

A joint model for (un)bounded longitudinal markers, competing risks, and recurrent events using patient registry data

Pedro Miranda Afonso^{1,2,*}, Dimitris Rizopoulos^{1,2}, Anushka K. Palipana^{3,4},
Emrah Gecili^{4,5}, Cole Brokamp^{4,5}, John P. Clancy⁶, Rhonda D.
Szczesniak^{4,5,7} and Eleni-Rosalina Andrinopoulou^{1,2}

¹Department of Biostatistics, Erasmus University Medical Center, the Netherlands; ²Department of Epidemiology, Erasmus University Medical Center, the Netherlands, ³School of Nursing, Duke University, USA; ⁴Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center, USA; ⁵Department of Pediatrics, University of Cincinnati, USA; ⁶Division of Statistics and Data Science, University of Cincinnati, USA; ⁷Division of Pulmonary Medicine, Cincinnati Children's Hospital Medical Center, USA

Abstract

Joint models for longitudinal and survival data have become a popular framework for studying the association between repeatedly measured biomarkers and clinical events. Nevertheless, addressing complex survival data structures, especially handling both recurrent and competing event times within a single model, remains a challenge. This causes important information to be disregarded. Moreover, existing frameworks rely on a Gaussian distribution for continuous markers, which may be unsuitable for bounded biomarkers, resulting in biased estimates of associations. To address these limitations, we propose a Bayesian shared-parameter joint model that simultaneously accommodates multiple (possibly bounded) longitudinal markers, a recurrent event process, and competing risks. We use the beta distribution to model responses bounded within any interval (a, b) without sacrificing the interpretability of the association. The model offers various forms of association, discontinuous risk intervals, and both gap and calendar timescales. A simulation study shows that it outperforms simpler joint models. We utilize the US Cystic Fibrosis Foundation Patient Registry to study the associations between changes in lung function and body mass index, and the risk of recurrent pulmonary exacerbations, while accounting for the competing risks of death and lung transplantation. Our efficient implementation allows fast fitting of the model despite its complexity and the large sample size from this patient registry. Our comprehensive approach provides new insights into cystic fibrosis disease progression by quantifying the relationship between the most important clinical markers and events more precisely than has been possible before. The model implementation is available in the R package `JMbayes2`.

Keywords: bounded outcomes, competing risks, cystic fibrosis, joint model, multivariate longitudinal data, recurrent events.

*Correspondence at: Department of Biostatistics, Erasmus University Medical Center, PO Box 2040, 3000 CA Rotterdam, the Netherlands. E-mail address: p.mirandaafonso@erasmusmc.nl.

1 Introduction

In clinical research, joint models for longitudinal and survival data have become a popular framework for studying biomarkers measured over time and their association with clinical events (Henderson et al. 2000; Tsiatis and Davidian 2004; Rizopoulos 2012). Several extensions have been developed to the basic framework for a single event time and a continuous longitudinal biomarker proposed by Faucett and Thomas (1996) and Wulfsohn and Tsiatis (1997). The literature is extensive, with recent comprehensive reviews by Hickey et al. (2016, 2018), Papageorgiou et al. (2019), and Alsefri et al. (2020). These reviews reflect the ongoing efforts to enhance the versatility of the framework and its ability to address the intricate features often found in longitudinal and survival data.

Cystic fibrosis (CF) is a severe genetic disorder that primarily affects the lungs and digestive system, leading to respiratory impairment and malnutrition (Farrell et al. 2008). Patients with CF often experience recurrent lung infections, known as pulmonary exacerbations (PEX), which can cause permanent lung damage and increase the risks of lung transplantation and death. The body mass index (BMI) and the percentage of predicted forced expiratory volume in one second (ppFEV₁) are routinely measured to monitor disease progression. CF care teams are interested in understanding the associations between ppFEV₁ decline, BMI changes, recurrent PEX, and the competing risks of death and lung transplantation using the US Cystic Fibrosis Foundation Patient Registry (CFFPR, Knapp et al. 2016). Previous studies that aimed to investigate such associations using joint models were hampered by the lack of an appropriate framework.

The joint modeling framework has previously been extended to incorporate complex survival data structures, such as recurrent (Liu et al. 2008; Liu and Huang 2009; Kim et al. 2012; Król et al. 2016) and competing event time data (Elashoff et al. 2008; Williamson et al. 2008; Andrinopoulou et al. 2014). However, the integration of both recurrent events and competing risks within a unified model remains a challenge, leading researchers to omit important information despite availability in patient registries. For example, Andrinopoulou et al. (2020) limited their analysis to the period up to the first PEX event, disregarding subsequent occurrences and informative censoring due to transplantation or death. When investigating the association between ppFEV₁ and the risks of death and lung transplantation, Miranda Afonso et al. (2023) treated these two events as a composite endpoint rather than as competing risks, assuming that they indicate the same prior health status, which is not clinically accurate.

An additional limitation of existing frameworks is their tendency to rely exclusively on the Gaussian distribution to model continuous markers. An important aspect of joint modeling is the appropriate parameterization of longitudinal submodels to ensure accurate extrapolation of unobserved biomarker evolution up to the event time. A Gaussian parameterization can be problematic for a bounded biomarker with many observations close to the boundaries, such as ppFEV₁, as it can cause the model to yield biologically implausible values, resulting in biased estimates of the marker evolution and its associations. Existing CF studies have modeled ppFEV₁ mostly using a Gaussian distribution. Szczesniak et al. (2023) explored the use of other distributions; however, deriving a meaningful clinical interpretation from the association in the linear predictor scale was challenging.

We address these collective limitations by introducing a comprehensive joint modeling

framework that can (i) effectively accommodate competing risk and recurrent event processes together with multiple longitudinal outcomes, and (ii) appropriately model bounded longitudinal markers with constrained distributions, without compromising the interpretability of their association. Our model captures the complex dynamics of CF by simultaneously considering recurrent PEx and the competing risks of death and lung transplantation, and by appropriately parameterizing the longitudinal markers ppFEV₁ and BMI using beta and Gaussian distributions, respectively. The choice of a beta distribution ensures that ppFEV₁ remains within the feasible range. The model allows for the use of various functional forms to link time-to-event and longitudinal processes, and accommodates discontinuous risk intervals and both gap and calendar timescales. The model has been made available in the user-friendly R package for joint models, `JMbayes2` (Rizopoulos et al. 2022), which is available in the Comprehensive R Archive Network (CRAN). The implementation approach emphasized versatility and efficiency to streamline the package’s adoption in complex settings with large sample sizes.

The remainder of this article is organized into four sections. Section 2 describes the proposed joint modeling framework in detail. In Section 3, a simulation study demonstrates the added value of our approach over simpler joint models. In Section 4, we apply the proposed model in a real-world setting using the CFFPR dataset. Lastly, Section 5 summarizes the main findings and outlines directions for future research.

2 Joint modeling framework

We propose a joint model with J longitudinal markers that can follow different distributions, K competing events, and one recurrent event process. Joint models assume a full joint distribution of the longitudinal and time-to-event processes that can be factorized in different ways (Sousa 2011). We focus on the shared-parameter joint models in this work; we assume that the time-to-event and longitudinal processes depend on an unobserved process defined by random effects. The observed processes are assumed independent conditional on the random effects. Below we present the submodels that make up the proposed joint model.

2.1 Longitudinal outcomes

To describe the subject-specific time evolution of the j th longitudinal outcome, we consider a mixed-effects regression model

$$\begin{cases} \mathbf{Y}_{j,i} \mid \mathbf{b}_{j,i} \sim \mathcal{F}_{j,\Psi_j} \\ \mathbf{b}_{j,i} \sim \mathcal{N}(\mathbf{0}, \mathbf{D}_j), \end{cases}$$

where $\mathbf{Y}_{j,i}$ is the j th response for the i th individual, $\mathbf{b}_{j,i}$ is the corresponding vector of random effects and \mathcal{F}_j is a set of discrete and continuous distributions (not restricted to the exponential family). The random effects follow a zero-mean multivariate normal distribution with unstructured variance-covariance matrix \mathbf{D}_j . The expected value of the j th outcome at time t conditional on the random effects, $\mu_{j,i}(t) = \text{E}\{Y_{j,i}(t) \mid \mathbf{b}_{j,i}\}$, has the form

$$\mu_{j,i}(t) = \mathcal{G}_j^{-1} \{ \eta_{j,i}(t) \} = \mathcal{G}_j^{-1} \{ \mathbf{x}_{j,i}^\top(t) \boldsymbol{\beta}_j + \mathbf{z}_{j,i}^\top(t) \mathbf{b}_{j,i} \}, \quad (1)$$

where $\eta_{j,i}(t)$ is the linear predictor, $\mathbf{x}_{j,i}(t)$ and $\mathbf{z}_{j,i}(t)$ are the design vectors of (possibly time-varying) covariates for the fixed effects $\boldsymbol{\beta}_j$ and the subject-specific random effects $\mathbf{b}_{j,i}$, respectively, and $\mathcal{G}_j(\cdot)$ is the link function. In this work, given the motivating case study, we focus our attention on two particular continuous distributions: Gaussian and beta.

Let $Y_{j,i}(t)$ be a random sample drawn from the distribution Beta(p, q) with nonnegative shape parameters p and q . We follow the beta density reparameterization proposed by Ferrari and Cribari-Neto (2004), which is indexed by the mean $\mu_{j,i} = p/(p + q)$ and a precision parameter $\phi = p + q$, which satisfies $0 < \mu_{j,i}(t) < 1$ and $\phi > 0$. This choice stems from the difficulty of interpreting shape parameters in terms of conditional expectations. The flexibility of the beta density enables it to adopt a plethora of distinctive shapes ranging from symmetric bell-shaped curves to flat, skewed, or U-shaped curves within the open interval $(0, 1)$ (Gupta and Nadarajah 2004). This versatility makes the beta distribution an appealing choice for modeling a continuous outcome that takes values within a known interval, such as in the case of ppFEV₁. We focus on the logit link $\log\{\mu/(1 - \mu)\}$ in this work, but other link functions can be used. For the logit link, the submodel's regression parameters $\boldsymbol{\beta}_j$ are interpretable in terms of expected changes in $\logit\{\mu_{j,i}(t)\}$. Effects plots can be employed to retrieve these interpretations to the original scale.

The model is heteroscedastic because the variance of $Y_{j,i}(t)$ is a function of its expected value, $\text{Var}\{Y_{j,i}(t)\} = \mu_{j,i}(t)\{1 - \mu_{j,i}(t)\}/(1 + \phi)$. Thus, the model intrinsically accommodates non-constant response variances.

When considering a normally distributed outcome, we use the identity link function in (1), such that $\mu_{j,i}(t) = \eta_{j,i}(t)$, and we account for the measurement error by including the term $\varepsilon_{j,i}(t)$ in $Y_{j,i}(t) = \eta_{j,i}(t) + \varepsilon_{j,i}(t)$, where $\varepsilon_{j,i}(t) \sim \mathcal{N}(0, \sigma_{y_j}^2)$. We assume the measurement errors $\varepsilon_{j,i}(t)$ to be mutually independent and independent of the random effects $\mathbf{b}_{j,i}$. Multiple longitudinal outcomes are associated through the variance-covariance matrix \mathbf{D} , which encompasses the J variance-covariance matrices \mathbf{D}_j . Joint models using the Gaussian distribution have been extensively discussed in the literature (see, for example, Rizopoulos et al. 2014).

2.2 Recurrent event times

For the risk of the recurring event, we rely on a proportional hazards risk model. The hazard function for the l th event at time t is modeled by

$$h_i^R(t) = h_0^R(t - t_{0_{l,i}}) \exp \left[\mathbf{w}_i^{R\top}(t) \boldsymbol{\gamma}^R + \sum_{j=1}^J \sum_{m=1}^{M_j} \mathcal{H}_{j,m}^R \{ \eta_{j,i}(t) \} \alpha_{j,m}^R + v_i^R \right],$$

for $t > t_{0_{l,i}} \geq 0$, where $t_{0_{l,i}}$ is the starting time of the risk interval for the l th recurrent event, and $v_i^R \sim \mathcal{N}(0, \sigma_v^2)$. For the baseline hazard function $h_0^R(t - t_{0_{l,i}})$, we use penalized B-spline functions, P-splines (Eilers and Marx 1996). Specifically, we use $\log h_0^R(t - t_{0_{l,i}}) = \sum_{q=1}^Q \gamma_{0_q}^R \text{bs}_q^R(t - t_{0_{l,i}})$, where $\text{bs}_q^R(t)$ are the P-splines' q th basis functions of degree d , and $\gamma_{0_q}^R$ are the corresponding unknown coefficients. In the relative risk component of the model, the design vector $\mathbf{w}_i^R(t)$ contains the measured characteristics with the corresponding vector of regression coefficients $\boldsymbol{\gamma}^R$; the design vector may incorporate baseline or time-varying

exogenous covariates.

The hazard of an event for individual i at time t is associated with the j th subject-specific marker trajectory through the latent association structure $\mathcal{H}_{j,m}^R \{\eta_{j,i}(t)\} = \mathcal{H}_{j,m}^R \{\eta_{j,i}(u)\}$, $0 \leq u \leq t$, which include the random effects $\mathbf{b}_{j,i}$. The longitudinal and recurrent event processes are assumed to be conditionally independent given $(\mathbf{b}_{1,i}^\top, \dots, \mathbf{b}_{J,i}^\top)$. The function $\mathcal{H}_{j,m}^R(\cdot)$ determines the form of association between the longitudinal and time-to-event processes. The available functional forms are elaborated upon in Section 2.4. The association parameter $\alpha_{j,m}^R$ measures the strength of the association between the m th functional form of the j th longitudinal outcome and the risk of the next event. The quantity $\exp\{\alpha_{j,m}^R\}$ is the hazard ratio (HR) for a one-unit increase in the value of $\mathcal{H}_{j,m}^R \{\eta_{j,i}(t)\}$ while the rest of the variables are kept constant.

We incorporate the random effect v_i^R to capture the correlation among event times within the same individual. Hereafter, we refer to the random effect terms in the risk models as frailties to distinguish them from the random effects in the longitudinal submodels. We assume that the subject-specific frailties and random effects are independent of each other, and that the event times from the same individual are independent conditional on v_i^R .

Our approach allows the recurrent event process to be modeled under the gap or calendar timescales, which use different zero-time references (Duchateau et al. 2003). As shown in the illustrative example in Figure 1, the calendar timescale uses a shared reference time for all events (e.g., study entry), $t_{0_l,i} = 0, \forall l$, while the gap timescale uses the end of the previous event, assuming a renewal after each event and resetting the time to zero. Furthermore, our model accommodates non-risk periods in which a patient is still experiencing the previous event and so is not yet at risk of experiencing the next one. For example, if we are interested in modeling the time to the next hospitalization, then a patient who is currently hospitalized is not at risk of being hospitalized again.

2.3 Competing risks

To model the risks associated with each of the competing events, we consider a cause-specific hazard, allowing for distinct specific forms of association between the longitudinal outcomes and each cause of failure. The instantaneous rate for failures of cause k at any time $t > 0$ is modeled by

$$h_{k,i}^T(t) = h_{0_k}^T(t) \exp \left[\mathbf{w}_{k,i}^{\top}(t) \boldsymbol{\gamma}_k^T + \sum_{j=1}^J \sum_{m=1}^{M_j} \mathcal{H}_{k,j,m}^T \{\eta_{j,i}(t)\} \alpha_{k,j,m}^T + v_{k,i}^T \right],$$

by censoring all other causes. Here, $h_{0_k}^T(t)$ is the cause-specific P-splines baseline hazard function, given by $\log h_{0_k}^T(t) = \sum_{q=1}^Q \gamma_{0_{kq}}^T \text{bs}_{k_q}^T(t)$, while $\mathbf{w}_{k,i}^T(t)$ is the vector of observed (baseline or time-varying exogenous) explanatory variables, and $\boldsymbol{\gamma}_k^T$ is the corresponding vector of regression coefficients.

The j th longitudinal response influences the risk of failure of cause k through $\mathcal{H}_{k,j,m}^T \{\eta_{j,i}(t)\}$. The association parameters $\alpha_{k,j,m}^T$ measure the strength of the association between each longitudinal outcome and the risk of the corresponding event. For a one-unit increase in $\mathcal{H}_{k,j,m}^T \{\eta_{j,i}(t)\}$, the HR for cause k is $\exp(\alpha_{k,j,m}^T)$. The longitudinal measurements and event

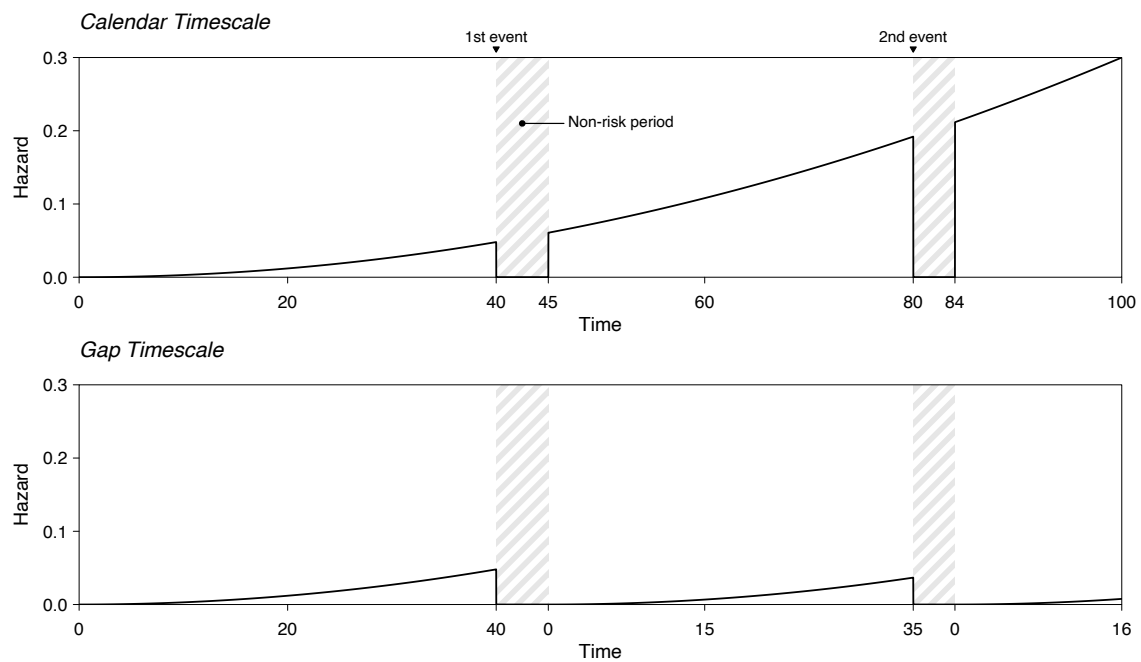


Figure 1: The hazard function for a hypothetical recurrent event process, assuming the calendar (top panel) or gap (bottom panel) timescale. During the study period, from time 0 to 100, the displayed individual experienced two recurrent events (e.g., hospitalizations) at times 40 and 80. These events lasted five and four time units, respectively; during these periods, the individual was not at risk of a new event.

times are assumed to be conditionally independent given $(\mathbf{b}_{1,i}^\top, \dots, \mathbf{b}_{J,i}^\top)$.

The k th competing event is associated with the recurrent event process through a zero-mean Gaussian random variable $v_{k,i}^\top$. We assume that the frailties $v_{k,i}^\top$ and v_i^R are proportional, $v_i^\top = v_i^R \alpha_k^v$, reflecting the common underlying factors that affect their risk. The magnitude of the association between each pair of processes is quantified by α_k^v , the log HR for a one-unit increase in the frailty term. We assume that correlations among different competing risks are driven by the shared frailty v_i^R . Conditional on v_i^R , the competing risks are independent of themselves and of the recurrent event times.

2.4 Forms of association

It has been recognized that the functional form used to link the longitudinal and event processes plays an important role in joint models (Rizopoulos et al. 2014; Mauff et al. 2017). As discussed in Sections 2.2 and 2.3, the hazards $h_i^R(t)$ and $h_i^\top(t)$ of an event for patient i at time t are associated with the j th subject-specific marker trajectory through $\mathcal{H}_{j,m}^R\{\eta_{j,i}(t)\}$ and $\mathcal{H}_{k,j,m}^\top\{\eta_{j,i}(t)\}$, respectively. Our model allows the specification of various forms of association between the longitudinal and time-to-event processes, such as underlying value, $\eta_{j,i}(t)$; slope, $d\eta_{j,i}(t)/dt$; standardized cumulative effect, $\frac{1}{t} \int_0^t \eta_{j,i}(s) ds$; and combinations of these regarding the same longitudinal outcome. Different forms can be assumed for each risk model.

When a nonlinear link function $\mathcal{G}(\cdot)$ is applied to the mean of the longitudinal outcome in (1), it may be challenging to interpret the associations $\exp(\alpha_{k,j,m}^\top)$ and $\exp(\alpha_{j,m}^R)$ in the linear predictor scale. In such situations, it is more convenient to transform the subject-specific linear predictor back to the outcome’s original scale before applying the functional form of interest, that is, $\mathcal{H}_{j,m}\{\mu_{j,i}(t)\} = \mathcal{H}_{j,m}[\mathcal{G}_j^{-1}\{\eta_{j,i}(t)\}]$, where $\mathcal{G}_j^{-1}(\cdot)$ is the inverse link function. For example, when considering the logit link, we can use the expit function $\mathcal{G}^{-1}(x) = \text{expit}(x) = \exp(x)/\{1+\exp(x)\}$ so that the association parameters are interpretable in terms of the mean $\mu_{j,i}(t)$ of $y_{j,i}(t)$, and not in terms of logit $\{\mu_{j,i}(t)\}$. Supplementary Table S2 lists the functional forms that can be used in our model to link the longitudinal and time-to-event outcomes, along with the corresponding transformation functions.

2.5 Inference and software

Inference on the joint model parameters is carried out under the Bayesian framework. The corresponding posterior probability distribution does not have a closed form, so we resort to the Metropolis–Hastings algorithm with adaptive optimal scaling using the Robbins–Monro algorithm (Garthwaite et al. 2016) to approximate it. Our C++ implementation of the posterior sampling algorithms allows fast model fitting despite its complexity and sample size, which have resulted in long computing times in previous analyses of the CFFPR (Andrinopoulou et al. 2020). The full and conditional posterior distributions, along with the prior specification, and additional details about the sampling heuristic, are available in Supplementary Section A.

We have made our model publicly available in the CRAN R package `JMbayes2` (Rizopoulos et al. 2022). In Supplementary Section B, we present an example of the use of the proposed joint model with `JMbayes2`. Our implementation allows the longitudinal processes to follow different distributions, such as the Student’s t, gamma, unit-Lindley, censored normal, binomial,

Poisson, negative binomial, and beta-binomial distributions. Furthermore, the flexibility of our `JMbayes2` implementation allows users to fit simpler joint models that only consider the competing risks or the recurrent event processes.

3 Simulation study

3.1 Design

The objective of our simulation study is twofold: to validate the proposed model and explore the bias introduced by model misspecification. We present two simulation scenarios, named A and B. Scenario A is designed to validate the implementation of the model by demonstrating its ability to recover the parameters' true values. This scenario considers two longitudinal outcomes, two competing risks, and one recurrent process. The model structures for the data generation and fitting processes are identical. In Scenario B, we examine the bias in the association parameter introduced by modeling a bounded outcome using a Gaussian distribution. This scenario involves a joint model with one longitudinal outcome and one terminal event. Two modeling strategies for the longitudinal submodel are considered: one using a beta distribution (the true model) and the other a Gaussian distribution (the misspecified model). The beta variant is used to assess the model under ideal conditions in which it is accurately specified, providing benchmark estimates for the Gaussian model. When considering the beta distribution, we include the longitudinal outcome in the hazards' linear predictors at its original scale, rather than the linear predictor scale, to ensure the comparability of association coefficients between the two models.

Supplementary Table S3 provides the full definitions of the joint models employed for the data generation process and the corresponding models fitted to the generated data for both scenarios, and Supplementary Table S4 lists the parameter values considered. We replicate each scenario 100 times. Supplementary Tables S5 and S6 detail the data generation process for each scenario, and Supplementary Table S7 summarizes the characteristics of the simulated datasets.

The joint models are fit using `JMbayes2` (v0.4.5). For each model, we use three Markov chains with 10,000 or 5,000 iterations per chain, discarding the first 7,500 and 2,500 iterations as a warm-up for Scenarios A and B, respectively. Details of the prior distributions assumed are available in the Supplementary Table S1. The convergence of the chains is assessed using the convergence diagnostic \hat{R} (Gelman and Rubin 1992) aiming for values below 1.10, and by visual inspection of the posterior traceplots of randomly chosen datasets within each scenario. The code used to perform the simulation study is publicly available at <https://github.com/pedromafonso/bounded-jm-simulation>.

3.2 Results

Table 1 summarizes the simulation results, listing the bias and mean squared error values obtained. Supplementary Figures S1 and S2 depict the distributions of estimated posterior means for both scenarios. In Scenario A, the estimates closely align with the true values, confirming the accuracy of the model. In Scenario B, the limitations of the Gaussian distribution become evident when dealing with inherently bounded longitudinal outcomes. Despite

apparent convergence (see Supplementary Figure S3), the Gaussian model extrapolates the longitudinal model to values outside the response domain, introducing bias in the estimation of the target association (bias: -5.9; mean squared error [MSE]: 34.7) and, consequently, in the remaining independent variables present in the risk model. These findings underscore both the critical role of model selection and the suitability of the beta regression model for scenarios involving constrained response variables.

Table 1: Bias and mean squared error for the joint model estimates obtained under the two simulated scenarios for 100 simulated datasets. Scenario A: the joint model comprises one bounded and one unbounded longitudinal marker, two competing risks, and one recurrent event process; the fitted model is equal to the data generation model. Scenario B: the joint model comprises one bounded longitudinal marker and one terminal event; of the two fitted models, the one that models the bounded marker with a Gaussian distribution is different from the data generation model. Abbreviations: M₁, 1st longitudinal marker; M₂, 2nd longitudinal marker; MSE, mean squared error; PEx, pulmonary exacerbation; R, Recurrent event; T₁, 1st terminal event; T₂, 2nd terminal event.

Submodel	Param.	Scenario A			Scenario B				
		True	Bias	MSE	True	Beta		Gaussian	
						Bias	MSE	Bias	MSE
M ₁	$\beta_{1,0}$	2.00	-0.001	0.000	2.00	-0.001	0.000	-1.235	1.526
	$\beta_{1,t}$	-1.50	0.001	0.000	-1.00	0.001	0.000	0.881	0.777
M ₂	$\beta_{2,0}$	0.80	0.000	0.000	—	—	—	—	—
	$\beta_{2,t}$	-0.05	0.000	0.000	—	—	—	—	—
R	γ^R	0.25	-0.010	0.002	—	—	—	—	—
	α_1^R	-2.00	-0.008	0.006	—	—	—	—	—
	α_2^R	-1.00	-0.003	0.003	—	—	—	—	—
T ₁	γ_1^T	0.25	-0.016	0.015	0.25	-0.004	0.006	-0.036	0.009
	$\alpha_{1,1}^T$	-2.00	-0.079	0.378	-2.00	-0.066	0.122	-5.870	34.696
	$\alpha_{1,2}^T$	-1.00	-0.019	0.018	—	—	—	—	—
	α_1^v	1.00	0.020	0.034	—	—	—	—	—
T ₂	γ_2^T	0.25	-0.013	0.010	—	—	—	—	—
	$\alpha_{2,1}^T$	-2.00	-0.026	0.199	—	—	—	—	—
	$\alpha_{2,2}^T$	-1.00	-0.020	0.012	—	—	—	—	—
	α_2^v	1.00	-0.005	0.046	—	—	—	—	—

4 Application

4.1 The CFFPR dataset

The CFFPR is one of the largest and most comprehensive databases of its kind, containing longitudinal clinical and demographic information on individuals living with CF in the US (Knapp et al. 2016). Supplementary Figure S4 outlines the exclusion process applied to address data quality issues, such as missing data or data entry errors. The remaining data describe 23,543 individuals, who collectively contributed 1,315,586 observations between January 1, 2000, and December 31, 2017. The demographic, social, and clinical characteristics of the individuals analyzed are summarized in Supplementary Table S8. The baseline characteristics are ethnicity, genotype, birth cohort, and sex. The time-varying characteristics include pancreatic enzyme intake—implying pancreatic insufficiency—and environmental influences such as neighborhood material deprivation index (as defined by Brokamp et al. 2019), percentage of green space, and moving-truck density. Previous research demonstrated that environmental and community characteristics, alongside clinical and demographic factors, are critical to comprehensively understand CF progression (Gecili et al. 2023; Palipana et al. 2023).

BMI and ppFEV₁ are commonly measured in routine checkups and registered in the CFFPR. BMI is an important clinical marker used to assess the nutritional status of individuals with CF, who are at increased risk of malnutrition and poor growth due to impaired nutrient absorption, pancreatic insufficiency, and increased energy requirements. FEV₁ measures the maximum volume of air that a person can forcefully exhale in the first second of expiration after taking a deep breath. ppFEV₁ compares a patient’s measured FEV₁ to the expected value for a person of the same age, sex, and height with normal lung function (Stanojevic et al. 2015). We assume that ppFEV₁ ranges from 0% to 150%, with a value of 100% meaning that the patient’s FEV₁ is equal to the expected value for a healthy individual. While it is uncommon, there are instances in which the ppFEV₁ is reported as above 100% owing to early intervention and treatment. Lower BMI and ppFEV₁ levels are associated with worse clinical outcomes (Liou et al. 2001). The median numbers of ppFEV₁ and BMI measurements per individual are 47 (interquartile range [IQR] 27–69) and 48 (IQR 28–72), respectively, with corresponding median follow-up times per individual of 11.92 (IQR 6.97–16.76) and 11.72 (IQR 6.85–16.61) years. Figure 2 displays the ppFEV₁ (left panel) and BMI (center panel) evolution experienced by nine randomly selected individuals over time. The profiles exhibit different follow-up durations and diverse nonlinear trends.

The most common cause of death in cystic fibrosis patients is respiratory failure, often due to lung damage caused by chronic PEx. For individuals with end-stage lung disease, lung transplantation is a treatment option. Data acquired after lung transplantation were excluded. In this study, we treated death by respiratory failure and lung transplantation as competing events. However, formally, these events are semi-competing, as an individual can still die after receiving a double-lung transplant. Time-to-event data record the ages at which individuals experienced these events. During the follow-up period, 10.88% of the individuals received a lung transplant, 17.97% died from respiratory failure, and the remaining 71.15%

Percentage of greenspace, impervious, and tree canopy areas within the Zone Improvement Plan Code Tabulation Area (ZCTA) derived from the National Land Cover Database (Jin et al. 2019).

were right-censored. The median (IQR) ages at lung transplantation, death, and censoring were 28.52 (22.84–36.55), 26.57 (21.36–35.93), and 23.50 (17.07–32.15) years, respectively. The right panel in Figure 2 shows the cumulative incidence functions for the competing risks of death and lung transplantation. We note that both of these events can cause nonignorable missing data in the measurements of ppFEV₁ and BMI.

A PEx is a sudden worsening of CF respiratory symptoms usually caused by an infection or inflammation in the airways (Flume et al. 2009). In this study, we define the recurrent PEx event as an episode of care documented in the CFFPR with intravenous antibiotic use. If a new PEx episode is recorded during an ongoing exacerbation, it is treated as the same event. This implies the existence of non-risk periods during the episode of care that must be accounted for during the modeling process. The median number of PEx per individual is 7 (IQR 3–14), with a median interval between consecutive PEx of 0.34 (IQR 0.15–0.77) years.

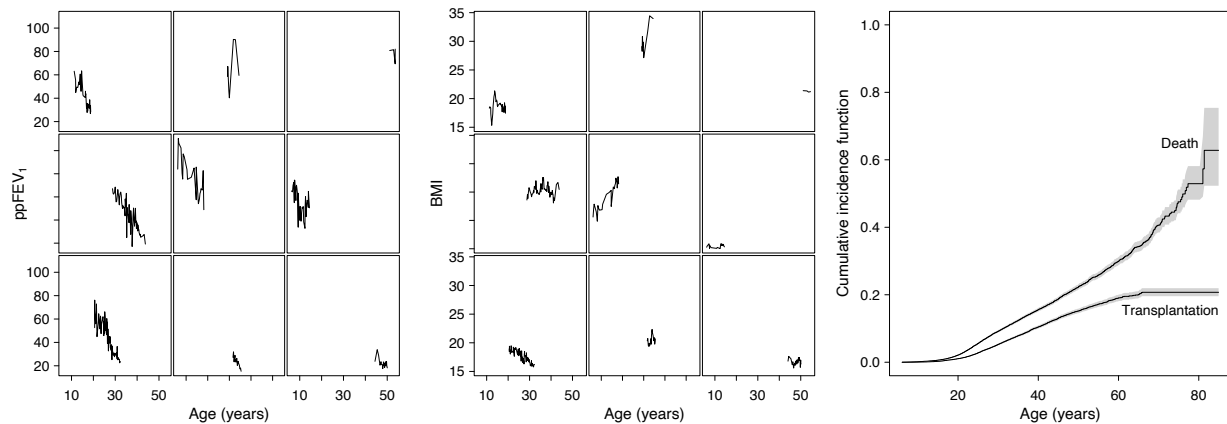


Figure 2: Longitudinal and survival outcomes of interest. Left: ppFEV₁ measurements against age for nine randomly selected individuals. Center: BMI measurements against age for the same individuals. Right: Cumulative incidence functions for the competing events of death and lung transplantation, with associated 95% confidence intervals.

4.2 Analysis

We fitted the joint model described in Section 2, considering two longitudinal outcomes ($J = 2$), one recurrent event process, and two competing events ($K = 2$). The longitudinal ppFEV₁ and BMI measurements are described using mixed-effects models assuming a beta and normal distribution, respectively. The formulations for these models are given as follows:

$$\begin{aligned} \text{logit} \{ \text{ppFEV}_{1,i}^{**}(t) \} = & (\beta_{1,0} + b_{1,0,i}) + (\beta_{1,t} + b_{1,t,i}) t + \beta_{1,\text{male}} \text{sex}_{\text{male},i} + \beta_{1,[93,98]} \text{YOB}_{[93,98],i} + \\ & + \beta_{1,\geq 98} \text{YOB}_{\geq 98,i} + \beta_{1,\text{htz}} \text{F508del}_{\text{htz},i} + \beta_{1,\text{oth}} \text{F508del}_{\text{oth},i} + \\ & + \beta_{1,\text{ethn}} \text{ethn}_{\text{hisp},i} + \beta_{1,\text{truck}} \text{truck}_i(t) + \beta_{1,\text{depr}} \text{depr}_i(t) + \beta_{1,\text{pgrn}} \text{pgrn}_i(t), \end{aligned}$$

and

$$\begin{aligned}
\text{BMI}_i(t) &= \tilde{\text{BMI}}_i(t) + \varepsilon_i(t) = \\
&= (\beta_{2,0} + b_{2,0,i}) + \sum_{q=1}^2 (\beta_{2,q} + b_{2,q,i}) \text{ns}_{2,q}(t) + \beta_{2,\text{male}} \text{sex}_{\text{male},i} + \beta_{2,[93,98]} \text{YOB}_{[93,98],i} + \\
&\quad + \beta_{2,\geq 98} \text{YOB}_{\geq 98,i} + \beta_{2,\text{htz}} \text{F508del}_{\text{htz},i} + \beta_{2,\text{oth}} \text{F508del}_{\text{oth},i} + \beta_{2,\text{ethn}} \text{ethn}_{\text{hispanic},i} + \\
&\quad + \beta_{2,\text{depr}} \text{depr}_i(t) + \beta_{2,\text{enzy}} \text{enzy}_i(t) + \varepsilon_i(t),
\end{aligned}$$

for $t > 0$, where $(b_{1,0,i}, b_{1,t,i}, b_{2,0,i}, b_{2,1,i}, b_{2,2,i}, b_{2,t^2,i})^\top \sim \mathcal{N}(\mathbf{0}, \mathbf{D})$, and $\varepsilon_i(t) \sim \mathcal{N}(0, \sigma_{y_2}^2)$, with the two random variables assumed independent of each other. Here, $\tilde{\text{BMI}}_i(t)$ is the BMI response without error, and $\text{ppFEV}_{1i}^{**}(t)$ is the ppFEV₁ response scaled to the interval (0, 1).

For ppFEV₁, we assume a linear average evolution over time, while for BMI, we assume a nonlinear evolution. More specifically, for BMI, we employ natural cubic splines with two degrees of freedom, denoted by $\text{ns}_{2,q}(t)$, $q = 1, 2$, with knots located at the 0%, 50% and 95% percentiles of the observed follow-up times.

The average ppFEV₁ and BMI responses are adjusted for baseline and time-varying individual characteristics including sex (male vs. female), $\text{sex}_{\text{male},i}$; birth cohort (< 93 , $[93, 98)$, or ≥ 98), $\text{YOB}_{<93,i}$ and $\text{YOB}_{[93,98],i}$; genotype (F508del homozygous, homozygous, or other/unknown), $\text{F508del}_{\text{htz},i}$ and $\text{F508del}_{\text{oth},i}$; ethnicity (hispanic vs. non-hispanic), $\text{ethn}_{\text{hispanic},i}$; and neighborhood deprivation index, $\text{depr}_i(t)$. Additionally, the average ppFEV₁ is adjusted for the percentage of green space, $\text{pgrn}_i(t)$, and the annual average daily moving-truck density in the ZCTA, $\text{truck}_i(t)$, while the BMI response is adjusted for enzyme intake $\text{enzy}_i(t)$. The birth cohort variable aims to account for the evolution in CF care over the years, including approvals of new therapeutics. For the random effects structure, we assume a subject-specific random intercept and the same nonlinear effect of time as for the fixed effects.

We are interested in investigating how individual characteristics affect the risk of death separately from how they affect the risk of transplantation. Therefore, we postulate two cause-specific risk models, one for each of these competing events. The hazard functions for the clinical events of PEx, transplantation, and death are denoted by $h_i^{\text{R}}(t)$, $h_{1,i}^{\text{T}}(t)$, and $h_{2,i}^{\text{T}}(t)$, respectively, and are defined as follows

$$\begin{aligned}
h_i^{\text{R}}(t) &= h_0^{\text{R}}(t - t_{0,i}) \exp \left[\gamma_{\text{PEx}}^{\text{R}} \text{nPEX}_i(t) + \text{ppFEV}_{1i}^{**}(t) \alpha_{1,1}^{\text{R}} + \frac{1}{t} \int_0^t \tilde{\text{BMI}}_i(s) \text{d}s \alpha_{2,1}^{\text{R}} + v_i^{\text{R}} \right], \\
h_{1,i}^{\text{T}}(t) &= h_{0_1}^{\text{T}}(t) \exp \left[\text{ppFEV}_{1i}^{**}(t) \alpha_{1,1,1}^{\text{T}} + \frac{\text{d ppFEV}_{1i}^{**}(t)}{\text{d}t} \alpha_{1,1,2}^{\text{T}} + \frac{1}{t} \int_0^t \tilde{\text{BMI}}_i(s) \text{d}s \alpha_{1,2,1}^{\text{T}} + v_i^{\text{R}} \alpha_1^{\text{v}} \right],
\end{aligned}$$

A response restricted to a closed interval between known theoretical limits a and b , so that $y \in [a, b]$, can be mapped to the interval (0, 1) by transforming the observed value y using $y^{**} = \{y^* \times (N - 1) + 0.5\} / N$, where $y^* = (y - a) / (b - a)$ and N is the sample size (Smithson and Verkuilen 2006).

and

$$h_{2,i}^T(t) = h_{0_2}^T(t) \exp \left[\text{ppFEV}_{1i}^{**}(t) \alpha_{2,1,1}^T + \frac{d \text{ppFEV}_{1i}^{**}(t)}{dt} \alpha_{2,1,2}^T + \frac{1}{t} \int_0^t \tilde{\text{BMI}}_i(s) ds \alpha_{2,2,1}^T + v_i^R \alpha_2^v \right],$$

for $t > 0$, where $v_i^R \sim \mathcal{N}(0, \sigma_v^2)$, $v_i^R \perp\!\!\!\perp (b_{1,0,i}, b_{1,t,i}, b_{2,0,i}, b_{2,t,i}, b_{2,t^2,i})$ and $v_i^R \perp\!\!\!\perp \varepsilon_i(t)$. Changes in BMI over time occur relatively slowly, whereas ppFEV₁ can experience sudden declines. Therefore, guided by clinical insights, we include in the hazards' linear predictors the ppFEV₁'s value, $d \text{ppFEV}_{1i}^{**}(t)/dt$, and rate of change, $d \text{ppFEV}_{1i}^{**}(t)/dt$, evaluated at its original scale—applying the $\text{expit}(\cdot)$ transformation to the linear predictor described in Section 2.1—and the standardized cumulative effect of BMI's underlying value, $\frac{1}{t} \int_0^t \tilde{\text{BMI}}_i(s) ds$. In the PEx model we include the number of previous PEx events, $\text{nPEX}_i(t)$ and consider the gap timescale. Regarding the baseline hazards, we consider 10 quadratic P-spline basis functions defined over a grid of equally spaced knots over the domain of the observed event times. We consider second-order differences in the penalty matrices.

We generated three Markov chains in `JMbayes2` (v0.4.5) with 20,000 iterations each, of which 10,000 were discarded for warm-up. We use the package's default prior distributions (see Supplementary Table S1). The traceplots and the \hat{R} (Gelman and Rubin 1992), with $\hat{R} < 1.10$, showed satisfactory convergence of the Markov chains.

4.3 Results

The effects plots in Figure 3 show the estimated evolution of BMI and ppFEV₁ with age. The results in the left panel suggest an increase in BMI up to early adulthood, followed by a gradual decrease. The right panel shows a period of rapid ppFEV₁ decline during childhood and adolescence, and a more gradual decline thereafter. When modeling ppFEV₁ with a Gaussian distribution and allowing for flexible temporal evolution, the resulting model produces non-feasible negative values (Figure 3, right panel).

The model parameter estimates are listed in Table 2. The risk of a PEx increases with the number of previous episodes. The results suggest that both ppFEV₁ and BMI are associated with the risks of experiencing PEx, transplantation, and death. A one-unit decrease in value and one-unit increase in the rate of ppFEV₁ decline increases the hazard of death by 11.58% (95% CI 11.34–11.82) and 9.15% (95% CI 7.51–10.83), respectively. A one-unit increase in the standardized cumulative effect of BMI increases the hazard of PEx by 7.06% (95% CI 5.42–8.70). The incidence of PEx is positively associated with transplantation and death. Frailer individuals are at a higher risk of PEx and are more likely to receive a lung transplant or die. A one-standard-deviation increase in the frailty term increases the hazards of death by 202.71% (95% CI 187.69–219.03). In Supplementary Section D, the reader can find a detailed explanation of how these conclusions were derived from the association parameters estimates in Table 2. The estimates for the association between ppFEV₁ and the risk of transplantation are different from that between ppFEV₁ and death, illustrating the value of modeling both events individually, rather than as a composite endpoint.

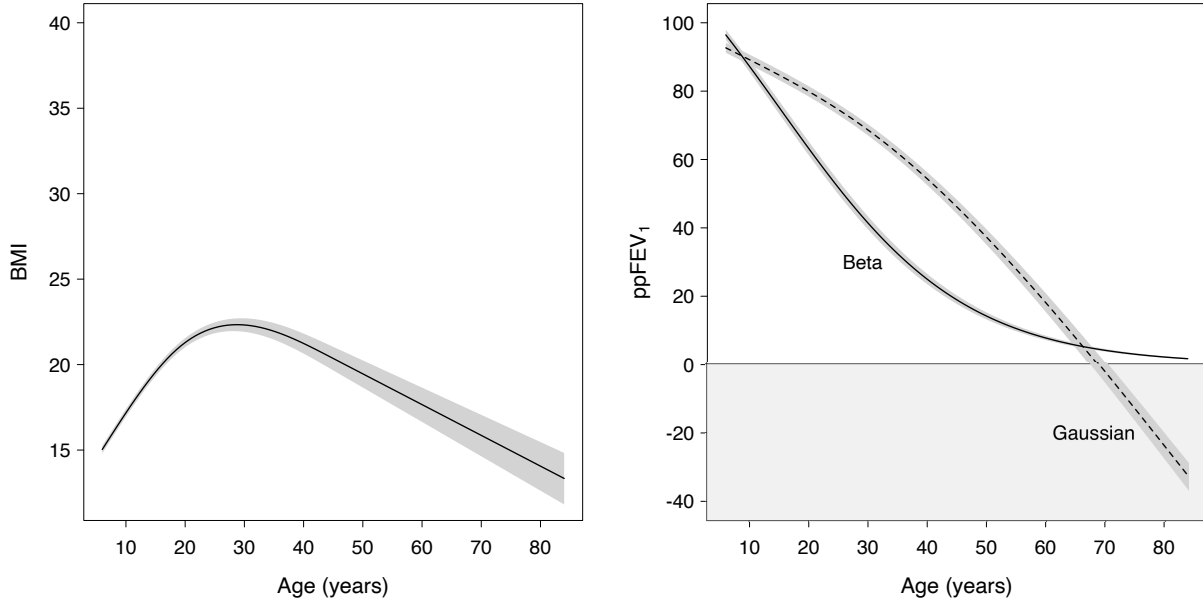


Figure 3: Left: Estimated BMI evolution with age, with associated 95% credible interval, for Hispanic females with CF, F508del homozygotes, who were born before 1993, did not take pancreatic enzymes, and lived in a community with deprivation index of 0.5. Right: Estimated ppFEV₁ evolution with age, with associated 95% credible interval, when assuming either a beta or Gaussian distribution for Hispanic females with CF, F508del homozygotes, who were born before 1993, and lived in a community with a deprivation index of 0.5, in which the percentage of green space is 50%, and in which the moving-truck density is 0.18 $\mu\text{truck-meters}/\text{m}^2$. For a Gaussian distribution, the model generates non-feasible negative values, despite incorporating flexible temporal evolution via natural cubic splines.

Table 2: Posterior means and 95% credible intervals for some of the joint model parameters fitted to the CFFPR dataset. Abbreviations: BMI, body mass index; CI, credible interval; HR, hazard ratio; PEx, pulmonary exacerbation; ppFEV₁, percent predicted forced expiratory volume in one second.

Model	Parameter/HR	Mean	95% CI	
ppFEV ₁				
	$\beta_{1,0}$	0.591	(0.554,	0.629)
	$\beta_{1,t}$	-0.065	(-0.065,	-0.064)
	$\beta_{1,\text{male}}$	0.001	(-0.017,	0.018)
	$\beta_{1,[93,98]}$	-0.157	(-0.180,	-0.133)
	$\beta_{1,\geq 98}$	-0.125	(-0.147,	-0.103)
	$\beta_{1,\text{htz}}$	0.019	(0.001,	0.038)
	$\beta_{1,\text{oth}}$	-0.024	(-0.050,	0.002)
	$\beta_{1,\text{ethn}}$	0.223	(0.191,	0.256)
	$\beta_{1,\text{depr}}$	-0.003	(-0.010,	0.005)
	$\beta_{1,\text{truck}}$	$-4.44e^{-5}$	($-3.16e^{-4}$,	$3.34e^{-4}$)
	$\beta_{1,\text{pgrn}}$	-0.266	(-0.519,	0.029)
BMI				
	$\beta_{2,0}$	15.053	(14.858,	15.244)
	β_{2,ns_1}	12.867	(12.585,	13.143)
	β_{2,ns_2}	1.881	(1.424,	2.330)
	$\beta_{2,\text{male}}$	-0.465	(-0.548,	-0.378)
	$\beta_{2,[93,98]}$	0.242	(0.127,	0.356)
	$\beta_{2,\geq 98}$	0.633	(0.523,	0.743)
	$\beta_{2,\text{htz}}$	0.170	(0.080,	0.259)
	$\beta_{2,\text{oth}}$	0.269	(0.140,	0.398)
	$\beta_{2,\text{ethn}}$	-0.191	(-0.348,	-0.032)
	$\beta_{2,\text{depr}}$	-0.038	(-0.101,	-0.021)
	$\beta_{2,\text{enzy}}$	0.021	(0.016,	0.026)
Recurrent PEx				
	$\exp(\gamma_{\text{PEx}}^{\text{R}})$	1.010	(1.009,	1.011)
	σ_v	0.835	(0.822,	0.849)
	$\exp(\alpha_{1,1}^{\text{R}}/150)$	0.962	(0.961,	0.962)
	$\exp(\alpha_{2,1}^{\text{R}})$	1.000	(1.000,	1.000)
Transplantation				
	$\exp(\alpha_{1,1,1}^{\text{T}}/150)$	0.830	(0.825,	0.835)
	$\exp(\alpha_{1,1,2}^{\text{T}}/150)$	0.863	(0.839,	0.891)
	$\exp(\alpha_{1,2,1}^{\text{T}})$	1.060	(1.044,	1.076)
	$\exp(\alpha_1^v)$	1.203	(1.122,	1.287)
Death				
	$\exp(\alpha_{2,1,1}^{\text{T}}/150)$	0.884	(0.882,	0.887)
	$\exp(\alpha_{2,1,2}^{\text{T}}/150)$	0.909	(0.892,	0.925)
	$\exp(\alpha_{2,2,1}^{\text{T}})$	1.071	(1.054,	1.087)
	$\exp(\alpha_2^v)$	1.326	(1.266,	1.389)

5 Discussion

Motivated by a clinical study on CF, we have developed the first Bayesian shared-parameter joint model that accommodates multiple continuous (possibly bounded) longitudinal markers, a recurrent event process, and multiple competing terminal events. Compared with previous frameworks, our comprehensive joint model enables more efficient use of all available information in scenarios with multiple markers and event times. In addition, by modeling a continuous and bounded longitudinal outcome using a beta distribution, we ensure that the longitudinal submodel predicts feasible values and provides meaningful insights into the association between the biomarker and the clinical event. This modeling framework can be particularly valuable for markers in pediatric populations expressed in percentiles or z-scores. The model is available in the R package `JMbayes2` (Rizopoulos et al. 2022) and is flexible enough to handle a wide range of applications.

The efficient implementation of the Markov chain Monte Carlo sampling algorithms in C++ ensures fast model fitting. Nonetheless, applying multivariate joint models to large datasets may require extended computing times. One can speed up model fitting by employing consensus Monte Carlo methods. Interested readers can find more details on how this approach can be implemented using `JMbayes2` in Miranda Afonso et al. (2023).

It can be argued that all biomarkers are inherently bounded, as they signify measurable quantities within biological systems and are typically constrained by physiological limits. In the context of this study, BMI could be seen as inherently bounded like ppFEV₁, making it a suitable candidate for modeling with a beta distribution. However, the normal distribution continues to be an effective approximation for BMI, as it will be for many other biomarkers, as the underlying distribution of the outcome lacks extreme skewness or heavy tails.

Although the proposed joint model exhibits great potential for advancing our understanding of complex disease dynamics, there remain opportunities for future research. We initially mapped the ppFEV₁ observations to the interval $[0, 1]$ and subsequently to the open interval $(0, 1)$ using the transformation proposed by Smithson and Verkuilen (2006). In future research, it may be worthwhile to explore the application of a zero-and-one inflated beta distribution to eliminate the need for the second transformation. Additionally, the derivation of individualized dynamic predictions (Andrinopoulou et al. 2021) represents an important research direction. Developing appropriate predictive assessment tools is also imperative for evaluating the model’s performance and enabling its translation into clinical practice.

Our findings shed new light on the progression of CF, and we hope they will contribute to the effective management of PEx, reducing the frequency and severity of episodes. By making our model publicly available, we hope to assist applied statisticians and epidemiologists in performing joint analyses of longitudinal and time-to-event data in other complex settings.

Acknowledgments

The authors would like to thank the Cystic Fibrosis Foundation for the use of CF Foundation Patient Registry data to conduct this study. Additionally, we would like to thank the patients, care providers, and clinic coordinators at CF Centers throughout the United States for their contributions to the CF Foundation Patient Registry.

Funding

This work was supported by grants from the National Institutes of Health (R01 HL141286).

Data Availability Statement

The data that support the findings of this study are available from the Cystic Fibrosis Foundation. Restrictions apply to the availability of these data, which were used under license for this study. Requests for data may be sent to datarequests@cff.org.

References

- Alsefri, M., Sudell, M., García-Fiñana, M., and Kolamunnage-Dona, R. (2020). Bayesian joint modelling of longitudinal and time to event data: A methodological review. *BMC Medical Research Methodology* **20**, 1–17.
- Andrinopoulou, E.-R., Clancy, J. P., and Szczesniak, R. (2020). Multivariate joint modeling to identify markers of growth and lung function decline that predict cystic fibrosis pulmonary exacerbation onset. *BMC Pulmonary Medicine* **20**, 1–11.
- Andrinopoulou, E.-R., Harhay, M. O., Ratcliffe, S. J., and Rizopoulos, D. (2021). Reflection on modern methods: Dynamic prediction using joint models of longitudinal and time-to-event data. *International Journal of Epidemiology* **50**, 1731–1743.
- Andrinopoulou, E.-R., Rizopoulos, D., Takkenberg, J. J., and Lesaffre, E. (2014). Joint modeling of two longitudinal outcomes and competing risk data. *Statistics in Medicine* **33**, 3167–3178.
- Bender, R., Augustin, T., and Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine* **24**, 1713–1723.
- Brokamp, C., Beck, A. F., Goyal, N. K., Ryan, P., Greenberg, J. M., and Hall, E. S. (2019). Material community deprivation and hospital utilization during the first year of life: An urban population-based cohort study. *Annals of Epidemiology* **30**, 37–43.
- Duchateau, L., Janssen, P., Kezic, I., and Fortpied, C. (2003). Evolution of recurrent asthma event rate over time in frailty models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **52**, 355–363.
- Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* **11**, 89–121.
- Elashoff, R. M., Li, G., and Li, N. (2008). A joint model for longitudinal measurements and survival data in the presence of multiple failure types. *Biometrics* **64**, 762–771.
- Farrell, P. M., Rosenstein, B. J., White, T. B., Accurso, F. J., Castellani, C., Cutting, G. R., Durie, P. R., LeGrys, V. A., Massie, J., Parad, R. B., et al. (2008). Guidelines for diagnosis of cystic fibrosis in newborns through older adults: Cystic Fibrosis Foundation consensus report. *The Journal of Pediatrics* **153**, S4–S14.

- Faucett, C. L. and Thomas, D. C. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: A Gibbs sampling approach. *Statistics in Medicine* **15**, 1663–1685.
- Ferrari, S. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics* **31**, 799–815.
- Flume, P. A., Mogayzel Jr, P. J., Robinson, K. A., Goss, C. H., Rosenblatt, R. L., Kuhn, R. J., Marshall, B. C., and Clinical Practice Guidelines for Pulmonary Therapies Committee (2009). Cystic fibrosis pulmonary guidelines: Treatment of pulmonary exacerbations. *American Journal of Respiratory and Critical Care Medicine* **180**, 802–808.
- Garthwaite, P. H., Fan, Y., and Sisson, S. A. (2016). Adaptive optimal scaling of metropolis–hastings algorithms using the robbins–monro process. *Communications in Statistics-Theory and Methods* **45**, 5098–5111.
- Gecili, E., Brokamp, C., Rasnick, E., Afonso, P. M., Andrinopoulou, E.-R., Dexheimer, J. W., Clancy, J. P., Keogh, R. H., Ni, Y., Palipana, A., et al. (2023). Built environment factors predictive of early rapid lung function decline in cystic fibrosis. *Pediatric Pulmonology* .
- Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1995). Efficient parametrisations for normal linear mixed models. *Biometrika* **82**, 479–488.
- Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* **7**, 457–472.
- Gupta, A. K. and Nadarajah, S. (2004). *Handbook of Beta Distribution and Its Applications*. CRC Press.
- Henderson, R., Diggle, P., and Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics* **1**, 465–480.
- Hickey, G. L., Philipson, P., Jorgensen, A., and Kolamunnage-Dona, R. (2016). Joint modelling of time-to-event and multivariate longitudinal outcomes: Recent developments and issues. *BMC Medical Research Methodology* **16**, 1–15.
- Hickey, G. L., Philipson, P., Jorgensen, A., and Kolamunnage-Dona, R. (2018). Joint models of longitudinal and time-to-event data with more than one event time outcome: A review. *The International Journal of Biostatistics* **14**,.
- Jin, S., Homer, C., Yang, L., Danielson, P., Dewitz, J., Li, C., Zhu, Z., Xian, G., and Howard, D. (2019). Overall methodology design for the United States national land cover database 2016 products. *Remote Sensing* **11**, 2971.
- Kim, S., Zeng, D., Chambless, L., and Li, Y. (2012). Joint models of longitudinal data and recurrent events with informative terminal event. *Statistics in Biosciences* **4**, 262–281.

- Knapp, E. A., Fink, A. K., Goss, C. H., Sewall, A., Ostrenga, J., Dowd, C., Elbert, A., Petren, K. M., and Marshall, B. C. (2016). The Cystic Fibrosis Foundation Patient Registry. Design and methods of a national observational disease registry. *Annals of the American Thoracic Society* **13**, 1173–1179.
- Król, A., Ferrer, L., Pignon, J.-P., Proust-Lima, C., Ducreux, M., Bouché, O., Michiels, S., and Rondeau, V. (2016). Joint model for left-censored longitudinal data, recurrent events and terminal event: Predictive abilities of tumor burden for cancer evolution with application to the FFC2 2000–05 trial. *Biometrics* **72**, 907–916.
- Lang, S. and Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics* **13**, 183–212.
- Liou, T. G., Adler, F. R., FitzSimmons, S. C., Cahill, B. C., Hibbs, J. R., and Marshall, B. C. (2001). Predictive 5-year survivorship model of cystic fibrosis. *American Journal of Epidemiology* **153**, 345–352.
- Liu, L. and Huang, X. (2009). Joint analysis of correlated repeated measures and recurrent events processes in the presence of death, with application to a study on acquired immune deficiency syndrome. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **58**, 65–81.
- Liu, L., Huang, X., and O’Quigley, J. (2008). Analysis of longitudinal data in the presence of informative observational times and a dependent terminal event, with application to medical cost data. *Biometrics* **64**, 950–958.
- Mauff, K., Steyerberg, E. W., Nijpels, G., van der Heijden, A. A., and Rizopoulos, D. (2017). Extension of the association structure in joint models to include weighted cumulative effects. *Statistics in Medicine* **36**, 3746–3759.
- Miranda Afonso, P., Rizopoulos, D., Palipana, A. K., Zhou, G. C., Brokamp, C., Szczesniak, R. D., and Andrinopoulou, E.-R. (2023). Efficiently analyzing large patient registries with bayesian joint models for longitudinal and time-to-event data. *arXiv preprint arXiv:2310.03351* .
- Palipana, A. K., Vancil, A., Gecili, E., Rasnick, E., Ehrlich, D., Pestian, T., Andrinopoulou, E.-R., Afonso, P. M., Keogh, R. H., Ni, Y., et al. (2023). Social-environmental phenotypes of rapid cystic fibrosis lung disease progression in adolescents and young adults living in the united states. *Environmental Advances* page 100449.
- Papageorgiou, G., Mauff, K., Tomer, A., and Rizopoulos, D. (2019). An overview of joint modeling of time-to-event and longitudinal outcomes. *Annual Review of Statistics and Its Application* **6**, 223–240.
- Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. CRC Press.
- Rizopoulos, D. and Ghosh, P. (2011). A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Statistics in Medicine* **30**, 1366–1380.

- Rizopoulos, D., Hatfield, L. A., Carlin, B. P., and Takkenberg, J. J. (2014). Combining dynamic predictions from joint models for longitudinal and time-to-event data using Bayesian model averaging. *Journal of the American Statistical Association* **109**, 1385–1397.
- Rizopoulos, D., Papageorgiou, G., and Miranda Afonso, P. (2022). *JM-bayes2: Extended Joint Models for Longitudinal and Time-to-Event Data*. <https://drizopoulos.github.io/JMbayes2/>, <https://github.com/drizopoulos/JMbayes2>.
- Smithson, M. and Verkuilen, J. (2006). A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods* **11**, 54.
- Sousa, I. (2011). A review on joint modelling of longitudinal measurements and time-to-event. *Revstat Statistical Journal* **9**, 57–81.
- Stanojevic, S., Bilton, D., McDonald, A., Stocks, J., Aurora, P., Prasad, A., Cole, T. J., and Davies, G. (2015). Global Lung Function Initiative equations improve interpretation of FEV1 decline among patients with cystic fibrosis. *European Respiratory Journal* **46**, 262–264.
- Szczesniak, R., Andrinopoulou, E.-R., Su, W., Afonso, P. M., Burgel, P.-R., Cromwell, E., Gecili, E., Ghulam, E., Goss, C. H., Mayer-Hamblett, N., et al. (2023). Lung function decline in cystic fibrosis: Impact of data availability and modeling strategies on clinical interpretations. *Annals of the American Thoracic Society* .
- Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica* pages 809–834.
- Williamson, P. R., Kolamunnage-Dona, R., Philipson, P., and Marson, A. G. (2008). Joint modelling of longitudinal and competing risks data. *Statistics in Medicine* **27**, 6426–6438.
- Wulfsohn, M. S. and Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics* **53**, 330–339.

Supplementary material

A Posterior distribution

We denote the j th longitudinal marker measured at time t for the i th individual by $Y_{j,i}(t)$, $i = 1, \dots, n$, $j = 1, \dots, J$. The longitudinal responses are collected for each subject at intermittent time points $\{t_{j,i,g}, j = 1, \dots, J, i = 1, \dots, n, g = 1, \dots, n_{j,i}\}$, where $n_{j,i}$ is the number of measurements of the longitudinal outcome j for individual i , generating the vector of repeated measurements $\mathbf{Y}_{j,i} = (Y_{j,i,1}, \dots, Y_{j,i,n_{j,i}})^\top$, with $Y_{j,i,g} \equiv Y_{j,i}(t_{j,i,g})$. That is, $Y_{j,i,g}$ is the value of the j th longitudinal outcome for individual i at time $t_{j,i,g}$. The number of measurements and the time points at which measurements are taken can differ between individuals, and a given individual can have different outcomes measured at different time points. Each individual may either experience one of the K distinct competing terminal events or be right-censored during follow-up. Let T_i denote the observed failure time for the i th individual, taken as $T_i = \min(T_{1,i}^*, \dots, T_{K,i}^*, C_i)$, where $T_{k,i}^*$ is their true failure time for each event $k = 1, \dots, K$, and C_i is the corresponding independent censoring time. The event indicator takes values $\delta_i^T \in \{0, 1, \dots, K\}$, with 0 corresponding to censoring and $1, \dots, K$ to the competing terminal events. We assume that the missing values in the longitudinal measurements, aside from those caused by the K events, are missing at random. Regarding the recurrent event process, let $R_{l,i}$ denote the time of the l th recurrent event experienced by the i th individual, $l = 1, \dots, L_i$, treated as $R_{l,i} = \min(R_{l,i}^*, T_i)$, with $R_{l,i}^*$ being the l th true failure time. The event indicator $\delta_{l,i}^R$ is 1 if $R_{l,i}^* < T_i$ and 0 otherwise. Joint models assume a full joint distribution of the longitudinal and time-to-event processes $(\mathbf{Y}_i, T_i, \mathbf{R}_i)$, where $\mathbf{Y}_i = (\mathbf{Y}_{1,i}^\top, \dots, \mathbf{Y}_{J,i}^\top)^\top$ and $\mathbf{R}_i = (R_{1,i}, \dots, R_{L_i,i})^\top$.

Let $\mathcal{D}_n = \{(\mathbf{Y}_i, T_i, \delta_i^T, \mathbf{R}_i, \delta_i^R), i = 1, \dots, n\}$ denote the observed information from a random sample of n individuals of the target population, where $\delta_i^R = (\delta_{1,i}^R, \dots, \delta_{L_i,i}^R)^\top$. The unknown parameters $\boldsymbol{\theta}$, the subject-specific random effects $\mathbf{b} = (\mathbf{b}_1^\top, \dots, \mathbf{b}_n^\top)^\top$, and the frailty terms $\mathbf{v} = (v_1, \dots, v_n)^\top$ are estimated from the posterior distribution

$$p(\boldsymbol{\theta}, \mathbf{b}, \mathbf{v} \mid \mathcal{D}_n) \propto p(\mathcal{D}_n \mid \boldsymbol{\theta}, \mathbf{b}, \mathbf{v}) p(\boldsymbol{\theta}, \mathbf{b}, \mathbf{v}),$$

where $p(\mathcal{D}_n \mid \boldsymbol{\theta}, \mathbf{b}, \mathbf{v})$ is the full likelihood of the model and $p(\boldsymbol{\theta}, \mathbf{b}, \mathbf{v})$ is the prior distribution. To evaluate the joint likelihood of the longitudinal and time-to-event outcome data, we assume that, given all observed covariates and the unobserved random effects, the longitudinal and survival processes are independent of each other, as are any given subject's longitudinal responses. Under this conditional independence assumption, the full likelihood can be written as

$$\begin{aligned} p(\mathcal{D}_n \mid \boldsymbol{\theta}, \mathbf{b}, \mathbf{v}) &= \prod_{i=1}^n \prod_{j=1}^J \prod_{g=1}^{n_{j,i}} p(Y_{j,i,g} \mid \boldsymbol{\theta}_j^Y, \boldsymbol{\theta}_j^b, \mathbf{b}_{j,i}) p(T_i, \delta_i^T \mid \boldsymbol{\theta}^T, \boldsymbol{\theta}^Y, \boldsymbol{\theta}^b, \theta^v, \mathbf{b}_i, v_i^R) \\ &\quad \times \prod_{l=1}^{L_i} p(R_{l,i}, \delta_{l,i}^R \mid \boldsymbol{\theta}^R, \boldsymbol{\theta}^Y, \boldsymbol{\theta}^b, \theta^v, \mathbf{b}_i, v_i^R) p(\mathbf{b}_i \mid \boldsymbol{\theta}^b) p(v_i^R \mid \theta^v), \end{aligned} \quad (2)$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}^{Y^\top}, \boldsymbol{\theta}^{T^\top}, \boldsymbol{\theta}^{R^\top}, \boldsymbol{\theta}^{b^\top}, \theta^v)^\top$ is the combined vector of unknown parameters. The

longitudinal model parameters are denoted by $\boldsymbol{\theta}^Y = (\boldsymbol{\theta}_1^{Y\top}, \dots, \boldsymbol{\theta}_J^{Y\top})^\top$, with $\boldsymbol{\theta}_j^Y = (\boldsymbol{\beta}_j^\top, \sigma_{y_j})^\top$. The parameter vector of the competing risks models is represented by $\boldsymbol{\theta}^T = (\boldsymbol{\theta}_1^{T\top}, \dots, \boldsymbol{\theta}_K^{T\top})^\top$, with $\boldsymbol{\theta}_k^T = (\boldsymbol{\gamma}_k^{T\top}, \boldsymbol{\alpha}_k^{T\top}, \boldsymbol{\gamma}_{0k}^{T\top}, \alpha_k^v)^\top$. Finally, $\boldsymbol{\theta}^R = (\boldsymbol{\gamma}^{R\top}, \boldsymbol{\alpha}^{R\top}, \boldsymbol{\gamma}_0^{R\top})^\top$ is the parameter vector of the recurrent time-to-event model, $\boldsymbol{\theta}^b \equiv \mathbf{D}$ and $\theta_j^b \equiv \mathbf{D}_j$ are the random-effects covariance matrices, and $\theta^v \equiv \sigma_v$ is the frailty standard deviation.

In (2), the likelihood contribution $p(Y_{j,i,g} | \boldsymbol{\theta}_j^Y, \boldsymbol{\theta}_j^b, \mathbf{b}_{j,i})$ of the g th observation from the j th outcome of the i th individual is the probability mass or density function of the distribution considered. For a normal distribution, the contribution takes the form

$$p(Y_{j,i,g} | \boldsymbol{\theta}_j^Y, \boldsymbol{\theta}_j^b, \mathbf{b}_{j,i}) \propto \sigma_{y_j}^{-1/2} \exp \left\{ -\frac{(Y_{j,i,g} - \mathbf{x}_{j,i,g}^\top \boldsymbol{\beta}_j - \mathbf{z}_{j,i,g}^\top \boldsymbol{\theta}_j^b)^2}{2\sigma_{y_j}^2} \right\}.$$

When assuming a beta distribution, this is instead

$$p(Y_{j,i,g}^* | \boldsymbol{\theta}_j^Y, \boldsymbol{\theta}_j^b, \mathbf{b}_{j,i}) = \frac{\Gamma(\phi)}{\Gamma(\mu_{j,i,g} \phi) \Gamma((1 - \mu_{j,i,g}) \phi)} Y_{j,i,g}^{*\mu_{j,i,g} \phi - 1} (1 - Y_{j,i,g}^*)^{(1 - \mu_{j,i,g}) \phi - 1},$$

where $\mu_{j,i,g} = \mathcal{G}^{-1}(\mathbf{x}_{j,i,g}^\top \boldsymbol{\beta}_j + \mathbf{z}_{j,i,g}^\top \boldsymbol{\theta}_j^b)$, $\Gamma(\cdot)$ denotes the gamma function, and $Y_{j,i,g}^*$ is the observed response $Y_{j,i,g}$ transformed to the standard unit interval, so that $Y_{j,i,g}^* \in (0, 1)$.

The i th likelihood contribution of the K competing terminal events in (2) takes the form

$$\begin{aligned} p(T_i, \delta_i^T | \boldsymbol{\theta}^T, \boldsymbol{\theta}^Y, \boldsymbol{\theta}^b, \theta^v, \mathbf{b}_i, v_i^R) &= \prod_{k=1}^K \exp \left[\gamma_{0,k_0}^T + \sum_{q=1}^Q \gamma_{0,k_q}^T \text{bs}_{k_q}^T(T_i) + \mathbf{w}_{k,i}^{T\top}(T_i) \boldsymbol{\gamma}_k^T \right. \\ &\quad \left. + \sum_{j=1}^J \sum_{m=1}^{M_j} \mathcal{H}_{k,j,m}^T \{ \eta_{j,i}(T_i) \} \alpha_{k,j,m}^T + v_i^R \alpha_k^v \right]^{I(\delta_i^T = k)} \\ &\times \exp \left(- \sum_{k=1}^K \exp(v_i^R \alpha_k^v) \int_0^{t_{j,i}^T} \exp \left[\gamma_{0,k_0}^T + \right. \right. \\ &\quad \left. \left. + \sum_{q=1}^Q \gamma_{0,k_q}^T \text{bs}_{k_q}^T(T_i) + \mathbf{w}_{k,i}^{T\top}(T_i) \boldsymbol{\gamma}_k^T + \right. \right. \\ &\quad \left. \left. + \sum_{j=1}^J \sum_{m=1}^{M_j} \mathcal{H}_{k,j,m}^T \{ \eta_{j,i}(T_i) \} \alpha_{k,j,m}^T \right] ds \right), \end{aligned} \quad (3)$$

where $I(\cdot)$ is the indicator function, $\text{bs}_{k_q}^T(t)$ is the P-splines' q th basis function of degree d , and γ_{0,k_q}^T are the corresponding unknown coefficients for the baseline hazard. The likelihood

contribution of the l th recurrent event experienced by the i th individual in (2) is given by

$$\begin{aligned}
p(R_{l,i}, \delta_{l,i}^R | \boldsymbol{\theta}^R, \boldsymbol{\theta}^Y, \boldsymbol{\theta}^b, \theta^v, \mathbf{b}_i, v_i^R) &= \exp \left[\gamma_{0_0}^R + \sum_{q=1}^Q \gamma_{0_q}^R \text{bs}_q^R(R_{l,i} - t_{0_{l,i}}) + \mathbf{w}_i^{R\top}(R_{l,i}) \boldsymbol{\gamma}^R \right. \\
&\quad \left. + \sum_{j=1}^J \sum_{m=1}^{M_j} \mathcal{H}_{j,m}^R \{\eta_{j,i}(R_{l,i})\} \alpha_{j,m}^R + v_i^R \right]^{\delta_{l,i}^R} \\
&\quad \times \exp \left(- \exp(v_i^R) \int_0^{t_{j,i}^R} \exp \left[\gamma_{0_0}^R + \right. \right. \\
&\quad \left. \left. + \sum_{q=1}^Q \gamma_{0_q}^R \text{bs}_q^R(R_{l,i} - t_{0_{l,i}}) + \mathbf{w}_i^{R\top}(R_{l,i}) \boldsymbol{\gamma}^R + \right. \right. \\
&\quad \left. \left. + \sum_{j=1}^J \sum_{m=1}^{M_j} \mathcal{H}_{j,m}^R \{\eta_{j,i}(R_{l,i})\} \alpha_{j,m}^R \right] ds \right), \tag{4}
\end{aligned}$$

where $\text{bs}_q^R(t)$ is the P-splines' q th basis function of degree d , and $\gamma_{0_q}^R$ is the corresponding unknown coefficient for the baseline hazard.

The integrals in (3) and (4) do not have analytical solutions. Thus we evaluate them using a 15-point Gauss–Kronrod quadrature rule, following Rizopoulos and Ghosh (2011). The random effects $p(\mathbf{b}_i | \boldsymbol{\theta}^b)$ and $p(v_i^R | \theta^v)$ contribute with the probability density functions of zero-mean multivariate and univariate Gaussian distributions, respectively.

The prior distributions considered for each parameter are listed in Table S1. We assume normal distributions for the fixed effects in both the longitudinal and survival submodels $\{\boldsymbol{\beta}_j, \boldsymbol{\gamma}_k^T\}$, and for the association coefficients $\{\alpha_{j,m}^R, \alpha_{k,j,m}^T, \alpha_k^v\}$. We use gamma distributions for the standard deviations $\{\sigma_{y_j}, \sigma_v\}$ of the frailty and error terms. For the covariance matrix \mathbf{D} , we assume a Lewandowski–Kurowicka–Joe distribution. For the P-spline coefficients in the baseline hazards, we consider multivariate Gaussian distributions,

$$\begin{aligned}
\boldsymbol{\gamma}_{0_k}^T | \tau_k^T &\sim N(\mathbf{0}, \tau_k^T \mathbf{M}_k^T), \quad \tau_k^T \sim \text{Gam}(k_{0_k}^T, \lambda_{0_k}^T), \\
\boldsymbol{\gamma}_0^R | \tau^R &\sim N(\mathbf{0}, \tau^R \mathbf{M}^R), \quad \tau^R \sim \text{Gam}(k_0^R, \lambda_0^R),
\end{aligned}$$

where $\mathbf{M}_k^T = \boldsymbol{\Delta}_{k,u}^{T\top} \boldsymbol{\Delta}_{k,u}^T + \mathbf{I} \epsilon_k^T$ and $\mathbf{M}^R = \boldsymbol{\Delta}_u^{R\top} \boldsymbol{\Delta}_u^R + \mathbf{I} \epsilon^R$ are the penalty matrices such that $\boldsymbol{\Delta}_{k,u}^T$ and $\boldsymbol{\Delta}_u^R$ form u th-order differences of adjacent B-splines, and the terms $\mathbf{I} \epsilon_k^T$ and $\mathbf{I} \epsilon^R$ introduce a small ridge penalty. The smoothness of the splines is controlled by the gamma hyperpriors on τ_k^T and τ^R . For more details on Bayesian P-splines, see the seminal work by Lang and Brezger (Lang and Brezger 2004).

The conditional posterior distributions for the parameters $\boldsymbol{\theta}_j^Y = (\boldsymbol{\beta}_j, \sigma_{y_j})$ of the j th

Table S1: Prior distributions considered for each parameter in the proposed joint model. Abbreviations: LKJ, Lewandowski–Kurowicka–Joe.

Outcome	Parameter	Prior
<i>j</i> th longitudinal marker	β_j	Normal($\mu_{\beta_j}, \Sigma_{\beta_j}$)
	\mathbf{D}	LKJ(η)
	σ_{y_j}	Gamma(k_{y_j}, λ_{y_j})
Recurrent event	$\gamma_0^R \mid \tau^R$	Normal($\mathbf{0}, \tau^R \mathbf{M}^R$)
	τ^R	Gamma(k_0^R, λ_0^R)
	$\gamma_{0_q}^R$	Normal($\mu_{\gamma_0^R}, \sigma_{\gamma_0^R}^2$)
	γ^R	Normal($\mu_{\gamma^R}, \sigma_{\gamma^R}^2$)
	$\alpha_{j,m}^R$	Normal($\mu_{\alpha_{j,m}^R}, \sigma_{\alpha_{j,m}^R}^2$)
	σ_v	Gamma(k_v, λ_v)
	$\gamma_{0_k}^T \mid \tau_k^T$	Normal($\mathbf{0}, \tau_k^T \mathbf{M}_k^T$)
<i>k</i> th competing terminal event	τ_k^T	Gamma($k_{0_k}^T, \lambda_{0_k}^T$)
	γ_k^T	Normal($\mu_{\gamma_k^T}, \Sigma_{\gamma_k^T}$)
	$\alpha_{k,j,m}^T$	Normal($\mu_{\alpha_{k,j,m}^T}, \sigma_{\alpha_{k,j,m}^T}^2$)
	α_k^v	Normal($\mu_{\alpha_k^v}, \sigma_{\alpha_k^v}^2$)

longitudinal outcome and the covariance matrix \mathbf{D} are

$$\begin{aligned}
p(\boldsymbol{\beta}_j \mid \mathcal{D}_n, \mathbf{b}, \mathbf{v}) &\propto \prod_{i=1}^n \prod_{g=1}^{n_{j,i}} p(Y_{j,i,g} \mid \boldsymbol{\theta}_j^Y, \boldsymbol{\theta}_j^b, \mathbf{b}_{j,i}) p(T_i, \delta_i^T \mid \boldsymbol{\theta}^T, \boldsymbol{\theta}^Y, \boldsymbol{\theta}^b, \theta^v, \mathbf{b}_i, v_i^R) \\
&\quad \times \prod_{l=1}^{L_i} p(R_{l,i}, \delta_{l,i}^R \mid \boldsymbol{\theta}^R, \boldsymbol{\theta}^Y, \boldsymbol{\theta}^b, \theta^v, \mathbf{b}_i, v_i^R) p(\mathbf{b}_i \mid \boldsymbol{\theta}^b) p(\boldsymbol{\beta}_j), \\
p(\sigma_{y_j} \mid \mathcal{D}_n, \mathbf{b}, \mathbf{v}) &\propto \prod_{i=1}^n \prod_{g=1}^{n_{j,i}} p(Y_{j,i,g} \mid \boldsymbol{\theta}_j^Y, \boldsymbol{\theta}_j^b, \mathbf{b}_{j,i}) p(\sigma_{y_j})
\end{aligned}$$

and

$$p(\mathbf{D} \mid \mathcal{D}_n, \mathbf{b}, \mathbf{v}) \propto \prod_{i=1}^n p(\mathbf{b}_i \mid \boldsymbol{\theta}^b) p(\mathbf{D}).$$

The conditional posterior distributions for the parameters $\boldsymbol{\theta}^R = (\boldsymbol{\gamma}^R, \boldsymbol{\alpha}^R, \boldsymbol{\gamma}_0^R)$ of the recurrent time-to-event outcome are

$$\begin{aligned}
p(\boldsymbol{\gamma}^R \mid \mathcal{D}_n, \mathbf{b}, \mathbf{v}) &\propto \prod_{i=1}^n \prod_{l=1}^{L_i} p(R_{l,i}, \delta_{l,i}^R \mid \boldsymbol{\theta}^R, \boldsymbol{\theta}^Y, \boldsymbol{\theta}^b, \theta^v, \mathbf{b}_i, v_i^R) p(\boldsymbol{\gamma}^R), \\
p(\boldsymbol{\alpha}^R \mid \mathcal{D}_n, \mathbf{b}, \mathbf{v}) &\propto \prod_{i=1}^n \prod_{l=1}^{L_i} p(R_{l,i}, \delta_{l,i}^R \mid \boldsymbol{\theta}^R, \boldsymbol{\theta}^Y, \boldsymbol{\theta}^b, \theta^v, \mathbf{b}_i, v_i^R) p(\boldsymbol{\alpha}^R), \\
p(\boldsymbol{\gamma}_0^R \mid \mathcal{D}_n, \mathbf{b}, \mathbf{v}) &\propto \prod_{i=1}^n \prod_{l=1}^{L_i} p(R_{l,i}, \delta_{l,i}^R \mid \boldsymbol{\theta}^R, \boldsymbol{\theta}^Y, \boldsymbol{\theta}^b, \theta^v, \mathbf{b}_i, v_i^R) p(\boldsymbol{\gamma}_0^R)
\end{aligned}$$

and

$$p(\sigma_v \mid \mathcal{D}_n, \mathbf{b}, \mathbf{v}) \propto \prod_{i=1}^n p(v_i^R \mid \theta^v) p(\sigma_v).$$

The conditional posterior distributions for the parameters $\boldsymbol{\theta}_k^T = (\boldsymbol{\gamma}_k^T, \boldsymbol{\alpha}_k^T, \boldsymbol{\gamma}_{0_k}^T, \alpha_k^v)$ of the k th competing time-to-event outcome are

$$\begin{aligned}
p(\boldsymbol{\gamma}_k^T \mid \mathcal{D}_n, \mathbf{b}, \mathbf{v}) &\propto \prod_{i=1}^n p(T_i, \delta_i^T \mid \boldsymbol{\theta}_k^T, \boldsymbol{\theta}^Y, \boldsymbol{\theta}^b, \theta^v, \mathbf{b}_i, v_i^R) p(\boldsymbol{\gamma}_k^T), \\
p(\boldsymbol{\alpha}_k^T \mid \mathcal{D}_n, \mathbf{b}, \mathbf{v}) &\propto \prod_{i=1}^n p(T_i, \delta_i^T \mid \boldsymbol{\theta}_k^T, \boldsymbol{\theta}^Y, \boldsymbol{\theta}^b, \theta^v, \mathbf{b}_i, v_i^R) p(\boldsymbol{\alpha}_k^T), \\
p(\boldsymbol{\gamma}_{0_k}^T \mid \mathcal{D}_n, \mathbf{b}, \mathbf{v}) &\propto \prod_{i=1}^n p(T_i, \delta_i^T \mid \boldsymbol{\theta}_k^T, \boldsymbol{\theta}^Y, \boldsymbol{\theta}^b, \theta^v, \mathbf{b}_i, v_i^R) p(\boldsymbol{\gamma}_{0_k}^T),
\end{aligned}$$

and

$$p(\alpha_k^v \mid \mathcal{D}_n, \mathbf{b}, \mathbf{v}) \propto \prod_{i=1}^n p(T_i, \delta_i^T \mid \boldsymbol{\theta}^T, \boldsymbol{\theta}^Y, \boldsymbol{\theta}^b, \theta^v, \mathbf{b}_i, v_i^R) p(\alpha_k^v).$$

The conditional posterior distributions for the random effects \mathbf{b}_i and the frailty term v_i^R are

$$\begin{aligned} p(\mathbf{b}_i \mid \mathcal{D}_i, \mathbf{b}_i, v_i^R) &\propto \prod_{j=1}^J \prod_{g=1}^{n_{j,i}} p(Y_{j,i,g} \mid \boldsymbol{\theta}_j^Y, \boldsymbol{\theta}_j^b, \mathbf{b}_{j,i}) p(T_i, \delta_i^T \mid \boldsymbol{\theta}^T, \boldsymbol{\theta}^Y, \boldsymbol{\theta}^b, \theta^v, \mathbf{b}_i, v_i^R) \\ &\quad \times \prod_{l=1}^{L_i} p(R_{l,i}, \delta_{l,i}^R \mid \boldsymbol{\theta}^R, \boldsymbol{\theta}^Y, \boldsymbol{\theta}^b, \theta^v, \mathbf{b}_i, v_i^R) p(\mathbf{b}_i \mid \boldsymbol{\theta}^b) \end{aligned}$$

and

$$\begin{aligned} p(v_i^R \mid \mathcal{D}_i, \mathbf{b}_i, v_i^R) &\propto p(T_i, \delta_i^T \mid \boldsymbol{\theta}^T, \boldsymbol{\theta}^Y, \boldsymbol{\theta}^b, \theta^v, \mathbf{b}_i, v_i^R) \\ &\quad \times \prod_{l=1}^{L_i} p(R_{l,i}, \delta_{l,i}^R \mid \boldsymbol{\theta}^R, \boldsymbol{\theta}^Y, \boldsymbol{\theta}^b, \theta^v, \mathbf{b}_i, v_i^R) p(v_i^R \mid \theta^v). \end{aligned}$$

We use hierarchical centering for the fixed effects of the longitudinal submodels (Gelfand et al. 1995), and we standardize the covariates of the survival submodels to facilitate the convergence of the MCMC algorithms. Additionally, to speed up the sampling process, we perform parallel sampling of the random effects from different individuals and run the Markov chains in parallel on multiple processor cores.

B An example with JMbayes2

To fit the joint model, users are required to structure their data into two distinct datasets: one dedicated to capturing information related to competing risks and recurrent events (survival dataset), and another focused on longitudinal markers (longitudinal dataset). Below, we provide a subsample of simulated datasets intended as an illustration.

The survival dataset encompasses details pertaining to both competing risks and recurrent events as shown below. Each subject is represented by multiple rows, corresponding to the number of recurrent risk periods, plus one additional row for each competing event. The strata variable is essential to differentiate between various event processes.

	id	tstart	tstop	status	strata	group
1	1	0.00	5.79	0	R	1
2	1	0.00	5.79	0	CR1	1
3	1	0.00	5.79	1	CR2	1
4	2	0.00	7.55	1	R	0
5	2	7.55	9.67	1	R	0
6	2	9.67	10.00	0	R	0
7	2	0.00	10.00	0	CR1	0
8	2	0.00	10.00	0	CR2	0

The longitudinal dataset describes repeated measurements taken on the same subjects, organized in a long format. As shown below, each row corresponds to a single observation, and there might be multiple rows for each subject, representing different measurements over various time points.

	id	time	y1	y2
1	1	0.00	0.89	0.77
2	1	0.26	0.84	0.76
4	1	2.09	0.29	0.69
5	2	0.00	0.93	NA
6	2	2.87	0.16	0.67
7	2	5.37	0.01	0.57
8	2	8.46	NA	0.46
9	2	8.85	0.02	0.46

To adjust the joint model, users must first fit the mixed effects and proportional hazards sub-models. Subsequently, these models are provided as arguments in the `jm()` function. Within the function call, users specify the preferred functional forms for the longitudinal outcomes in each relative-risk model, along with the chosen timescale. An illustrative example is presented below.

```

# 1. Load the package
library(JMbayes2)

# 2. Fit the longitudinal and survival submodels

# 2.1 Bounded longitudinal outcome (beta distribution)
beta_fit <- mixed_model(y2 ~ time treat, # fixed-effects formula
                        random = ~ time | id, # random-effects formula
                        family = beta.fam(), # distribution family
                        data = long) # longitudinal dataset

# 2.2 Unbounded longitudinal outcome (Gaussian distribution)
gaus_fit <- lme(y1 ~ time, # fixed-effects formula
                random = ~ time | id, # random-effects formula
                data = long) # longitudinal dataset

# 2.3 Proportional hazards model
ph_fit <- coxph(Surv(tsart, tstop, status) ~
                group : strata(strata), # model formula
                data = surv) # survival dataset

# 3. Fit the joint model that links the submodels
jm_fit <- jm(ph_fit, # survival submodel
             list(beta_fit, gaus_fit), # longitudinal submodels
             time_var = "time", # time variable in the longitudinal
             # submodels
             recurrent = "gap", # event timescale, or "calendar"
             functional_forms = ~ vexpit(value(y1)):strata # func-forms
             + value(y2)):strata) # formula

summary(jm_fit)

```

Further details about the package usage can be found on the dedicated website:
<https://drizopoulos.github.io/JMbayes2/>.

Table S2: Functional forms available in the R package `JMbayes2` to link the longitudinal and time-to-event outcomes, and the associated transformation functions. †: `velocity()` can be used as an alias for `slope()`.

Functional form	Function	Argument
Underlying value	$\eta_{j,i}(t)$	<code>value()</code>
	$\log \{\eta_{j,i}(t)\}$	<code>vlog(value())</code>
	$\log_2 \{\eta_{j,i}(t)\}$	<code>vlog2(value())</code>
	$\log_{10} \{\eta_{j,i}(t)\}$	<code>vlog10(value())</code>
	$\sqrt{\eta_{j,i}(t)}$	<code>vsqrt(value())</code>
	$\exp \{\eta_{j,i}(t)\}$	<code>vexp(value())</code>
	$\expit \{\eta_{j,i}(t)\}$	<code>vexpit(value())</code>
	$a + b\eta_{j,i}(t) + c\eta_{j,i}^2(t)$	<code>poly2(value())</code>
	$a + b\eta_{j,i}(t) + c\eta_{j,i}^2(t) + d\eta_{j,i}^3(t)$	<code>poly3(value())</code>
	$a + b\eta_{j,i}(t) + c\eta_{j,i}^2(t) + d\eta_{j,i}^3(t) + e\eta_{j,i}^4(t)$	<code>poly4(value())</code>
Slope	$d\eta_{j,i}(t)/dt$	<code>slope()</code> †
	$ d\eta_{j,i}(t)/dt $	<code>vabs(slope())</code>
	$d \exp \{\eta_{j,i}(t)\} / dt$	<code>Dexp(slope())</code>
	$d \expit \{\eta_{j,i}(t)\} / dt$	<code>Dexpit(slope())</code>
Acceleration	$d^2\eta_{j,i}(t)/dt^2$	<code>acceleration()</code>
Standardized cumulative effect	$\frac{1}{t} \int_0^t \eta_{j,i}(s) ds$	<code>area()</code>

C Simulation study

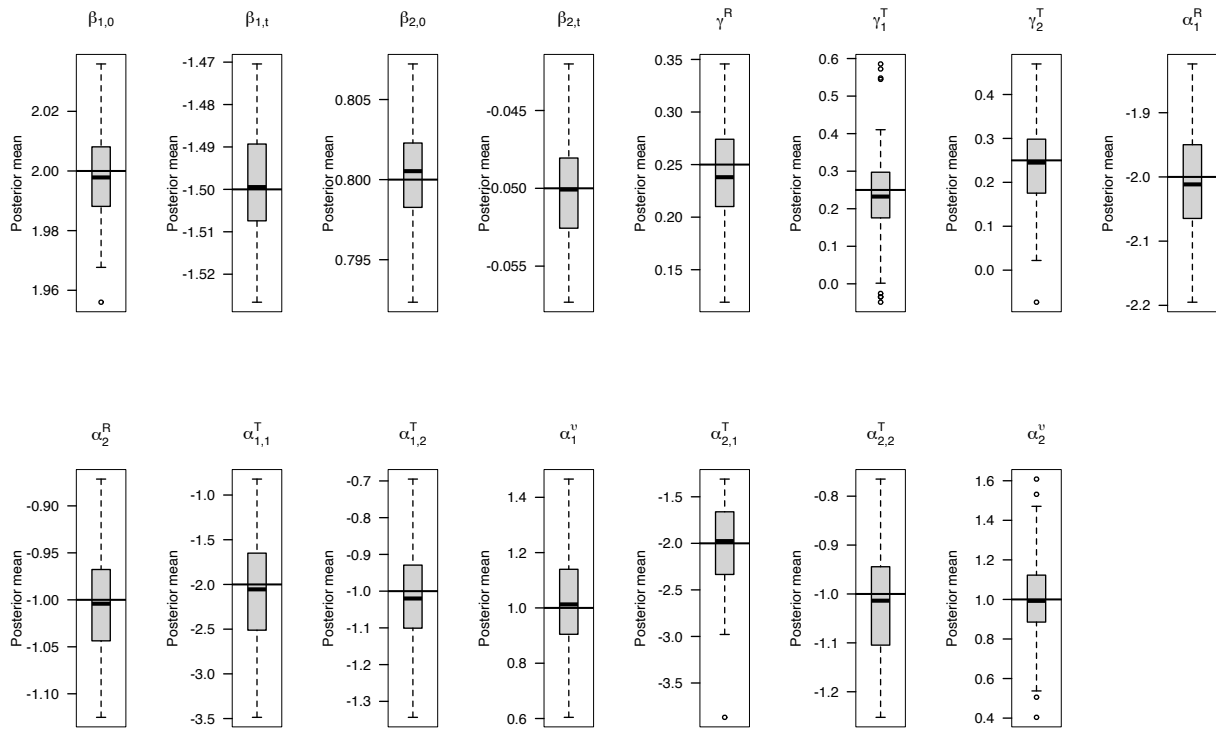


Figure S1: Estimated posterior means for joint model coefficients obtained in the simulation scenario A.

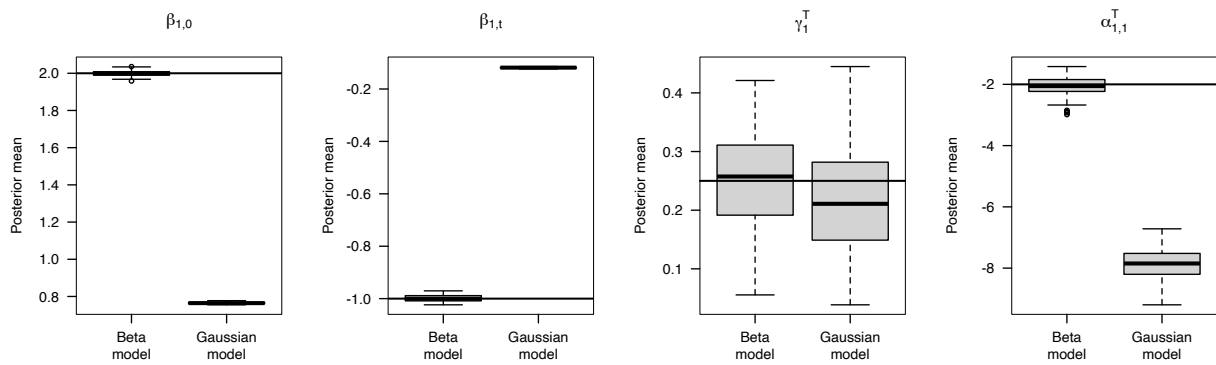


Figure S2: Estimated posterior means for joint model coefficients obtained in the simulation scenario B.

Table S3: Comparison of joint models considered under the simulation scenarios A and B. In scenario A, the fitted model is equal to the data generation model. In scenario B, the fitted model uses a Gaussian distribution to model the beta longitudinal outcome from the data generation model. We highlight the differences between the data and fitted models by enclosing varying elements within boxes in the model formulas. Abbreviations: M₁, 1st longitudinal marker; M₂, 2nd longitudinal marker; MSE, mean squared error; PEx, pulmonary exacerbation; R, recurrent event; T₁, 1st competing/terminal event; T₂, 2nd competing event.

	Scenario A	Scenario B	
	Data/Fit model	Data model	Fit model
M ₁	$\text{logit} \{ \mu_{1,i}(t) \} = \eta_{1,i}(t) =$ $= (\beta_{1,0} + b_{1,0,i}) + (\beta_{1,t} + b_{1,t,i})t$ $(b_{1,0,i}, b_{1,t,i}) \sim \mathcal{N} \left(\mathbf{0}, \begin{bmatrix} \sigma_{1,0}^2 & 0 \\ 0 & \sigma_{1,t}^2 \end{bmatrix} \right)$	$\boxed{\text{logit} \{ \mu_{1,i}(t) \}} = \eta_{1,i}(t) =$ $= (\beta_{1,0} + b_{1,0,i}) + (\beta_{1,t} + b_{1,t,i})t$ $(b_{1,0,i}, b_{1,t,i}) \sim \mathcal{N} \left(\mathbf{0}, \begin{bmatrix} \sigma_{1,0}^2 & 0 \\ 0 & \sigma_{1,t}^2 \end{bmatrix} \right)$	$\boxed{\mu_{1,i}(t)} = \eta_{1,i}(t) + \boxed{\varepsilon_{1,i}(t)} =$ $= (\beta_{1,0} + b_{1,0,i}) + (\beta_{1,t} + b_{1,t,i})t + \boxed{\varepsilon_{1,i}(t)}$ $(b_{1,0,i}, b_{1,t,i}) \sim \mathcal{N} \left(\mathbf{0}, \begin{bmatrix} \sigma_{1,0}^2 & 0 \\ 0 & \sigma_{1,t}^2 \end{bmatrix} \right)$ $\varepsilon_{1,i}(t) \sim \mathcal{N} (0, \sigma_{1,t}^2)$
M ₂	$\mu_{2,i}(t) = \eta_{2,i}(t) + \varepsilon_{2,i}(t) =$ $= (\beta_{2,0} + b_{2,0,i}) + (\beta_{2,t} + b_{2,t,i})t + \varepsilon_{2,i}(t)$ $(b_{2,0,i}, b_{2,t,i}) \sim \mathcal{N} \left(\mathbf{0}, \begin{bmatrix} \sigma_{2,0}^2 & 0 \\ 0 & \sigma_{2,t}^2 \end{bmatrix} \right)$ $\varepsilon_{2,i}(t) \sim \mathcal{N} (0, \sigma_{2,t}^2)$	-	-
R	$h_i^R(t) = h_0^R(t) \exp \left[w_i^R \gamma^R + v_i^R \right]$ $+ \text{expit} \{ \eta_{1,i}(t) \} \alpha_1^R + \eta_{2,i}(t) \alpha_2^R$	-	-
T ₁	$h_{1,i}^T(t) = h_{0_1}^T(t) \exp \left[w_{1,i}^T \gamma_1^T + v_i^R \alpha_1^v \right]$ $+ \text{expit} \{ \eta_{1,i}(t) \} \alpha_{1,1}^T + \eta_{2,i}(t) \alpha_{1,2}^T$	$h_{1,i}^T(t) = h_{0_1}^T(t) \exp \left[w_{1,i}^T \gamma_1^T + \boxed{\text{expit} \{ \eta_{1,i}(t) \}} \alpha_{1,1}^T \right]$	$h_{1,i}^T(t) = h_{0_1}^T(t) \exp \left[w_{1,i}^T \gamma_1^T + \boxed{\eta_{1,i}(t)} \alpha_{1,1}^T \right]$
T ₂	$h_{2,i}^T(t) = h_{0_2}^T(t) \exp \left[w_{2,i}^T \gamma_2^T + v_i^R \alpha_2^v \right]$ $+ \text{expit} \{ \eta_{1,i}(t) \} \alpha_{2,1}^T + \eta_{2,i}(t) \alpha_{2,2}^T$	-	-

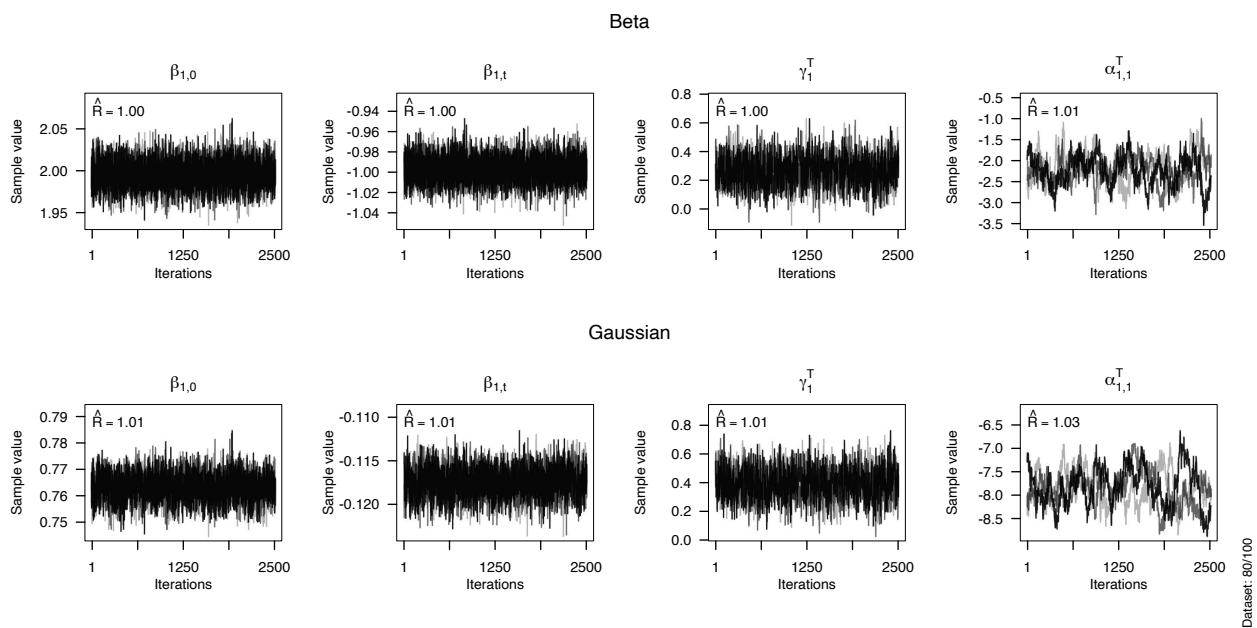


Figure S3: Traceplots for the joint model coefficients' Markov chains after warm-up, for a randomly chosen dataset under scenario B. Top: Joint model with a beta submodel. Bottom: Joint model with a Gaussian submodel.

Table S4: Parameter values employed in the joint model for generating data in the simulation study. Abbreviations: M₁, 1st longitudinal marker; M₂, 2nd longitudinal marker; MSE, mean squared error; PEx, pulmonary exacerbation; R, recurrent event; T₁, 1st competing/terminal event; T₂, 2nd competing event.

	Scenario A	Scenario B
M ₁		
$\beta_{1,0}$	2.000	1.00
$\beta_{1,t}$	-1.500	-1.50
$\sigma_{1,0}$	0.250	0.25
$\sigma_{1,t}$	0.150	0.15
ϕ_y	10 ⁴	10 ⁴
M ₂		
$\beta_{2,0}$	0.800	–
$\beta_{2,t}$	-0.050	–
$\sigma_{2,0}$	0.010	–
$\sigma_{2,t}$	0.010	–
σ_y	0.005	–
R		
$h_{0_1}^R$	0.200	–
γ^R	0.250	–
α_1^R	-2.000	–
α_2^R	-1.000	–
T ₁		
$h_{0_1}^T$	0.200	0.10
γ_1^T	0.250	0.25
$\alpha_{1,1}^T$	-2.000	-2.00
$\alpha_{1,2}^T$	-1.000	–
α_1^V	1.000	–
T ₂		
$h_{0_2}^T$	2.000	–
γ_2^T	0.250	–
$\alpha_{2,1}^T$	-2.000	–
$\alpha_{2,2}^T$	-1.000	–
α_2^V	1.000	–

Table S5: Outline of the data generation process for scenario A.

Longitudinal outcome (1/2):

- 1: Generate $n = 1000$ random samples from $\mathcal{N}(\mathbf{0}, \Sigma^{-1})$ for the individual-specific random effects, $\mathbf{b}_i = (\mathbf{b}_{1,i}^\top, \mathbf{b}_{2,i}^\top)^\top: \mathbf{b}$.

$$\begin{matrix} n \times 2 & n \times 2 & n \times 4 \end{matrix}$$
- 2: Generate $(n \times (n_i - 1))$ random samples from $\mathcal{U}(0, 10)$ for the individual visiting times and add the time 0, $\mathbf{t}_i: \mathbf{t}$.

$$(n \cdot n_i) \times 1$$
- 3: Generate the $(n \times 2)$ vectors of n_i individual underlying longitudinal responses, $\boldsymbol{\mu}_{1,i}$ and $\boldsymbol{\mu}_{2,i}$: $\boldsymbol{\mu}_{i,1} = \text{expit}(\boldsymbol{\eta}_{1,i})$ and $\boldsymbol{\mu}_{i,2} = \boldsymbol{\eta}_{2,i}$, where $\boldsymbol{\eta}_{j,i} = [\mathbf{1} \ \mathbf{t}_i] \boldsymbol{\beta}_j + [\mathbf{1} \ \mathbf{t}_i] \mathbf{b}_i$, $j = 1, 2$.

$$\begin{matrix} n_i \times 1 & n_i \times 2 & 2 \times 1 & n_i \times 2 & 2 \times 1 \end{matrix}$$
- 4: Generate $(n \times n_i)$ random samples from Beta (p, q) , where $p = \phi \times \mu_{1,i}(t)$ and $q = \phi \times \{1 - \mu_{1,i}(t)\}$, for the observed beta longitudinal responses: \mathbf{y}_1 .

$$(n \cdot n_i) \times 1$$
- 5: Generate $(n \times n_i)$ random samples from $\mathcal{N}(0, \sigma_y^2)$ for the observation measurement error, $\varepsilon_{2,i}(t): \boldsymbol{\varepsilon}_2$.

$$(n \cdot n_i) \times 1$$
- 6: Obtain the observed Gaussian longitudinal response by summing the vectors $\boldsymbol{\eta}_2$ and $\boldsymbol{\varepsilon}_2$:

$$\mathbf{y}_2$$
.

$$(n \cdot n_i) \times 1$$

Survival outcome:

- 7: Generate n random samples from Bern (0.5) for the individual's group, $w_i: \mathbf{w}$.

$$n \times 1$$
- 8: Generate $(n \times 2)$ random samples from $\mathcal{U}(0, 1)$, $u_{1,i}^\top$ and $u_{2,i}^\top: \mathbf{u}_1^\top$ and \mathbf{u}_2^\top .

$$\begin{matrix} n \times 1 & n \times 1 \end{matrix}$$
- 9: Define $H_{j,i}^\top(t) = \int_0^t h_{j,i}^\top(s) ds$, where $h_{j,i}^\top(t) = h_{j,0}^\top(t) \exp\{w_i \gamma_j^\top + \text{expit}\{\eta_{1,i}(t)\} \alpha_{j,1}^\top + \eta_{2,i}(t) \alpha_{j,2}^\top + v_i \alpha_j^v\}$, $j = 1, 2$.
- 10: Numerically solve $\exp(-H_{1,i}^\top(t_i^{T*})) = u_{j,i}^\top$ for $t_{j,i}^{T*}$ (Bender et al. 2005), for $j = 1, 2$, to obtain the individual true event times: \mathbf{t}_1^{T*} and \mathbf{t}_2^{T*} .

$$\begin{matrix} n \times 1 & n \times 1 \end{matrix}$$
- 11: Calculate the observed event time $t_i^T = \min(t_{1,i}^{T*}, t_{2,i}^{T*}, t_{\max})$, where t_{\max} is the deterministic maximum follow-up time: \mathbf{t}^T .

$$n \times 1$$
- 12: Define the censoring indicator δ_i^T as 1 if $t_i^T = t_{1,i}^{T*}$, 2 if $t_i^T = t_{2,i}^{T*}$, and 0 otherwise.

Longitudinal outcome (2/2):

- 13: Remove all $\mathbf{y}_{1,i}(t)$ and $\mathbf{y}_{2,i}(t)$ for $t > t_i$.

Recurrent outcome:

- 14: Generate n random samples from $\mathcal{U}(0, 1)$, $u_{l,i}^R: \mathbf{u}_l^R$.

$$n \times 1$$
- 15: Define $H_i^R(t) = \int_0^t h_i^R(s) ds$, where $h_i^R(t) = h_0^R(t) \exp\{w_i \gamma^R + \text{expit}\{\eta_{1,i}(t)\} \alpha_1^R + \eta_{2,i}(t) \alpha_2^R\}$.
- 16: Numerically solve $\exp(-H_i^R(t_{l,i}^{R*})) = u_{l,i}^R$ for $t_{l,i}^{R*}$ (Bender et al. 2005), to obtain the individual true l th recurrent event times: \mathbf{t}_l^{R*} .

$$n \times 1$$
- 17: Calculate the l th observed event time $t_{l,i}^R = \min(t_{l,i}^{R*}, t_i^T): \mathbf{t}_l^R$.

$$n \times 1$$
- 18: Define the censoring indicator as $\delta_{l,i}^R$ as 1 if $t_{l,i}^R = t_i^T$, and 0 otherwise.
- 19: Repeat steps 14–18 for each individual until $\sum_i t_{l,i}^R > t_i^T$.

Table S6: Outline of the data generation process for scenario B.

Longitudinal outcome (1/2):

- 1: Generate $n = 1000$ random samples from $\mathcal{N}(\mathbf{0}, \Sigma^{-1})$ for the individual-specific random effects, \mathbf{b}_i : \mathbf{b} .
 $n \times 2$
 - 2: Generate $(n \times (n_i - 1))$ random samples from $\mathcal{U}(0, 10)$ for the individual visiting times and add the time 0, \mathbf{t}_i : \mathbf{t} .
 $(n \cdot n_i) \times 1$
 - 3: Generate the n vectors of n_i individual underlying longitudinal responses, $\boldsymbol{\mu}_i$: $\boldsymbol{\mu}_i = \text{expit}(\boldsymbol{\eta}_i)$, where $\boldsymbol{\eta}_i = \begin{bmatrix} \mathbf{1} & \mathbf{t}_i \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{1} & \mathbf{t}_i \end{bmatrix} \mathbf{b}_i$.
 $n_i \times 1$ $n_i \times 2$ 2×1 $n_i \times 2$ 2×1
 - 4: Generate $(n \times n_i)$ random samples from Beta (p, q) , where $p = \phi \times \boldsymbol{\mu}_i$ and $q = \phi \times (1 - \boldsymbol{\mu}_i)$, for the observed longitudinal responses: \mathbf{y} .
 $(n \cdot n_i) \times 1$
-

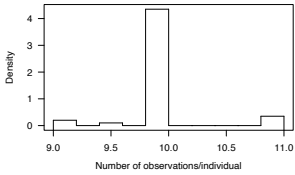
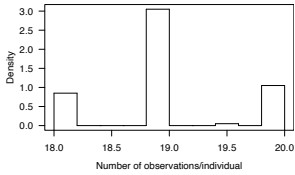
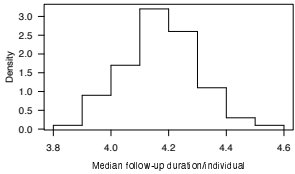
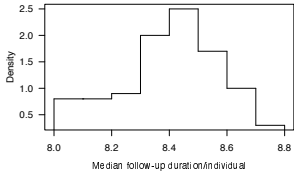
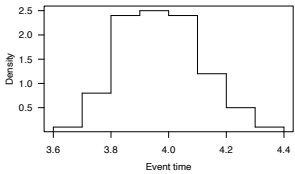
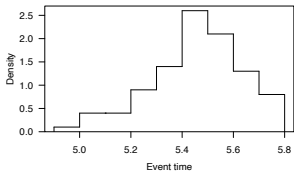
Survival outcome:

- 7: Generate n random samples from Bern (0.5) for the individual's group, w_i : \mathbf{w} .
 $n \times 1$
 - 8: Generate n random samples from $\mathcal{U}(0, 1)$, u_i : \mathbf{u} .
 $n \times 1$
 - 9: Define $H_i(t) = \int_0^t h_i(s) ds$, where $h_i(t) = h_0(t) \exp \{w_i \gamma + \text{expit} \{ \eta_{1,i}(t) \} \alpha_{1,1}^T \}$.
 - 10: Numerically solve $\exp(-H_i(t_i^*)) = u_i$ for t_i^* (Bender et al. 2005) to obtain the individual true event times: \mathbf{t}^* .
 $n \times 1$
 - 11: Calculate the observed event times $t_i = \min(t_i^*, t_{\max})$, where t_{\max} is the deterministic maximum follow-up time: \mathbf{t} .
 $n \times 1$
 - 12: Define the censoring indicator as $\delta_i = \begin{cases} 1 & t_i \leq t_{\max}, \\ 0 & t_i > t_{\max}. \end{cases}$
-

Longitudinal outcome (2/2):

- 13: Remove all $y_i(t)$ for $t > t_i$.
-

Table S7: Characteristics of the simulated datasets. Abbreviations: ind, individual; IQR, interquartile range; pct, percentile.

	Scenario A	Scenario B
Number of replicas	100	100
Number of individuals	1000	1000
Number of observations		
M ₁ , median (IQR)	10945.5 (10852–11087.75)	15334 (15211.5–15443.25)
M ₂ , median (IQR)	10945.5 (10852–11087.75)	–
Number of observations/individual		
M ₁ , median, median (IQR)	10 (10–10)	19 (19–19)
M ₂ , median, median (IQR)	10 (10–10)	–
		
Aggregated follow-up duration		
M ₁ , median (IQR)	4750.81 (4699.53–4823.44)	7041.97 (6975.63–7103.06)
M ₂ , median (IQR)	4750.81 (4699.53–4823.44)	–
Follow-up duration/individual		
M ₁ , median, median (IQR)	4.17 (4.09–4.26)	8.42 (8.31–8.53)
M ₂ , median, median (IQR)	4.17 (4.09–4.26)	–
		
Competing/terminal event time		
T ₁ , median, median (IQR)	3.97 (3.88–4.07)	5.48 (5.35–5.58)
T ₂ , median, median (IQR)	3.97 (3.88–4.07)	–
Censoring, median, median (IQR)	10 (10–10)	10 (10–10)
		
Competing/terminal event		
T ₁ , %, median (IQR)	0.42 (0.41–0.43)	0.54 (0.53–0.55)
T ₂ , %, median (IQR)	0.41 (0.4–0.42)	–
Censoring, %, median (IQR)	0.17 (0.16–0.18)	0.46 (0.45–0.47)
Number of recurrent events/individual		
Median, median (IQR)	3 (3–3)	–
Group		
1, %, median (IQR)	0.5 (0.49–0.51)	0.5 (0.49–0.51)

D CFFPR study

Here, we explain how to interpret the association parameter estimates from the proposed joint model.

A q -unit increase in the ppFEV₁'s expected value changes the hazard rate of PEx by a factor of $\exp\left\{\frac{\alpha_{1,1}^R}{150} \times q\right\}$. We divide $\alpha_{1,1}^R$ by 150 to rescale the parameter to the original marker scale. The same rationale applies to the association parameters $\alpha_{1,1,1}^T$ and $\alpha_{2,1,1}^T$ regarding the risks of transplantation and death, respectively.

A p -unit increase in the rate of decline of ppFEV₁'s expected value changes the hazard rate of transplantation by a factor of $\exp\left\{\frac{\alpha_{1,1,2}^T}{150} \times p\right\}$. We divide $\alpha_{1,1,2}^T$ by 150 because we transform the marker from the range of 0 to 150 to the desired range of 0 to 1, in which the beta distribution is defined. In other words, $\text{ppFEV}_{1i}(t) = \frac{\text{ppFEV}_{1i}^*(t)}{150}$, where $\text{ppFEV}_{1i}^*(t)$ and $\text{ppFEV}_{1i}(t)$ denote markers in the original and transformed scales, respectively. Given that $\frac{d}{dt} \text{ppFEV}_{1i}(t) = \frac{1}{150} \frac{d}{dt} \text{ppFEV}_{1i}^*(t)$ and noting that $\alpha_{1,1,2}^T \left(\frac{1}{150} \frac{d}{dt} \text{ppFEV}_{1i}^*(t)\right) = \frac{\alpha_{1,1,2}^T}{150} \left(\frac{d}{dt} \text{ppFEV}_{1i}^*(t)\right)$, we can obtain the association parameter in the original scale by doing $\alpha_{1,1,2}^{T*} = \frac{\alpha_{1,1,2}^T}{150}$. The same reasoning applies to the association parameter $\alpha_{2,1,2}^T$ regarding the risks of death.

An b -unit increase in the BMI's standardized cumulative value changes the hazard rate of PEx by a factor of $\exp\left\{\alpha_{2,1}^R \times b\right\}$. The same reasoning applies to the association parameters $\alpha_{1,2,1}^T$ and $\alpha_{2,2,1}^T$ regarding the risks of transplantation and death, respectively.

An f -unit increase in the frailty term v_i^R changes the hazard rate of transplantation by a factor of $\exp\left\{\alpha_1^v \times f\right\}$. The same reasoning applies to the association parameter α_2^v for the risk of death.

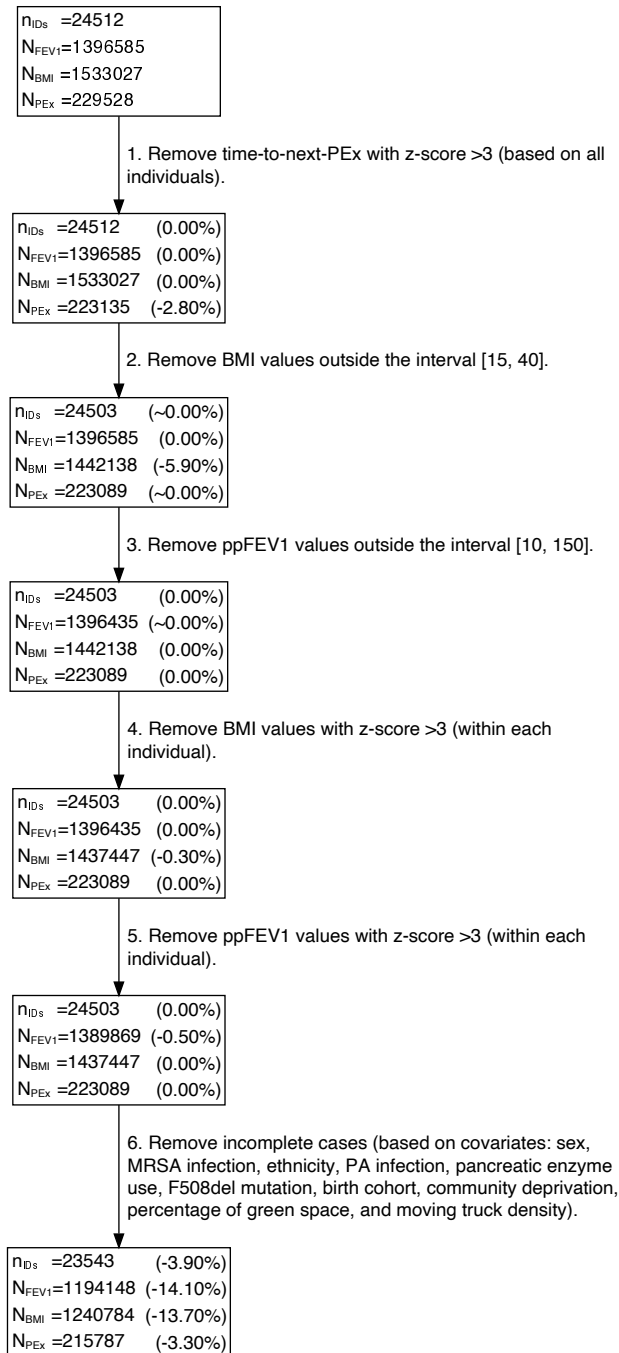


Figure S4: Cleaning process of CFFPR data sample. The sequence of steps 1–5 was employed to identify and remove atypical measurements likely arising from data entry errors. n_{IDS} : number of individuals; n_{FEV1} : number of ppFEV₁ measurements; n_{BMI} : number of BMI measurements; n_{PEX} : number of PEX.

Table S8: Follow-up, demographic, social, and clinical characteristics of the CF individuals analyzed. Abbreviations: BMI, body mass index; CF, cystic fibrosis; IQR, interquartile range; PEx, pulmonary exacerbation; ppFEV₁, percent predicted forced expiratory volume in one second. †: Percentage of greenspace, impervious, and tree canopy areas within the ZIP Code Tabulation Area (ZCTA) derived from the National Land Cover Database (Jin et al. 2019)

Characteristics		
Number of individuals		23,543
Number of measurements		
	ppFEV ₁	1,523,40
	BMI	1,523,40
Number of measurements/individual		
	ppFEV ₁ , median (IQR)	46.00 (27.00–69.00)
	BMI, median (IQR)	48.00 (27.00–72.00)
Aggregated follow-up duration (years)		
	ppFEV ₁	266,345.20
	BMI	262,875.00
Follow-up duration/individual (years)		
	ppFEV ₁ , median (IQR)	11.92 (6.97–16.76)
	BMI, median (IQR)	11.72 (6.85–16.61)
Baseline age (years)		
	ppFEV ₁ , median (IQR)	11.37 (6.36–20.19)
	BMI, median (IQR)	11.80 (6.66–20.15)
Age at end of follow-up (years)		
	Censoring, median (IQR)	23.50 (17.07–32.15)
	Lung transplantation, median (IQR)	28.52 (22.84–36.55)
	Death, median (IQR)	26.57 (21.36–35.93)
Competing terminal event		
	Censoring	16,751 (71.15%)
	Lung transplantation	2,562 (10.88%)
	Death	4,230 (17.97%)
Number of PEx/individual		
	Median (IQR)	7.00 (3.00–14.00)
Interval between consecutive PEx (years)		
	Median (IQR)	0.34 (0.15–0.77)
Baseline ppFEV ₁		
	Median (IQR)	80.30 (59.70–95.90)
Baseline BMI		
	Median (IQR)	17.17 (15.66–20.31)
Birth cohort		
	<1993	13,895 (59.02%)
	[1993, 1998)	3,672 (15.60%)
	≥1998	5,976 (25.38%)
Genotype (F508del)		
	Homozygous	11,236 (47.73%)
	Heterozygous	8,655 (36.76%)
	Neither	3,652 (15.51%)
Sex		
	Female	11,829 (50.24%)
Ethnicity		
	Hispanic	1,767 (7.51%)
	Other	21,776 (92.49%)
Neighborhood deprivation index		
	Median (IQR)	0.33 (0.27–0.40)
Percentage of green space [†]		
	Median (IQR)	89.81 (71.81–96.94)
Moving-truck density (truck-meters/m ²)		
	Median (IQR)	0.18 (0.00–0.94)
Pancreatic enzymes intake		
	At baseline	6,887 (29.25%)
	Throughout follow-up	1,868 (7.93%)
	Sometime during follow-up	22,564 (95.84%)