The New England Journal of Statistics in Data Science Volume 4, 5–17 (2026)
DOI: https://doi.org/10.51387/26-NEJSDS99

Modeling Disease Progression in the Presence of an Outcome-Dependent Visiting Process with Application to Cystic Fibrosis Clinical Data

Weiji SU, Xia WANG, Pedro MIRANDA-AFONSO,
Eleni-Rosalina ANDRINOPOULOU, and Rhonda D. SZCZESNIAK*

Abstract

The timing of longitudinal measurements may depend upon outcome or disease severity. In biomedical studies relying on clinical encounter data, patients often have dense, irregular collections of visit data when suffering a worse health condition. In parallel, the longitudinal measurements may be impacted by the period of irregular visiting. Ignoring the impact of the outcome-dependent visiting process when constructing a longitudinal disease progression model can produce biased results. We propose a Bayesian joint model linking a mixed-effects model for the longitudinal marker and Weibull proportional hazards model with a log frailty for the visiting process, adjusting both longitudinal marker and event processes with covariates. We examine different random effect structures and performance characterizing disease trajectory. Motivated by clinical data on cystic fibrosis lung disease, we estimate the longitudinal process for lung function decline. Individuals with lower lung function tend to have more frequent clinical visits than those with higher lung function. Simulation studies suggest that incorporating a time-dependent Gaussian process is more important for model fit than adding the survival model via joint modeling; the random intercepts model exhibits maximum bias, especially when there is an outcome-dependent visiting process.

KEYWORDS AND PHRASES: Bayesian, Frailty, Gaussian process, Irregular visits, Joint model, Longitudinal model, Medical monitoring.

1. INTRODUCTION

In longitudinal biomedical studies, each patient's process of disease progression is often measured over clinical visits; however, the frequency of and duration between visits can substantially differ between patients and within an individual patient over time. As a result, it becomes very unlikely that any two patients will have the same observed visiting times [19]. Moreover, a given patient might have irregular but dense visit patterns when experiencing a poorer health condition and have regular visits at other periods. On the other hand, patients could forego visits during periods of worsening health. The resulting missingness is not negligible and therefore must be modeled as part of the analysis. In addition, clinical patients might miss part of visits during the observational period due to a health condition, which is not negligible in the analysis. Ignoring the impact of the visiting process can yield biased results in estimating the effect of important clinical factors on the longitudinal measurements in analysis [14].

*Corresponding author: Rhonda D. Szczesniak. Division of Biostatistics & Epidemiology and Pulmonary Medicine, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA. Email: rhonda.szczesniak@cchmc.org

To overcome this challenge in longitudinal data analysis, two types of approaches have often been applied in the literature. The first type of approach is based on the idea of weighted generalized estimating equations (GEE), which is also widely used in addressing missing data when the longitudinal outcomes are missing not at random [18, 3]. The method relies on inverse-probability weighting (IPW) to estimate the outcomes through GEE models, which provide robust inference at the population level but lack individual-level estimation. A second common approach, which requires additional assumptions to gain individual-level estimation, is based on the shared-parameter joint model framework [21], which refers to jointly modeling the longitudinal measurements and the visiting process by sharing information between two outcomes. The underlying assumption in these joint models is, conditional on the shared information, the longitudinal measurements are independent of the visiting process, and the impact of irregular visits may be estimated by the shared information. Lipsitz and colleagues proposed a joint model linking a linear mixed-effects model for a longitudinal response and another model of the counting process for the number of visits [13]. The model assumed the study observed at random visit time and that the number of visits depends on the previous longitudinal response. Later

work by Sun et al. focused on the marginal mean of longitudinal measurements; they proposed a joint model, which used GEE to fit the longitudinal process and a nonhomogeneous Poisson process for the visiting process [28]. Broadly assessing strengths and weaknesses of approaches, Neuhaus et al. (2018) compared a single generalized linear mixed-effects model with random intercept and slope, a GEE with inverse intensity rate ratio and a joint model which includes the generalized linear model and a gamma-distributed visiting process for the cases of the binary and normal responses at designed times of visits [17]. They concluded that mixed-effects models have better performance in estimating the covariate effects. Gasparini et al. accounted for the informative visiting process by fitting a proportional hazard jointly with a linear mixed-effect model using a random intercept [5]. They showed that the joint model had a better fit than the GEE model with normalized inverse intensity weighting.

The aforementioned analysis of continuous longitudinal measurements has shown that mixed-effect models have a smaller bias in scenarios with informative outcomes [17]. Thus, we investigate mixed-effects models with different random effect structures and introduce a Gaussian process (GP) to explain the fluctuation and correlation over time within a subject. A parametric Weibull proportional hazard model is used to fit the visiting process. A log-normal distributed frailty term is estimated to account for the variation in visit behavior between subjects. Previous work proposed a joint model which inserts the exponential frailty term into the longitudinal mixed-effects model to share the component and link the two outcomes [5]. We employ a similar idea, assuming these two outcomes are associated through a hazard ratio (HR) of having a visit at the corresponding time point. To facilitate estimation, we utilize a Bayesian approach. Our main interest is on the longitudinal measurements and covariate:outcome associations in both the mixed-effect and survival models.

This work is motivated by studying the acquisition of lung function measurements at clinical encounters from pediatric patients with cystic fibrosis (CF) lung disease. CF is a life-threatening genetic disease which mainly affects lung function with persistent coughing and frequent lung infections. Based on the most recent reports from CF registries (2016-2018), an estimated 106,000 people live with CF worldwide [23]. In the clinical setting, monitoring lung function via the marker forced expiratory volume in 1 s of % predicted (hereafter, FEV1) is essential to survival in CF, as the primary cause of death is respiratory failure [4]. Attenuated decreases in lung function relative to patient- and/or center-level norms, clinically termed rapid decline, typically manifest during adolescence and early adulthood, but can persist throughout the lifespan [31]. Hence, the US Cystic Fibrosis Foundation guidelines recommend quarterly visits to assess pulmonary function in patients aged 6 years and older [15]; however, patients may complete clinic visits at different frequencies for various reasons. Attendance also varies within an individual patient over time. It's plausible that increased visit frequency is related to disease severity, but most CF observational studies of lung function aggregate FEV1 data. For example, past studies have aggregated subject-specific FEV1 data across visits into quarterly values prior to modeling [10, 30]. More recent work examining visit frequency as a time-varying covariate in longitudinal FEV1 modeling suggests increased visit intensity is associated with improved lung function [31], while other analyses with conventional mixed-effects modeling have found no evidence for this association [29].

Accurate estimation of CF patients' lung function can shed light on disease progression and enhance physician efforts to monitor a patient's pulmonary condition, but the extent to which estimation is confounded by visit frequency has yet to be examined. Figure 1 (a) shows observed FEV1 with highlighted lines marking individual trajectories for 6 representative subjects that were included in our analysis cohort, which is subsequently described. There is substantial between- and within-subject variability in FEV1. The patients also have different lengths of follow-up and distinct visit frequencies during the collection period. Figure 1 (b) is the histogram plot of the gap/waiting time between visits. The gap time is the duration of time (in years) between two visits. The typical (median) between-visit time is around 0.19 years or 2.3 months, shown as the dashed vertical line in Figure 1 (b). The majority of subjects have 0.5 years or 6 months between visits, and the maximum gap time is around 1.5 years. In this paper, we propose a Bayesian joint model for the longitudinal lung function and patients' visiting processes to determine the extent to which irregular visit patterns may impact lung function decline. We also fit a single mixed-effects model adjusted by the number of visits within the prior year as a time-varying covariate, to account for the impact of irregular visits, which is recommended by Sun and colleagues [28] and has been used in previous CF studies [31]. We compare the joint model with this single mixed-effects model by different Bayesian criteria, and we compare model performance in estimating the effect of clinically relevant factors, such as the coefficients of covariates, after accounting for the visiting process. We also examine the impact of model assumptions through simulation studies.

The paper is organized into the following sections. The notation and model set-ups with different random effects structures are introduced in Section 2. Bayesian priors, model estimation and comparison criteria are described in Section 3. An overview of the CF cohort data and its application, including model comparison, estimation, and inference, are presented in Section 4. Simulation studies are reported in Section 5. We conclude with a discussion of the approach and findings in Section 6. Implementation was performed in R with code available as supplemental files.

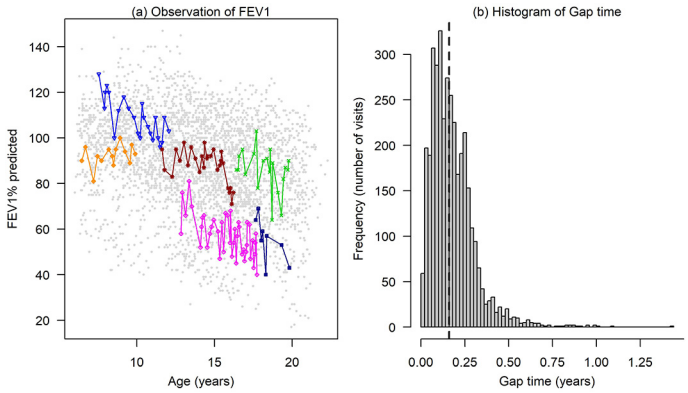


Figure 1: Observed FEV1 and Histogram Plot of Gap Time between Visits for Cystic Fibrosis Analysis Cohort. In (a), the gray dotted points are the observed lung function FEV1 over age in years, the solid-colored lines are the trajectory of FEV1 for six selected cystic fibrosis (CF) subjects; in (b), the vertical dashed line is the median value on the histogram of gap time in the CF data.

2. MODEL STRUCTURES

2.1 Survival Model

Let $t_{i,j}$ denote the follow-up time for the i^{th} subject observed at the j^{th} visit, where $i = 1, \ldots, N$ and $j = 1, \ldots, n_i$. Let $t^*_{i,j}$ denote the gap time between the j^{th} and $(j+1)^{\text{st}}$ visits, and the indicator variable $\delta_{i,j}$ denotes whether the gap time $t^*_{i,j}$ is observed, with $\delta_{i,j} = 1$ when a visit has occurred. Let $\mathbf{W}_{i,j}$ denote a $q \times 1$ vector of covariates at time $t_{i,j}$ in the survival model with the corresponding coefficient vector $\boldsymbol{\beta}$, where $\mathbf{W}_{i,j}$ include static and time-varying variables.

In order to estimate the visiting process in a longitudinal analysis, we assume that the gap time $t^*_{i,j}$ between visits follows a Weibull distribution. Conditional on the frailty term $e^{(\mu_i)}$, the Weibull proportional hazards (PH) model can be expressed as

$$\lambda_i\left(t^*_{i,j} \mid u_i\right) = \lambda_0\left(t^*_{i,j}\right) \exp\left(\mathbf{W}'_{i,j}\boldsymbol{\beta} + u_i\right), \tag{2.1}$$

where the $\lambda_0\left(t^*_{i,j}\right)$ is the Weibull baseline hazard with shape parameter κ and scale parameter $\exp(\beta_0)$ with the form

$$\lambda_0\left(t^*_{i,j}\right) = \kappa\left(t^*_{i,j}\right)^{(\kappa-1)} \exp(\beta_0);$$

the term u_i is the individual-specific random effect assumed to follow a normal distribution with mean 0 and variance σ^2_μ; that is, the frailty term $\exp(\mu_i)$ follows a lognormal distribution. The frailty term estimates the change in the hazard function for the i^{th} subject relatively to the population hazard, and thus the variance of this frailty term indicates heterogeneity associated with unobserved covariates.

In this PH survival model, the coefficient vector $\boldsymbol{\beta}$ contains the log HRs of observing an event, i.e., a patient visit; the HR formed as $\exp(\boldsymbol{\beta})$ indicates the change in the hazard function with a one-unit increase in the corresponding covariates $\mathbf{W}_{i,j}$. In our analysis, we assume that the hazard $\lambda_i\left(t^*_{i,j}\right)$ depends on both static and time-varying clinical covariates, and it depends on the previous visit time $t_{i,j-1}$.

2.2 Longitudinal Model

Let $y_{i,j}$ denote the longitudinal response, e.g., lung function measurement, for the i^{th} subject observed at time $t_{i,j}$, where $i = 1, \ldots, N$ and $j = 1, \ldots, n_i$. Let $\mathbf{X_{i,j}}$ denote a $p \times 1$ vector of covariates at time $t_{i,j}$ in the longitudinal LME model with the corresponding coefficient vector $\boldsymbol{\alpha}$, where $\mathbf{X_{i,j}}$ include static and time-dependent variables. The general form of the LME model with random effects can be expressed as

$$y_{i,j} = \mathbf{X'_{i,j}}\boldsymbol{\alpha} + \mathbf{Z'_{i,j}}\mathbf{b_i} + \epsilon_{i,j}. \tag{2.2}$$

Let $\mathbf{b_i}$ denote the vector of random effects, such as random intercept or random slope, and $\mathbf{Z_{i,j}}$ denote the corresponding design matrix; and the $\epsilon_{i,j}$ denotes the measurement error which is independent and identically distributed with $\epsilon_{i,j} \sim N(0, \sigma^2)$. In general, the random effects $\mathbf{b_i}$ are assumed to be normally distributed with mean zero and variance-covariance matrix $\mathbf{G}$. There are many potential structures for the variance-covariance matrix. In the following section, we introduce three LME models with different random effects structures.

2.2.1 Longitudinal Model with Random Intercept

To fit the continuous longitudinal response, e.g., our marker of lung function over time, we use an LME model with different covariance structures as we explain subsequently. We start with an LME with random intercepts. The model can be expressed as

$$y_{i,j} = \mathbf{X_{i,j}}'\boldsymbol{\alpha} + b_{0,i} + \epsilon_{i,j}, \tag{2.3}$$

where the term $b_{0,i}$ is the random intercept, and we assume that $b_{0,i} \sim N(0, \sigma^2_{b_0})$. All the other terms are defined as expressed previously.

2.2.2 Longitudinal Model with Random Intercept and Slope

To form an LME model with random intercepts and slopes, we add the random effect over observed time in Equation (2.3). The model can be expressed as

$$y_{i,j} = \mathbf{X_{i,j}}'\boldsymbol{\alpha} + b_{0,i} + b_{1,i}t_{i,j} + \epsilon_{i,j}, \tag{2.4}$$

where the term $b_{1,i}$ denotes the random slope over time $t_{i,j}$; the time index used in the analysis is the patient's age at a given visit (in years). We assume that the individual random intercept and random slope are correlated and follow a multivariate normal distribution (MVN) as

$$(b_{0,i}, b_{1,i})' \sim \text{MVN}\left(0, \begin{pmatrix} \sigma^2_{b_0} & \rho_b\sigma_{b_0}\sigma_{b_1} \\ \rho_b\sigma_{b_0}\sigma_{b_1} & \sigma^2_{b_1} \end{pmatrix}\right).$$

All the other terms are defined as expressed previously.

2.2.3 Longitudinal Model with Random Intercept and GP

The LME model shown in Section 2.2.2 assumes linear variation over time for each subject from the population in the longitudinal response, which may not be realistic for longer-term follow-up [32]. In this section, we assume this variation is nonlinearly changing over time and replace the random slope term by a Gaussian Process (GP) over observed time. The model can be expressed as

$$y_{i,j} = \mathbf{X_{i,j}}'\boldsymbol{\alpha} + b_{0,i} + \psi_i(t_{i,j}) + \epsilon_{i,j}, \tag{2.5}$$

where the term $\psi_i(t_{i,j})$ is from the stationary GP

$$\boldsymbol{\psi_i(t)} = (\psi_i(t_{i,1}), \ldots, \psi_i(t_{i,n_i}))'$$

with mean 0 and variance-covariance matrix Σ of two hyperparameters: marginal variance ϕ^2 and length scale ρ, written as

$$\boldsymbol{\psi_i(t)} \sim \mathrm{GP}(0, \Sigma(\phi^2, \rho)).$$

The variance-covariance matrix is assumed to be the squared exponential kernel with the form

$$k(t, t') = \phi^2 \exp\left(-\frac{1}{2\rho^2}|t - t'|^2 \right).$$

We can also view this model as equivalent to that in Equation (2.1) but assuming a nonlinear time-dependent measurement error term $\epsilon^*_{i,j}$. The term $\epsilon^*_{i,j}$ includes two parts with the form

$$\epsilon^*_{i,j} = \psi_i(t) + \epsilon_{i,j},$$

such that one part is the GP term $\boldsymbol{\psi_i(t)}$, which accounts for the correlation over time within the i^{th} subject, and the other part is a white noise term $\epsilon_{i,j}$, which is independent and identically distributed as

$$\epsilon_{i,j} \sim N(0, \sigma^2).$$

2.3 Model Structures in the Joint Model

In order to model jointly the longitudinal lung function and the visiting process, we combine the survival model in Section 2.1 with one of the LME models from Sections 2.2 as submodels in a shared-parameter joint model. Let b_i denote the vector of random effects, such as random intercept and/or random slope, and $Z_{i,j}$ the corresponding design matrix. The general form of our proposed joint model is

$$\begin{aligned} y_{i,j} &= \mathbf{X_{i,j}}'\boldsymbol{\alpha} + \mathbf{Z_{i,j}}'\mathbf{b_i} + \psi_{i,j}(t) + \gamma f(u_i, \boldsymbol{\beta} \mid \mathbf{W_{i,j}}) + \epsilon_{i,j}, \\ \lambda_i(t^*_{i,j} \mid u_i) &= \lambda_0(t^*_{i,j}) \exp\left(\mathbf{W_{i,j}}'\boldsymbol{\beta} + u_i\right), \end{aligned} \tag{2.6}$$

where $f(u_i, \boldsymbol{\beta} \mid \mathbf{W_{i,j}})$ is the shared component between the two submodels and γ is the corresponding association coefficient. The shared component can take many forms, such as the frailty term, the log HR, the HR, the time-specific hazard, or a function combining these quantities. In our analysis, we specify $f(u_i, \boldsymbol{\beta} \mid \mathbf{W_{i,j}})$ as the HR for having an observed visit, i.e.,

$$f(u_i, \boldsymbol{\beta} \mid \mathbf{W_{i,j}}) = \exp\left(\mathbf{W_{i,j}}'\boldsymbol{\beta} + u_i\right), \tag{2.7}$$

so that the longitudinal lung function measurements are associated with the HR for the visiting process. The association parameter γ and the functional form $f(u_i, \boldsymbol{\beta} \mid \mathbf{W_{i,j}})$ are specified in the longitudinal submodel, in contrast with the conventional shared-parameter joint modeling setup [21], where the association is specified in the event submodel. Under this specification, a one-unit increase in $\exp(\mathbf{W_{i,j}}'\boldsymbol{\beta} + u_i)$ (i.e., in the visit-specific HR) is associated with an average change of γ units in the expected lung function value. When $\gamma = 0$, there is no association between the two outcomes through the shared component.

For illustration purposes, assuming a linear time effect in both submodels with coefficients α_1 and β_1, respectively, the population rate of change of the longitudinal lung function under this joint model framework, $R(t)$, is of the form

$$R(t \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma, \mathbf{X_{i,j}}, \mathbf{W_{i,j}}) = \alpha_1 + \gamma\beta_1 \exp\left(\mathbf{W}'_{i,j}\boldsymbol{\beta}\right). \tag{2.8}$$

This is a dynamic rate of change due to the time-varying covariates used in the shared component in our proposed model structure. It could be a constant rate of change when applying a static shared component within subjects.

3. BAYESIAN PRIORS AND MODEL COMPARISON

3.1 Prior Estimation

The likelihood for the i^{th} subject at visit j under the proposed joint model structure, conditional on the random effects $\mathbf{b_i}$, u_i, and GP term $\psi_i(t_{i,j})$, is expressed as

$$\begin{aligned} &f(y_{i,j}, t^*_{i,j}, \delta_{i,j} \mid \mathbf{b_i}, \psi_i(t_{i,j}), u_i) \\ &= \prod_{j=1}^{n_i} f_{\text{Longi}}(y_{i,j} \mid \mathbf{b_i}, \psi_i(t_{i,j}), u_i) \times f_{\text{Surv}}(t^*_{i,j}, \delta_{i,j} \mid u_i), \end{aligned} \tag{3.1}$$

where $f_{\text{Longi}}(y_{i,j} \mid \mathbf{b_i}, \psi_i(t_{i,j}), u_i)$ is the density of the longitudinal measurement conditional on the random effects $\mathbf{b_i}$, u_i, and the GP term $\psi_i(t_{i,j})$, which follows a normal distribution with mean

$$\boldsymbol{X'_{i,j}\alpha} + \boldsymbol{Z'_{i,j}b_i} + \psi_i(t_{i,j}) + \gamma f(u_i, \boldsymbol{\beta} \mid \mathbf{W_{i,j}})$$

and variance σ^2.

The term $f_{\text{Surv}}(t^*_{i,j}, \delta_{i,j} \mid u_i)$ is the density of the gap time $t^*_{i,j}$ conditional on the frailty term u_i, which follows a Weibull distribution with shape parameter κ and scale

parameter $\exp(\beta_0)$, with the form

$$\left[\kappa t_{i,j}^{*(\kappa-1)} \exp(\beta_0) \exp(\mathbf{W}'_{i,j}\boldsymbol{\beta} + u_i)\right]^{\delta_{i,j}} \times \exp\left\{ - \int_0^{t^*_{i,j}} \kappa s_{i,j}^{*(\kappa-1)} \exp(\beta_0) \exp(\mathbf{W}'_{i,j}\boldsymbol{\beta} + u_i)\, ds \right\}.$$

The distributions of $\mathbf{b}$, u_i, and the GP term $\psi_i(t_{i,j})$ are defined as in the previous section. All the unknown parameters

$$\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\beta}', \gamma, \sigma^2_{b_0}, \sigma^2_{b_1}, \rho_b, \sigma^2_\mu, \beta_0, \kappa, \phi, \rho, \sigma^2)'$$

are estimated in the Bayesian framework through the Hamiltonian Monte Carlo (HMC) algorithm [16].

A normal prior $N(0, 100)$ is assumed for the coefficient parameters in $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and γ; in order to estimate $\mathbf{b_i}$, a Cholesky Lewandowski-Kurowicka-Joe (LKJ) correlation distribution is considered as a prior for the parameters $(\sigma^2_{b_0}, \sigma^2_{b_1}, \rho_b)'$; a truncated normal prior $N(0, 25)$ is assigned for parameter σ^2_μ and the Weibull shape parameter κ; for the GP term $\psi_{i,j}(t)$, the prior for hyperparameter ϕ^2 is a diffuse inverse-gamma distribution with shape $a = 2$ and scale $b = 1$, and the prior for hyperparameter ρ is a truncated normal $N(0, 25)$, as recommended by the Stan development team [25] and in work by Gelman [7].

Cholesky decomposition is used in estimating the covariance matrix in the GP $\psi_{i,j}(t)$ and the random effect $\mathbf{b_i}$; for the variance of measurement error, we assign a diffuse inverse-gamma distribution ($a = 2$, $b = 1$).

All results are examined for convergence using Gelman-Rubin diagnostics and trace plots of the HMC samplings [8].

3.2 Model Comparison Criteria

For model comparison, we use three Bayesian criteria. Our primary criterion is the log-pseudo marginal likelihood (LPML) [6], which is a leave-one-out-cross-validation based on the conditional predictive ordinate (CPO). We also consider the complete deviance information criterion (DIC) [24, 2] and the Watanabe-Akaike information criterion (WAIC) [34]. The model with the largest LPML, the smallest DIC and WAIC is preferred. Along with these Bayesian criteria, we also examine the predictive accuracy in estimating the longitudinal measurements by root mean squared error (RMSE) and mean absolute deviation (MAD). A model with minimum RMSE and MAD is preferable.

4. CYSTIC FIBROSIS CLINICAL STUDY

4.1 Data Introduction

The analysis cohort for this work consists of data from the Cystic Fibrosis Care Center at Cincinnati Children's Hospital Medical Center (2012 to 2017). Ethics approval for this study was obtained by the local institutional Review Board. Patients included in the analysis cohort had a valid CF diagnosis and were at least 6 years of age, the age at which pulmonary function testing standards can be reliably performed by children according to clinical guidelines. We utilized data from patients who had at least two recorded visits, in order to minimally ascertain visit behavior for modeling. The cohort consisted of 197 subjects who contributed a total of 3,652 records in the application. Patients' clinical and demographic characteristics at the initial visit are reported in Table 1.

Table 1. Characteristics of Cystic Fibrosis Cohort at the First Observed Visit (N = 197 Patients).

Characteristics	Analysis Cohort
Age in years, mean (SD)	12.03 (4.48)
Follow-up years per patient, mean (SD)	3.12 (1.65)
F508del alleles, n (%)	
Heterozygous	71 (36.04)
Homozygous	116 (58.88)
None/Unknown	10 (5.08)
Gender (Male), n (%)	94 (47.72)
FEV1 % predicted, mean (SD)	90.59 (19.23)
Medicaid insurance use, n (%)	70 (35.53)
Microbiology	
Pa, n (%)	42 (21.32)
MRSA, n (%)	37 (18.78)
CF-related diabetes mellitus, n (%)	42 (21.32)
Gap time in years, mean (SD)	0.25 (0.19)
Number of visits/year, mean (SD)	5.52 (2.70)
Died, n (%)	2 (1.02)

There were slightly more females than males, and the average gap time between the initial visit and the second visit is 0.25 years, around 4 months, with an average FEV1 of 90.6% predicted. Among all the observations, the average observed age is 13.9 years old, ranging from 6 to 22.3 years old. The average FEV1 is 88.0% predicted, and the range is from 17.0 to 147.0. The average gap time between two adjacent visits is 0.19 years, roughly 2.3 months, and the range of the gap time is from 0.01 to 1.4 years.

4.2 Model Application and Comparison

We apply the single longitudinal LME models (Section 2.2) and the joint models (Section 2.3). The following 9 models are investigated:

1. *Model* '*Int_null*': a single longitudinal LME model with random intercept and adjusted by the age at visits, gender, diagnosis of CF-related diabetes mellitus (CFRD), infection with Methicillin-resistant Staphylococcus aureus (MRSA), infection with Pseudomonas aeruginosa (Pa) and whether patients use Medicaid insurance.
2. *Model* '*Int_nvisit*': extension of the *Int_null*, including the number of visits within the prior calendar year and the corresponding interaction term with age at visit as covariates.

3. *Model* '*Int_JM*': a joint model which includes *Int_null* as the longitudunal submodel and a Weibull proportional hazards (Section 2.1) as the survival submodel. The survival submodel is adjusted by the following covariates: previous visit time, gender, CFRD, MRSA, Pa and whether patients use Medicaid insurance.
4. *Model* '*Slope_null*': based on *Int_null*, use random intercept and random slope as the random structure.
5. *Model* '*Slope_nvisit*': based on *Slope_null*, add the number of visits within the prior one calendar year and the corresponding interaction term with age at visits as covariates.
6. *Model* '*Slope_JM*': a joint model which includes *Slope_null* as the longitudinal submodel and a Weibull proportional hazards (Section 2.1) as the survival submodel.
7. *Model* '*GP_null*': based on *Int_null*, include a GP term over age at visits.
8. *Model* '*GP_nvisit*': based on *GP_null*, add the number of visits within the prior one calendar year and the corresponding interaction term with age at visits as covariates.
9. *Model* '*GP_JM*': a joint model which includes *GP_null* as the longitudinal submodel and the Weibull proportional hazards (Section 2.1) as the survival submodel.

Models 1), 4) and 7) represent the longitudinal analysis that has no ability to deal with the effect of the irregular visiting. These models have the same covariates with coefficient term $\mathbf{X}'_{i,j}\boldsymbol{\alpha}$, which is expressed as

$$\alpha_0 + \alpha_1 \text{Age}_{i,j} + \alpha_2 \text{Gender}_i + \alpha_3 \text{CFRD}_{i,j} + \alpha_4 \text{MRSA}_{i,j} + \alpha_5 \text{Pa}_{i,j} + \alpha_6 \text{Insurance}_{i,j}.$$

Models 2), 5), and 8) use the number of visits information as time-varying covariates to explain the irregular visiting. The covariates with coefficient term $\mathbf{X}'_{i,j}\boldsymbol{\alpha}$ are expressed as

$$\alpha_0 + \alpha_1 \text{Age}_{i,j} + \alpha_2 \text{Gender}_i + \alpha_3 \text{CFRD}_{i,j} + \alpha_4 \text{MRSA}_{i,j} + \alpha_5 \text{Pa}_{i,j} + \alpha_6 \text{Insurance}_{i,j} + \alpha_7 \text{nvisit}_{i,j} + \alpha_8 (\text{Age}_{i,j} \times \text{nvisit}_{i,j}).$$

Models 3), 6), and 9) use the joint model to monitor the patients' visiting process. The longitudinal submodel has the same covariates with coefficient term $\mathbf{X}'_{i,j}\boldsymbol{\alpha}$ as in model 1). The covariate structure with coefficient term $\mathbf{W}'_{i,j}\boldsymbol{\beta}$ for the survival submodel is expressed as

$$\beta_1 t_{i,j-1} + \beta_2 \text{Gender}_i + \beta_3 \text{CFRD}_{i,j} + \beta_4 \text{MRSA}_{i,j} + \beta_5 \text{Pa}_{i,j} + \beta_6 \text{Med_Insurance}_{i,j}.$$

In this application of CF data, we assume that the population lung function has a linear decline over time in the LME models 1) and 4). Thus, the population rate of change for lung function is assumed constant with value α_1. It is also known that the rate of change in lung function in CF is constant for earlier ages. In other applications, people could assume different forms of the population mean based on the particular disease progression or the longitudinal response evolution, leading to a different form of population rate of change.

All 9 models are fitted under Bayesian inference using the HMC sampler and all results passed the convergence diagnostic (with $\hat{R} < 1.05$). The trace plots for the HMC samplings are shown in Figure S1 in the Supplementary Material. The comparison results are shown in Table 2 based on the criteria LPML, DIC, WAIC, RMSE, and MAD.

Table 2. Model Comparisons from the Cystic Fibrosis Cohort Application.

Model[a]	LPML	DIC	WAIC	RMSE	MAD
Rand int					
Int_null	−12690.22	25360.13	25368.51	7.57	4.62
Int_nvisit	−12689.39	25358.25	25367.19	7.56	4.60
Int_JM	**−12559.25**	**25085.03**	**25104.62**	**7.25**	**4.47**
Rand int & slope					
Slope_null	−12441.44	24836.94	24863.76	6.86	4.19
Slope_nvisit	**−12438.67**	**24832.54**	**24862.03**	**6.85**	4.19
Slope_JM	−12438.98	24840.62	24866.82	6.88	**4.17**
Rand int & GP					
GP_null	−12345.77	**24621.30**	**24660.61**	6.29	3.78
GP_nvisit	−12346.33	24622.89	24661.44	**6.28**	**3.76**
GP_JM	**−12342.78**	24621.32	24661.32	6.29	3.79

[a]Model abbreviations are specified in Section 4.2. 'Int_', 'Slope_' and 'GP_' refer to random intercept, random intercept and slope, or random intercept and GP term; 'null', 'nvisit', and 'JM' indicate adjustment by age, gender, CFRD, MRSA, Pa, Medicaid insurance; additionally, 'nvisit' adjusts for number of visits and interaction; 'JM' is under a joint model framework. Abbreviations: DIC = deviance information criterion; LPML = log-pseudo marginal likelihood; MAD = mean absolute deviation; RMSE = root mean square error; WAIC = Watanabe-Akaike information criterion.

We first compare the models within the same random effect structure, in order to check the model improvement according to how irregular visiting is accounted for. Among the models with random intercept, there is a gradual improvement from model '*Int_null*' to '*Int_nvisit*' and '*Int_JM*' since the LPML increases, DIC, WAIC, RMSE and MAD decreases followed the order. In the models with random intercept or GP over time, there is an improvement from '*null*' model to '*nvisit*' or '*JM*' model, while '*nvisit*' and '*JM*' models perform similar since the difference of all criteria is negligible. These results indicate the irregular visiting has an impact on lung function evolution. Furthermore, for the models with random slope or GP over time, this structure accounts for part of the impact of the irregular visiting.

Next, we compare the models across the different random effects structures. A large difference is observed between models from each of the two random effects structures, and the largest improvement is found when moving from models with random intercepts to those models with

both random intercepts and slopes. The models with random intercept and GP have the best performance of those considered. We also notice, in this CF application, that the joint models perform slightly better than other single longitudinal models with the same random effects structure; however, the improvement by joint models is not remarkable, especially compared with the models 2), 5), 8) including number of visits information as time-varying covariates. The joint model with random intercept and GP has the best performance in this CF analysis in terms of the largest LPML value. Residual diagnostic plots of observed FEV1 vs. the fitted values, the standardized (std) residuals vs. the fitted values, quantile-quantile (QQ) plot and the density plot of the std residuals are presented in Figure S2 of the supplementary material. The plots show that residuals are independently distributed and have constant variance. The normal QQ plot is roughly linear with a light lower tail and the density is symmetric, which indicates the residual is normally distributed.

4.3 Posterior Inferential Results

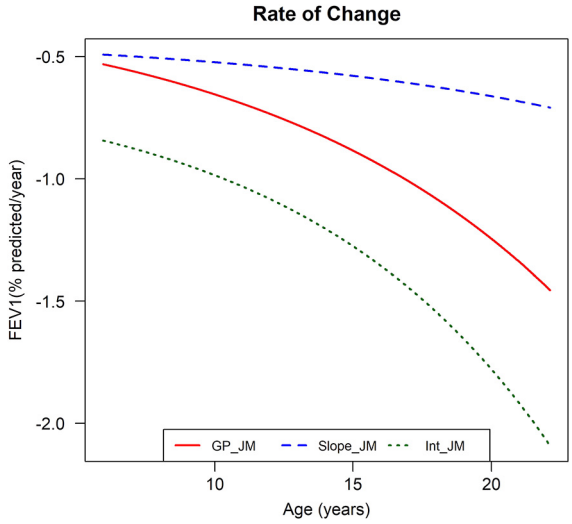


Figure 2: Rate of Change in Lung Function (FEV1) for Cystic Fibrosis Population. The rate of change over age in years by three joint models applied to cystic fibrosis data with the solid red line for GP_JM, dashed blue line for $Slope_JM$ and green dotted line for Int_JM (model structures are defined in Section 4.2).

Parameter estimates and their corresponding 95% highest posterior density credible intervals, hereafter referred to as CrIs, under the joint model with GP are shown in Table 3, and the joint model with a random slope and random intercept is reported in Table S1. The posterior mean value of FEV1 at baseline (i.e., at age 0 years) is estimated to be 102.11% and rate of change of −0.25% predicted/year. Higher initial FEV1 % predicted values (e.g., above 100) are common earlier in life for people with CF, but trajectories tend to decrease significantly during adolescence and early adulthood [12]. The rate of change in lung function for the three joint models is shown in Figure 2. The rate of change is negative, implying that lung function monotonically declines from age 6 to 20 years old, and the decline ranges from −0.5 to −2. The CF patients infected with CFRD typically experienced a 8.54% predicted drop in FEV1, and patients infected with Pa tended to have a 2.75% predicted drop. The coefficient of association implies a negative relationship between a patient's lung function and the HR of having a visit; that is, a one-unit increase in the HR of having a visit corresponds to lung function decreasing by 1.99% predicted. This indicates that patients who are likely to have more frequent visits have lower FEV1. This result is consistent with the models that include the number of visits within the calendar year, 'nvisit', as a time-dependent covariate. The parameter estimation of the model with 'nvisit' and 'age×nvisit' is in Table S2 in the supplementary file. Given the potential for collider bias with assessing effects related to past visit frequency, we omit interpretation of coefficients that include 'nvisit' effects [11].

Table 3. Joint Model Parameter Estimates with Gaussian Process Applied to the Cystic Fibrosis Cohort.

Parameter	Est (95% HPD interval)
LME with GP Submodel (Lung Function)	
α_0, Intercept	102.11 (96.38, 107.98)
α_1, Age	−0.25 (−0.71, 0.22)
α_2, Gender (Male)	0.31 (−4.06, 4.51)
α_3, CF-related diabetes mellitus	−8.54 (−13.84, −3.43)
α_4, MRSA	−0.71 (−2.06, 0.7)
α_5, Pa	−2.75 (−3.88, −1.61)
α_6, Medicaid insurance	−1.65 (−6, 2.83)
ϕ^2, GP marginal variance	5.06 (4.4, 5.83)
ρ, GP length scale	1.05 (0.75, 1.36)
$\sigma^2_{b_0}$, Variance of intercept	15.93 (14.26, 17.80)
σ^2, Variance of residual	6.76 (6.56, 6.94)
γ, Association	−1.99 (−3.47, −0.67)
Survival Submodel (Visiting Process)	
β_0, Intercept	1.71 (1.41, 2.02)
β_1, Age at previous visit	0.09 (0.07, 0.10)
β_2, Gender (Male)	−0.22 (−0.42, −0.02)
β_3, CF-related diabetes mellitus	0.09 (−0.18, 0.33)
β_4, MRSA	0.03 (−0.08, 0.12)
β_5, Pa	−0.15 (−0.29, −0.05)
β_6, Medicaid insurance	0.23 (0.02, 0.45)
κ, Weibull shape	1.89 (1.84, 1.94)
σ^2_μ, Variance of u_i	0.66 (0.58, 0.76)

Abbreviations: GP = Gaussian process; HPD = highest posterior density; LME = linear mixed effects model.

In the visiting process, the reference level is defined in the survival submodel as female, with no infection of CFRD, MRSA, or Pa, and not using Medicaid insurance. The Weibull shape parameter is 1.89 (95% CrI: 1.84, 1.94), which means the HR of observing a visit increases as the gap time (between visits) increases. Older patients are more likely to have visits than younger patients. Males are more likely to have fewer visits than females. Patients infected with CFRD or MRSA are likely to have more visits. Patients infected with Pa tend to have fewer visits than others. Patients who use Medicaid insurance are likely to have 0.23

(95% CrI: 0.02, 0.45) more visits than those who use other types of insurance. Figure 3 shows the probability of having a visit over the gap time. The joint model estimate for the median gap time is 0.19 years or 2.3 months. The probability is 80.5% for CF subjects to have a visit within 0.3 years (roughly 3.6 months) since their previous visit, and 68.6% have a visit within 0.25 years, which is the CF Foundation's recommended frequency.

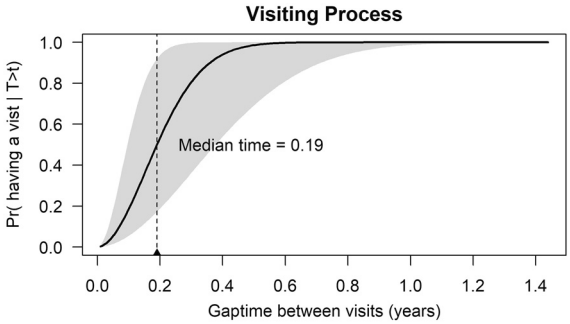


Figure 3: Probability of Having a Visit over the Gap Times between Visits for Cystic Fibrosis Cohort by Joint Model with GP. The solid black line is the estimated probability of having a visit, the grey shaded area is the 95% credible interval and the vertical dashed line is the median gap time between visits.

We present the predicted values of FEV1 using the proposed joint model for three CF subjects over different age periods and a range of lung function in Figure 4, where Subject 1 is a boy followed from 6.5 to 10 years old and with FEV1 from 81 to 100; Subject 2 is a girl followed from age 8.8 to 13.8 years old and with FEV1 ranging from 37 to 73; Subject 3 is a boy followed from 11.6 to 16.2 years old and with FEV1 from 71 to 98. The plots show subjects' lung function (FEV1) over time with grey points as observed FEV1, and the solid lines are the predicted values shaded with the 95% CrI. We find that the predicted values capture the individualized trends in lung function and the subjects' non-linear fluctuation is reasonably estimated by GP term from the proposed joint model.

5. SIMULATION STUDIES AND RESULTS

We conduct simulation studies in order to examine the performance of the proposed model and the ability of the proposed model to recover the true values of parameters. We investigate four scenarios combining the two factors, the frequency (sparse or dense) of visits and the extent of outcome-dependent or independent visits. The data with dense visiting mimics patients with a higher number of visits but shorter gap times between visits, compared to patients with sparse visits within the same length of time. The data with outcome-dependent visits mimics the outcome-dependent visiting process, compared to outcome-independent visit.

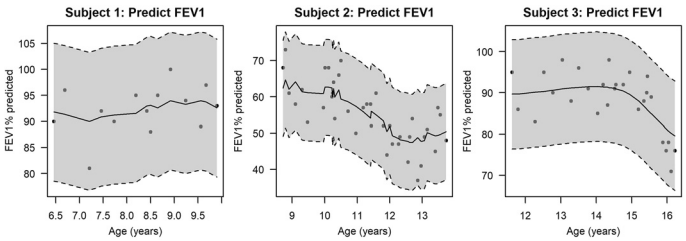


Figure 4: FEV1 Trajectories and Patients' Visiting Processes for three Selected Subjects with Cystic Fibrosis from the Joint Model with GP. Each of the three plots describes the longitudinal lung function in terms of FEV1% predicted, where the grey dots are the observed values, the black solid line is the predicted values, and the grey shaded area is the 95% credible interval. Estimated lung function trajectories are adjusted for the visiting process.

For each simulated dataset, we generate the observations for $N = 100$ subjects. Then we fit the data with the 9 models defined in Section 4.2, which include the single LME models with or without 'nvisit' and joint models under different random effects structures. Given the computational intensity, we replicated $M = 50$ simulated datasets under each scenario. All HMC chains passed convergence diagnostics.

In each simulation, we generate each patient's visit times using a Weibull distribution, assuming shape parameter of 1.85 and scale parameter of 1 for the scenario of sparse visits (i.e., wider gap time), and scale parameter 0.1 for the scenario of dense visits (i.e., narrower gap time). Average gap time between visits was either 1 or 0.3 years; these annual and quarterly visit scenarios correspond to less frequent vs. guidelines-based visiting processes in the CF example. We simulate two variables, one is a time-varying variable, representing age at visits; initial age ranges from 6 to 20 years old generated from $Uniform(6, 20)$; the other is a static variable, representing gender, which is generated from a $Bernoulli(0.5)$. The true regression parameters of age and gender in the continuous longitudinal response $y_{(i,j)}$ are assigned as $(\alpha_1, \alpha_2)' = (-0.5, 1)'$ and in the survival model as $(\beta_1, \beta_2)' = (0.1, -0.2)'$. The intercept α_0 is 85. We generate the random intercept $b_{(0,i)}$ from the distribution $N(0, 1^2)$. The GP terms $\psi_{(i,j)}$ are generated from a $GP(0, \Sigma(4, 0.5))$. The log of the frailty term u_i is generated from a $N(0, (0.5)^2)$. We set the association parameter equal to -2 for the outcome-dependent scenario and equal to 0 for the outcome-independent scenario. We then generate the longitudinal continuous response $y_{(i,j)}$ from a normal distribution with mean response function: $\alpha_0 + \alpha_1 \text{Age} + \alpha_2 \text{Gender} + b_{(0,i)} + \psi_{(i,j)}(t) + \gamma \exp(\beta_1 \text{Age} + \beta_2 \text{Gender} + u_i)$ and the variance σ^2 is set to 1.

After $M = 50$ replications of each simulation, we summarize the results by the percentage of best performances based on each of the comparison criteria LPML, DIC, WAIC, RMSE, and MAD among all the models compared. A model

Table 4. Simulation Results from $M = 50$ Replications for Data with Longitudinal Responses Dependent ($\gamma = -2$) on Dense Visits.

Model	Intercept			Age			Gender			Association		
	C_{95}	Bias (SD)	RMSE	C_{95}	Bias (SD)	RMSE	C_{95}	Bias (SD)	RMSE	C_{95}	Bias (SD)	RMSE
Random intercept												
Int_null	0	9.72 (1.39)	9.90	0	1.40 (0.36)	1.41	46	2.70 (1.14)	3.00	—	—	—
Int_nvisit	76	1.68 (1.10)	2.07	0	0.78 (0.38)	0.79	42	2.60 (1.13)	2.90	—	—	—
Int_JM	78	2.84 (2.53)	6.95	84	0.08 (0.26)	0.11	98	0.65 (0.94)	1.09	94	1.47 (1.90)	3.87
Random intercept and slope												
Slope_null	2	5.79 (1.13)	5.93	0	1.00 (0.30)	1.00	66	1.71 (0.89)	1.88	—	—	—
Slope_nvisit	22	3.97 (1.19)	4.22	0	0.89 (0.32)	0.90	68	1.64 (0.90)	1.82	—	—	—
Slope_JM	86	0.83 (0.83)	1.08	94	0.06 (0.21)	0.08	94	0.41 (0.60)	0.54	94	0.21 (0.40)	0.27
Random intercept and GP												
GP_null	4	5.73 (1.23)	5.92	0	1.15 (0.32)	1.15	46	2.66 (1.14)	2.95	—	—	—
GP_nvisit	28	3.59 (1.25)	3.91	0	0.99 (0.35)	1.00	46	2.61 (1.13)	2.90	—	—	—
GP_JM	90	0.69 (0.72)	0.86	92	0.05 (0.20)	0.07	96	0.38 (0.53)	0.48	100	0.21 (0.40)	0.26

See Section 4.2 for model abbreviations and descriptions. Abbreviations: RMSE = root mean square error; SD = standard deviation.

with a larger value of this percentage means that this model has more chances to perform better in the simulated scenario. To check the ability to recover the true parameters, we used three commonly used criteria, including the coverage probability of 95% credible intervals (CrIs), C_{95}, which is calculated as

$$C_{95} = \frac{\sum_{m=1}^{M} I_m(95\% \text{ CrI contains true parameter value})}{M},$$

the bias, which is calculated as

$$\text{Bias} = \frac{\sum_{m=1}^{M}(\hat{\theta}_m - \theta)}{M},$$

with $\hat{\theta}_m$ denoting the posterior mean estimate of the unknown parameter θ at the m^{th} replication, and RMSE, which is calculated as

$$\text{RMSE} = \left\{ \frac{\sum_{m=1}^{M}(\hat{\theta}_m - \theta)^2}{M} \right\}^{1/2}.$$

A parameter with a larger C_{95}, smaller bias, and smaller RMSE indicates a more accurate estimation of the true parameter in the model.

The simulation results for the scenarios of patients with dense visits and longitudinal measurements dependent on the vising process are shown in Table 4, and the independent case is reported in Table 5. Model selection results corresponding to the dependent and independent cases are reported in Tables S3 and S4, respectively. Results for the scenarios of patients with sparse visits are reported in Tables S5-S6.

We first check the models by the order '*_null*', '*_nvisit*', and '*_JM*' in all the scenarios. When the responses are dependent on patients' visits, the '*_null*' models across different random effect structures do not provide good estimation for 'Intercept' and time-varying covariates, such as 'Age', and the estimation for static covariates, such as 'Gender' is slightly better but still not satisfactory. When the responses are independent of patients' visits, the '*_null*' models perform better and provide reasonable estimates of covariate effects. The '*_nvisit*' models are similar to '*_null*' estimating 'Age' and 'Gender' and slightly better in estimating 'Intercept', especially in the models only with random intercept. We notice that, in the scenarios of outcome-dependent responses shown in Table 4 and Table S5, although the '*_nvisit*' model does not recover the true parameters well, the bias and RMSE in the '*_nvisit*' models are smaller than the '*_null*' models. This indicates that, in the analysis of outcome-dependent responses, adding the number of visits within a calendar year into models can help improve the model performance in estimating the effect of covariates and reduce the bias.

Despite aforementioned improvements, many of the estimates are still not satisfactory according to coverage (i.e., low C_{95}) and notably high bias, especially for time-varying covariates. More broadly, the consequences of misspecifying the random effects structure differ markedly between the dependent and independent scenarios. Under the independent scenario, even the clearly misspecified '*Int_*' and '*Slope_*' models do not deviate substantially from the true model and can sometimes offer comparable performance in terms of bias and RMSE (for example, *Slope_null* in Tables 5 and S6). In contrast, under the dependent scenario, the impact of misspecification becomes much more severe. The worst observed scenario is the intercept under the simplest model, *Int_null*, in which coverage drops to 0%, compared with 78% under the independent scenario.

The '*_JM*' models perform the best in estimating 'intercept', 'Age' and 'Gender' effects, compared to the '*_null*' and '*_nvisit*' models within each random effect structure in all four simulation scenarios. This result is especially enhanced in the scenarios when the longitudinal responses depend on the visiting process. For the association parameter $\gamma = -2$, '*_JM*' models provide accurate estimates, recovering the true value with small bias, which implies the joint models can detect the association between the longitudinal measurements and patient visits.

Table 5. Simulation Results from $M = 50$ Replications for Data with Longitudinal Responses Independent ($\gamma = 0$) of Dense Visits.

Model	Intercept			Age			Gender			Association		
	C_{95}	Bias (SD)	RMSE	C_{95}	Bias (SD)	RMSE	C_{95}	Bias (SD)	RMSE	C_{95}	Bias (SD)	RMSE
Random intercept												
Int_null	78	0.39 (0.55)	0.49	90	0.02 (0.13)	0.03	92	0.18 (0.38)	0.23	—	—	—
Int_nvisit	82	0.70 (0.78)	0.92	84	0.04 (0.20)	0.06	92	0.18 (0.39)	0.23	—	—	—
Int_JM	82	0.41 (0.55)	0.50	94	0.03 (0.14)	0.04	94	0.19 (0.39)	0.24	90	0.04 (0.20)	0.06
Random intercept and slope												
Slope_null	94	0.44 (0.56)	0.54	98	0.03 (0.13)	0.03	92	0.19 (0.39)	0.24	—	—	—
Slope_nvisit	80	0.76 (0.83)	1.03	82	0.05 (0.21)	0.06	92	0.19 (0.39)	0.24	—	—	—
Slope_JM	96	0.46 (0.55)	0.55	100	0.03 (0.13)	0.04	94	0.20 (0.40)	0.25	98	0.04 (0.19)	0.05
Random intercept and GP												
GP_null	98	0.33 (0.49)	0.40	96	0.02 (0.12)	0.02	92	0.15 (0.38)	0.21	—	—	—
GP_nvisit	90	0.62 (0.66)	0.76	94	0.04 (0.16)	0.05	92	0.15 (0.38)	0.21	—	—	—
GP_JM	98	0.35 (0.47)	0.41	100	0.02 (0.12)	0.03	90	0.16 (0.39)	0.22	98	0.03 (0.17)	0.05

See Section 4.2 for model abbreviations and descriptions. Abbreviations: RMSE = root mean square error; SD = standard deviation.

Comparison between the dependent and independent scenarios further highlights the advantages of the '*_JM*' models. Under the independent scenario, where the data generating process (DGP) follows the *GP_null* model, the *GP_null* model and the more complex *GP_nvisit* and *GP_JM* models yield similar inference in terms of bias and RMSE. Among the three GP-based models, '*GP_JM*' is most frequently identified as the best performing model (Table S4). Under the dependent scenario, where DGP is based on the '*GP_JM*' model, any model that excludes the joint modeling component results in zero coverage for the coefficient of the time-dependent covariate and poor coverage for the other two coefficients. Overall, while the '*_JM*' models perform well in both dependent and independent scenarios, model performance deteriorates rapidly when dependence is ignored. Although the '*_JM*' model is more complex, its additional flexibility allows it to better accommodate key features of the data.

This pattern also holds for the GP components of the random effects. Although our study does not include data replicates generated under a DGP without the GP component, Su et al. demonstrated that the GP model performs well for independent and identically distributed (iid) random effects, whereas fitting data that contain GP-structured random effects with an iid-only model leads to substantial loss of accuracy [27]. The consequences of such misspecification are similarly evident in our comparison of '*Int_null*', '*Slope_null*' and '*GP_null*' in Table 5.

By comparing the models across the different frequencies of visits, we notice that both '*_null*' and '*_nvisit*' models perform better in the sparse visits than the dense visits. Meanwhile, '*_JM*' models perform slightly better in the dense visits than in the sparse visits. By comparing across the random effect structures, when the longitudinal responses are dependent on the visits, we found a big improvement in parameter estimation of 'Intercept' and 'Gender' by including a random slope or GP term effects. The joint models provide a better estimation of 'Age' across the random effect structures by checking C_{95}, bias and RMSE, but it is worth noting that running a single simulation replicate could take up to four hours for the most complex setting. This approximate run time is consistent with the computational intensity that we have observed in prior simulation studies of similar model structures [26]. When the responses are independent of visits, the improvement in estimating the 'Intercept' and 'Gender' is reduced. This evidence implies that, when there is an outcome-dependent response, the models with random slope or the GP term partially account for the visiting effect and improves the model estimation of the intercept and static covariates. Meanwhile, by estimating the visiting process, the proposed joint models appear to take care of the effect of dependency on the outcome.

6. DISCUSSION

In this paper, we have proposed a joint model for the continuous longitudinal measurement in the presence of an outcome-dependent visiting process. We examined the model performance among three different random effects structures with an application to CF data and simulation studies. Based on our simulation studies, we conclude that the joint models with a visiting process provide more accurate parameter inference in longitudinal analysis, compared to a single model structure. Furthermore, this is evident no matter how the longitudinal response pattern varies with respect to patients' visiting process. In the studies with dense visits, this proposed joint model is preferred over the models with random intercept or random slope. We found that models adjusted by patient's visiting behavior, such as the number of visits within the prior year, can help improve the model performance and reduce the bias in parameter estimation when the longitudinal measurements are dependent on patient's visits; however, the coefficient estimate of a given time-varying covariate needs to be carefully examined in these settings. The random slope structure and a GP term can explain some part of patients' visiting effect and improve the overall model fitting, but may not aid parameter estimation. We also observed that an improper assumption in the random effect structure impairs the model performance.

In the application to the CF study, the random intercept models do not work well compared with other models. Furthermore, the models with random intercept and GP over time have the best performance of those considered. This result is likely due to the large heterogeneity in CF data [30, 32, 1], which can be estimated by the nonlinear correlated structure from the GP term. The improvement from a single longitudinal model to the proposed joint model is large when only using a random intercept in modeling the CF FEV1 data. However, when adding the random slope or a GP term, though the joint model is the best fitting in terms of LPML, this improvement is reduced. From model estimation, the lung function measurement FEV1 and the patient's visiting process have a negative association with estimation -1.99 $(-3.47, -0.67)$, which shows that irregular visiting does affect the FEV1 trajectory estimation. This indicates patients who have lower FEV1 are likely to have more frequent visits. This result could be because patients at a certain period need more frequent visits due to a sudden lung function decline or more severe patients are recommended with more visits than usual. The average gap time between visits in this CF data is 0.18 years (2.16 months) with a range from 0.1 to 1.44 years; from the estimation of the joint model, the median gap time is estimated as 0.19 years (2.3 months).

Covariates were selected for each submodel in our real-data application based on information from prior CF studies. Our proposed joint modeling framework allows the same covariates to be included in both the longitudinal and event submodels, but additional care is needed when interpretating their effects. In the longitudinal submodel, a given covariate serves as a predictor of the lung function trajectory, describing systematic differences in the mean level and evolution of lung function. In the hazard submodel, the same covariate can exert a direct effect on the instantaneous risk of an observed visit that is not mediated by the longitudinal lung function. In our specification, the association structure is introduced through the longitudinal submodel, which captures how covariates affect lung function indirectly through their impact on the HR for having an observed visit. Consequently, when a covariate such as sex is included in both submodels, its overall effect on lung function decomposes into a direct effect (through the fixed effects in the longitudinal submodel) and an indirect effect mediated by the visiting process; that is, through $f(u_i, \boldsymbol{\beta} \mid \mathbf{W}_{(i,j)})$. The association parameter γ summarizes the strength and direction of this indirect effect of the visiting process on lung function, conditional on the covariates and random effects. Adaptations to allow causal inference under the proposed model would require a different specification of the functional form and shared parameter to mimic the "reverse" shared parameter setup assumed in our proposed joint model [33].

Despite the insights gained on modeling properties and the application, our approach has some limitations. In the CF application, observations may be left truncated due to calendar time, since our visit records began in 2012. In larger analyses with CF registries, left truncation has been accounted for by using age at diagnosis or birth cohort [30, 32]; however, we did not consider these covariates, given the sample size of our center-level cohort. In addition, observations are truncated at the time of death or lung transplant, and we assume these dropouts are not informative. For pediatric applications, however, there are few deaths or lung transplants during visits (Table 1). Informative dropout would need to be considered if the CF analysis cohort includes older patients. Acute drops in a patient's lung function may precede these events; thus, it might be helpful to consider them in estimating lung disease progression.

Extensions of our current work with application to a larger data source could be considered to account for the aforementioned types of censoring. For example, building the model as a multivariate joint model could account for a terminal event, such as death. We assumed in the CF application that population-level lung function declines linearly and that the GP captures nonlinear progression within an individual patient. This assumption is appropriate in the CF context, since our analysis cohort is pediatric. However, in longitudinal analysis, long-term sequences of visits, such as those over a life span, population-level linearity may not be a valid assumption. Potential extensions of our application are to use a nonlinear form for the mean structure, such as a semiparametric form in the continuous longitudinal model [30], incorporate non-Gaussian terms [1], or characterize survivorship via a multivariate joint model [22]. Furthermore, we assume the lung function measurements are associated with the hazard of having a visit, but different forms of association structures could be investigated in future applications. Those may include additional random effects to account for clustering in multi-center cohort studies [20]. Additional research is needed to extend the proposed model to settings with non-proportional hazards. Currently, we use the standard parametric Weibull distribution to model the gap times. If other parametric survival models are more suitable for the data, or if greater flexibility is needed, it would be worthwhile to explore alternative parametric distributions or even nonparametric approaches for modeling the gap times. Finally, another extension of our current work is jointly modeling two dependent outcomes via a latent process, wherein a general form of a joint model is linked through a latent stochastic process; this has been discussed by Henderson and colleagues [9].

DATA AVAILABILITY STATEMENT

The authors do not have permission to share the clinical data, but all code and the simulated data set have been provided as supplemental material for reproducibility purposes. For those interested in acquiring the clinical data, please contact the corresponding author for additional information.

SUPPLEMENTARY MATERIAL

ESM1.pdf: Extended results on parameter estimates of the joint models with random slopes or intercepts only (Tables S1-S2); performance across models according to fit statistics (Tables Tables S3-S4); simulation study results on model parameter estimates and performance (Tables S5-S6); trace plots and residual diagnostics from the real data application (Figures S1-S2).

ESM2: Collection of files that includes the implementation code for the models and a simulated dataset.

ACKNOWLEDGEMENTS

The authors would like to thank the patients, families and clinicians who contribute with data, care, and research in cystic fibrosis, specifically from the Cystic Fibrosis Center within the Division of Pulmonary Medicine at Cincinnati Children's Hospital.

DISCLOSURE STATEMENT

The authors report that there are no competing interests to declare with respect to the research, authorship, and/or publication of this article.

FUNDING

The authors received financial support for this research from the National Institutes of Health (NIH) under Grants R01HL141286 and K25HL125954.

Accepted 26 January 2026

REFERENCES

[1] Asar, O., Bolin, D., Diggle, P. and Wallin, J. (2018). *Linear Mixed-Effects Models for Non-Gaussian Repeated Measurement Data*. Pre-print arXiv:180402592v1. https://doi.org/10.1111/rssc.12405. MR4166856

[2] Celeux, G., Forbes, F., Robert, C. and Titterington, D. (2006). Deviance Information Criteria for Missing Data Models. *Bayesian Analysis* **1**(4) 651–674. https://doi.org/10.1214/06-BA122. MR2282197

[3] Cole, S. and Hernan, M. (2008). Constructing inverse probability weights for marginal structural models. *Am J Epidemiol* **168**(6) 656–664.

[4] Cystic Fibrosis Foundation (2023). *Cystic Fibrosis Foundation Patient Registry*.

[5] Gasparini, A., Abrams, K., Barrett, J., Major, R., Sweeting, M., Brunskill, N. et al. (2020). Mixed-effects models for health care longitudinal data with an informative visiting process: A Monte Carlo simulation study. *Stat Neerl* **74**(1) 5–23. https://doi.org/10.1111/stan.12188. MR4050397

[6] Geisser, S. and Eddy, W. (1979). A Predictive Approach to Model Selection. *J Am Stat Assoc* **74**(365) 153–160. MR0529531

[7] Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* **1** 515–534. https://doi.org/10.1214/06-BA117A. MR2221284

[8] Gelman, A. and Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* **7** 457–472.

[9] Henderson, R., Diggle, P. and Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics* **1**(4) 465–480.

[10] Konstan, M., Schluchter, M., Xue, W. and Davis, P. (2007). Clinical use of Ibuprofen is associated with slower FEV1 decline in children with cystic fibrosis. *Am J Respir Crit Care Med* **176**(11) 1084–1089.

[11] Lederer, D., Bell, S., Branson, R. et al. (2018). Control of Confounding and Reporting of Results in Causal Inference Studies: Guidance for Authors from Editors of Respiratory, Sleep, and Critical Care Journals. *Ann Am Thorac Soc*.

[12] Liou, T., Elkin, E., Pasta, D., Jacobs, J., Konstan, M., Morgan, W. et al. (2010). Year-to-year changes in lung function in individuals with cystic fibrosis. *J Cyst Fibros* **9**(4) 250–256.

[13] Lipsitz, S., Fitzmaurice, G., Ibrahim, J., Gelber, R. and Lipshultz, S. (2002). Parameter estimation in longitudinal studies with outcome-dependent follow-up. *Biometrics* **58**(3) 621–630. https://doi.org/10.1111/j.0006-341X.2002.00621.x. MR1933535

[14] McCulloch, C., Neuhaus, J. and Olin, R. (2016). Biased and unbiased estimation in longitudinal studies with informative visit processes. *Biometrics* **72**(4) 1315–1324. https://doi.org/10.1111/biom.12501. MR3591616

[15] Mogayzel, P. J., Naureckas, E., Robinson, K., Mueller, G., Hadjiliadis, D., Hoag, J. et al. (2013). Cystic fibrosis pulmonary guidelines. Chronic medications for maintenance of lung health. *Am J Respir Crit Care Med* **187**(7) 680–689.

[16] Neal, R. (2011) *MCMC using Hamiltonian Dynamics*. CRC Press, Boca Raton. MR2858447

[17] Neuhaus, J., McCulloch, C. and Boylan, R. (2018). Analysis of longitudinal data from outcome-dependent visit processes: Failure of proposed methods in realistic settings and potential improvements. *Stat Med* **37**(29) 4457–4471. https://doi.org/10.1002/sim.7932. MR3879439

[18] Preisser, J., Lohman, K. and Rathouz, P. (2002). Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random. *Stat Med* **21**(20) 3035–3054.

[19] Pullenayegum, E. and Lim, L. (2016). Longitudinal data subject to irregular observation: A review of methods with a focus on visit processes, assumptions, and study design. *Stat Methods Med Res* **25**(6) 2992–3014. https://doi.org/10.1177/0962280214536537. MR3572895

[20] Pullenayegum, E., Birken, C., Maguire, J. and Collaboration, T. (2021). Clustered longitudinal data subject to irregular observation. *Stat Methods Med Res* **30**(4) 1081–1100. https://doi.org/10.1177/0962280220986193. MR4259889

[21] Rizopoulos, D. (2012) *Joint models for longitudinal and time-to-event data: with applications in R*. CRC Press, Boca Raton. xiv, 261 p.

[22] Rizopoulos, D. and Ghosh, P. (2011). A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Stat Med* **30**(12) 1366–1380. https://doi.org/10.1002/sim.4205. MR2828959

[23] Scotet, V., L'Hostis, C. and Ferec, C. (2020). The Changing Epidemiology of Cystic Fibrosis: Incidence, Survival and Impact of the CFTR Gene Discovery. *Genes (Basel)* **11**(6).

[24] Spiegelhalter, D., Best, N., Carlin, B. and Van der Linde, A. (2002). Bayesian Measures of Model Complexity and Fit (with Discussion). *Journal of the Royal Statistical Society, Series B* **64**(4) 583–616. https://doi.org/10.1111/1467-9868.00353. MR1979380

[25] Stan Development Team (2017). *Stan Modeling Language Users Guide and Reference Manual*.

[26] Su, W. (2020). *Flexible Joint Hierarchical Gaussian Process Model for Longitudinal and Recurrent Event Data*. University of Cincinnati. MR4533229

[27] Su, W., Wang, X. and Szczesniak, R. (2021). Flexible link functions in a joint hierarchical Gaussian process model. *Biometrics* **77**(2) 754–764. https://doi.org/10.1111/biom.13291. MR4307670

[28] Sun, J. -D., Sun, L. and Zhao, X. (2005). Semiparametric re-

gression analysis of longitudinal data with informative observation times. *J Am Stat Assoc* **100**(471) 882–889. https://doi.org/10.1198/016214505000000060. MR2201016

[29] Szczesniak, R., Andrinopoulou, E., Su, W., Afonso, P., Burgel, P., Cromwell, E. et al. (2023). Lung Function Decline in Cystic Fibrosis: Impact of Data Availability and Modeling Strategies on Clinical Interpretations. *Ann Am Thorac Soc*.

[30] Szczesniak, R., McPhail, G., Duan, L., Macaluso, M., Amin, R. and Clancy, J. (2013). A semiparametric approach to estimate rapid lung function decline in cystic fibrosis. *Ann Epidemiol* **23**(12) 771–777.

[31] Szczesniak, R., Su, W., Brokamp, C., Keogh, R., Pestian, J., Seid, M. et al. (2020). Dynamic predictive probabilities to monitor rapid cystic fibrosis disease progression. *Stat Med* **39**(6) 740–756. https://doi.org/10.1002/sim.8443. MR4067763

[32] Taylor-Robinson, D., Whitehead, M., Diderichsen, F., Olesen, H., Pressler, T., Smyth, R. et al. (2012). Understanding the natural progression in with cystic fibrosis: a longitudinal study. *Thorax* **67**(10) 860–866.

[33] van Oudenhoven, F., Swinkels, S., Ibrahim, J. and Rizopoulos, D. (2020). A marginal estimate for the overall treatment effect on a survival outcome within the joint modeling framework. *Stat Med* **39**(28) 4120–4132. https://doi.org/10.1002/sim.8713. MR4175019

[34] Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning* **11** 3571–3594. MR2756194

Weiji Su. Eli Lilly and Company, Indianapolis, IN, USA.

Xia Wang. Division of Statistics and Data Science, University of Cincinnati, Cincinnati, OH, USA.

Pedro Miranda-Afonso. Department of Biostatistics and Department of Epidemiology, Erasmus Medical Center, Rotterdam, The Netherlands.

Eleni-Rosalina Andrinopoulou. Statistics and Data Science Innovation Hub, GSK, The Netherlands.

Rhonda D. Szczesniak. Division of Biostatistics & Epidemiology and Pulmonary Medicine, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA and Department of Pediatrics, University of Cincinnati, Cincinnati, OH, USA. E-mail address: rhonda.szczesniak@cchmc.org