Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–15

DOI: 10.1177/0962280220951834
journals.sagepub.com/home/smm
SAGE

Pattern discovery of health curves using an ordered probit model with Bayesian smoothing and functional principal component analysis

Shijia Wang[1], Yunlong Nie[2], Jason M Sutherland[3] and Liangliang Wang[2]

Abstract

This article is motivated by the need for discovering patterns of patients' health based on their daily settings of care to aid the health policy-makers to improve the effectiveness of distributing funding for health services. The hidden process of one's health status is assumed to be a continuous smooth function, called the health curve, ranging from perfectly healthy to dead. The health curves are linked to the categorical setting of care using an ordered probit model and are inferred through Bayesian smoothing. The challenges include the nontrivial constraints on the lower bound of the health status (death) and on the model parameters to ensure model identifiability. We use the Markov chain Monte Carlo method to estimate the parameters and health curves. The functional principal component analysis is applied to the patients' estimated health curves to discover common health patterns. The proposed method is demonstrated through an application to patients hospitalized from strokes in Ontario. Whilst this paper focuses on the method's application to a health care problem, the proposed model and its implementation have the potential to be applied to many application domains in which the response variable is ordinal and there is a hidden process. Our implementation is available at https://github.com/liangliangwangsfu/healthCurveCode.

Keywords

Bayesian smoothing, functional principal component analysis, Markov chain Monte Carlo, B-spline, stroke

1 Introduction

A stroke is a debilitating event that may cause temporary or permanent impairment of cognitive or physical function. Risk factors for stroke include high blood pressure, smoking, obesity, and diabetes. People experiencing strokes are often hospitalized for their injuries, occasionally for lengthy periods, where they receive sophisticated diagnostics, imaging, intensive medical management, and specialized treatments or therapies that initiate the rehabilitative phase of treatment.

Survival from stroke is common, and patients discharged alive immediately begin treatment for the event's sequelae. Patients receive health services that vary in intensity, frequency, setting, and provider type matched to their health needs. For instance, some portion of patients are admitted into inpatient rehabilitation, hospital-based programs where they receive medical care and several hours of intense rehabilitative therapies each day to

[1]School of Statistics and Data Science, LPMC and KLMDASR, Nankai University, China
[2]Department of Statistics and Actuarial Science, Simon Fraser University, Canada
[3]Centre for Health Services and Policy Research, School of Population and Public Health, University of British Columbia, Canada

Corresponding author:
Liangliang Wang, Simon Fraser University, Room SC K10550, 8888 University Drive, Burnaby, BC, Canada V5A 1S6.
Email: liangliang_wang@sfu.ca

regain cognitive or physical function. Patients with mild symptoms may be discharged to their place of residence, where they may receive occupational or speech therapies several times a week. Underlying variability in post-stroke care is the assumption that where patients receive health care is a function of their health, an unobserved construct that integrates patients' physical and cognitive function, daily activities, pain, depression, or anxiety.

The focus of this research is to use the relationship between (observable) health care settings and intensity of care to estimate patients' health trajectories and then to identify common trajectories in the population. The motivation for our research is a lack of understanding patients' unobserved health condition. While patients' health is unobservable, treatment settings are observable. The second stage of this research, after identifying common trajectories, provides novel insight into health outcomes following hospital treatment for stroke. The contribution of the second stage is to identify potential opportunities for intervening in order to avoid adverse patient outcomes. Moreover, identifying the most common trajectories of health provides the cornerstone for future research on the distribution of funding for health services to ultimately improve the effectiveness of spending on post-stroke care.

This study is based on a retrospective analysis of an existing population-based longitudinal cohort of ischemic and hemorrhagic stroke patients in the province of Ontario, Canada. For the individuals in the cohort, the setting of care is observable on each day following discharge from hospital for 90 days. Based on linking administrative datasets using anonymous identifiers, patients are observed in six settings where health services are provided: *Acute Care (Hospital), Emergency Department, Hospital-based chronic care, Long-term Care, Hospital-based rehabilitation*, and *Home*. Patients' gender and age at discharge from hospital are also recorded. This cohort is ideal to study longitudinal health since most of Ontario's health services are publicly funded and centrally reported, facilitating linkage between settings of care. There are minor gaps in the coverage of health services, as some hospitals do not collect or report outpatient rehabilitation services.[1] The analysis of anonymized data for this study is approved by the Behavioral Research Ethics Board of the University of British Columbia.

In this paper, we estimate health status, represented by a curve from the date of discharge to 90 days afterward. Although patients' health status is unobservable, we assume that it is a continuously valued scale that ranges from 0 (death) to 1 (perfect health) after normalization (see Section 5). The rationale for constructing a continuously valued and bounded scale of health is borrowed from the approach of health preference weights or ordinal values that represent individuals' preferences for health-related outcomes. These values are derived from patients' health states or their responses to patient-reported outcome measures (PROMs), such as the EQ-5D.[2,3]

The strategy used in this study is to first estimate patients' health status based on where they received their health care each day following their stroke hospitalization. An ordered probit model[4–7] is used to model the settings of health care over the study's 90-day follow-up period. We treat the health care setting as an ordinal variable to reflect the intensity of care provided. For example, patients in acute care are assumed to be in worse health than patients in long-term care. The following ordering of health settings is assumed: death is the worst health; patients in acute care have worse health than patients seen and discharged from an emergency department; patients in the emergency department have worse health than patients in hospital-based chronic care; patients in hospital-based chronic care have worse health than patients in long-term care; patients in long-term care have worse health than patients in hospital-based rehabilitation; and patients in hospital-based rehabilitation have worse health than patients at home-receiving home care services. These relationships are summarized as: $Death < Acute\ \ Care\ \ (Hospital) < Emergency\ \ Department < Hospital\text{-}based\ \ chronic\ \ care < Long\text{-}term\ \ Care < Hospital\text{-}based\ rehabilitation < Home$.

Using Bayesian smoothing,[8,9] each patient's health curve is represented by a linear combination of a set of basis functions, for example, the B-spline basis functions.[10] The coefficients of basis functions are assigned a prior distribution. We then develop a Markov chain Monte Carlo (MCMC) algorithm to estimate the posterior distributions of the covariate coefficients, basis function coefficients, and hyperparameters of the model. The Bayesian approach is used because it can fit our complicated model with nontrivial constraints on the lower bound of the health status (death) and on the model parameters to ensure model identifiability. In addition, it can take into account the uncertainty arising from the smoothing parameter selection. Our implementation of the Metropolis–Hastings (MH) within Gibbs is related to the MCMC methods for ordered probit models applied in the literature.[4,6,11]

The second stage is to apply functional principal component analysis (FPCA) to the estimated individual health curves to uncover common patterns of health among post-hospital stroke patients. FPCA plays a significant role as a tool for dimensionality reduction in functional data analysis, where the individual datum is a random curve defined on a common bounded interval.[12,13] FPCA projects the patients' health curves into simple functional

principal component (FPC) scores. In this application of FPCA, the scores are analyzed using clustering methods to identify clusters of health curves that share similar patterns.

The rest of the paper is organized as follows: Section 2 describes the ordered probit model with Bayesian smoothing. Section 3 presents the MCMC algorithm for the model. Section 4 focuses on the FPCA. Then, Section 5 illustrates the application of the method to the stroke data. In Section 6, we evaluate the proposed method through simulation studies. We finish with the concluding remarks and directions for future research work in Section 7.

2 Ordered probit model with Bayesian smoothing

We first introduce some notation for the observed data. Assume that we have observations for n-independent individuals at m time points, $t_1, t_2, \ldots, t_m$, in the time interval $\mathcal{T} = [0, t_m]$. We use Y_{ij} to denote the health care setting for the ith individual at time t_{ij}, where $i = 1, \ldots, n$, and $j = 1, \ldots, m$. Each Y_{ij} assumes values from $J+1$ ordered categorical values: $\{0, \ldots, J\}$. In the application of the stroke data, $Y_{ij} = 0$ is denoted for *Death*; we assume that there are $J = 6$ hierarchical levels of health care settings: $Y_{ij} = 1$ for *Acute Care (Hospital)*, $Y_{ij} = 2$ for *Emergency Department*, $Y_{ij} = 3$ for *Hospital-based chronic care*, $Y_{ij} = 4$ for *Long-term Care*, $Y_{ij} = 5$ for *Hospital-based rehabilitation*, and $Y_{ij} = 6$ for *Home*. Let m_i denote the last observed day or the death day for the ith patient and let $\mathcal{T}_i = [0, t_{i,m_i}]$.

The health curve for the ith patient is denoted as $X_i(t)$, for $t \in \mathcal{T}_i$. Since *Death* is a special health status, it is nonsense to use various values for deaths of different individuals. Therefore, we fix the value of $X_i(t)$ to a predetermined small value ω_l when the ith individual is dead for $i = 1, \ldots, n$. We put the constraint that $X_i(t) \geq \omega_l$.

We will incorporate q covariates, $W_{i1}, \ldots, W_{iq}$, for the ith patient in the model. For example, two covariates, *gender* and *age*, are considered in this paper. More specifically, $W_{i1} = 0$ when the ith subject is female and $W_{i1} = 1$ when the ith subject is male; W_{i2} denotes a continuous value for age.

Furthermore, we assume that the setting of health care, Y_{ij}, is determined by a latent continuous variable Z_{ij}. The latent continuous variable Z_{ij} is assumed to have a normal distribution, $N(g_i(t_{ij}), \sigma^2)$, with mean $g_i(t_{ij}) = X_i(t_{ij}) + \mathbf{w}_i^T\boldsymbol{\beta}$ and variance σ^2. Here, $X_i(t_{ij})$ represents the health status of the ith subject at time t_{ij}, $\mathbf{w}_i = (1, W_{i1}, \ldots, W_{iq})^T$ mainly includes the q covariates, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_q)^T$ is the vector of the intercept and the associated coefficients for the q covariates.

We assume that $Y_{ij} = k$ when $Z_{ij} \in (\tau_{k-1}, \tau_k]$, for $k = 0, \ldots, J$, where $\{\tau_k\}$ is a series of ordered thresholds that satisfy $\tau_{-1} < \tau_0 < \tau_1 < \cdots < \tau_{J-1} < \tau_J$ with the ending thresholds $\tau_{-1} = -\infty$ and $\tau_J = \infty$. Then, we have

$$P(Y_{ij} = k | X_i(t_{ij}), \mathbf{w}_i, \boldsymbol{\beta}, \sigma^2) = \Phi\left(\frac{\tau_k - X_i(t_{ij}) - \mathbf{w}_i^T\boldsymbol{\beta}}{\sigma}\right) - \Phi\left(\frac{\tau_{k-1} - X_i(t_{ij}) - \mathbf{w}_i^T\boldsymbol{\beta}}{\sigma}\right), \quad k = 1, \ldots, J,\ i = 1, \ldots, n,\ j = 1, \ldots, m \tag{1}$$

where $\Phi(\cdot)$ is the cumulative distribution function (cdf) of the standard normal distribution. In other words, the cdf of $N(X_i(t_{ij}) + \mathbf{w}_i^T\boldsymbol{\beta}, \sigma^2)$ is divided into a sequence of partitions that correspond to the ordinal categories. From equation (1), the setting where one patient receives health care depends on one's health status, $X_i(t_{ij})$, covariates $\mathbf{w}_i$, coefficients $\boldsymbol{\beta}$, variance σ^2, and the thresholds $\tau_0, \ldots, \tau_{J-1}$.

Note that the model in equation (1) is unidentifiable. To ensure identifiability, we need to fix two values out of the three parameters, τ_0, τ_{J-1}, and σ^2. We choose to let σ^2 be a free parameter and predetermine τ_0 and τ_{J-1}.[14] We use $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_{J-2})^T$ to denote the vector of unknown thresholds between health state categories.

We assume that the health curve for the ith patient has the form $X_i(t) = b_i + \mu_i(t)$, where b_i is the random intercept and $\mu_i(\cdot)$ denotes the nonparametric effect of time. We assume $b_i \sim N(0, \sigma_b^2)$ and $\mu_i(t) = \sum_{k=1}^{K_i} c_{ik}\psi_{ik}(t) = \boldsymbol{\psi}_i(t)^T\mathbf{c}_i$, where $\boldsymbol{\psi}_i(t) = (\psi_{i1}(t), \ldots, \psi_{iK_i}(t))^T$ denotes the vector of B-spline basis functions for the ith subject, $\mathbf{c}_i = (c_{i1}, \ldots, c_{iK_i})^T$ is the corresponding vector of basis function coefficients, and K_i is the number of basis functions for the ith subject. We let $\mu_i(t_{ij}) = \mu_i(t_{i,m_i})$ for $j = m_i + 1, \ldots, m$ to reflect the fact that the health status does not change after death. To ensure identifiability, we require $\int_{\mathcal{T}_i} \mu_i(t)\mathrm{d}t = 0$, where $\mathcal{T}_i = [0, t_{i,m_i}]$. In this paper, we use cubic B-splines[10] as basis functions.

We use $\mathbf{y}$ to denote the vector of all the observations $\{Y_{ij}\}$ and use $\mathbf{z}$ to denote the vector of all the latent variables $\{Z_{ij}\}$, where $i = 1, \ldots, n, j = 1, \ldots, m$. Let $\mathbf{c} = (\mathbf{c}_1^T, \ldots, \mathbf{c}_n^T)^T$ be a long vector of all basis function coefficients and let $\mathbf{b} = (b_1, \ldots, b_n)^T$. The complete likelihood is expressed as

$$L(\mathbf{c}, \boldsymbol{\tau}, \sigma^2, \mathbf{b}, \sigma_b^2, \boldsymbol{\beta}|\mathbf{y}, \mathbf{z}) = \prod_{i=1}^{n}\prod_{j=1}^{m} \phi(Z_{ij}; X_i(t_{ij}) + \mathbf{w}_i^T\boldsymbol{\beta}, \sigma^2) I(Z_{ij} \in B_{ij})$$

where $I(\cdot)$ is the indicator function and $B_{ij} = (\tau_{k-1}, \tau_k]$ if $Y_{ij} = k$. In the next section, we propose to use the MCMC method to estimate the unknown parameters in the model.

3 Bayesian inference using MCMC

In the Bayesian framework, we need to assign appropriate priors for the model's parameters. The unknown parameters in our ordered probit model include the vector of basis function coefficients $\mathbf{c}$, the covariate coefficients $\boldsymbol{\beta}$, the random effects $\mathbf{b}$, the threshold vector $\boldsymbol{\tau}$, the variance σ^2 of Z_{ij}, and the variance σ_b^2 of the random intercept. In this section, we propose priors for these parameters and the hyperparameters, followed by the details of the MCMC algorithm.

The ith patient's health curve, $X_i(t) = b_i + \mu_i(t)$, can be estimated with the smoothing spline method by minimizing the penalized sum of squares

$$\sum_{j=1}^{m}\{Z_{ij} - X_i(t_{ij}) - \mathbf{w}_i^T\boldsymbol{\beta}\}^2 + \lambda^* \int_{\mathcal{T}_i} \{\mu_i^{(2)}(t)\}^2 \mathrm{d}t \tag{2}$$

where $\mathcal{T}_i = [0, t_{i,m_i}]$, λ^* is a positive smoothing parameter, and $\mu_i^{(2)}(t)$ is the second derivatives of $\mu_i(t)$ with respect to t. The integral term of equation (2) represents a roughness penalty on $\mu_i(t)$, and the smoothing parameter λ^* controls the roughness of the fitted curve $\mu_i(t)$. Note that we consider the roughness penalty up to the last day of observation or the day of death to ensure the health curve for a living person is a smooth function.

The Bayesian counterpart of the above smoothing spline method is to assume a prior density for $\mu_i(t)$ proportional to the "partially improper" Gaussian process

$$\left(\frac{\lambda^*}{2\sigma^2}\right)^{M_i/2} \exp\left\{-\frac{\lambda^*}{2\sigma^2}\int_{\mathcal{T}_i}\{\mu_i^{(2)}(t)\}^2 \mathrm{d}t\right\}$$

where $M_i = K_i - 2$ and σ^2 is the variance of Z_{ij}. Since $\mu_i(t) = \boldsymbol{\psi}_i(t)^T\mathbf{c}_i$ for $t \in \mathcal{T}_i$, we have $\mu_i^{(2)}(t) = \boldsymbol{\psi}_i^{(2)}(t)^T\mathbf{c}_i$, where $\boldsymbol{\psi}_i^{(2)}(t)$ is the second derivative of $\boldsymbol{\psi}_i(t)$ with respect to t. Therefore

$$\int_{\mathcal{T}_i}\{\mu_i^{(2)}(t)\}^2 \mathrm{d}t = \mathbf{c}_i^T\left[\int_{\mathcal{T}_i}\boldsymbol{\psi}_i^{(2)}(t)\boldsymbol{\psi}_i^{(2)}(t)^T\mathrm{d}t\right]\mathbf{c}_i$$

To ease the notation, let $\mathbf{P}_i = \int_{\mathcal{T}_i}\boldsymbol{\psi}_i^{(2)}(t)\boldsymbol{\psi}_i^{(2)}(t)^T\mathrm{d}t$, which is a known, symmetric and positive semi-definite $K_i \times K_i$ matrix with the rank M_i after we determine the basis functions.[15] Now the prior distribution for $\mathbf{c}_i$ can be rewritten as

$$\pi(\mathbf{c}_i|\lambda^*) \propto \left(\frac{\lambda^*}{2\sigma^2}\right)^{M_i/2} \exp\left\{-\frac{\lambda^*}{2\sigma^2}\mathbf{c}_i^T\mathbf{P}_i\mathbf{c}_i\right\}$$

where the hyperparameter λ^* is the smoothing parameter. Let $\lambda = \lambda^*/\sigma^2$ and set the prior for λ, denoted $\pi(\lambda)$, to $Gamma(a_\lambda, b_\lambda)$, where a_λ and b_λ are the shape parameter and scale parameter, respectively, in the Gamma distribution. The prior for σ^2 is set to $\mathrm{IG}(a_\epsilon, b_\epsilon)$, with the density $f(\sigma^2|a_\epsilon, b_\epsilon) = 1/(\Gamma(a_\epsilon)b_\epsilon{}^{a_\epsilon}(\sigma^2)^{a_\epsilon+1})$ $\exp\left(-1/(b_\epsilon\sigma^2)\right)$, $\sigma^2 \geq 0$.

For simplicity, let $\mathbf{z}_i = (Z_{i1}, \ldots, Z_{im})^T$ and define $\boldsymbol{\Psi}_i = (\boldsymbol{\psi}_i(t_{i1}), \ldots, \boldsymbol{\psi}_i(t_{im_i}))^T$, which is an $m_i \times K_i$ matrix. Let $\boldsymbol{\mu}_i = (\mu_i(t_{i1}), \ldots, \mu_i(t_{im}))^T$. Then, we have the vector for the ith subject's health $\mathbf{x}_i = b_i + \boldsymbol{\mu}_i$. Let $\mathbf{W}_i$ be a matrix with m rows with each row being $\mathbf{w}_i^T$. Hence, we have $\mathbf{z}_i = \mathbf{x}_i + \mathbf{W}_i\boldsymbol{\beta} + \varepsilon_i,\ \ i = 1, \cdots, n$, where $\varepsilon_i \sim N(0, \sigma^2\mathbf{I}_m)$ and $\mathbf{I}_m$ is an identity matrix of dimension m.

Letting $\pi(\boldsymbol{\tau})$ be the prior distribution for $\boldsymbol{\tau}$, the posterior is proportional to the product of the likelihood function and priors, which can be written as

$$\begin{aligned}\pi(\mathbf{c}, \tau, \lambda, \sigma^2, \mathbf{b}, \sigma_b^2, \beta|\mathbf{y}, \mathbf{z}) \propto \pi(\tau)\prod_{i=1}^{n}\Bigg\{ &\frac{1}{(\sigma^2)^{1/2}} \cdot \exp\left[-\frac{1}{2\sigma^2}(\mathbf{z}_i - \mathbf{x}_i - \mathbf{W}_i\beta)^T(\mathbf{z}_i - \mathbf{x}_i - \mathbf{W}_i\beta)\right] \\ \cdot (\lambda)^{M_i/2}\exp(-\frac{\lambda}{2}\mathbf{c}_i^T\mathbf{P}_i\mathbf{c}_i) \cdot &\frac{1}{(\sigma_b^2)^{1/2}}\exp\left(-\frac{b_i^2}{2\sigma_b^2}\right)\Bigg\} \cdot (\lambda)^{a_\lambda - 1}\exp\left(-\frac{\lambda}{b_\lambda}\right) \\ \cdot \frac{1}{(\sigma_b^2)^{a_b+1}}\exp\left(-\frac{1}{b_b\sigma_b^2}\right) \cdot &\frac{1}{(\sigma^2)^{a_\epsilon+1}}\exp\left(-\frac{1}{b_\epsilon\sigma^2}\right) \cdot |\Sigma_{\beta 0}|^{-1/2}\exp\left(-\frac{(\beta - \mu_{\beta 0})^T\Sigma_{\beta 0}^{-1}(\beta - \mu_{\beta 0})}{2}\right)\end{aligned}$$

We also require the nontrivial constraint that the value of health remains ω_l when death happens and afterward. That is, $\boldsymbol{\psi}_i(t_{i,m_i})^T\mathbf{c}_i + b_i = \omega_l$, if $m_i \leq m$, which is simply equivalent to setting $c_{iK_i} = \omega_l - b_i$.

The MCMC algorithm, which is a Gibbs sampler with a MH step, is derived by identifying the full conditional distributions for λ, σ^2, σ_b^2, $\mathbf{c}_i$, b_i, $\boldsymbol{\beta}$, Z_{ij}, and incorporating one MH step for the threshold parameters $\boldsymbol{\tau}$. The details are shown below.

The full conditional distribution for λ is

$$\lambda|\mathbf{c} \sim Gamma\left(\sum_{i=1}^{n} M_i/2 + a_\lambda, \frac{1}{\left(1/b_\lambda + \sum_{i=1}^{n}\mathbf{c}_i^T\mathbf{P}_i\mathbf{c}_i/2\right)}\right) \tag{3}$$

The full conditional distribution for σ^2 is

$$\sigma^2|\mathbf{z}, \mathbf{c}, \mathbf{b}, \boldsymbol{\beta} \sim \mathrm{IG}(\tilde{a}_\epsilon, \tilde{b}_\epsilon) \tag{4}$$

where $\tilde{a}_\epsilon = a_\epsilon + n/2$, and $\tilde{b}_\epsilon = \left[1/b_\epsilon + (1/2)\sum_{i=1}^{n}(\mathbf{z}_i - \mathbf{g}_i)^T(\mathbf{z}_i - \mathbf{g}_i)\right]^{-1}$, denoting $\mathbf{g}_i = \mathbf{x}_i + \mathbf{W}_i\boldsymbol{\beta}$.

The full conditional distribution for σ_b^2 is

$$\sigma_b^2|\mathbf{b} \sim \mathrm{IG}\left(n/2 + a_b, \frac{1}{\sum_{i=1}^{n} b_i^2/2 + 1/b_b}\right) \tag{5}$$

The full conditional distribution for $\mathbf{c}_i$ is

$$\mathbf{c}_i|\mathbf{z}_i, \lambda, \sigma^2, \mathbf{b}, \boldsymbol{\beta} \sim N(\tilde{\boldsymbol{\mu}}_{c_i}, \boldsymbol{\Sigma}_{c_i}) \tag{6}$$

where $\boldsymbol{\Sigma}_{c_i} = (\boldsymbol{\Psi}_i^T\boldsymbol{\Psi}_i\sigma^{-2} + \lambda\mathbf{P}_i)^{-1}$, and $\tilde{\boldsymbol{\mu}}_{c_i} = \Sigma_{c_i}\boldsymbol{\Psi}_i^T(Z_{i,1:m_i} - b_i - \mathbf{w}_i^T\boldsymbol{\beta})\sigma^{-2}$. Here $Z_{i,1:m_i}$ denotes the vector $(Z_{i,1}, \ldots, Z_{i,m_i})^T$.

The full conditional distribution for b_i is

$$b_i|\mathbf{z}_i, \sigma^2, \mathbf{c}_i, \boldsymbol{\beta} \sim N(\tilde{\mu}_b, \tilde{\sigma}_b^2) \tag{7}$$

where

$$\begin{aligned}\tilde{\mu}_b &= \tilde{\sigma}_b^2\sigma^{-2} \cdot \sum_{j=1}^{m}(Z_{ij} - \mu_i(t_{ij}) - \mathbf{w}_i^T\boldsymbol{\beta})^T \\ \tilde{\sigma}_b^2 &= \left(m\sigma^{-2} + 1/\sigma_b^2\right)^{-1}\end{aligned}$$

The full conditional distribution for $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta}|\mathbf{z}, \sigma^2, \mathbf{b}, \mathbf{c} \sim N(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \tag{8}$$

where

$$\boldsymbol{\Sigma}_\beta = \left(\sum_{i=1}^{n} \mathbf{W}_i \mathbf{W}_i^T \sigma^{-2} + \Sigma_{\beta 0}^{-1} \right)^{-1}$$

and

$$\boldsymbol{\mu}_\beta = \Sigma_\beta \left[\sum_{i=1}^{n} (\mathbf{z}_i - \mathbf{x}_i) \mathbf{W}_i^T \sigma^{-2} + \boldsymbol{\mu}_{\beta 0}^T \Sigma_{\beta 0}^{-1} \right]$$

We group $\mathbf{z}$ and $\boldsymbol{\tau}$ as a block in the Gibbs sampler. The full conditional distribution of $[\mathbf{z}, \boldsymbol{\tau}]$ is

$$\pi(\mathbf{z}, \boldsymbol{\tau}|\mathbf{y}, \mathbf{c}, \lambda, \sigma^2, b_i, \sigma_b^2, \boldsymbol{\beta}) = \pi(\mathbf{z}|\boldsymbol{\tau}, \mathbf{y}, \mathbf{c}, \lambda, \sigma^2, b_i, \sigma_b^2, \boldsymbol{\beta}) \pi(\boldsymbol{\tau}|\mathbf{y}, \mathbf{c}, \lambda, \sigma^2, b_i, \sigma_b^2, \boldsymbol{\beta})$$

The full conditional distribution for Z_{ij} is a truncated normal distribution, i.e.

$$Z_{ij}|\boldsymbol{\tau}, Y_{ij}, \mathbf{c}_i, \lambda, \sigma^2, b_i, \sigma_b^2, \boldsymbol{\beta} \sim N(g_i(t_{ij}), \sigma^2) I(Z_{it} \in B_{it}) \tag{9}$$

where $g_i(t_{ij}) = X_i(t_{ij}) + \mathbf{w}_i^T \boldsymbol{\beta}$ and $B_{ij} = (\tau_{k-1}, \tau_k]$, if $Y_{ij} = k$. Here, $X_i(t_{ij}) = b_i + \boldsymbol{\psi}_i(t_{ij})^T \mathbf{c}_i$ for $j \leq m_i$; otherwise, $X_i(t_{ij}) = \omega_l$.

The full conditional density for $\boldsymbol{\tau}$ is

$$\pi(\boldsymbol{\tau}|\mathbf{y}, \mathbf{c}, \lambda, \sigma^2, \mathbf{b}, \sigma_b^2, \boldsymbol{\beta}) \propto \pi(\boldsymbol{\tau}) \prod_{i=1}^{n} \prod_{j=1}^{m} \left[\Phi\left(\frac{\tau_{Y_{ij}} - g_i(t_{ij})}{\sigma} \right) - \Phi\left(\frac{\tau_{Y_{ij}-1} - g_i(t_{ij})}{\sigma} \right) \right] \tag{10}$$

Note that there exist order restrictions on the threshold parameter $\boldsymbol{\tau}$. To preserve the order, we reparameterize the threshold parameters by $\gamma_j = \ln\{(\tau_j - \tau_{j-1})/(1 - \tau_j)\}$, for $1 \leq j \leq J-2$.[16] Thus, $\tau_j = (\tau_{j-1} + e^{\gamma_j})/(1 + e^{\gamma_j})$, for $1 \leq j \leq J-2$. Let $\pi(\boldsymbol{\gamma}|\mathbf{y}, \mathbf{c}, \lambda, \sigma^2, \mathbf{b}, \sigma_b^2, \boldsymbol{\beta})$ be the full conditional density for $\boldsymbol{\gamma}$ obtained by reparametrizing (equation (10)). The prior on $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{J-2})$ is set to $\mathbf{N}(\boldsymbol{\mu}_\gamma, \boldsymbol{\Sigma}_\gamma)$, where we assume $\boldsymbol{\mu}_\gamma = (0, \ldots, 0)^T$ and $\boldsymbol{\Sigma}_\gamma$ is a diagonal matrix with σ_γ^2 on the diagonal.

As the full conditional distribution of $\boldsymbol{\gamma}$ does not have a closed form, an MH step in the Gibbs sampler is required.[6] We choose the proposal distribution $q(\boldsymbol{\gamma}^*|\boldsymbol{\gamma}) = N(\boldsymbol{\gamma}, \mathbf{V}_\gamma)$, a normal distribution with mean $\boldsymbol{\gamma}$ and variance $\mathbf{V}_\gamma$. The chain accepts the proposed value $\boldsymbol{\gamma}^*$ with probability

$$\alpha(\boldsymbol{\gamma}, \boldsymbol{\gamma}^*) = \min\left\{ 1, \frac{\pi(\boldsymbol{\gamma}^*|\mathbf{y}, \mathbf{c}, \lambda, \sigma^2, \mathbf{b}, \sigma_b^2, \boldsymbol{\beta}) q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^*)}{\pi(\boldsymbol{\gamma}|\mathbf{y}, \mathbf{c}, \lambda, \sigma^2, \mathbf{b}, \sigma_b^2, \boldsymbol{\beta}) q(\boldsymbol{\gamma}^*|\boldsymbol{\gamma})} \right\} \tag{11}$$

Algorithm 1 summarizes the MCMC algorithm of N iterations for the ordered probit model with Bayesian smoothing.

Algorithm 1 The MCMC algorithm for the ordered probit model with Bayesian smoothing.

1. Set $j = 0$; initialize $\mathbf{c}^{(0)}, \mathbf{z}_i^{(0)}, \lambda^{(0)}, \sigma^{2(0)}$, and $\boldsymbol{\gamma}^{(0)}$.
2. **for** $j = 1, 2, \ldots, N$ **do**
3. sample $\lambda^{(j+1)}|\mathbf{c}^{(j)}$ using equation (3);
4. sample $\sigma^{2(j+1)}|\mathbf{z}^{(j)}, \mathbf{c}^{(j)}, \mathbf{b}^{(j)}, \boldsymbol{\beta}^{(j)}$ using equation (4);

5. sample $\sigma_b^{2(j+1)}|\mathbf{b}^{(j)}$ using equation (5);
6. sample $\mathbf{c}_i^{(j+1)}|\mathbf{z}_i^{(j)}, \lambda^{(j+1)}, \sigma^{2(j+1)}, \mathbf{b}^{(j)}, \boldsymbol{\beta}^{(j)}$ using equation (6);
7. sample $b_i^{(j+1)}|\mathbf{z}_i^{(j)}, \sigma^{2(j+1)}, \mathbf{c}_i^{(j+1)}, \boldsymbol{\beta}^{(j)}$ using equation (7);
8. sample $\boldsymbol{\beta}^{(j+1)}|\mathbf{z}^{(j)}, \sigma^{2(j+1)}, \mathbf{b}^{(j+1)}, \mathbf{c}^{(j+1)}$ using equation (8);
9. sample $Z_{ij}^{(j+1)}|\boldsymbol{\tau}^{(j)}, Y_{ij}, \mathbf{c}_i^{(j+1)}, \sigma^{2(j+1)}, b_i^{(j+1)}, \boldsymbol{\beta}^{(j+1)}$ using Equation (9);
10. sample $\boldsymbol{\gamma}^* \sim N(\boldsymbol{\gamma}^{(j)}, \mathbf{V}_\gamma)$. Accept the proposed value $\boldsymbol{\gamma}^*$ with probability $\alpha(\boldsymbol{\gamma}^{(j)}, \boldsymbol{\gamma}^*)$ using equation (11). If $\boldsymbol{\gamma}^*$ is accepted, $\boldsymbol{\gamma}^{(j+1)} = \boldsymbol{\gamma}^*$; otherwise, $\boldsymbol{\gamma}^{(j+1)} = \boldsymbol{\gamma}^{(j)}$.

4. Pattern discovery via FPCA

After running the MCMC algorithm, we can obtain smooth curves representing patients' health by $\hat{X}_i(t) = \hat{b}_i + \boldsymbol{\psi}_i(t)^T \hat{\mathbf{c}}_i$, where $\hat{b}_i$ and $\hat{\mathbf{c}}_i$ are the posterior samples of the parameters. These estimated health curves can be further used to discover patterns of patients' health using FPCA.

FPCA is widely used to explain major variations in curves. Suppose we have a square integrable stochastic process $X(t)$, $t \in \mathcal{T}$, with the mean $\mathrm{E}(X(t)) = \mu(t)$ and the covariance function $Cov(X(s), X(t)) = G(s, t)$. Mercer's theorem[17] states that $G(s, t)$ has an orthogonal expansion in $L^2(\mathcal{T})$

$$G(s,t) = \sum_{k=1}^{\infty} \lambda_k \phi_k(s) \phi_k(t)$$

where $\phi_k(t)$ and λ_k are the kth eigenfunctions and eigenvalues, respectively, of the covariance function with the order $\lambda_1 \geq \lambda_2 \geq \cdots$. The eigenfunctions $\{\phi_k(t)\}_{k=1}^{\infty}$ satisfy

$$\int \phi_k^2(t)\mathrm{d}t = 1, \text{ and } \int \phi_j(t)\phi_k(t)\mathrm{d}t = 0 \text{ for any } j \neq k$$

Let $X_i(t), i = 1, \ldots, n$ be one realization of the stochastic process $X(t)$. Let $\mu(t)$ be the overall mean function. Then, the Karhunen–Loève expansion of $X_i(t)$ is

$$X_i(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_{ik} \phi_k(t)$$

where $\phi_k(t)$ is also called the kth FPC and ξ_{ik} is called the kth FPC score which can be expressed as

$$\xi_{ik} = \int_{\mathcal{T}} [X_i(t) - \mu(t)] \phi_k(t) \mathrm{d}t \tag{12}$$

Usually, $X_i(t)$ can be sufficiently well approximated by the top L FPCs, where L is a small number relative to the number of time points when $X(t)$ is observed. These top L FPCs explain the most variation of the sample curves. All sample curves can be projected to the finite L-dimensional space expanded by the first L FPCs.

We assume that the overall mean function and covariance surface are smooth, and consequently, the eigenfunctions are smooth. Therefore, FPCs are usually represented as a linear combination of a set of smooth and flexible basis functions. It would be more appealing if we accommodated smoothness when estimating FPCs.[15] There are two approaches to obtain smooth estimates for FPCs. One method is directly smoothing FPCs via penalizing their roughness; the second method is first smoothing the functional data and then estimating the corresponding FPCs based on the smoothed data. In this paper, we adopt the second approach: we apply FPCA based on the estimated health curves to explore the major variations of all health curves using the publicly available fda package on the CRAN project of R (http://cran.r-project.org/web/packages/fda/fda.pdf).

5 Real data analysis

We apply our proposed method to examine unobserved patterns of health curves in a cohort of stroke patients discharged from acute care alive. In the description of our method, we have included two important covariates: gender and age. Besides gender and age, the geographic regions where patients live have been shown to have a significant impact not only on the setting where patients receive treatment but also on the intensity of the treatment or frequency that patients receive treatment.[1] Therefore, in our data analysis, we will apply our method to one community in Ontario. Patients' ages range from 1 to 102 years old with mean 72.4 years and standard deviation 14.4 years. In total, our cohort consists of 2370 patients, of which 1202 (50.7%) are female and 1168 (49.3%) are male patients.

Each patient is observed in at least one setting of health care from Day 1 to Day 90 ($m = 90$), where the first day is defined as the day of discharge from acute care from their stroke treatment. A detailed description of the cohort can be found in the previous work.[1] Figure 1 shows the number of patients in each setting of health care throughout 90 days after the discharge. The majority of the patients received their health care at *Home* (setting 6), whereas only approximately 5% of patients died (setting 0) during the 90 days.

Without prior knowledge about the parameters, flat noninformative priors are applied. To achieve this, in the prior $\pi(\lambda) \sim Gamma(\lambda_a, \lambda_b)$, we set the shape parameter λ_a to 0.01 and the scale parameter λ_b to 100 to have mean 1 and variance 100. In the prior $\sigma_\epsilon^2 \sim \mathrm{IG}(a_\epsilon, b_\epsilon)$, we use the shape parameter $a_\epsilon = 1$ and the scale parameter $b_\epsilon = 0.005$ for the same purpose. We choose the priors for the reparameterized threshold parameters: $\pi(\boldsymbol{\gamma}) \sim \mathbf{N}(0_{4\times1}, 100\mathbf{I}_{4\times4})$, where $\mathbf{I}_{4\times4}$ is an identity matrix.

To ensure identifiability, we predetermine the values of τ_0 and τ_{J-1} to be –5 and 1, respectively. Note that the choice for these values is arbitrary. In Section 6.2, we conduct a sensitivity analysis for different choices of τ_0 and τ_{J-1}. Our analysis indicates that these values will only affect the parameter estimates, but will not affect the patterns of health curves and the explanation of the estimated coefficients.

In Bayesian smoothing, it is essential to decide where to locate the knots: a large number of knots can lead to more accurate estimates at the cost of a high computational burden, while a smaller number of knots can alleviate the computation but may underfit the data. In this study, we put one knot on each day of observation until Day 90 or the day of death to allow flexibility in fitting curves at the cost of heavy computation. The roughness of health curves is controlled by the smoothing parameter.

We arbitrarily choose initial values of the unknown parameters and run the MCMC chain for 6000 iterations. All parameters converge after 1800 iterations by visualizing the trace plots. After discarding the 1800 "burn-in" samples, we obtain 4200 sample draws. We apply thinning (discarding two out of every three samples) to the resulting converged chain, resulting samples of size 4200 to size 1400. The analysis takes 24.2 h on an Intel Xeon CPU E5-2683v4@2.10 GHz.

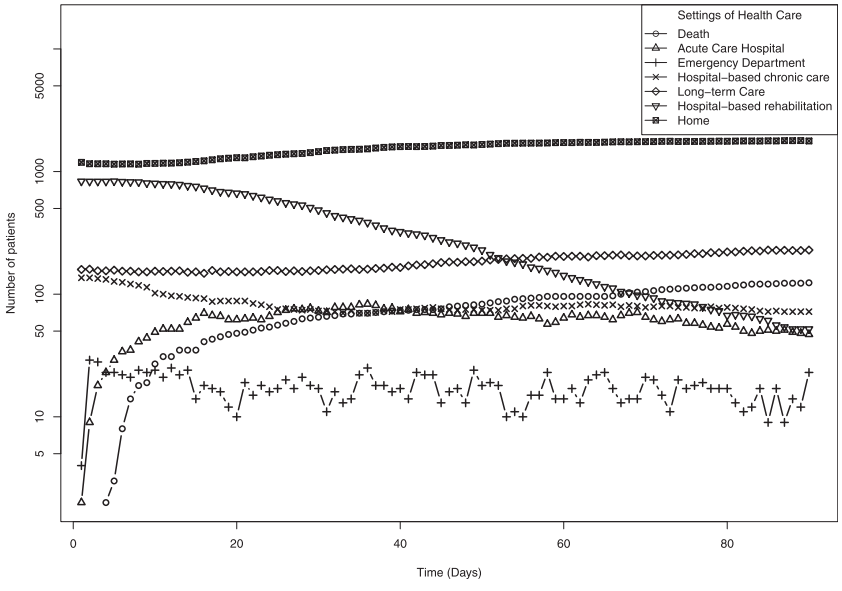

Figure 1. Numbers of patients in settings of health care versus days.

The posterior means and 95% credible intervals of β_0, β_1, and β_2 are 2.121 (2.001, 2.229), 0.066 (0.043, 0.091), and –0.033 (–0.035, –0.032), respectively. Note here female is coded 0 and male 1. A positive estimate of β_1 indicates that males tend to receive care in less intensive settings than females, conditional on everything else is the same. A negative estimate of β_2 means that older patients tend to receive health care in more intensive settings. More results of the parameter $\boldsymbol{\beta}$ and the other parameters are provided in the Supplementary material.

For an easier interpretation, the estimated health curve $\hat{X}_i(t)$ is transformed so that the curve values are within the range [0,1], in which 0 and 1 represent death and perfect health, respectively. Note that we only constrain the minimum value of the unnormalized health, corresponding to death, to be ω_l, and there is no prefixed value for the perfect health. In our application, the value of the perfect health, denoted ω_u, is set to the 99.5% quantile of the $\{\hat{X}_i(t_{ij})\}$, $i = 1, \ldots, n, j = 1, \ldots, m_i$, leading to $\omega_u = 3.92$; the value of ω_l is set to -5. Then the normalized health curve is computed as

$$\tilde{X}_i(t) = \min\left\{\frac{\hat{X}_i(t) - \omega_l}{\omega_u - \omega_l}, 1\right\} \tag{13}$$

To illustrate the typical health curves, Figure 2 shows the raw data and the estimated normalized health curves of four patients. The upper panel displays the raw data of health care settings throughout the 90 days of follow-up. The majority of patients receive most of their care at home, represented by the solid black trajectory. Some patients have a stable health condition but at a lower level such as the one represented by the dashed dotted blue curve. Some patients experience deteriorating health and die during the observed period, represented by the dotted green curve. In contrast, other patients, represented by the dashed read curve, generally have improving health, though short periods of health decline might exist. The lower panel depicts the corresponding estimated health curves after normalization using equation (13). It shows that the estimated health curves well resemble the shapes of the raw data of health care settings.

The first plot in Figure 3 shows the estimated mean health curve (normalized) and its 95% credible bands. The estimated mean health curve suggests that the average health status trend is improving over the observed period. It starts from approximately 0.77 at Day 0, declines slightly for about eight days, and then increases to approximately 0.85 at the end of the 90-day period.

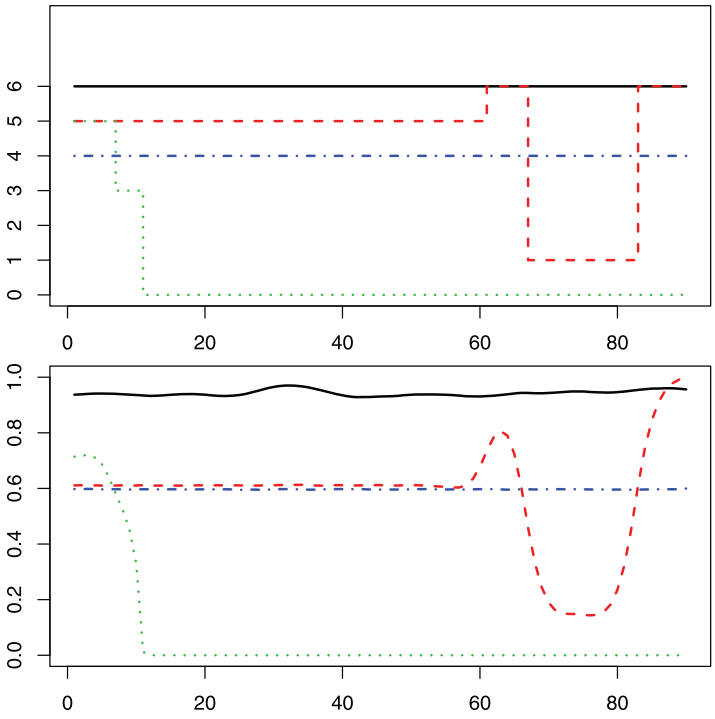

Figure 2. Health care settings versus the follow-up period for four typical individuals (upper panel) and the corresponding normalized estimated health curves (bottom panel).

The FPCA is used to explore the major variation of the estimated health curves. Figure 3 also shows the estimated top three FPCs for the health curves. The first FPC explains 83.3% of the total variation of all the health curves. The first FPC is positive for the entire follow-up period and increases with time until Day 66, which indicates that the curves have a larger variation in the second half of the observed period compared to the first half. The second FPC explains 10.9% of the total variation of all health curves. The second FPC is positive in [0,44] and negative in [45,90], indicating that the second source of variations of health curves comes from the change of magnitude in health from the first 44 days to the remaining days. The third FPC explains 3.7% of the total variation of health curves. The third FPC is negative in [21,66] and positive elsewhere. The third FPC can be interpreted as the change of the health curves between in the middle stage [21,66] from the beginning stage [0,20] and the end stage [67,90].

After obtaining the top FPCs, we can calculate the FPC scores using equation (12). Figure 4 shows the biplot of the first two FPC scores, which represents a projection of the 2370 health curves to the two-dimensional space. We then use the k-means clustering method[18] to partition the two FPC scores into four clusters. The four clusters are distinguished by four different symbols in Figure 4.

Figure 5 shows the mean curve, its 95% credible bands, and a sample of typical health curves for each cluster. About 25.4% of the total patients, illustrated in the top left panel, have improving health. Then, about 4.9% of the total patients, shown in the top right panel, have poor health—some have declining health and die. Around 19.7% of the patients, plotted in the bottom left panel, have stable moderate health. About 50.0% of the total patients, shown on the bottom right panel, have been in stable good health condition during the 90 days with some fluctuations.

The estimated health curves and the discovered patterns can be further used for health care practitioners and policy-makers to more effectively distribute the funding for health services. Suppose we have a function of medical expense $\mathcal{C}(h)$ for the health status $h \in [0, 1]$. We can compute the posterior mean and its 95% CI for the ith patient's expense at time t using $\mathcal{C}(\hat{X}_i(t))(\mathcal{C}(\hat{X}_{i,0.025}(t)), \mathcal{C}(\hat{X}_{i,0.975}(t)))$. Health care practitioners can use this to access if they are overtreating patients by comparing the observed expenses to the estimated expenses. Similarly, we can estimate the average medical expenses for all individuals and for each cluster using the mean health curve

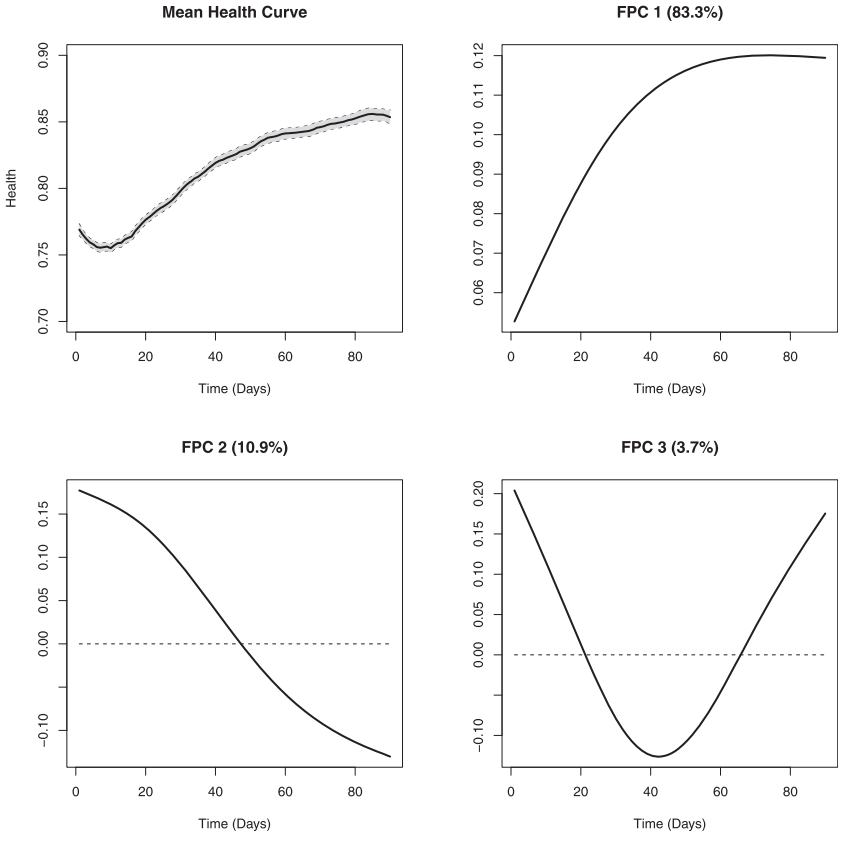

Figure 3. The estimated mean health curve and its 95% credible bands and the estimated top three FPCs for the health curves.

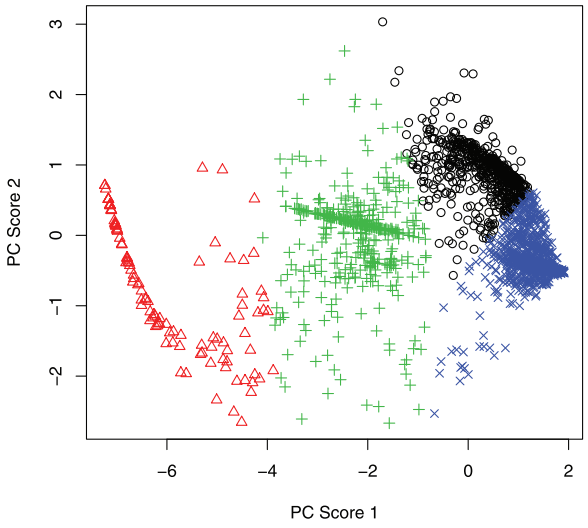

Figure 4. Biplot of the first two FPC scores for 2370 health curves. Four symbols and colors represent four clusters based on the two FPC scores.

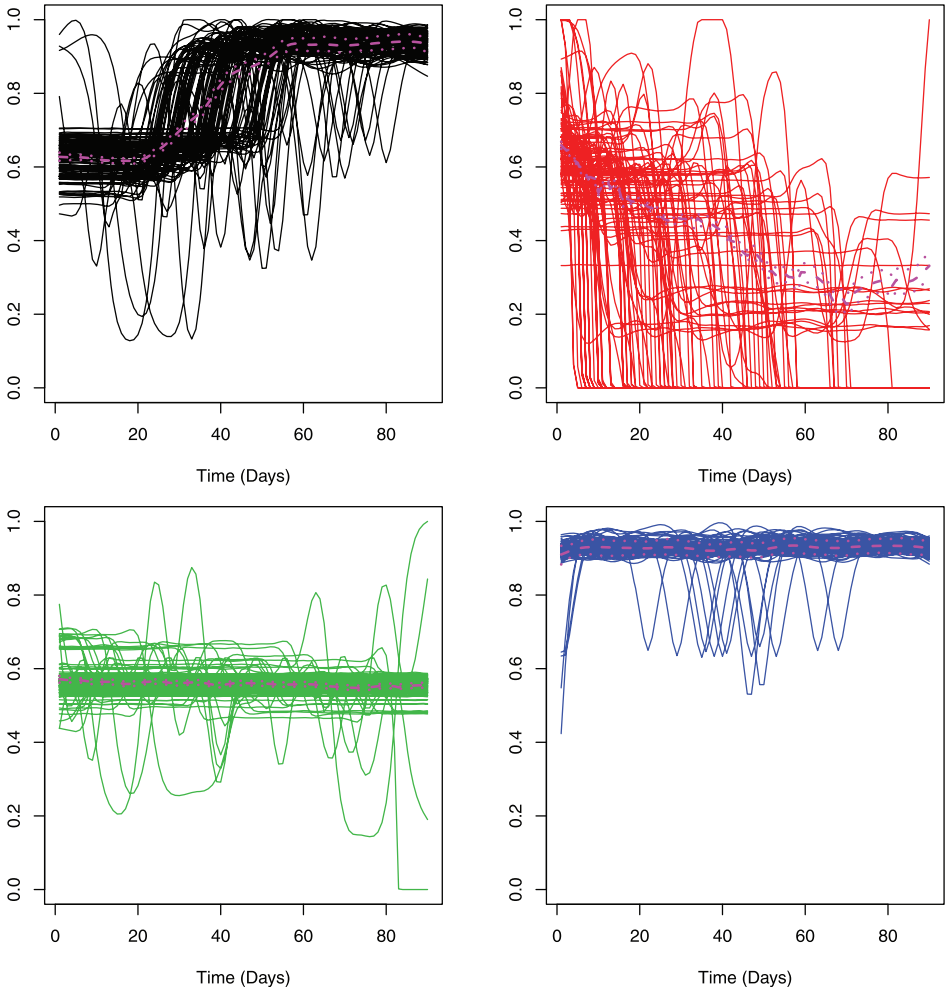

Figure 5. A sample of the estimated health curves (normalized) in four clusters based on the *k*-means clustering using the first two functional principal component scores. The mean curve and its 95% credible bands of each cluster are displayed by dash and dot lines, respectively.

in Figure 3 and the curves for each cluster in Figure 4, respectively. By carefully studying the relationship between the medical expenses and patients' health, it will save billions of dollars for unnecessary treatments and save more lives for who are in need. However, these studies will require collecting and analyzing data about medical expenses, which is out of the scope of this paper.

6 Simulation studies and sensitivity analysis

6.1 Simulation studies

In this section, we examine the performance of the proposed method via simulation studies.

We simulate 100 replicates of data sets under the same simulation setting to estimate bias, standard deviation, and mean square error for each parameter estimate using our proposed method. In each simulated data set, we simulate data of $n = 200$ subjects and $m = 15$ repeated measurements for each subject by mimicking the real stroke data. The normalized health curves are generated as $\tilde{X}_i(t) = \mu(t) + \sum_{k=1}^{3} \xi_{ik}\phi_k(t)$, where the true mean function $\mu(t)$ and the FPCs $\phi_k(t)$ are set to the estimates from the real data, as shown in Figure 3. The FPC scores are generated as $\xi_{ik} \sim N(0, \lambda_k)$, where $\lambda_1 = 3$, $\lambda_2 = 0.5$, and $\lambda_3 = 0.3$. We obtain the unnormalized health curves by $X_i(t) = \omega_l + \tilde{X}_i(t)(\omega_u - \omega_l)$, where ω_l and ω_u are chosen to be the values used in the real data analysis. The latent continuous variable Z_{ij} is generated using a normal distribution with mean, $g_i(t_{ij}) = X_i(t_{ij}) + \beta_0 + \beta_1 \cdot gender + \beta_2 \cdot age$, and variance σ^2 for $i = 1, \ldots, n$, and $j = 1, \ldots, m$. The covariate *gender* is generated from a Bernoulli distribution with parameter 0.5, and the covariate *age* is generated from a uniform distribution from 35 to 100. The random effects b_i are generated independently from a normal distribution with mean 0 and variance 16. Then, we simulate the health care setting Y_{ij} for the ith individual at time t_{ij} as $Y_{ij} = k$ if $Z_{ij} \in (\tau_{k-1}, \tau_k]$, for $k = 0, \ldots, J$. Here, the threshold $\tau_j = (\tau_{j-1} + e^{\gamma_j})/(1 + e^{\gamma_j})$, for $1 \le j \le J - 2$. The true parameter values are set to $J = 6$, $\gamma_1 = -1, \gamma_2 = -1, \gamma_3 = -0.5, \gamma_4 = -0.5$, $\sigma^2 = 1.1$, $\beta_0 = 0$, $\beta_1 = 1.5$, and $\beta_2 = -0.05$, where β_1 and β_2 are coefficients of gender and age, respectively.

For each data set, we set the priors as described in Section 5 and use the MCMC method introduced in Section 3 to estimate the model parameters. Table 1 provides a summary of the bias, standard deviation (SD), and mean square error (MSE) for each parameter estimate. As shown in the table, almost all of the parameters are reasonably well estimated. The relatively large bias of γ_2 is attributable to the fact that only a very small number of data points are generated for the first and second health care settings. We expect that the accuracy of estimation will increase with the increase of the sample size.

In another scenario, sample curves are generated for $n = 200$ subjects; for each subject, there are $m = 50$ repeated measurements. We generate the normalized health curves from $\tilde{X}_i(t) = \mu(t) + \sum_{k=1}^{3} \xi_{ik}\phi_k(t)$, where $\xi_{i1} \sim N(0, 3)$, $\xi_{i2} \sim N(0, 0.5)$ and $\xi_{i3} \sim N(0, 0.3)$. The mean curve is $\mu(t) = (t - 0.3m)^2/(2m) + 0.4$, and the three FPCs are set to $\phi_1(t) = 0.1 - (t - 0.5m)^2/(400m) + 0.001t$, $\phi_2(t) = 0.15\cos(3.14t/m)$, and $\phi_3(t) = 0.15\cos(6.28t/m) + 0.05$. We generate the two covariates, *gender* and *age*, in the same way as in the first scenario. The random effects b_i are generated independently from a normal distribution with mean 0 and variance 4. The true parameter values that are used to generate data are $\boldsymbol{\gamma} = (-1, -2, -0.4, -0.5)$, $\sigma^2 = 0.64$, $\beta_1 = 0.2$, and $\beta_2 = -0.1$. We simulated the data $\{Y_{ij}\}$ in the same way as in the first scenario. Figure 6 shows the estimated health curves (black solid) for several randomly selected individuals and their 95% credible bands, where the true health curves are indicated by the red-dashed curves, and the corresponding simulated observations of health setting. The estimated curves are close to the true ones. More results are provided in the Supplementary material.

Table 1. The biases, SDs), and MSEs for estimated parameters in the simulation study.

Parameter	True	Bias	SD	MSE
β_1	1.5	0.291	0.890	0.878
B_2	–0.05	–0.00424	0.0241	0.000599
Γ_1	–1	–0.594	0.103	0.364
Γ_2	–1	–0.131	0.0528	0.0200
Γ_3	–0.5	–0.0122	0.0389	0.00166
Γ_4	–0.5	0.0344	0.0382	0.00265
σ^2	1.1	0.0632	0.0608	0.00770
σ_b^2	16	0.160	1.357	1.867

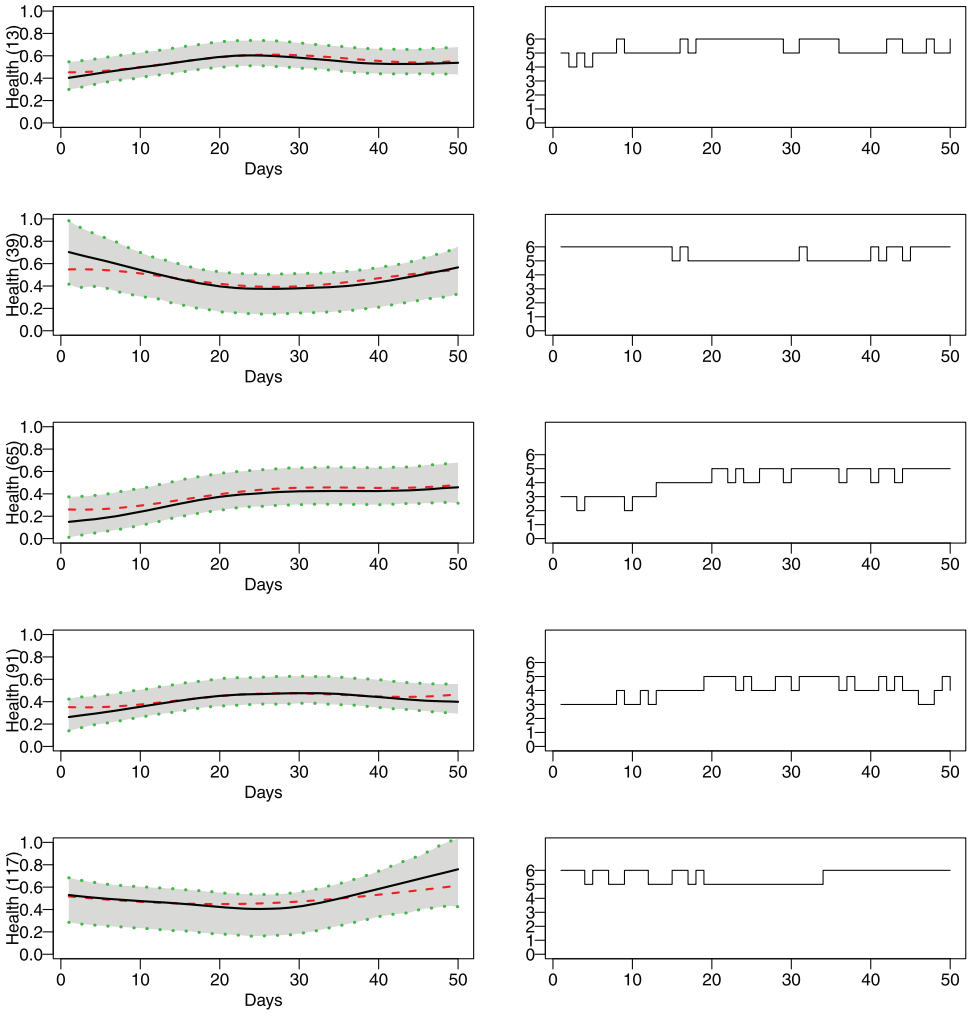

Figure 6. Left panel: the estimated health curves (black solid) for some randomly selected individuals and their 95% credible bands; the true health curves are indicated by the red dashed curves. Right panel: the corresponding simulated observations of health setting.

Table 2. Settings of τ_0, τ_{J-1}, ω_u, and ω_l in the sensitivity analysis and some estimated parameters.

Setting	τ_0	τ_{J-1}	ω_l	ω_u	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
1	–1	1	–2	1.49	1.41 (1.18, 1.65)	0.19 (0.14, 0.27)	–0.01 (–0.01, –0.01)
2	–5	1	–5	3.89	1.13 (0.71, 1.62)	0.25 (0.13, 0.41)	–0.02 (–0.03, –0.02)
3	–10	1	–10	7.55	1.99 (1.21, 2.82)	0.70 (0.44, 0.94)	–0.05 (–0.06, –0.04)

6.2 Sensitivity analysis

In this section, we perform some sensitivity analysis to show that the results are robust to the choice of τ_0, τ_{J-1}, ω_u, and ω_l. Table 2 displays three settings of τ_0, τ_{J-1}, ω_u, and ω_l used in the sensitivity analysis. The estimated-parameter values are different; some of them are shown in Table 2. But the estimated health curves are almost identical. Figure 7 shows health care settings versus the follow-up period for four randomly selected individuals and the corresponding normalized estimated health curves under the three settings of τ_0, τ_{J-1}, ω_u, and ω_l.

7 Conclusion and discussion

We have proposed to estimate health curves by linking the smooth curves to the categorical settings of care using an ordered probit model with Bayesian smoothing. Although the motivating problem is patients' post-stroke

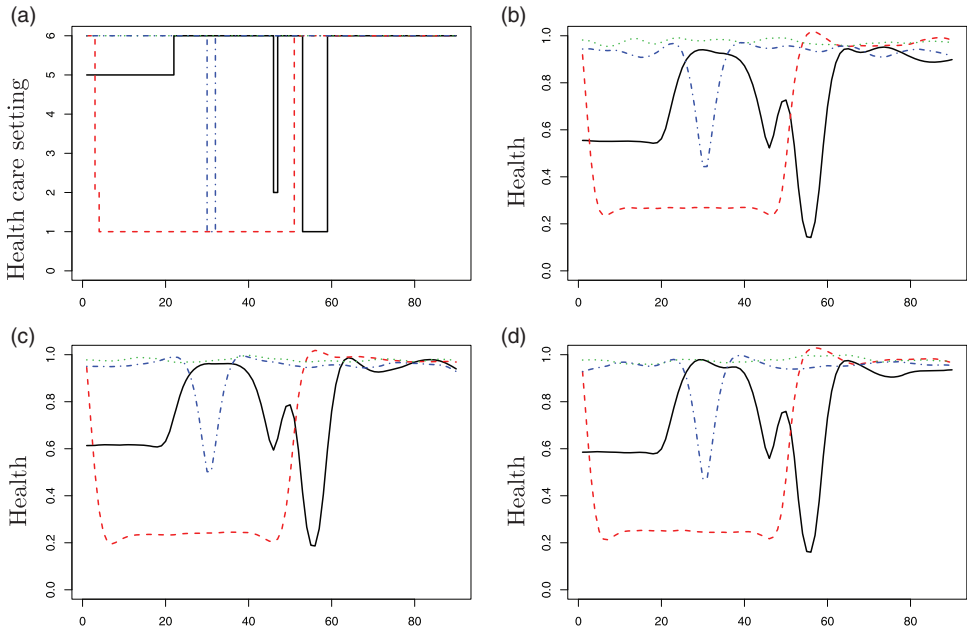

Figure 7. Health care settings versus the follow-up period for four randomly selected individuals (a) and the corresponding normalized estimated health curves (b–d) from settings 1–3, respectively.

health, the proposed approach has the potential to be applied to many applications with ordinal response variable and an underlying hidden continuous process. For example, PROMs are becoming increasingly common as a patient-centered measure of health. PROMs are brief surveys that ask respondents to answer a number of questions regarding their self-assessed health, potentially administered on a longitudinal basis. In this potential application, patients' health is unobserved. However, numeric outcome scores generated by the surveys over time may help to identify the risk of health deterioration.

Our method is not a trivial application of the MCMC approach for a probit model due to the identifiability issue and the constraint by the special health status, death. Although it seems more complicated than some alternative methods, it is more statistically sound at a little more price paid for the computation. One could choose to analyze the ordinal data with metric models, but the real ordinal data do not have equal distance between levels, and they are often strongly skewed, heavily tailed, or multi-model, which may violate the assumptions of a metric model. Liddell and Kruschke[19] have demonstrated that ordinal models can provide better description of the data and suggest to prefer ordinal models over metric model for ordinal data. In addition, the outcome variable, settings for health care, not only depends on health but also depends on several important confounding covariates, such as gender, age, and community. Therefore, direct smoothing will wrongly treat the settings of health care as patients' health and lead to misleading results.

After we obtain the estimated health curves, we could use them for all kinds of further analyses. We have explained in detail how to use FPCA, a dimension reduction technique, to identify major variations in the estimated health curves and to conduct clustering analysis based on the FPC scores. These results can be combined with other information about the medical expenses to assist the health care practitioners and policy-makers to smartly distribute the resources to avoid overtreat or undertreat the patients. For further work, alternative parameterizations based on covariate-adjusted FPCA[20–23] might provide additional insights into the trajectories of patients' health. We can conduct FPCA on a window sliding across time for prediction. Another direction is to develop a spatial model that considers all the communities and their geographic variation to assist the policy-makers to reallocate resources to areas with poorer health outcomes.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship and/or publication of this article: This research was supported by Discovery Grant (RGPIN-2019-06131) of L. Wang from the National Science and Engineering Research Council of Canada. Research was enabled in part by support provided by WestGrid (www.westgrid.ca) and Compute Canada (www.computecanada.ca).

ORCID iD

Liangliang Wang https://orcid.org/0000-0002-8509-7985

Supplemental material

Supplemental material for this article is available online.

References

1. Hellsten E, Chu S, Crump RT, et al. New pricing approaches for bundled payments: leveraging clinical standards and regional variations to target avoidable utilization. *Health Policy* 2016; **120**: 316–326.
2. EuroQoL Group. EuroQol – a new facility for the measurement of health-related quality of life. *Health Policy* 1990; **16**: 199–208.
3. Brooks R. EuroQol: the current state of play. *Health Policy* 1996; **37**: 53–72.
4. Albert JH and Chib S. Bayesian analysis of binary and polychotomous response data. *J Am Stat Assoc* 1993; **88**: 669–679.
5. Weiss AA. Specification tests in ordered logit and probit models. *Econom Rev* 1997; **16**: 361–391.
6. Munkin MK and Trivedi PK. Bayesian analysis of the ordered probit model with endogenous selection. *J Econom* 2008; **143**: 334–348.
7. Cameron AC and Trivedi PK. *Regression analysis of count data*. 2nd ed. Cambridge, UK: Cambridge University Press, 2013.
8. Berry SM, Carroll RJ and Ruppert D. Bayesian smoothing and regression splines for measurement error problems. *J Am Stat Assoc* 2002; **97**: 160–169.
9. Campbell D and Steele RJ. Smooth functional tempering for nonlinear differential equation models. *Stat Comput* 2012; **22**: 429–443.
10. De Boor C. A practical guide to splines. In: Marsden JE and Sirovich L (eds) *Applied mathematical sciences*. rev. ed. Berlin: Springer, 2001.
11. Ronning G and Kukuk M. Efficient estimation of ordered probit models. *J Am Stat Assoc* 1996; **91**: 1120–1129.
12. Ramsay JO and Silverman BW. *Applied functional data analysis: methods and case studies*. vol. **77**. New York: Springer, 2002.
13. Ferraty F and Vieu P. Nonparametric models for functional data, with application in regression, time series prediction and curve discrimination. *Nonparam Stat* 2004; **16**: 111–125.
14. Jeliazkov I, Graves J and Kutzbach M. Fitting and comparison of models for multivariate ordinal outcomes. In: Chib S, Griffiths W, Koop G and Terrell D (eds) *Advances in econometrics*, Volume 23, 2008, pp.115–156.
15. Ramsay JO and Silverman BW. *Functional data analysis*. New York: Springer, 2005.
16. Chen MH and Dey DK. Bayesian analysis for correlated ordinal data models. In: D Dey, SK Ghosh and BK Mallick (eds) *Generalized linear models: a Bayesian perspective*. New York: Marcel Dekker, 2000, pp.133–159.
17. Mercer J. Functions of positive and negative type, and their connection with the theory of integral equations. *Philos Trans R Soc Lond A: Math Phys Eng Sci* 1909; **209**: 415–446.
18. Hartigan JA and Wong MA. Algorithm as 136: a k-means clustering algorithm. *J R Stat Soc Ser C (Appl Stat)* 1979; **28**: 100–108.
19. Liddell TM and Kruschke JK. Analyzing ordinal data with metric models: what could possibly go wrong? *J Exp Social Psychol* 2018; **79**: 328–348.
20. Chiou JM, Müller HG and Wang JL. Functional quasi-likelihood regression models with smooth random effects. *J R Stat Soc: Ser B (Stat Methodol)* 2003; **65**: 405–423.
21. Cardot H. Conditional functional principal components analysis. *Scand J Stat* 2007; **34**: 317–335.
22. Chiou J-M and Müller HG. Modeling hazard rates as functional data for the analysis of cohort lifetables and mortality forecasting. *J Am Stat Assoc* 2009; **104**: 572–585.
23. Jiang CR and Wang JL. Covariate adjusted functional principal components analysis for longitudinal data. *Ann Stat* 2010; **38**: 1194–1226.