Ferede *et al.*
BMC Medical Research Methodology (2024) 24:56
https://doi.org/10.1186/s12874-024-02164-y

BMC Medical Research Methodology

RESEARCH **Open Access**

Flexible Bayesian semiparametric mixed-effects model for skewed longitudinal data

Melkamu M. Ferede[1*], Getachew A. Dagne[2], Samuel M. Mwalili[3], Workagegnehu H. Bilchut[4], Habtamu A. Engida[5] and Simon M. Karanja[6]

Abstract

Background In clinical trials and epidemiological research, mixed-effects models are commonly used to examine population-level and subject-specific trajectories of biomarkers over time. Despite their increasing popularity and application, the specification of these models necessitates a great deal of care when analysing longitudinal data with non-linear patterns and asymmetry. Parametric (linear) mixed-effect models may not capture these complexities flexibly and adequately. Additionally, assuming a Gaussian distribution for random effects and/or model errors may be overly restrictive, as it lacks robustness against deviations from symmetry.

Methods This paper presents a semiparametric mixed-effects model with flexible distributions for complex longitudinal data in the Bayesian paradigm. The non-linear time effect on the longitudinal response was modelled using a spline approach. The multivariate skew-t distribution, which is a more flexible distribution, is utilized to relax the normality assumptions associated with both random-effects and model errors.

Results To assess the effectiveness of the proposed methods in various model settings, simulation studies were conducted. We then applied these models on chronic kidney disease (CKD) data and assessed the relationship between covariates and estimated glomerular filtration rate (eGFR). First, we compared the proposed semiparametric partially linear mixed-effect (SPPLM) model with the fully parametric one (FPLM), and the results indicated that the SPPLM model outperformed the FPLM model. We then further compared four different SPPLM models, each assuming different distributions for the random effects and model errors. The model with a skew-t distribution exhibited a superior fit to the CKD data compared to the Gaussian model. The findings from the application revealed that hypertension, diabetes, and follow-up time had a substantial association with kidney function, specifically leading to a decrease in GFR estimates.

Conclusions The application and simulation studies have demonstrated that our work has made a significant contribution towards a more robust and adaptable methodology for modeling intricate longitudinal data. We achieved this by proposing a semiparametric Bayesian modeling approach with a spline smoothing function and a skew-t distribution.

Keywords Bayesian inference, Semiparametric mixed-models, Longitudinal data, Skew-distributions, Chronic kidney disease

*Correspondence:
Melkamu M. Ferede
melkamum2m@gmail.com
Full list of author information is available at the end of the article

Introduction

Longitudinal data are present in numerous clinical and other follow-up studies that involve monitoring subjects over time to understand the impact of exposures, processes, or characteristics on outcomes. These studies involve tracking a group of subjects and recording data at different time points throughout the study duration. For example, one or more renal functional progress biomarkers (e.g., serum creatinine, albuminuria, glomerular filtration rate, and other biomarkers) of a chronic kidney disease (CKD) patient can be measured repeatedly until end-stage renal disease and/or other events of interest occur.

This research was driven by longitudinal data on CKD, a significant global health issue that affects approximately 500 million individuals worldwide [1]. Around 80 percent of these CKD cases are found in low- and middle-income countries. A prevalence of approximately 35.52 percent of CKD was observed among people with diabetes in Ethiopia [2]. To comprehend how CKD progresses within individuals and across populations, as well as to assess the impact of treatments over time, conducting longitudinal data analysis is necessary.

Longitudinal data can show a variety of features over time and across subjects in many real-world situations during follow-up studies. A suitable choice of methods for analysing such complex longitudinal data is therefore sought. The most popular method proposed is the linear mixed-effects (LME) model [3–6] with a Gaussian response. The generalized linear mixed-effects models [7–9] and non-linear mixed-effects models [10] have been also used as an extension of LME model.

Despite the increasing popularity of LME models in applications, the specification and statistical inference of these models may necessitate much attention when treating and analysing longitudinal data with many features. One of these features is that longitudinal data can exhibit nonlinear, irregular patterns over time, along with asymmetry. Thus, to model and analyse longitudinal data with this feature, LME (fully parametric) models may not be flexible enough.

Another feature is that, unlike linear models, mixed models make assumptions regarding the distribution of model errors as well as random-effects. In the literature, it is usually assumed that the model errors and/or random-effects follow a multivariate normal distribution. In practice, longitudinal data might exhibit asymmetric distributions, leading to biased statistical results [11, 12]. Because of this, employing a normal distribution for model errors may lack robustness against deviations from normality and may be too limited to accurately describe the among- and within-subject variability of longitudinal outcomes [13]. Many previous studies suggest considering a more flexible distribution for model errors to make a valid statistical inference [13, 14]. There are different suggestions in the literature concerning the impact of misspecification of a random-effect distribution on parameter estimation and inference. For instance, Molenberghs and Verbeke [15] suggest that misspecification of the random-effects distribution can lead to biased parameter estimates in nonlinear and generalised linear mixed models; in linear mixed models, however, deviations from the normality assumption may have very little impact on parameter estimation. McCulloch and Neuhaus [16] considered a generalised linear mixed model using a maximum likelihood estimation technique to evaluate the misspecification of the distribution of a random effect. Their findings demonstrated the robustness of most aspects of statistical inferences to the normality of random effects. Other authors in the recent literature, however, suggest that future research should accept more flexible distributional assumption for random-effects in addition to model errors [17, 18]. As a result, skew distributions have recently been used in the literature to handle asymmetry and model longitudinal data more flexibly [18–20].

Thus, in this study, we propose a flexible Bayesian mixed-effects model in a semiparametric setting with a smoothing spline specification and skew distributions for longitudinal data with the aforementioned features. To assess the effectiveness of the proposed methods in various model specifications, simulation studies were conducted. Finally, the proposed model was applied to data on CKD.

Methods

Motivating CKD data and longitudinal outcome trajectories

This paper utilizes a dataset spanning eight years, from June 2014 to June 2022, in the context of chronic kidney disease (CKD). The CKD data was gathered from the University of Gondar Comprehensive Specialized Hospital in Ethiopia, primarily extracted from patients' profiles (or charts) and medical records. Only patients with three or more follow-ups are included in the analysis. The dataset encompasses repeatedly recorded renal function biomarkers, comorbidities, and baseline characteristics of 198 CKD patients. On average, the patients were approximately 55 years old, with 56.6% being male. Around one-third (34.4%) of the CKD patients in the study population had baseline hypertension. Furthermore, the baseline prevalence of diabetes among the CKD patients was determined to be 23.81%.

The estimated glomerular filtration rate (eGFR), which estimates the rate at which the kidneys filter blood, is utilized as a longitudinal response variable. Thus, the analysis of this study considered 1,425 eGFR measurements from 189 patients. The minimum, maximum and average number of measurements per patient were 3, 18 and 8, respectively. 63.5% (120) patients had six and above number of measurements, and out of them 43% patients had ten and above measurements. Of the total 1,425 measurements, based on the National Kidney Foundation guidelines [21], 39.7% indicated CKD Stage 3 (moderate kidney disease), 32.9% indicated Stage 4 (severe kidney disease), and 14.6% indicated Stage 5 (end-stage renal disease). To accurately represent the diverse patterns of renal function progression and create an appropriate model, the analysis includes patients with an eGFR value below ninety. Figure 1 displays the eGFR profiles of patients with CKD. The figure depicts the presence of non-linear trajectories and a positively skewed distribution of eGFR over time.

Bayesian modelling

The semiparametric mixed-effects longitudinal outcome model

In this paper, the longitudinal variable is denoted as y_{ij}, which represents the value of the response eGFR for subject i at the j^{th} time point t_{ij}. The indices i and j range from 1 to m and 1 to m_i respectively, indicating the total number of subjects and the number of measurements for each subject. Let $\boldsymbol{x}_{ij} = (x_{1ij}, \ldots, x_{pij})^T$ denotes a $1 \times p$ vector of associated p covariates. Most previous studies on chronic kidney disease have taken a parametric approach, like utilizing linear mixed-effects models, to model the longitudinal response variable y_{ij} and the associated covariates $\boldsymbol{x}_{ij}$. However, as demonstrated in the presentation of the motivating CKD data above, the outcome eGFR exhibits irregular (non-linear) trajectories over time. Therefore, this paper introduces a semiparametric mixed-effects model that considers the non-linear trajectories of $\boldsymbol{y}_i$ by employing a spline approach.

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{N}_i(\boldsymbol{t}_i) + \boldsymbol{H}_i\boldsymbol{\xi}_i + \boldsymbol{\varepsilon}_i, i = 1, \ldots, m, \tag{1}$$

where $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{imi})^T$ represent the vector of longitudinal response variable, $\boldsymbol{X}_i = (\boldsymbol{x}_{1i}, \ldots, \boldsymbol{x}_{pi})^T$ denote the design matrix of fixed-effects, and $\boldsymbol{H}_i = (\boldsymbol{h}_{1i}, \ldots, \boldsymbol{h}_{qi})^T$ represent the design matrix of random-effects. $\boldsymbol{\beta}$ and $\boldsymbol{\xi}_i$ represent parameter vectors that are associated with the covariates of fixed and random effects. In Eq. (1), the effect of measurement time $\boldsymbol{t}_i = (t_{i1}, \ldots, t_{imi})^T$ on the response $\boldsymbol{y}_i$ is modelled using a non-parametric approach. This is achieved by employing a smoothing function $\boldsymbol{N}_i(\boldsymbol{t}_i)$, which can be defined as follows:

$$\mathbf{N}_i(\mathbf{t}_i) = f(\mathbf{U}(\mathbf{t}_i), \mathbf{V}_i(\mathbf{t}_i)) = \mathbf{U}(\mathbf{t}_i) + \mathbf{V}_i(\mathbf{t}_i), \tag{2}$$

where $\mathbf{U}(\mathbf{t}_i)$ and $\mathbf{V}_i(\mathbf{t}_i)$ represent unknown smoothing functions for the population and subject-specific variations of the longitudinal response $\boldsymbol{y}_i$ due to time effects $\mathbf{t}_i$, respectively. The random vectors $\boldsymbol{\xi}_i$, $\mathbf{V}_i(\mathbf{t}_i)$, and $\boldsymbol{\varepsilon}_i$ are assumed to be independent one another.

A regression spline method is utilized to specify the unknown functions $\mathbf{U}(\mathbf{t}_i)$ and $\mathbf{V}_i(\mathbf{t}_i)$ in Eq. (2), and can be defined as a linear combination of spline basis functions,

$\boldsymbol{\Phi}_{ki}(\mathbf{t}_i) = (\boldsymbol{\Phi}_k(t_{ij}), \ldots, \boldsymbol{\Phi}_k(t_{imi}))^T$
and $\boldsymbol{\Lambda}_{li}(\mathbf{t}_i) = (\boldsymbol{\Lambda}_l(t_{ij}), \ldots, \boldsymbol{\Lambda}_l(t_{imi}))^T$. Where.

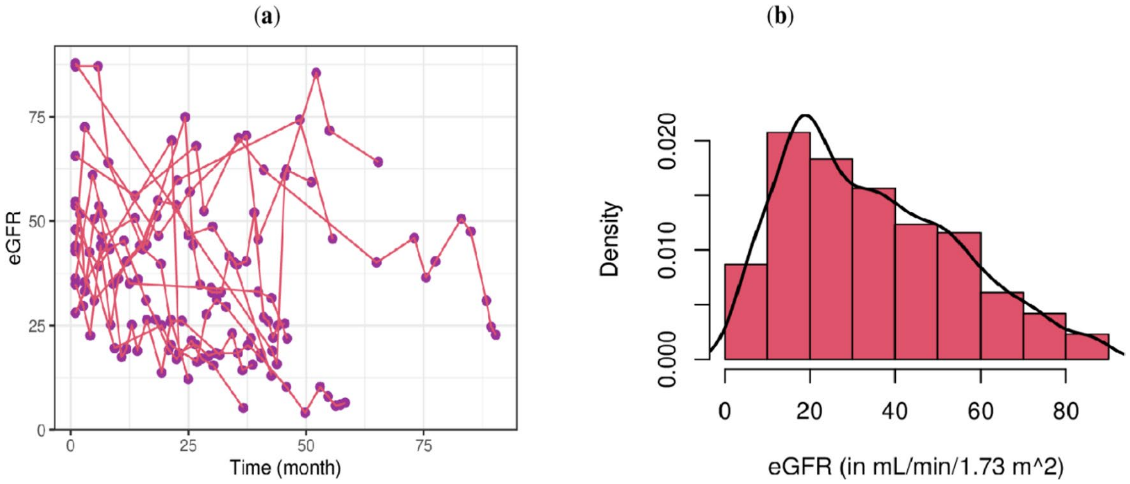


Fig. 1 The trajectories and distribution of the outcome eGFR: (**a**) the line-plots of eGFR over time for some randomly selected patients, indicating non-linear patterns in the trajectories of eGFR; and (**b**) the histogram with density for all the patients, indicating that eGFR has a distribution that is skewed towards the left

$\mathbf{\Phi}_k\left(t_{ij}\right) = (\phi_0(t_{ij}), \phi_1(t_{ij}) \ldots, \phi_{k-1}(t_{ij}))^T$ and.
$\mathbf{\Lambda}_l\left(t_{ij}\right) = (\lambda_0(t_{ij}), \lambda_1(t_{ij}) \ldots, \lambda_{l-1}(t_{ij}))^T; j = 1, \ldots, m_i$.
Mathematically, the specification can be given by

$$\begin{aligned} \mathbf{U}(\mathbf{t}_i) &\approx \sum_{k=0}^{k-1} \mathbf{\Phi}_k(\boldsymbol{t}_i)^T \eta_k = \mathbf{\Phi}_k(\mathbf{t}_i)\boldsymbol{\eta}_k \\ \mathbf{V}_i(\mathbf{t}_i) &\approx \sum_{l=0}^{l-1} \mathbf{\Lambda}_l(\boldsymbol{t}_i)^T \vartheta_{il} = \mathbf{\Lambda}_l(\mathbf{t}_i)\boldsymbol{\vartheta}_{il} \end{aligned} \tag{3}$$

where $\boldsymbol{\eta}_k = (\eta_0, \eta_1, \ldots, \eta_{k-1})^T$ and $\boldsymbol{\vartheta}_{il} = (\vartheta_{i0}, \vartheta_{i1}, \ldots, \vartheta_{i(l-1)})^T$ are $k \times 1$ and $l \times 1$ parameter vectors of the fixed-effect spline basis $\mathbf{\Phi}_k(\mathbf{t}_i)$ and random spline basis effects$\mathbf{\Lambda}_l(\mathbf{t}_i)$, respectively. The B-spline, truncated power or natural cubic spline basis can be used to construct the bases ($\mathbf{\Phi}_k(\mathbf{t}_i)$ and$\mathbf{\Lambda}_l(\mathbf{t}_i)$) in (3). In this study, natural cubic spline with percentile-based knots is considered to approximate the bases. By using Eq. (3), model (1) can be rewritten as:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \mathbf{\Phi}_k(\mathbf{t}_i)\boldsymbol{\eta}_k + \boldsymbol{H}_i\boldsymbol{\xi}_i + \mathbf{\Lambda}_l(\mathbf{t}_i)\boldsymbol{\vartheta}_{il} + \boldsymbol{\varepsilon}_i, i = 1, \ldots, m \tag{4}$$

Let $\boldsymbol{Z}_i = (\boldsymbol{X}_i, \mathbf{\Phi}_k(\mathbf{t}_i))$ and $\boldsymbol{R}_i = (\boldsymbol{H}_i, \mathbf{\Lambda}_l(\mathbf{t}_i))$ be the fixed-effect (population) and random effects design matrices, respectively. Furthermore, let $\boldsymbol{\alpha} = (\boldsymbol{\beta}_p^T, \boldsymbol{\eta}_k^T)^T$ and $\boldsymbol{\varphi}_i = (\boldsymbol{\xi}_{iq}^T, \boldsymbol{\vartheta}_{il}^T)^T$ be the associated parameter vectors. Then, model (1) can be reformulated as

$$\begin{aligned} \mathbf{y}_i &= \mathbf{Z}_i\boldsymbol{\alpha} + \mathbf{R}_i\boldsymbol{\varphi}_i + \boldsymbol{\varepsilon}_i, i = 1, \ldots, m \\ \boldsymbol{\varphi}_i &\sim ST_{q+l,\rho_\varphi}\left(0, \mathbf{\Sigma}_\varphi, \boldsymbol{\delta}_\varphi\right) \\ \boldsymbol{\varepsilon}_i &\sim ST_{m_i,\rho_\varepsilon}\left(0, \sigma_\varepsilon^2\mathbf{I}_{m_i}, \delta_\varepsilon\mathbf{I}_{m_i}\right) \end{aligned} \tag{5}$$

Most previous studies assumed a Gaussian distribution for the random-effects $\boldsymbol{\varphi}_i$ (representing inter-subject variation of $\mathbf{y}_i$) as well as for the model errors $\boldsymbol{\epsilon}_i$ (representing within-subject variation) due to its computational convenience. However, in this study, we considered multivariate skew-t distributions [22] for both $\boldsymbol{\varphi}_i$ and $\boldsymbol{\varepsilon}_i$. That is, $\boldsymbol{\varphi}_i \sim ST_{q+l,\rho_\varphi}\left(0, \mathbf{\Sigma}_\varphi, \boldsymbol{\delta}_\varphi\right)$ and $\boldsymbol{\varepsilon}_i \sim ST_{m_i,\rho_\varepsilon}\left(0, \sigma_\varepsilon^2\mathbf{I}_{m_i}, \delta_\varepsilon\mathbf{I}_{m_i}\right)$. Where $ST(.)$ is a skew-t distribution; ρ_φ and ρ_ε denote degrees of freedom; $\mathbf{\Sigma}_\varphi$ and $\mathbf{\Sigma}_\varepsilon = \sigma_\varepsilon^2\mathbf{I}_{m_i}$ denote covariance matrices; and $\boldsymbol{\delta}_\varphi$ and $\boldsymbol{\delta}_\varepsilon = \delta_\varepsilon\mathbf{I}_{m_i}$ are skewness vectors of the random-effects $\boldsymbol{\varphi}_i$ and model errors $\boldsymbol{\varepsilon}_i$, respectively.

Hierarchical reformulation of the model

The statistical inference from a semiparametric mixed-effects model with multivariate skew-t distributions using the likelihood approach can be computationally demanding. Hence, to overcome this challenge, we adopted the Bayesian approach, which offers computational efficiency. This approach not only reduces the computational load but also allows for more accurate parameter estimation by leveraging existing information (prior knowledge) for parameter estimation. By employing Markov Chain Monte Carlo (MCMC) algorithms, the Bayesian approach enables us to estimate the parameters more efficiently while obtaining posterior distributions that provide a comprehensive quantification of parameter uncertainty.

In order to carry out the MCMC, it is crucial to reformulate the model (5) by rep-resenting the skew-t distributions using the stochastic representation considered by [22] (See Appendix B). To achieve this, we introduced random vectors $\boldsymbol{W}_{\varphi_i} = (W_{\varphi_i 1}, \ldots, W_{\varphi_i (q+1)})^T$ and $\boldsymbol{W}_{\boldsymbol{\varepsilon}_i} = (W_{\varepsilon_i 1}, \ldots, W_{\varepsilon_i m_i})^T$, as well as random variables ν_φ and ν_ε to represent the skew-t distributions associated with the random effects $\boldsymbol{\varphi}_i$ and model errors $\boldsymbol{\varepsilon}_i$, respectively. Consequently, we present the hierarchical reformulation of model (5) as follows:

$$\begin{aligned} \mathbf{y}_i|\boldsymbol{\varphi}_i, \mathbf{W}_{\varepsilon i}, \nu_{\varepsilon i}; \boldsymbol{\alpha}, \sigma_\varepsilon^2, \mathbf{\Sigma}_\varphi, \delta_\varepsilon, \rho_\varepsilon &\sim N_{m_i}\left(\mathbf{Z}_i\boldsymbol{\alpha} + \mathbf{R}_i\boldsymbol{\varphi}_i + \delta_\varepsilon\mathbf{W}_{\varepsilon i}, \nu_{\varepsilon i}^{-1}\sigma_\varepsilon^2 1_{m_i}\right), \\ \varphi_i|\mathbf{W}_{\varphi i}, \nu_{\varphi i}, \mathbf{\Sigma}_\varphi, \boldsymbol{\delta}_\varphi, \rho_\varphi &\sim N_{q+l}\left(\boldsymbol{\delta}_\varphi\mathbf{W}_{\varphi i}, \nu_{\varphi i}^{-1}\mathbf{\Sigma}_\varphi\right), \\ \mathbf{W}_{\varphi i}|\nu_{\varphi i} &\sim N_{q+l}\left(0, \nu_{\varphi i}^{-1}\mathbf{I}_{q+l}\right)I\left(\mathbf{W}_{\varphi i} > 0\right), \\ \mathbf{W}_{\varepsilon i}|\nu_{\varepsilon i} &\sim N_{m_i}\left(0, \nu_{\varepsilon i}^{-1}\mathbf{I}_{m_i}\right)I(\mathbf{W}_{\varepsilon i} > 0), \\ \nu_{\varepsilon i}\big|\rho_\epsilon \sim \Gamma(\rho_\varepsilon/2, \rho_\varepsilon/2), \nu_{\varphi i}\big|\rho_\varphi &\sim \Gamma\left(\rho_\varphi/2, \rho_\varphi/2\right) \end{aligned} \tag{6}$$

Where $N_b()$, in general, stands for a multivariate normal distribution with a dimension of b, and $\Gamma()$ denotes a gamma distribution.

Prior specification and the posterior distribution

Let $\mathbf{\Omega} = \left\{\boldsymbol{\alpha}, \sigma_\varepsilon^2, \mathbf{\Sigma}_\varphi, \boldsymbol{\delta}_\varphi, \delta_\epsilon, \rho_\varphi, \rho_\varepsilon\right\}$ represent the set of all parameters in the hierarchical model (6). We specify the prior distributions for each parameter in $\mathbf{\Omega}$ as follows:

- The fixed-effects and skewness parameters $\boldsymbol{\alpha}$, $\boldsymbol{\delta}_\varphi$, and δ_ϵ are assumed to follow independent normal prior distributions $N_\mathrm{p}(\boldsymbol{\alpha}_0, \mathbf{\Theta}_{\boldsymbol{\alpha}})$, $N_{\mathrm{q}+1}(0, \boldsymbol{\kappa}_{\delta_\varphi})$, and $N(0, \kappa_{\delta_\varepsilon})$, respectively.
- The scale parameters $\mathbf{\Sigma}_\varphi$ and σ_ε^2 follow an inverse-Wishart and inverse-Gamma prior distributions, $IW_{\mathrm{q}+1}(\mathbf{D}_\varphi, \nu_\varphi)$ and $IG(\varrho_{\varepsilon 1}, \varrho_{\varepsilon 2})$, respectively.

- The degrees of freedom parameters ρ_φ and ρ_ε are assumed to follow truncated exponential prior distributions, $Exp(\rho_{\vartheta 0})I(\rho_\vartheta > 3)$ and $Exp(\rho_{\varepsilon 0})I(\rho_\varepsilon > 3)$, respectively.

The hyperparameter matrices $\mathbf{\Theta}_{\boldsymbol{\alpha}}$ and $\mathbf{D}_\varphi$ are assumed diagonal for convenient implementation. Then, the prior distribution of all the parameters, denoted as $\pi(\mathbf{\Omega})$, can be defined as the product of the individual prior distributions of each parameter.

Suppose $\mathbf{G} = \left\{\mathbf{y}_i, \mathbf{Z}_i, \mathbf{W}_{\varphi_i}, \mathbf{W}_{\varepsilon_i}, \nu_\varphi, \nu_\varepsilon\right\}$ be the observed data. An approximation of the posterior density of $\mathbf{\Omega}$ given $\mathbf{G}$ can be obtained as follows:

$$
\begin{aligned}
\pi(\mathbf{\Omega}|\mathcal{G}) \quad & \propto f(\mathcal{G}|\mathbf{\Omega}) \times \pi(\mathbf{\Omega}) \\
& \propto \prod_{i=1}^{m} \int_{\varphi_i} \left\{ \left(\sigma_\epsilon^2\right)^{-\frac{m_i}{2}} \exp\left(-\tfrac{1}{2}\left(\mathbf{y}_i - \boldsymbol{\mu}_y\right)^T \left(\frac{\sigma_\epsilon^2 \mathbf{I}_{m_i}}{\nu_{\epsilon i}}\right)^{-1} \left(\mathbf{y}_i - \boldsymbol{\mu}_y\right) \right) \right\} \\
& \times \left|\mathbf{\Sigma}_\varphi\right|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}\left(\boldsymbol{\varphi}_i - \boldsymbol{\delta}_\varphi \mathbf{W}_\varphi\right)^T \nu_{\varphi i} \mathbf{\Sigma}_\varphi^{-1} \left(\boldsymbol{\varphi}_i - \boldsymbol{\delta}_\varphi \mathbf{W}_{\varphi i}\right) \right) \\
& \times \exp\left(-\tfrac{1}{2} \nu_{\varphi i} \mathbf{W}_{\varphi i}^T \mathbf{W}_{\varphi i} \right) \times \exp\left(-\tfrac{1}{2} \nu_{\epsilon i} \mathbf{W}_{\epsilon i}^T \mathbf{W}_{\epsilon i} \right) \\
& \left\{ \frac{1}{\Gamma(\rho_\varphi/2)(\rho_\varphi/2)^{\rho_\varphi/2}} \nu_{\varphi i}^{\frac{\rho_\varphi}{2}-1} \right\} \exp\left(-\tfrac{2}{\rho_\varphi} \nu_{\varphi i} \right) \\
& \left\{ \frac{1}{\Gamma(\rho_\epsilon/2)(\rho_\epsilon/2)^{\rho_\epsilon/2}} \nu_{\epsilon i}^{\frac{\rho_\epsilon}{2}} - 1 \right\} \exp\left(-\tfrac{2}{\rho_\epsilon} \nu_{\epsilon i} \right) \Bigg\} d\varphi_i \\
& \times \exp\left(-\tfrac{1}{2} (\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^T \mathbf{\Theta}_\alpha^{-1} (\boldsymbol{\alpha} - \boldsymbol{\alpha}_0) \right) \\
& \times \left(\sigma_\epsilon^2\right)^{-\varrho_{\epsilon 1} - 1} \exp\left(-\varrho_{\epsilon 2}/\sigma_\epsilon^2 \right) \\
& \times \left|\mathbf{\Sigma}_\varphi\right|^{-\frac{(\nu_\varphi + q + l + 1)}{2}} \exp\left(-\tfrac{1}{2} \mathrm{tr}\left(\mathbf{D}_\varphi \mathbf{\Sigma}_\varphi^{-1} \right) \right) \\
& \times \exp\left(-\tfrac{1}{2} \boldsymbol{\delta}_\varphi^T \left|\kappa_{\delta_\varphi}\right|^{-1} \boldsymbol{\delta}_\varphi \right) \times \exp\left(-\tfrac{1}{2\kappa_{\delta_\epsilon}} \delta_\epsilon^2 \right) \\
& \times \exp\left(-\rho_\varphi \rho_\varphi \right) \times \exp\left(-\rho_{\epsilon 0} \rho_\epsilon \right)
\end{aligned}
\tag{7}
$$

where $f(\mathcal{G}|\mathbf{\Omega})$ is the joint likelihood function of $\mathbf{G}$ given $\mathbf{\Omega}$ and $\boldsymbol{\mu}_y = \mathbf{Z}_i \boldsymbol{\alpha} + \mathbf{R}_i \boldsymbol{\varphi}_i + \delta_\varepsilon \mathbf{W}_{\varepsilon i}$.

The Metropolis–Hastings algorithm within Gibbs sampler can be used to draw samples from the full conditional posterior distributions of the parameters and to estimate their posterior means and standard deviations. For all models, the Markov chain Monte Carlo (MCMC) procedure was implemented using WinBUGS14 software, which simplifies the implementation of the MCMC algorithm by eliminating the need to derive full conditionals and specify the algorithm explicitly.

Model comparison and diagnostics checking

The specification and implementation of the proposed model in the Bayesian approach may require to conduct convergence diagnostic checks and thoroughly examine the distributional assumptions before drawing any statistical inferences about the parameters. Failure to do so may result in biased estimates and invalid statistical inference. Thus, in this study, the Brooks-Gelman-Rubin (BGR) plot [23], trace plot, ACF plot and the Geweke's test of convergence are all used to evaluate convergence. After confirming convergence, we proceed to evaluate the effectiveness of the proposed semiparametric mixed-effects model (5) by exploring various distributional assumptions for the random-effects and model errors. This model comparison involves considering different distributional specifications and examining their performance in capturing the underlying characteristics of the data. These specifications are given below:

MoSTST: A semiparametric partially linear mixed-effects model (SPPLMEM) with multivariate skew-t (ST) distributions for both the random- effects $\boldsymbol{\varphi}_i$ and model errors $\boldsymbol{\varepsilon}_i$.
MoNST: An SPPLMEM with multivariate normal (N) distribution of $\boldsymbol{\varphi}_i$ and ST-distribution of $\boldsymbol{\varepsilon}_i$.
MoSNSN: An SPPLMEM with both $\boldsymbol{\varphi}_i$ and $\boldsymbol{\varepsilon}_i$ follow multivariate skew-normal (SN) distributions.
MoNN: An SPPLMEM with multivariate normal (N) distributions of $\boldsymbol{\varphi}_i$ and $\boldsymbol{\varepsilon}_i$. MoNN model is the standard choice in longitudinal data analysis.

In order to evaluate the performance of the estimators and make comparisons between different models, we utilised several statistical measures. Additionally, to compare and select the best-fitting Bayesian model for the skewed longitudinal response, we employed the deviance information criterion, which takes into account both the goodness of fit and model complexity.

Deviance information criterion (DIC)

In this paper, DIC [24] is used to choose the best-fitting Bayesian semiparametric mixed-effects model. DIC is the most popular Bayesian model comparison tool in the literature: the smaller this value, the better the model fit. The DIC for the hierarchical Bayesian model (6) with parameters vector Ω and observed longitudinal data D can be defined as

$$DIC = Dev\left(\overline{\Omega}\right) + 2\mathcal{P}_{\mathcal{D}} \tag{8}$$

where

$$Dev\left(\overline{\Omega}\right) = \mathrm{Dev}(E(\Omega|\boldsymbol{D})) \tag{9}$$

is the deviance computed at the posterior mean of model parameters. And

$$\mathcal{P}_{\mathcal{D}} = \overline{Dev(\Omega)} - Dev\left(\overline{\Omega}\right) \tag{10}$$

is effective number of parameters. Where $\overline{Dev}(\Omega)$ represents the expected deviance; $Dev(\Omega) = -2\log(f(\boldsymbol{D}|\boldsymbol{\Omega}))$ is the deviance function; and $f(\boldsymbol{D}|\boldsymbol{\Omega})$ is the likelihood of the parameters in Eq. (6).

Results

Simulation studies

Simulation studies were conducted to assess and compare the effectiveness of the proposed semiparametric mixed-effects model in various model settings. In these simulations, a sample of 400 individuals was considered, each having eleven equally spaced measurement times, resulting in a total of 4,400 observations. The longitudinal data was simulated using the semiparametric mixed-effects model (5). The specifications of this general model can be given as follows:

$$\begin{aligned} y_{ij} &= \alpha_1 + \alpha_2 * Z_{1ij} + \alpha_3 * Z_{2ij} + \varphi_{i1} + (\lambda_1 + \varphi_{i2}) * \phi_1\left(t_{ij}\right) \\ &+(\lambda_2 + \varphi_{i3}) * \phi_2\left(t_{ij}\right) + (\lambda_3 + \varphi_{i4}) * \phi_3\left(t_{ij}\right) + \varepsilon_{ij} \end{aligned} \tag{11}$$

where y_{ij} and Z_{pij} are the longitudinal response and binary covariates, $p = 1, 2$. $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)^T$ and $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)^T$ denote parameter vectors of the fixed effects and $\boldsymbol{\varphi}_i = (\varphi_{i1}, \varphi_{i2}, \varphi_{i3}, \varphi_{i4})^{\mathrm{T}}$ denote parameter vector of the random-effects. $\boldsymbol{\Phi}\left(t_{ij}\right) = (\phi_1\left(t_{ij}\right), \phi_2\left(t_{ij}\right),\phi_3\left(t_{ij}\right))^{\mathrm{T}}$ is a vector of natural cubic spline bases used in the regression spline method. We used eleven equally spaced time points (t_{ij} = 0, 1, 2, 3, …, 10) with percentile based knots to generate the spline bases [14, 25, 26].

To create longitudinal data with a skewed distribution, the components of the random effects $\boldsymbol{\varphi}_i$ and the error terms $\boldsymbol{\varepsilon}_i$ are simulated from a gamma distribution $\gamma(2, 1)$. These generated values are then subtracted by two [27, 28]. The vectors $\boldsymbol{\alpha} = (27.5, -5, -4)^{\mathrm{T}}$ and $\boldsymbol{\lambda} = (-9, -25, -5)$ are set accordingly.

Furthermore, $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ are generated using Bernoulli distributions with probabilities (proportions) 0.24 and 0.44, respectively.

While performing the Bayesian inference, we considered weakly informative priors for the parameters. Specifically, each component of $\boldsymbol{\alpha}, \boldsymbol{\lambda}, \boldsymbol{\delta}_\varphi$, and δ_ε was assumed to follow a normal prior distribution, N (0, 100). Furthermore, inverse Wishart $IW(0.01\mathbf{I}_4, 4)$, inverse gamma IG(0.01, 0.01), Exp(0.5), and Exp(0.5) priors are considered for $\boldsymbol{\Sigma}_\varphi, \sigma_\varepsilon^2, \rho_\varphi$, and ρ_ε, respectively.

Three MCMC chains were run using R2WinBUGS in R. Each chain consisted of 90,000 iterations, and a burn-in of 45,000 iterations was applied. After thinning, we retained a total of 4,500 posterior estimates for each parameter from each model.

In assessing convergence, Figure A.1 (Appendix A) displays the trace plots, while Figure A.2 (Appendix A) exhibits the plots of ACF (autocorrelation function) and BGR diagnostic plots of the parameters derived from the proposed semiparameteric mixed-effects model [5]. These figures clearly demonstrate convergence. In addition, none of the absolute values of Geweke's test statistics results (Appendix A) for the parameters exceeded the 95% critical value of 1.96, demonstrating strong evidence of convergence.

We computed the relative bias (RB), which indicates the extent of bias in the estimators; the 95% coverage probability (CP) to assess the accuracy of credible intervals; and the root-mean-square (RMS) error to measure the overall prediction accuracy. The results presented in Table 1 provide an evaluation of different models in terms of their posterior mean estimates, along with RB, RMSE, CP, and DIC based on simulation studies. Specifically, the evaluation focuses on semiparametric mixed-effect models (SPMEMs) with skew distributions in comparison to a Gaussian SPME model for skewed longitudinal data. The findings indicate that the proposed Bayesian SPMEMs with skew distributions (MoSTST, MoNST, and MoSNSN) outperformed the Gaussian model (MoNN). The DIC values of the skewed models MoNST, MoSNSN, and MoSTST are comparatively smaller (11,674, 12,997, and 13,249, respectively) than those of the normal model MoNN (DIC = 14,528). The two models with skew-t and skew-normal distributions of model errors and random effects (MoSTST and MoSNSN) have relatively closer DIC values. Specifically, the model with a skew-t distribution for model errors and a normal distribution of random effects (MoNST) has the smallest DIC value and exhibited better performance compared to the other models. In terms of relative bias (RB) and RMSE,

Table 1 Simulation Results: Parameter Estimates (Est) with True Value (TV), RB, RMS Error, CP, and DIC for Each Model

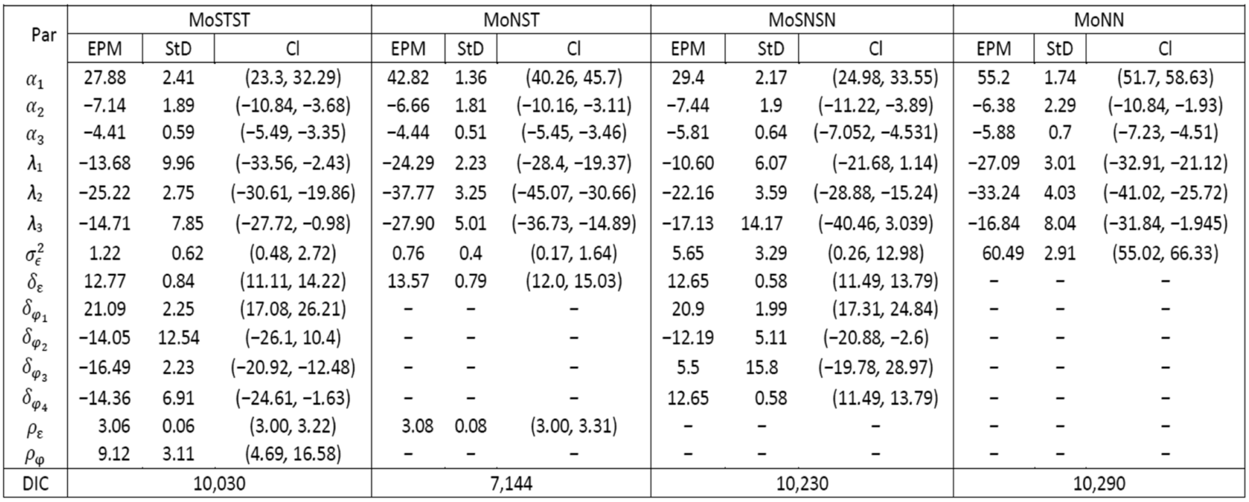
Par	**TV**	**Method**	**Models**			
			MoSTST	**MoNST**	**MoSNSN**	**MoNN**
α_1	27.5	Est	27.505	27.506	27.456	29.981
		RB	0.000	0.036	-0.002	0.090
		RMS	0.123	0.502	0.564	1.984
		CP	94.76	94.91	93.71	79.34
α_2	-5.0	Est	-4.870	-4.867	-4.891	-4.874
		RB	-0.026	-0.027	-0.022	-0.025
		RMS	0.155	0.159	0.134	0.160
		CP	67.27	66.80	72.09	53.44
α_3	-4.0	Est	-3.945	-3.919	-3.939	-3.845
		RB	-0.014	-0.020	-0.015	-0.039
		RMS	0.085	0.114	0.088	0.175
		CP	86.47	87.13	84.18	85.24
λ_1	-9.0	Est	-8.933	-8.923	-8.972	-8.851
		RB	-0.007	-0.009	-0.003	-0.017
		RMS	0.399	0.198	0.439	0.225
		CP	95.24	95.00	95.89	87.93
λ_2	-25.0	Est	-25.520	-25.454	-25.566	-25.208
		RB	0.021	0.018	0.023	0.008
		RMS	0.605	0.541	0.653	0.348
		CP	60.04	62.96	59.60	87.93
λ_3	-5.0	Est	-5.056	-4.960	-5.047	-4.856
		RB	0.011	-0.008	0.010	-0.029
		RMS	0.521	0.140	0.561	0.193
		CP	96.62	94.96	97.58	79.38
σ^2_{ϵ}	0.5	Est	0.471	0.487	0.458	4.783
		RB	0.058	0.026	0.084	8.565
		RMS	0.07	0.064	0.077	4.285
		CP	92.69	94.93	90.64	32.51
$\sigma^2_{\varphi_1}$	0.1	Est	0.087	0.113	0.103	0.197
		RB	-0.131	0.130	0.025	0.969
		RMS	0.035	0.061	0.038	0.145
		CP	96.53	94.49	95.40	86.29
$\sigma^2_{\varphi_2}$	0.3	Est	0.329	0.397	0.367	0.499
		RB	0.097	0.323	0.222	0.662
		RMS	0.172	0.237	0.209	0.378
		CP	94.29	91.29	92.96	90.36
$\sigma^2_{\varphi_3}$	0.3	Est	0.29	0.391	0.321	0.67
		RB	−0.032	0.304	0.07	1.232
		RMS	0.178	0.324	0.189	0.655
		CP	95.96	94.56	94.42	89.33
$\sigma^2_{\varphi_4}$	0.4	Est	0.4	0.662	0.476	0.815
		RB	0	0.655	0.19	1.038
		RMS	0.17	0.413	0.241	0.549
		CP	95.64	87.47	92.91	80.22
DIC			13,249.70	11,674.80	12,997.80	14,528.40

Table 2 Comparison of Parameter Estimates (PE) between the Proposed Semiparametric Mixed-Effects Model (SPPLMEM) and the Fully Parametric Mixed-Effects Model (FPLMEM)

Model	SPPLMEM			FPLMEM		
Par	PE	StD	CI	PE	StD	CI
α_1	55.29	1.58	(52.03, 58.09)	55.86	1.722	(52.28, 58.00)
α_2	− 6.38	2.36	(− 11.01, − 1.80)	− 7.58	2.17	(− 11.94, − 3.40)
α_3	− 5.69	0.67	(− 7.00, − 4.32)	− 6.25	0.67	(− 7.60, − 4.90)
σ_ϵ^2	60.19	3.03	(54.74, 66.61)	72.3	3.25	(66.14, 79.14)
DIC	10,290			10,430		

StD Standard Deviation, *CI* 95% Credible Interval

Table 3 Summary results of CKD data analysis based on four Bayesian models with different distributional specifications



Par	MoSTST			MoNST			MoSNSN			MoNN		
	EPM	StD	CI	EPM	StD	CI	EPM	StD	CI	EPM	StD	CI
α_1	27.88	2.41	(23.3, 32.29)	42.82	1.36	(40.26, 45.7)	29.4	2.17	(24.98, 33.55)	55.2	1.74	(51.7, 58.63)
α_2	−7.14	1.89	(−10.84, −3.68)	−6.66	1.81	(−10.16, −3.11)	−7.44	1.9	(−11.22, −3.89)	−6.38	2.29	(−10.84, −1.93)
α_3	−4.41	0.59	(−5.49, −3.35)	−4.44	0.51	(−5.45, −3.46)	−5.81	0.64	(−7.052, −4.531)	−5.88	0.7	(−7.23, −4.51)
λ_1	−13.68	9.96	(−33.56, −2.43)	−24.29	2.23	(−28.4, −19.37)	−10.60	6.07	(−21.68, 1.14)	−27.09	3.01	(−32.91, −21.12)
λ_2	−25.22	2.75	(−30.61, −19.86)	−37.77	3.25	(−45.07, −30.66)	−22.16	3.59	(−28.88, −15.24)	−33.24	4.03	(−41.02, −25.72)
λ_3	−14.71	7.85	(−27.72, −0.98)	−27.90	5.01	(−36.73, −14.89)	−17.13	14.17	(−40.46, 3.039)	−16.84	8.04	(−31.84, −1.945)
σ_ϵ^2	1.22	0.62	(0.48, 2.72)	0.76	0.4	(0.17, 1.64)	5.65	3.29	(0.26, 12.98)	60.49	2.91	(55.02, 66.33)
δ_ε	12.77	0.84	(11.11, 14.22)	13.57	0.79	(12.0, 15.03)	12.65	0.58	(11.49, 13.79)	−	−	−
δ_{φ_1}	21.09	2.25	(17.08, 26.21)	−	−	−	20.9	1.99	(17.31, 24.84)	−	−	−
δ_{φ_2}	−14.05	12.54	(−26.1, 10.4)	−	−	−	−12.19	5.11	(−20.88, −2.6)	−	−	−
δ_{φ_3}	−16.49	2.23	(−20.92, −12.48)	−	−	−	5.5	15.8	(−19.78, 28.97)	−	−	−
δ_{φ_4}	−14.36	6.91	(−24.61, −1.63)	−	−	−	12.65	0.58	(11.49, 13.79)	−	−	−
ρ_ε	3.06	0.06	(3.00, 3.22)	3.08	0.08	(3.00, 3.31)	−	−	−	−	−	−
ρ_φ	9.12	3.11	(4.69, 16.58)	−	−	−	−	−	−	−	−	−
DIC	10,030			7,144			10,230			10,290		

EPM Estimated Posterior Mean, *StD* Standard Deviation, *CI* 95% Credible Interval

however, the model with a skew-t distribution for both random effects and model errors (MoSTST) demonstrated superior performance. This suggests that incorporating skewness in modelling the longitudinal data and proposing a more flexible distributional assumption (skew distribution) allows for better capturing the inherent asymmetries and heavy tails present in the data, leading to more accurate estimates. Overall, these results emphasize the advantages of employing Bayesian SPMEM with skew distribution over the conventional Gaussian model, offering greater flexibility and improved performance in accurately modelling complex longitudinal data.

Results of the CKD data analysis

In this paper, we included diabetes and hypertension as binary covariates based on the real CKD dataset and three spline basis functions of time with four random-effects to model and analyse the longitudinal response, the estimated glomerular filtration rate (eGFR). Accordingly, we reformulate the general semiparametric mixed-effects model (6) as follows:

$$\begin{aligned} eGFR_{ij} &= \alpha_1 + \alpha_2 * Diabetes_{ij} + \alpha_3 * Hypertension_{ij} + \varphi_{i1} + (\lambda_1 + \varphi_{i2}) * \phi_1\left(Time_{ij}\right) \\ &\quad + (\lambda_2 + \varphi_{i3}) * \phi_2\left(Time_{ij}\right) + (\lambda_3 + \varphi_{i4}) * \phi_3\left(Time_{ij}\right) + \varepsilon_{ij} \end{aligned} \tag{12}$$

where the parameter vectors $\boldsymbol{\alpha}$, $\boldsymbol{\lambda}$, $\boldsymbol{\varphi}_i$, and $\boldsymbol{\Phi}(Time_{ij})$ are as defined as in the simulation section study. In order to obtain an approximation of the spline bases, we considered two internal knots at 9 and 25 months and two boundary knots at 0 and 96 months. The locations of these knots were determined based on the quantiles of the distribution of observed measurement time points. We proceed to analyse the CKD data using the proposed model with varying distributional assumptions, and subsequently compare and interpret the results. We begin by initially comparing the performance of two models: the proposed semiparametric (partially linear) mixed-effects model (SPPLMEM) specified in Eq. (12), and a fully parametric (linear) mixed-effects model (FPLMEM) that assumes Gaussian distributions for both the random effects and model errors. The FPLMEM is specifically defined as follows:

$$eGFR_{ij} = \alpha_1 + \alpha_2 * Diabetes_{ij} + \alpha_3 * Hypertension_{ij} + b_{i1} + (\alpha_4 + b_{i2}) * Time_{ij} + \varepsilon_{ij} \tag{13}$$

where $Time_{ij}$ denotes the observed measurement time of the longitudinal biomarkers for the ith subject at the jth visit. The results (Table 2) show that the estimates of some parameters become large from FPLMEM compared to SPPLMEM. For instance, the estimates of α_2 and σ^2 from SPPLMEM were − 6.38 and 60.19, while from FPLMEM they became − 7.58 and 72.30, respectively. In addition, in order to select the most suitable Bayesian model that accurately represents the CKD data, we also compute the deviance information criterion (DIC) [24]. Our analysis reveals that the SPPLMEM gives a lower DIC value (DIC = 10, 290) in comparison to the FPLMEM (DIC = 10, 430).

After selecting the SPPLMEM as the most suitable model that accurately represents the data, we proceed to further compare four different SPPLMEMs by taking into account different distributional specifications. For model errors and random-effects as described in the simulation study. We fitted four Bayesian semiparametric mixed-effects models to the CKD data. The MCMC setup, computations, and convergence diagnostic methods employed were identical to those described in the simulation study. Table 3. displays a summary of the data analysis results and estimates for the parameters (Par) obtained from the four models with different distributional specifications.

As shown in Table 3., the CKD data analysis results reveal that each model produces slightly varied yet statistically significant estimates of most of the parameters. When comparing the models, the findings reveal that the utilization of the 4th model (MoNN), which employs a multivariate normal distribution for random effects and model errors, may result in an overestimation of some of the parameters. Specifically, the population parametersα_1, α_2, α_3, λ_1, λ_2, and λ_3 are prone to being overestimated. Notably, as can be clearly seen, the estimated scale parameter (the variance) of model errors (σ_ε^2) is significantly larger in MoNN compared to the other models. The 3rd model (MoSNSN) also gives larger parameter estimates (e.g.,$\widehat{\sigma}_\varepsilon^2$) compared to the first two models. Furthermore, the estimated skewness parameter of the outcome eGFR (δ_ε) is significantly different from zero in the first three models: MoSTST, MoNST, and MoSNSN. Some of the skewness parameters of the random effects ($\boldsymbol{\delta}_\varphi$) are also significantly different from zero in MoSTST and MoSNSN. Thus, the significantly different from zero positive estimates of δ_ε and the subject-specific random intercept δ_{φ_1} confirm the presence of positive skewness in the longitudinal eGFR data. In other words, the non-zero estimates of the skewness parameters and relatively small estimates of the variances may indicate that the proposed Bayesian models with skew-t distribution of model errors and/or random effects (MoSTST and MoNST) fit the CKD data well. This is in line with the results of the simulation studies.

In general, the proposed models (MoSTST and MoNST) outperform and the standard MoNN. In particular, MoNST has been chosen as the best model for further in-depth interpretation and discussion of the results because it has a relatively small DIC value, despite the fact that both MoSTST and MoSNSN have some significant skewness parameter estimates for the random effects. As can be seen from the simulation studies, MoNST also has a lower DIC value. This finding, a mixed model (skewed in our case) with a normal distribution of random-effects, is consistent with the study [15].

The results of all models indicate that the variables examined in this study, namely hypertension, diabetes, and follow-up time (the spline bases), are statistically significant factors contributing to the decline of patients' kidney function. This is attributed to the negative and significant association between these covariates and the response variable, eGFR. In other words, it is evident that these covariates have a substantial association with the decrease in GFR estimates. For example, the diabetes coefficient ($\widehat{\alpha}_2 = -6.66$, 95% CI: [− 10.16, − 3.11]) from MoNST (the best-fitting model) can be interpreted as the eGFR value of a CKD patient with diabetes being reduced by 6.66 units compared to a CKD patient without diabetes, while holding the same covariates and random effects. Additionally, a hypertensive CKD patient is associated with a 4.44 unit lower eGFR value ($\widehat{\alpha}_3 = -4.44$, 95% CI: [− 5.45, − 3.46]) compared to a non-hypertensive CKD patient, with the same covariates and random effects.

Discussion

In recent years, there has been a growing emphasis in the literature on effectively modeling longitudinal data with many features. This includes giving careful consideration to the functional forms of longitudinal markers and the assumptions made about the distribution of random effects and model errors. With this in mind, the main objective of this study was to develop a flexible Bayesian mixed-effects model that addresses the problems commonly observed in longitudinal CKD data, encompassing characteristics such as skewness, non-linear effects over time, and flexible distributions for both random effects and model errors. The ultimate goal was to establish a robust statistical methodology that enables accurate and reliable inference in complex longitudinal data analysis.

We therefore proposed a Bayesian semiparametric mixed-effects model for the longitudinal response eGFR that addresses the above issues. To capture the non-linear effects of time and the flexibility of eGFR, regression splines were employed in the model. Additionally, multivariate skew distributions were incorporated to account for skewness in eGFR and to relax the assumptions about its distribution. Simulation studies were first conducted to provide a comprehensive description and evaluation of the performance of the proposed model.

We applied the proposed model by analysing data on chronic kidney disease (CKD) and assessing the relationship between covariates and estimated glomerular filtration rate (eGFR). The model comparison process in this study involved two steps. Firstly, we compared the proposed semiparametric partially linear mixed-effect (SPPLM) model with the fully parametric one (FPLM), and our results indicated that the SPPLM model outperformed the FPLM model. In the second step, we further compared four different SPPLM models, each assuming different distributions for the random-effects and model errors. As described in the data analysis and results, the SPPLM models with skew-t distribution exhibited a superior fit to the CKD data in comparison to the Gaussian SPPLM model.

The findings from the application revealed that hypertension, diabetes, and follow-up time had a substantial association with kidney function, specifically leading to a decrease in eGFR. These factors were identified as important predictors and exhibited a negative correlation with kidney function.

Additionally, the results of this study imply that when dealing with longitudinal data characterized by the aforementioned features, it is useful to incorporate nonparametric smoothing functions (splines) to capture non-linear time-effects and utilize skew distributions for model errors and/or random effects. In particular, accounting for skewness in the longitudinal data analysis by utilizing a more flexible distribution, the skew-t distribution, is crucial to handle asymmetry in the data and get unbiased results. By doing so, we can obtain less biased results and draw valid statistical inferences. Additionally, employing a flexible distributional assumption for the random-effects can lead to a more accurate explanation of subject-specific variations.

Apart from the CKD follow-up data that served as motivation, our methodology has broader applicability in cases where the longitudinal data have similar characteristics and the fundamental model requirements (or settings) are satisfied.

Conclusion

In conclusion, we have proposed a semiparametric Bayesian modeling approach with flexible distributions for complex longitudinal data. The results of the simulation and application studies have demonstrated that our work has made a significant contribution towards a more robust and adaptable methodology for modeling intricate longitudinal data. We recommend paying special attention to the specifications of the functional forms of longitudinal biomarkers and distributional assumptions of model errors when modeling complex longitudinal data.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-024-02164-y.

Additional file 1: Appendix A. Convergence diagnostic checking results. **Figure A.1.** Trace plots of some representative parameters from the chosen model. **Figure A.2.** Autocorrelation function plots (a) and BGR plots (b) of some representative parameters. **Table A.1.** Results of the Geweke's test of convergence. The computed value of the test statistic for each parameter from the chosen model. **Appendix B.** Skew Distributions.

Acknowledgements
Not applicable.

Authors' contributions

MMF, GAD, SMM, and HAE conceived and designed the study. MMF and WHB participated in the data collection. MMF and HAE carried out the data analysis and wrote the manuscript. SMM, GAD, HAE, WHB and SMK reviewed drafts of the manuscript and offered interpretation and critical comments. All authors participated in the review and approval of the final manuscript.

Funding
Not applicable.

Availability of data and materials

The actual CKD data utilized to exemplify the proposed model can be obtained from the corresponding author upon a substantial request.

Declarations

Ethics approval and consent to participate

The study was reviewed and approved by the Institutional Ethical Review Board of the University of Gondar, Ethiopia (Ref. VP/RTT/05/777/2022). All methods were carried out in accordance with relevant guidelines and regulations (the Helsinki Declaration). The Institutional Ethical Review Board of the University of Gondar also waived the need for written informed consent from individuals due to the retrospective nature of the study and its major focus on model development. The data were therefore anonymized before the analysis.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Statistics, University of Gondar, Gondar, Ethiopia. [2]Department of Epidemiology and Biostatistics, College of Public Health, University of South Florida, Tampa FL 33612, USA. [3]Department of Statistics and Actuarial Sciences, Jomo Kenyatta University of Agriculture and Technology (JKUAT), Nairobi, Kenya. [4]Department of Internal Medicine, College of Medicine and Health Sciences, University of Gondar, Gondar, Ethiopia. [5]Department of Mathematics, Debre Markos University, Debre Markos, Ethiopia. [6]School of Public Health, Jomo Kenyatta University of Agriculture and Technology (JKUAT), Nairobi, Kenya.

Received: 11 July 2023 Accepted: 29 January 2024
Published online: 01 March 2024

References

1. Stanifer JW, Muiru A, Jafar TH, Patel UD. Chronic kidney disease in low-and middle-income countries. Nephrol Dial Transplant. 2016;31(6):868–74.
2. Shiferaw WS, Akalu TY, Aynalem YA. Chronic Kidney Disease among Diabetes Patients in Ethiopia: A Systematic Review and Meta-Analysis. Int J Nephrol. 2020;2020:15.
3. Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics. 1982;1:963–74.
4. Diggle PJ, Heagerty P, Liang K, Zegger SL. Analysis of longitudinal data. 2nd ed. Oxford: Oxford University Press; 2002.
5. Hedeker D, Gibbons RD. Longitudinal data analysis. Hoboken, NJ: John Wiley & Sons; 2006.
6. Nguyen DV, S¸entu¨rk D, Carroll RJ. Covariate-adjusted linear mixed effects model with an application to longitudinal data. J Nonparametr Stat. 2008;20(6):459–81.
7. Wu H, Ding AA, De Gruttola V. Estimation of HIV dynamic parameters. Stat Med. 1998;17(21):2463–85.
8. Nelder JA, Wedderburn RW. Generalized linear models. Journal of the Royal Statistical Society: Series A (General). 1972;135(3):370–84.
9. Tang NS, Tang AM, Pan DD. Semiparametric Bayesian joint models of multivariate longitudinal and survival data. Comput Stat Data Anal. 2014;77:113–29.
10. Lu X, Huang Y. Bayesian analysis of non-linear mixed-effects mixture models for longitudinal data with heterogeneity and skewness. Stat Med. 2014;33(16):2830–49.
11. Sahu SK, Dey DK, Branco MD. A new class of multivariate skew distributions with applications to Bayesian regression models. Canadian Journal of Statistics. 2003;31(2):129–50.
12. Huang X, Li G, Elashoff RM. A joint model of longitudinal and competing risks survival data with heterogeneous random effects and outlying longitudinal measurements. Statistics and its interface. 2010;3(2):185.
13. Arellano-Valle R, Bolfarine H, Lachos V. Bayesian inference for skew-normal linear mixed models. J Appl Stat. 2007;34(6):663–82.
14. Ariyo OS, Adeleke MA. Simultaneous Bayesian modelling of skew-normal longitudinal measurements with non-ignorable dropout. Comput Statistics. 2022;37(1):303–25.
15. Molenberghs G, Verbeke G. Models for Discrete Longitudinal Data. New York: Springer Series in Statistics. 2005. p. 419–435. https://sci-hub.se/https://link.springer.com/10.1007/0-387-28980-1.
16. McCulloch CE, Neuhaus JM. Misspecifying the Shape of a Random Effects Distribution: Why Getting It Wrong May Not Matter. Stat Sci. 2011;26(3):388–402.
17. Baghfalaki T, Kalantari S, Ganjali M, Hadaegh F, Pahlavanzadeh B. Bayesian joint modeling of ordinal longitudinal measurements and competing risks survival data for analysing Tehran Lipid and Glucose Study. J Biopharm Stat. 2020;30(4):689–703.
18. Zhang H, Huang Y. Bayesian joint modeling for partially linear mixed-effects quantile regression of longitudinal and time-to-event data with limit of detection, covariate measurement errors and skewness. J Biopharm Stat. 2021;31(3):295–316.
19. Lu X, Huang Y, Chen J, Zhou R, Yu S, Yin P. Bayesian joint analysis of heterogeneous-and skewed-longitudinal data and a binary outcome, with application to AIDS clinical studies. Stat Methods Med Res. 2018;27(10):2946–63.
20. Azarbar A, Wang Y, Nadarajah S. Simultaneous Bayesian modeling of longitudinal and survival data in breast cancer patients. Communications in Statistics-Theory and Methods. 2021;50(2):400–14.
21. Goolsby MJ. National Kidney Foundation Guidelines for chronic kidney disease: evaluation, classification, and stratification. J Am Acad Nurse Pract. 2002;14(6):238–42.
22. Lee S, McLachlan GJ. Finite mixtures of multivariate skew t-distributions: some recent and new results. Stat Comput. 2014;24(2):181–202.
23. Brooks SP, Gelman A. General methods for monitoring convergence of iterative simulations. J Comput Graph Stat. 1998;7(4):434–55.
24. Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A. Bayesian measures of model complexity and _t. Journal of the royal statistical society: Series b (statistical methodology). 2002;64(4):583–639.
25. Dagne GA, Huang Y. Bayesian semiparametric mixture Tobit models with left censoring, skewness, and covariate measurement errors. Stat Med. 2013;32(22):3881–98.
26. Andrinopoulou ER, Rizopoulos D, Takkenberg JJ, Lesare E. Combined dynamic predictions using joint models of two longitudinal outcomes and competing risk data. Stat Methods Med Res. 2017;26(4):1787–801.
27. Zhang H, Huang Y. Quantile regression-based Bayesian joint modeling analysis of longitudinal-survival data, with application to an AIDS cohort study. Lifetime Data Anal. 2020;26:339–68.
28. Ferede MM, Mwalili S, Dagne G, Karanja S, Hailu W, El-Morshedy M, et al. A Semiparametric Bayesian Joint Modelling of Skewed Longitudinal and Competing Risks Failure Time Data: With Application to Chronic Kidney Disease. Mathematics. 2022;10(24):4816.

Publisher's Note