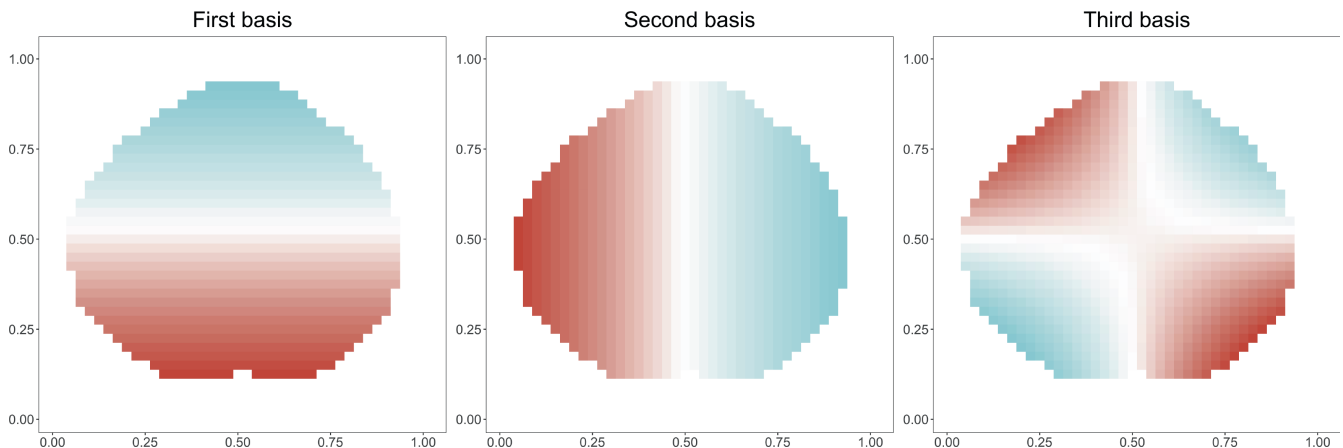


FIGURE 1 | Illustrations of the true basis functions used in the simulation study. [Colour figure can be viewed at wileyonlinelibrary.com]

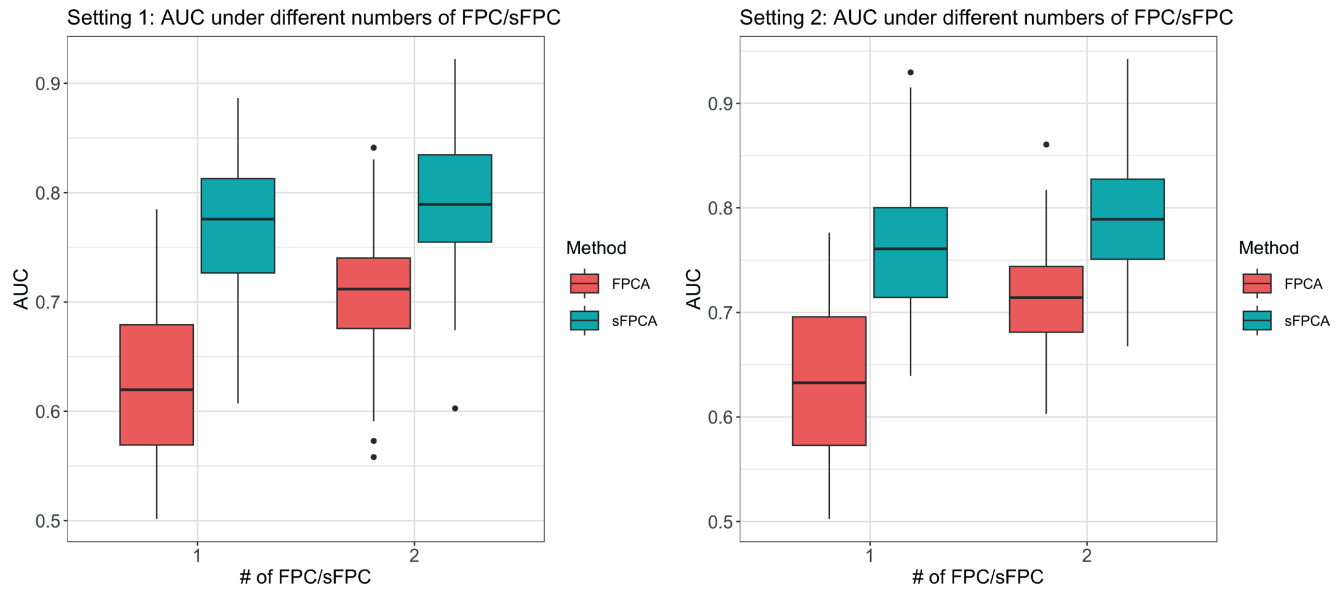


FIGURE 2 | Boxplots for the estimated AUC under FPCA and supervised FPCA with one and two FPCs/sFPCs, respectively, under a censoring rate of 45% in simulation. [Colour figure can be viewed at wileyonlinelibrary.com]

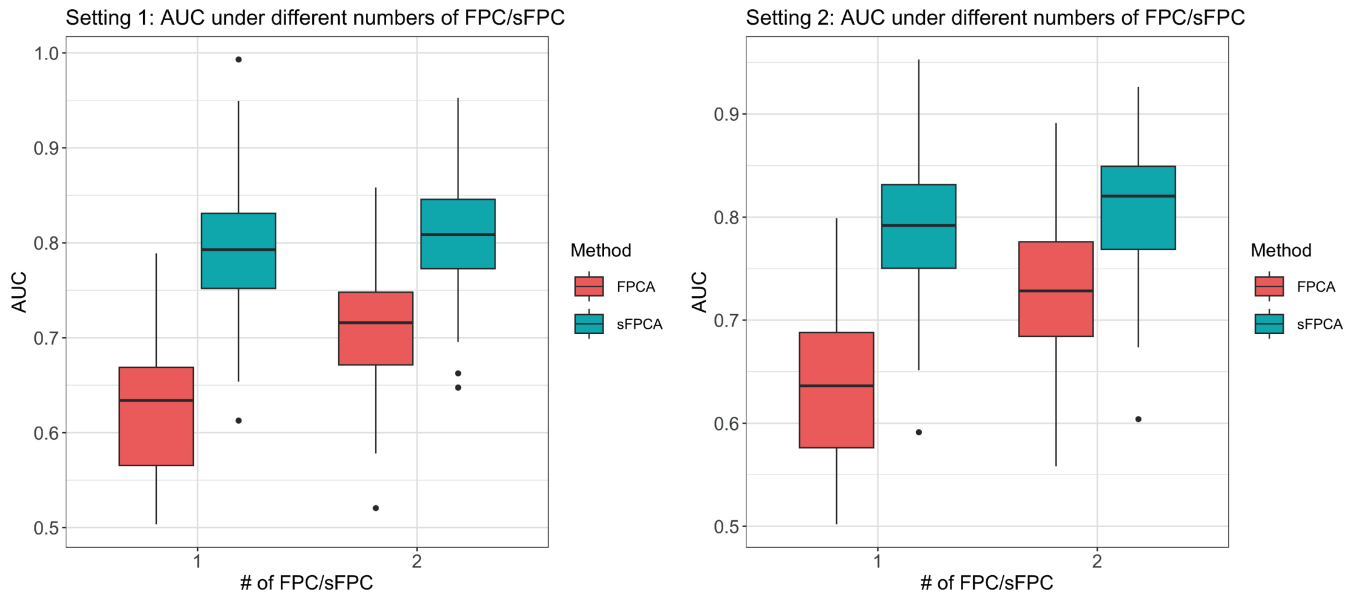


FIGURE 3 | Boxplots for the estimated AUC under FPCA and supervised FPCA with one and two FPCs/sFPCs, respectively, under a censoring rate of 25% in simulation. [Colour figure can be viewed at wileyonlinelibrary.com]

higher AUC value under the mixture cure rate model compared to conventional FPCA, regardless of the number of FPCs/sFPCs used in the model. Note that an increase in the number of FPCs utilized leads to improved performance for the conventional FPCA. This is probably because more variations contained in the imaging data are captured with more FPCs. Under Setting 2, where the coefficient function $c(s)$ is a linear combination of all three basis functions, the results are analogous to those observed in Setting 1. The proposed method demonstrates superior performance over the conventional FPCA in terms of AUC.

Figures S1 and S2 contrast the first and second sFPC' with FPCs under a censoring rate of 45%. This comparison is conducted against the true basis function ψ_3 from a randomly selected

simulation run under Setting 1. It is apparent that the first sFPC obtained from the proposed method closely aligns with the true underlying basis function, while the first FPC does not capture all the variations contained in the basis function.

4 | Alzheimer's Disease Neuroimaging Initiative Data Analysis

We applied the proposed method to the ADNI dataset [2], which contains MRI imaging data of 1724 subjects at baseline. The subjects' demographic information, including sex and age at the first scan, are available. The diagnosis of each subject is classified into three categories: cognitively normal (CN), mild cognitive

impairment (MCI), and Alzheimer's disease (AD). Each subject underwent periodic examinations and re-diagnoses around every 6 months during the first 2 years and subsequently every 12 months until the end of their follow-up periods. In this study, a subject is considered to be low-risk (cured) if the subject was diagnosed with either CN or MCI at the end of follow-up, to be an event if the subject was diagnosed with AD before the end of follow-up. Note that all low-risk subjects are right-censored. In this analysis, we focused on 1299 subjects for whom MRI imaging data were available at baseline. Individuals diagnosed with AD at the beginning of the study were excluded. Out of the 1299 subjects, 297 were diagnosed with AD before the end of the follow-up period, which is about 23% of the patients under study. The age distribution of the subjects is illustrated in the histogram shown in Figure S3, with an average age of 73.53 years. The average follow-up time is 36.65 months.

The MRI scans from the ADNI study need to undergo a series of preprocessing steps to ensure that the images are all reconstructed into the same stereotaxic space for consistency. We register all the scans affinely using the same template based on processing pipelines on FreeSurfer version 5.3.0, the details of which can be found in Ma et al. [27]. Such a pipeline helps to achieve uniform isotropic resolution, that is, the slices extracted from each scan correspond to the same anatomical brain regions and can be interpreted consistently. We select the middle slice among all the slices because it contains several important brain regions that are known to be associated with Alzheimer's disease progression, for example, the ventricle, hippocampus, and neocortex, and so forth. Such a strategic anatomical positioning allows us to associate any magnitude in values in the coefficient functions with the key brain regions for further qualitative analysis.

We aimed to use the baseline brain imaging data as functional data to predict the time to disease progression to AD. In addition to the functional data, the patient's sex and age at enrollment were included in the model. The probability of not being cured

was modeled by the following logistic regression model:

$$\pi_i = \frac{\exp \left[\alpha_0 + \alpha_1 \text{age}_i + \alpha_2 \text{sex}_i + \int_{s \in \Omega} c(s) \{Z_i(s) - \mu(s)\} ds \right]}{1 + \exp \left[\alpha_0 + \alpha_1 \text{age}_i + \alpha_2 \text{sex}_i + \int_{s \in \Omega} c(s) \{Z_i(s) - \mu(s)\} ds \right]} \\ \approx \frac{\exp \left(\alpha_0 + \alpha_1 \text{age}_i + \alpha_2 \text{sex}_i + \boldsymbol{\eta}^T \hat{\boldsymbol{\xi}}_i \right)}{1 + \exp \left(\alpha_0 + \alpha_1 \text{age}_i + \alpha_2 \text{sex}_i + \boldsymbol{\eta}^T \hat{\boldsymbol{\xi}}_i \right)}$$

where the coefficient function $c(s)$ represents the effect of brain imaging data $Z_i(s)$ on the uncured probability. Given that $Z_i(s) = \mu(s) + \sum_{k=1}^{K^*} \xi_{i,k} \phi_k(s)$, $s \in \Omega$ with $\phi_k(s)$ being the basis function over the triangulation, we can write the coefficient function in the form of $c(s) = \sum_{k=1}^{K^*} \eta_k \phi_k(s)$, where $\boldsymbol{\eta} = (\eta_1, \dots, \eta_{K^*})^T$ is the vector of coefficients corresponding to the first K^* supervised scores. Similarly, the proportional hazards model for the uncured group is

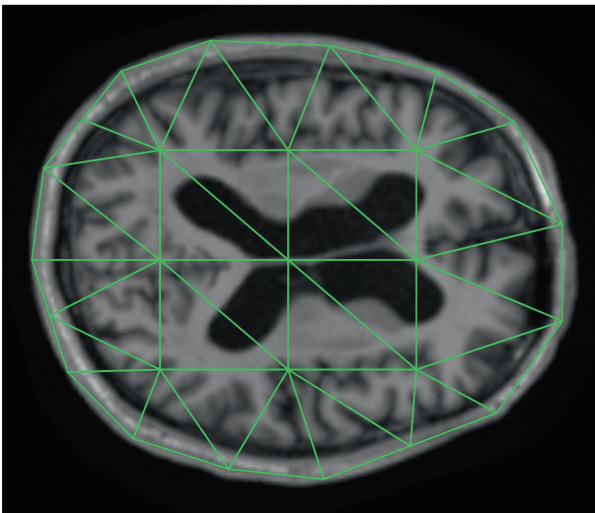
$$h_i(t) = h_0(t) \exp \left(\beta_1 \text{age}_i + \beta_2 \text{sex}_i + \boldsymbol{\zeta}^T \hat{\boldsymbol{\xi}}_i \right)$$

where $\boldsymbol{\zeta} = (\zeta_1, \dots, \zeta_{K^*})^T$ is a vector of coefficients for the first K^* supervised scores.

We employed two different triangulations for partitioning the irregular region of the brain image, resulting in configurations of 33 and 53 triangles, respectively, as shown in Figure 4. Each triangle utilizes a degree 3 polynomial Bernstein basis function.

The brain imaging data and coefficient functions were estimated as a linear combination of the bivariate spline basis functions over triangulation. To avoid over-optimism, a 5-fold nested cross-validation was conducted. Within each of the primary folds, a secondary level of 5-fold cross-validation was performed specifically for selecting the tuning parameter θ with fine increments on the grid $(0, 1]$. To determine the optimal number of sFPCs, denoted as K^* , we repeatedly fitted the model with incrementally increasing values of K^* until no significant improvement in AUC

33 triangles



53 triangles

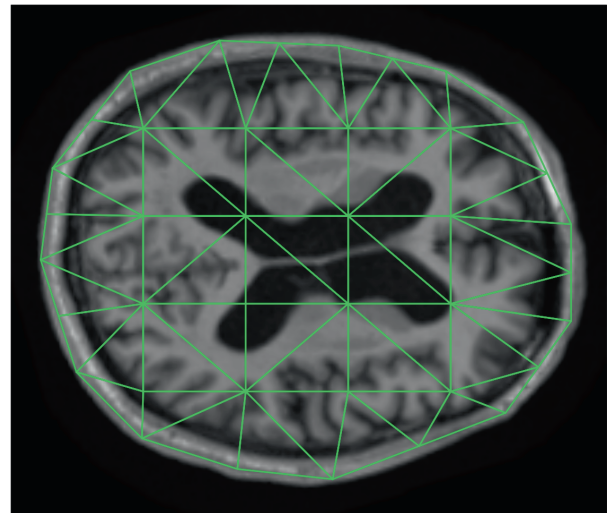


FIGURE 4 | Illustrations of the triangulation grid used for the ADNI dataset with irregular boundary. [Colour figure can be viewed at wileyonlinelibrary.com]

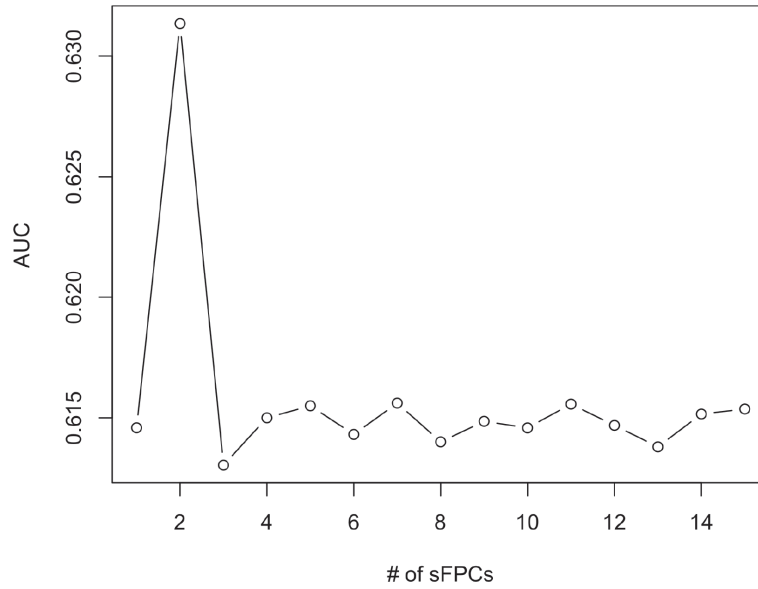


FIGURE 5 | The estimated AUC under supervised FPCA with different numbers of sFPCs in the ADNI study.

TABLE 1 | Estimated AUC with the standard error in parentheses, under the 5-fold cross-validation with the benchmark model, and the inclusion of conventional and supervised FPC scores in the ADNI study.

Triangulation	Benchmark	sFPCA	FPCA
33	0.542 (0.023)	0.645 (0.031)	0.640 (0.033)
53	0.542 (0.023)	0.645 (0.029)	0.639 (0.031)

was observed. Figure 5 presents the estimated AUC values under the proposed method with the number of sFPCs ranging from 1 to 15. Based on this result, we opted for $K^* = 2$.

For comparison, we have also included the results under conventional FPCA and the results computed with only the demographic variables, specifically baseline age and sex. Note that the latter can be viewed as the benchmark, which serves as a reference to evaluate the benefits of integrating function data into the model. The results are summarized in Table 1. The performance remains consistent across both triangulations. Both the sFPCA and conventional FPCA achieved higher AUC values compared with that of the benchmark study. The sFPCA method retained the best predictive performance in terms of AUC. This suggests that the addition of functional scores can provide complementary information to the established risk factors for AD.

The coefficient function in the proportional hazards model can be formulated as $\sum_{k=1}^{K^*} \zeta_k \phi_k(s)$, which can be interpreted as the effect of the brain MRI image at a specific location $s \in \Omega$ on the hazard function. Similarly, in the logistic regression model, the coefficient function $\sum_{k=1}^{K^*} \eta_k \phi_k(s)$ characterizes the image effects on the probability of a subject being uncured. Figures 6 and 7 display the estimated coefficient functions of the ADNI dataset under sFPCA and FPCA with 33 and 53 triangles, respectively. The top row, showcasing results from sFPCA, and the bottom row, displaying results from FPCA, each contain two panels. The left panels represent the estimated coefficient functions in the proportional hazards model, highlighting the regions of the brain that significantly

impact survival outcomes. The right panels show the estimated coefficient functions in the logistic regression model, identifying areas of the brain associated with the logistic components of the model. The coefficient functions exhibit negative values in the red regions and positive values in the blue regions. Note that the coefficient functions estimated under sFPCA demonstrate more complex boundary delineations between red and blue regions compared to those estimated under FPCA. While there are areas of overlap in the effects represented by both methods, the magnitudes of these effects vary. As shown in Figures 6 and 7, the estimated coefficient function $\eta(s)$ in the logistic regression model (right panel) from sFPCA has a darker contrast in the central region (as compared with FPCA), which roughly coincides with the ventricle region. The increase in the size of the ventricle is known to be strongly linked with cognitive decline and dementia as a result of brain atrophy. An enlargement of the ventricle region has been proven to be a key indicator for AD risks [28, 29].

5 | Discussion

In this paper, we propose an sFPCA method to extract important features from high-dimensional imaging data within the context of a mixture cure rate model. This study primarily focuses on capturing the variability in brain images among different patients. By extracting image-based features and incorporating them into survival analysis, we aim to improve the accuracy of risk prediction. To address the challenges posed by the irregular shapes commonly encountered in imaging data, we employ piecewise polynomial bivariate splines over triangulation. Additionally, we implement the IPCW technique to refine the objection functions, thereby accounting for the missing information contained in the censored outcomes.

One limitation of the ADNI study is the short follow-up period: on average, the median follow-up time is 24 months for AD patients and 48 months for CN/MCI patients. This will prevent a plateau from being observable in the data. Under such cases, we may

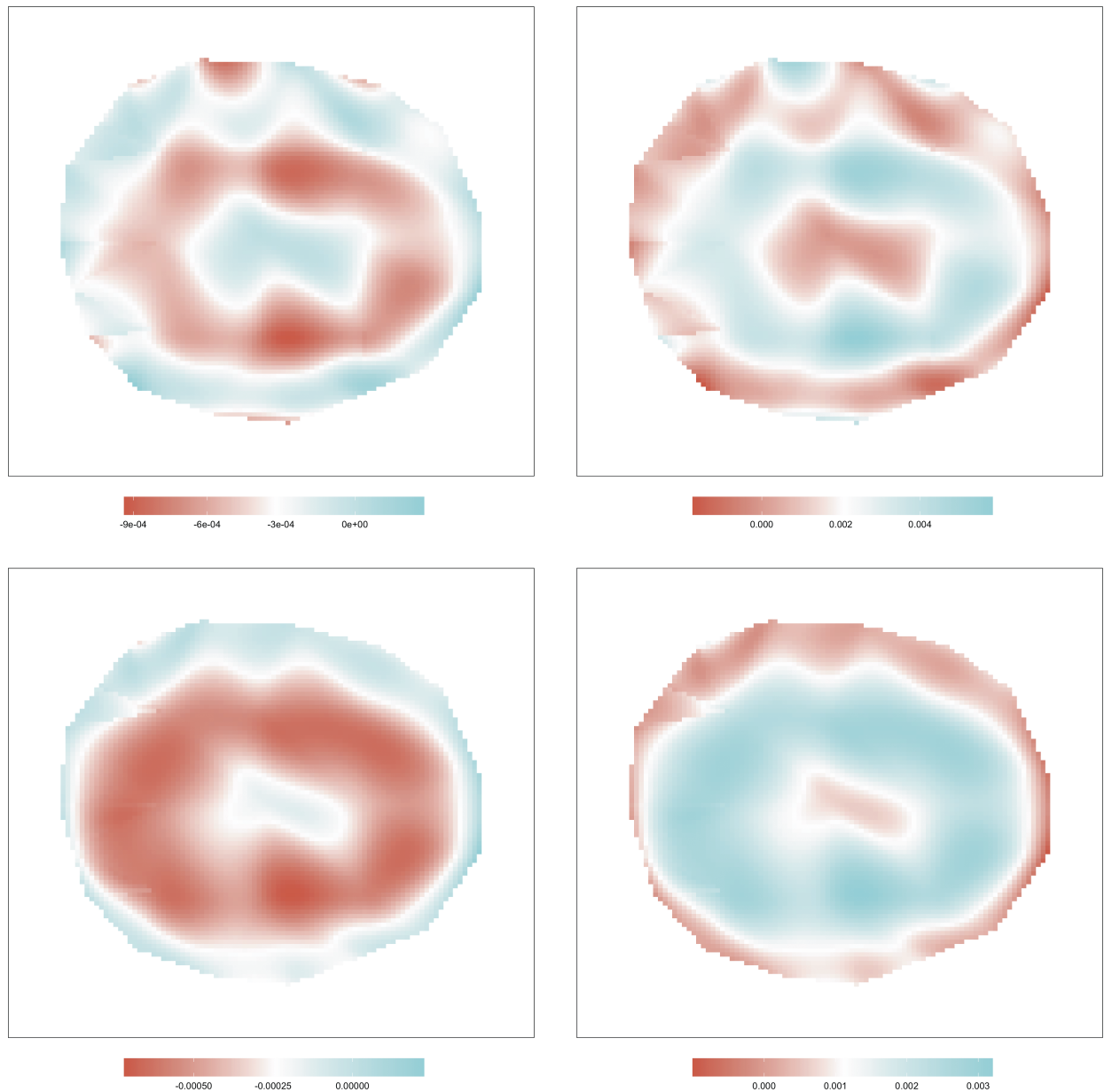


FIGURE 6 | The estimated coefficient functions under sFPCA (first row) and FPCA (second row) for a randomly chosen fold within cross-validation with 33 triangles. Left panel: the estimated coefficient function $\zeta(s)$ in the proportional hazards model. Right panel: the estimated coefficient function $\eta(s)$ in the logistic regression model. [Colour figure can be viewed at wileyonlinelibrary.com]

still apply a cure rate model under the assumption that some individuals have inherently significantly lower risk of developing AD. Various avenues of research suggest the existence of a low-risk subgroup among the MCI/CN population, and a cure rate model can serve as a valuable tool for identifying and understanding such a subgroup.

Our work can be extended in several directions. Firstly, one may be interested in extending our method to high-dimensional longitudinal imaging data. For instance, in datasets like the ADNI, patients undergo multiple MRI scans over the follow-up period. Fully leveraging these longitudinal imaging data can enhance the discrimination power of our model, particularly in identifying individuals at higher risk of disease progression. For such data, an approach building upon multivariate FPCA,

as discussed in Li and Luo [30], may be useful. Secondly, our method can be extended to accommodate a population consisting of multiple sub-groups. For diseases like AD, subjects may present different trajectories, including progression to AD, stability in their condition, or the development of other types of neurodegenerative diseases. In such situations, we may consider a latent-class model [31] to capture the heterogeneous nature of AD. Such models assume that the population consists of several subgroups, each described by a distinct regression model. For instance, Wu et al. [32] and You et al. [33] considered a logistic-Cox mixture model for analyzing lung cancer and ovarian cancer datasets, respectively. This modeling approach allows for a more nuanced understanding of the different progression patterns within AD, highlighting the complex nature of its development and progression.

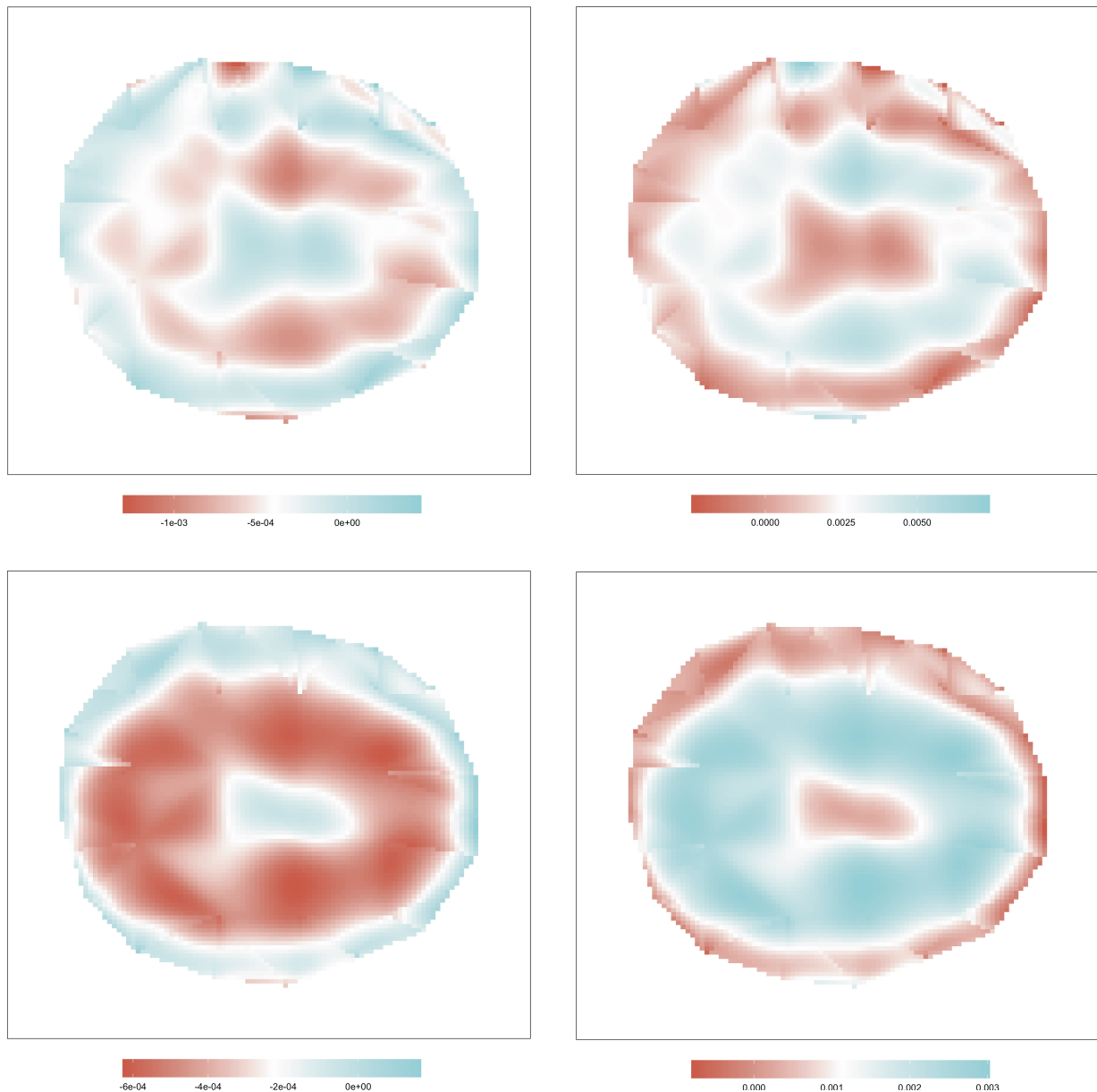


FIGURE 7 | The estimated coefficient functions under sFPCA (first row) and FPCA (second row) for a randomly chosen fold within cross-validation with 53 triangles. Left panel: the estimated coefficient function $\zeta(s)$ in the proportional hazards model. Right panel: the estimated coefficient function $\eta(s)$ in the logistic regression model. [Colour figure can be viewed at wileyonlinelibrary.com]

Acknowledgments

This project is partially supported by the Discovery Grants (RGPIN-2021-02963 to H. Shi and RGPIN-2023-04057 to J. Cao) of the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canada Research Chair program. As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in the analysis or writing of this report. A complete listing of ADNI investigators can be found at: <http://adni.loni.usc.edu/wpcontent/uploads/howtoapply/ADNIAcknowledgementList.pdf>.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available from the Alzheimer's Disease Neuroimaging Initiative: ADNI. Restrictions apply to the availability of these data, which were used under license for this study. Data are available from the author(s) with the permission of the Alzheimer's Disease Neuroimaging Initiative: ADNI.

References

1. C. Patterson, *World Alzheimer Report 2018. The State of the Art of Dementia Research: New Frontiers* (London, UK: Alzheimer's Disease International, 2018).
2. C. R. Jack, Jr., M. A. Bernstein, N. C. Fox, et al., "The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI Methods," *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine* 27, no. 4 (2008): 685–691.

3. R. Sperling, "The Potential of Functional MRI as a Biomarker in Early Alzheimer's Disease," *Neurobiology of Aging* 32 (2011): S37–S43.
4. R. Wolz, V. Julkunen, J. Koikkalainen, et al., "Multi-Method Analysis of MRI Images in Early Diagnostics of Alzheimer's Disease," *PLoS One* 6, no. 10 (2011): e25446.
5. M. Seto, R. L. Weiner, L. Dumitrescu, and T. J. Hohman, "Protective Genes and Pathways in Alzheimer's Disease: Moving Towards Precision Interventions," *Molecular Neurodegeneration* 16 (2021): 29, <https://doi.org/10.1186/s13024-021-00452-5>.
6. T. T. Perls, "Cognitive Trajectories and Resilience in Centenarians—Findings From the 100-Plus Study," *JAMA Network Open* 4 (2021): e2032538, <https://doi.org/10.1001/jamanetworkopen.2020.32538>.
7. M. Zhang, A. B. Ganz, S. Rohde, et al., "Resilience and Resistance to the Accumulation of Amyloid Plaques and Neurofibrillary Tangles in Centenarians: An Age-Continuous Perspective," *Alzheimer's & Dementia* 19 (2022): 2831–2841, <https://doi.org/10.1002/alz.12899>.
8. K. F. Lam, D. Y. T. Fong, and O. Y. Tang, "Estimating the Proportion of Cured Patients in a Censored Sample," *Statistics in Medicine* 24, no. 12 (2005): 1865–1879.
9. R. A. Maller and X. Zhou, *Survival Analysis With Long-Term Survivors* (New York: John Wiley, 1996).
10. J. Z. Huang, H. Shen, and A. Buja, "The Analysis of Two-Way Functional Data Using Two-Way Regularized Singular Value Decompositions," *Journal of the American Statistical Association* 104, no. 488 (2009): 1609–1620.
11. V. Zupnik, B. Caffo, D. M. Yousem, C. Davatzikos, B. S. Schwartz, and C. Crainiceanu, "Functional Principal Component Model for High-Dimensional Brain Imaging," *NeuroImage* 58, no. 3 (2011): 772–784.
12. G. Li, H. Shen, and J. Z. Huang, "Supervised Sparse and Functional Principal Component Analysis," *Journal of Computational and Graphical Statistics* 25, no. 3 (2016): 859–878.
13. X. Zhang, Q. Sun, and D. Kong, "Supervised Principal Component Regression for Functional Responses With High Dimensional Predictors," *Journal of Computational and Graphical Statistics* 33 (2023): 1–8.
14. S. N. Wood, M. V. Bravington, and S. L. Hedley, "Soap Film Smoothing," *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 70, no. 5 (2008): 931–955.
15. M. J. Lai and L. L. Schumaker, *Spline Functions on Triangulations* (Cambridge, UK: Cambridge University Press, 2007).
16. M. J. Lai and L. Wang, "Bivariate Penalized Splines for Regression," *Statistica Sinica* 23, no. 3 (2013): 1399–1417.
17. S. Yu, G. Wang, L. Wang, C. Liu, and L. Yang, "Estimation and Inference for Generalized Geosadditive Models," *Journal of the American Statistical Association* 115, no. 530 (2020): 761–774.
18. L. Wang, G. Wang, M. J. Lai, and L. Gao, "Efficient Estimation of Partially Linear Models for Data on Complicated Domains by Bivariate Penalized Splines Over Triangulations," *Statistica Sinica* 30, no. 1 (2020): 347–369.
19. S. Jiang, J. Cao, G. A. Colditz, and B. Rosner, "Predicting the Onset of Breast Cancer Using Mammogram Imaging Data With Irregular Boundary," *Biostatistics* 24, no. 2 (2023): 358–371.
20. D. Kong, J. G. Ibrahim, E. Lee, and H. Zhu, "FLCRM: Functional Linear Cox Regression Model," *Biometrics* 74, no. 1 (2018): 109–117.
21. E. Lee, H. Zhu, D. Kong, Y. Wang, K. S. Giovanello, and J. G. Ibrahim, "BFLCRM: A Bayesian Functional Linear Cox Regression Model for Predicting Time to Conversion to Alzheimer's Disease," *Annals of Applied Statistics* 9, no. 4 (2015): 2153–2178.
22. S. Jiang, J. Cao, B. Rosner, and G. A. Colditz, "Supervised Two-Dimensional Functional Principal Component Analysis With Time-To-Event Outcomes and Mammogram Imaging Data," *Biometrics* 79, no. 2 (2023): 1359–1369.
23. K. Fukunaga and W. L. Koontz, "Representation of Random Processes Using the Finite Karhunen-Loeve Expansion," *Information and Control* 16, no. 1 (1970): 85–101.
24. B. Efron, "Logistic Regression, Survival Analysis, and the Kaplan-Meier Curve," *Journal of the American Statistical Association* 83, no. 402 (1988): 414–425.
25. A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood From Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 39, no. 1 (1977): 1–22.
26. Y. Zhang, X. Han, and Y. Shao, "The ROC of Cox Proportional Hazards Cure Models With Application in Cancer Studies," *Lifetime Data Analysis* 27, no. 2 (2021): 195–215.
27. D. Ma, K. Popuri, M. Bhalla, et al., "Quantitative Assessment of Field Strength, Total Intracranial Volume, Sex, and Age Effects on the Goodness of Harmonization for Volumetric Analysis on the ADNI Database," *Human Brain Mapping* 40, no. 5 (2019): 1507–1527.
28. S. H. Guptha, E. Holroyd, and G. Campbell, "Progressive Lateral Ventricular Enlargement as a Clue to Alzheimer's Disease," *Lancet* 359 (2002): 2040, [https://doi.org/10.1016/s0140-6736\(02\)08806-2](https://doi.org/10.1016/s0140-6736(02)08806-2).
29. S. M. Nestor, R. Rupsingh, M. Borrie, et al., "Ventricular Enlargement as a Possible Measure of Alzheimer's Disease Progression Validated Using the Alzheimer's Disease Neuroimaging Initiative Database," *Brain* 131 (2008): 2443–2454, <https://doi.org/10.1093/brain/awn146>.
30. K. Li and S. Luo, "Dynamic Prediction of Alzheimer's Disease Progression Using Features of Multiple Longitudinal Outcomes and Time-To-Event Data," *Statistics in Medicine* 38, no. 24 (2019): 4804–4818.
31. G. J. McLachlan, S. X. Lee, and S. I. Rathnayake, "Finite mixture models," *Annual Review of Statistics and Its Application* 6 (2019): 355–378.
32. W. Rf, M. Zheng, and W. Yu, "Subgroup Analysis With Time-To-Event Data Under a Logistic-Cox Mixture Model," *Scandinavian Journal of Statistics* 43, no. 3 (2016): 863–878.
33. N. You, S. He, X. Wang, J. Zhu, and H. Zhang, "Subtype Classification and Heterogeneous Prognosis Model Construction in Precision Medicine," *Biometrics* 74, no. 3 (2018): 814–822.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.