Statistics in Medicine

WILEY

Statistics in Medicine

RESEARCH ARTICLE OPEN ACCESS

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. Clancy[6] | Rhonda D. Szczesniak[4,5,7] | Eleni-Rosalina Andrinopoulou[1,2]

[1]Department of Biostatistics, Erasmus University Medical Center, Rotterdam, the Netherlands | [2]Department of Epidemiology, Erasmus University Medical Center, Rotterdam, the Netherlands | [3]Duke University School of Nursing, Durham, NC, USA | [4]Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA | [5]Department of Pediatrics, University of Cincinnati, Cincinnati, OH, USA | [6]Cystic Fibrosis Foundation, Bethesda, MD, USA | [7]Division of Pulmonary Medicine, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

Correspondence: Pedro Miranda Afonso (p.mirandaafonso@erasmusmc.nl)

Received: 24 May 2024 | **Revised:** 22 January 2025 | **Accepted:** 27 February 2025

Funding: This work was supported by National Institutes of Health (R01 HL141286).

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

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`.

Abbreviations: BMI, body mass index; CF, cystic fibrosis; CFFPR, Cystic Fibrosis Foundation patient registry; CI, credible interval; HR, hazard ratio; IQR, interquartile range; MSE, mean squared error; PEx, pulmonary exacerbations; $ppFEV_1$, percentage of predicted forced expiratory volume in one second; ZCTA, zone improvement plan code tabulation area.

1 | Introduction

Cystic fibrosis (CF) is a severe genetic disorder that primarily affects the lungs and digestive system, leading to respiratory impairment and malnutrition [1]. 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 ($ppFEV_1$) are routinely measured to monitor disease progression. CF care teams are interested in using the US Cystic Fibrosis Foundation Patient Registry (CFFPR) [2] to understand the associations between $ppFEV_1$ decline, BMI changes, recurrent PEx, and the competing risks of death and lung transplantation.

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 [3–5]. Several extensions have been developed to the basic framework for a single event time and a continuous longitudinal biomarker proposed by Faucett and Thomas [6] and Wulfsohn and Tsiatis [7]. The literature is extensive, with recent comprehensive reviews by Hickey et al. [8, 9] Papageorgiou et al. [10] and Alsefri et al. [11] The joint modeling framework has previously been extended to incorporate complex survival data structures, such as recurrent [12–15] and competing [16–18] event time data. However, integrating both recurrent events and competing risks within a unified model remains challenging, leading researchers to omit important information available in patient registries. For example, Andrinopoulou et al. [19] 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 $ppFEV_1$ and the risks of death and lung transplantation, Miranda Afonso et al. [20] 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 Gaussian distributions to model continuous markers. An important aspect of joint modeling is appropriately parameterizing longitudinal submodels to ensure accurate extrapolation of unobserved biomarkers 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 $ppFEV_1$, 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 primarily modeled $ppFEV_1$ using a Gaussian distribution. Szczesniak et al. [21] considered other distributions; however, it proved challenging to derive a meaningful clinical interpretation of the associations in the linear predictor scale.

We address these limitations by introducing a comprehensive joint modeling framework that can (i) effectively accommodate competing risks and recurrent event processes together with multiple longitudinal outcomes, and (ii) use the beta distribution to model bounded longitudinal markers without compromising the interpretability of their associations. 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 $ppFEV_1$ and BMI using beta and Gaussian distributions, respectively. The model allows for the use of various functional forms to link time-to-event and longitudinal processes, and it accommodates discontinuous risk intervals and both gap and calendar timescales. We extended `JMbayes2`, [22] which is an R package for joint models available in the Comprehensive R Archive Network (CRAN), to incorporate the proposed model.

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 is used to demonstrate 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. 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 [23]. 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 jth longitudinal outcome, we consider a mixed-effects regression model

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

where $\mathbf{Y}_{j,i}$ is the jth response for the ith 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 jth outcome at time t conditional on the random effects, $\mu_{j,i}(t) = \mathrm{E}\{\mathrm{Y}_{j,i}(t)|\mathbf{b}_{j,i}\}$, has the form

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

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 non-negative shape parameters p and q. We

follow the beta density reparameterization proposed by Ferrari and Cribari-Neto, [24] 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 $\phi > 0$. For fixed $\mu_{j,i}$, the larger the value of ϕ, the smaller the variance of $Y_{j,i}$. In the context of our application, ϕ can be regarded as a nuisance parameter. 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 $(0, 1)$ [25]. 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 ppFEV_1. 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 $\text{logit}\{\mu_{j,i}(t)\}$. 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 equation (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 $\varepsilon_{j,i}(t) \sim \mathcal{N}(0, \sigma^2_{y_j})$. 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 $\mathbf{D}_j$ along its diagonal. These J matrices may be correlated or independent from each other. Joint models using the Gaussian distribution have been extensively discussed in the literature (see, e.g., Rizopoulos et al. [26]).

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 lth event at time t is modeled by

$$h^{\text{R}}_i(t) = h^{\text{R}}_0\left(t - t_{0_{l,i}}\right)\exp\left[\mathbf{w}^{\text{R}\top}_i(t)\boldsymbol{\gamma}^{\text{R}} + \sum_{j=1}^{J}\sum_{m=1}^{M_j}\mathcal{H}^{\text{R}}_{j,m}\{\eta_{j,i}(t)\}\alpha^{\text{R}}_{j,m} + \upsilon^{\text{R}}_i\right]$$

for $t > t_{0_{l,i}} \geq 0$, where $t_{0_{l,i}}$ is the starting time of the risk interval for the lth recurrent event, and $\upsilon^{\text{R}}_i \sim \mathcal{N}\left(0, \sigma^2_{\upsilon}\right)$. For the baseline hazard function $h^{\text{R}}_0\left(t - t_{0_{l,i}}\right)$, we use penalized B-spline functions (P-splines) [27]. Specifically, we use $\log h^{\text{R}}_0\left(t - t_{0_{l,i}}\right) = \sum_{q=1}^{Q}\gamma^{\text{R}}_{0_q}\text{bs}^{\text{R}}_q\left(t - t_{0_{l,i}}\right)$, where $\text{bs}^{\text{R}}_q(t)$ are the P-splines' qth basis functions of degree d, and $\gamma^{\text{R}}_{0_q}$ are the corresponding unknown coefficients. In the relative risk component of the model, the design vector $\mathbf{w}^{\text{R}}_i(t)$ contains the measured characteristics with the corresponding vector of regression coefficients $\boldsymbol{\gamma}^{\text{R}}$; 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 jth subject-specific marker trajectory through the M_j latent association structures $\mathcal{H}^{\text{R}}_{j,m}\{\eta_{j,i}(t)\} = \mathcal{H}^{\text{R}}_{j,m}\{\eta_{j,i}(u)\}, 0 \leq u \leq t$, which include the random effects $\mathbf{b}_{j,i}$. The function $\mathcal{H}^{\text{R}}_{j,m}(\cdot)$ determines the mth ($m = 1, \ldots, M_j$) form of association between the jth longitudinal marker and the time-to-event process. The longitudinal and recurrent event processes are assumed to be conditionally independent given $(\mathbf{b}^{\top}_{1,i}, \ldots, \mathbf{b}^{\top}_{J,i})$. The available functional forms are elaborated upon in Section 2.4. The association parameter $\alpha^{\text{R}}_{j,m}$ measures the strength of the association between the mth functional form of the jth longitudinal outcome and the risk of the next event. The quantity $\exp\left\{\alpha^{\text{R}}_{j,m}\right\}$ is the hazard ratio (HR) for a one-unit increase in the value of $\mathcal{H}^{\text{R}}_{j,m}\{\eta_{j,i}(t)\}$ while the rest of the variables are kept constant.

We incorporate the random effect υ^{R}_i 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 υ^{R}_i.

Our approach allows the recurrent event process to be modeled under the gap or calendar timescales, which use different zero-time references, $t_{0_{l,i}}$ [28]. 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, $t_{0_{l,i}} = (1-\delta_{l,0})t_{1_{l-1,i}}, \forall\, l$, where $\delta_{l,0}$ is the Kronecker delta ($\delta_{l,0} = 1$ if $l = 0$, and $\delta_{l,0} = 0$ otherwise) and $t_{1_{l,i}}$ is the observed event time for the lth recurrent event, assuming a renewal after each event and resetting the time to zero. For example, in the context of hospital readmissions, using the calendar timescale, the HR reflects how the risk of readmission changes over absolute time since the study began; while with the gap timescale, the HR measures how the risk of readmission depends on the time elapsed since the last admission. 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, $t_{0_{l,i}} = (1-\delta_{l,0})(t_{1_{l-1,i}} + d_{l-1,i}), \forall\, l$, where $d_{l,i}$ denotes the duration of the lth recurrent event. For example, if we are interested in modeling the time to the next hospital readmission, 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 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

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

by censoring all other causes. Here, $h^{\text{T}}_{0_k}(t)$ is the cause-specific P-splines baseline hazard function, given by $\log h^{\text{T}}_{0_k}(t) =$

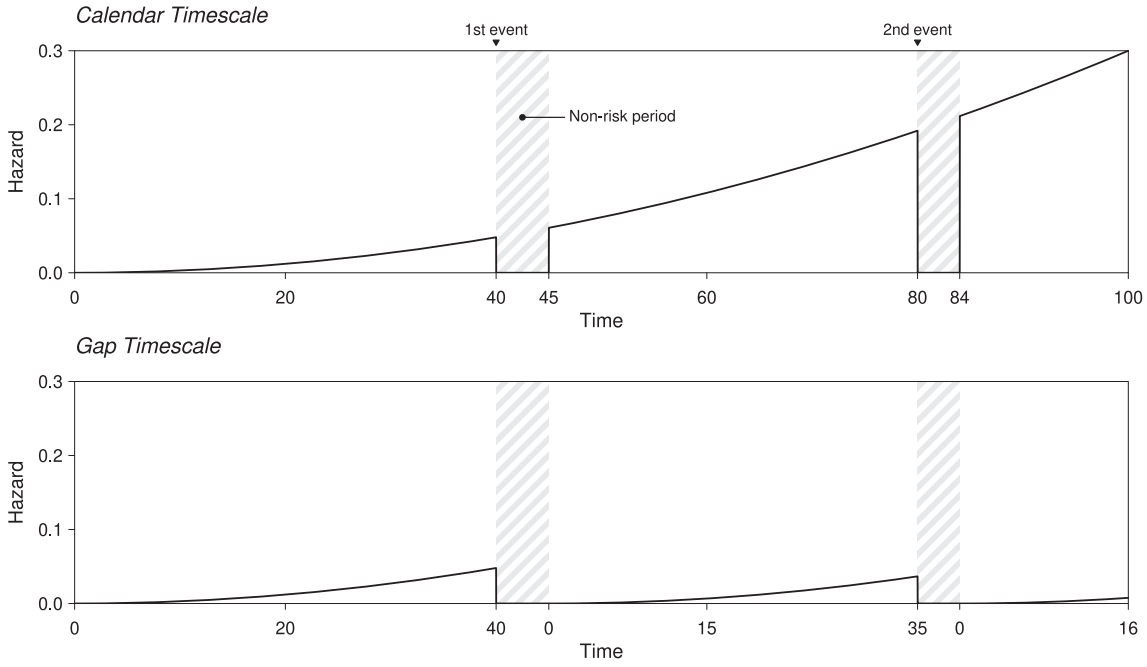

FIGURE 1 | The hazard function for a hypothetical recurrent event process, assuming the calendar (top panel) or gap (bottom panel) timescale. During the study period, from time 0 to 100, the displayed individual experienced two recurrent events (e.g., hospitalizations) at times 40 and 80. These events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

$\sum_{q=1}^{Q} \gamma_{0_{k_q}}^{\top} \text{bs}_{k_q}^{\top}(t)$, while $w_{k,i}^{\top}(t)$ is the vector of observed (baseline or time-varying exogenous) explanatory variables, and $\boldsymbol{\gamma}_k^{\top}$ is the corresponding vector of regression coefficients.

The jth longitudinal response influences the risk of failure due to cause k through $\mathcal{H}_{k,j,m}^{\top}\{\eta_{j,i}(t)\}$. The association parameters $\alpha_{k,j,m}^{\top}$ measure the strength of the association between each longitudinal outcome and the risk of the corresponding event. For a one-unit increase in $\mathcal{H}_{k,j,m}^{\top}\{\eta_{j,i}(t)\}$, the HR for cause k is $\exp(\alpha_{k,j,m}^{\top})$. The longitudinal measurements and event times are assumed to be conditionally independent given $(\mathbf{b}_{1,i}^{\top}, \ldots, \mathbf{b}_{J,i}^{\top})$.

The kth competing event is associated with the recurrent event process through a zero-mean Gaussian random variable $\upsilon_{k,i}^{\top}$. We assume that the frailties $\upsilon_{k,i}^{\top}$ and υ_i^{R} are proportional, $\upsilon_i^{\top} = \upsilon_i^{\text{R}}\ \alpha_k^{\upsilon}$, reflecting the common underlying factors that affect their risk. The magnitude of the association between each pair of processes is quantified by α_k^{υ}, 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 υ_i^{R}. Conditional on υ_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 [26, 29]. As discussed in Sections 2.2 and 2.3, the hazards $h_i^{\text{R}}(t)$ and $h_i^{\top}(t)$ of an event for patient i at time t are associated with the jth subject-specific marker trajectory through $\mathcal{H}_{j,m}^{\text{R}}\{\eta_{j,i}(t)\}$ 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, $\text{d}\eta_{j,i}(t)/\text{d}t$; standardized cumulative effect, $\frac{1}{t}\int_0^t \eta_{j,i}(s)\ \text{d}s$; and combinations of these regarding the same longitudinal outcome. Different forms can be assumed for each risk model. The choice of functional form should align with the biological understanding of the relationship between the biomarker and the risk of the event. For example, if recent biomarker values are expected to strongly influence the risk, the underlying current value might be most appropriate. On the other hand, if the cumulative exposure of the biomarker over time is thought to affect the risk, a summary measure of its history, such as the standardized cumulative effect, may be better suited. This approach ensures that model selection is grounded in clinical insights, thereby supporting the interpretability and relevance of the results.

When a non-linear link function $\mathcal{G}(\cdot)$ is applied to the mean of the longitudinal outcome in equation (1), it may be challenging to interpret the associations $\exp(\alpha_{k,j,m}^{\top})$ and $\exp(\alpha_{j,m}^{\text{R}})$ 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, $\mathcal{H}_{j,m}\{\mu_{j,i}(t)\} = \mathcal{H}_{j,m}\left[\mathcal{G}_j^{-1}\{\eta_{j,i}(t)\}\right]$, 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 $\text{logit}\{\mu_{j,i}(t)\}$. 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 [30] 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 dataset [19]. 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 extended the CRAN R package `JMbayes2` [22] to incorporate the proposed joint model. An example of its application is provided in Supplementary Section B. To facilitate adaptation to other applications, our implementation supports distributions beyond the Gaussian and Beta for the longitudinal processes and allows for simpler joint models that consider only 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 (no model misspecification). 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 on 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. For the longitudinal processes, we assume models with a random intercept and a linear random slope. The two random effects are normally distributed and are assumed to be independent. Baseline hazard functions are estimated using a penalized splines approximation, as detailed in Sections 2.2 and 2.3. Supplementary Tables S5 and S6 detail the data generation process for each scenario, and Supplementary Table S7 summarizes the characteristics of the simulated datasets. We replicate each scenario 500 times. For each dataset, we considered 1 000 individuals. In scenario A, individuals collectively contributed a median of 10 935.5 observations for each longitudinal outcome (IQR 10 807.75–11 087.5), while in scenario B, the median was 15 311.5 (IQR 15 195.75–15 430). Individuals in scenario A experienced a median of three recurrent events per dataset, with median rates for the two competing events of 0.42 (IQR 0.41–0.43) and 0.41 (IQR 0.40–0.43), and a median censoring rate of 0.17 (IQR 0.16–0.18).

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, using `JMbayes2` version 0.4.5. The difference in the number of iterations reflects the varying complexity of the two models. The greater complexity of the model in Scenario A requires additional iterations to ensure adequate convergence and accuracy in parameter estimation. Details of the prior distributions assumed are available in Supplementary Table S1. The convergence of the chains is assessed using the convergence diagnostic $\hat{R}$, [31] 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 on the bias, mean squared error (MSE), empirical standard error (ESE), mean estimated posterior standard deviation (MSDev), and coverage probability (CP) of the 95% credible interval. The definitions of these quantities are provided in Supplementary Section C. 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. The median computation time for Scenario A was 21.77 min (IQR 22.04–22.28) on a machine with an AMD Ryzen Threadripper PRO 3975WX 32-core 64-thread processor running at 3.49 GHz, using 256 GB of RAM, running Windows 11 Pro (v21H2). 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: −5.5; MSE: 30.2) and, consequently, in the remaining independent variables present in the risk model. The misspecified model underestimates the true effect of the longitudinal process on the hazard. In clinical practice, this underestimation could lead to inaccurate risk assessments, potentially downplaying the importance of the marker in guiding treatment decisions. 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 | Summary of performance for the joint model estimates obtained under the two simulated scenarios for 500 simulated datasets.

		Scenario A						Scenario B									
									Beta					Gaussian			
Submodel	Param.	True	Bias	MSE	ESE	MSDev	CP	True	Bias	MSE	ESE	MSDev	CP	Bias	MSE	ESE	MSDev	CP
M_1																		
	$\beta_{1,0}$	2.00	−0.002	0.000	0.000	0.016	0.934	2.00	−0.001	0.000	0.000	0.016	0.946	−1.238	1.532	0.000	0.005	0.000
	$\beta_{1,t}$	−1.50	0.001	0.000	0.000	0.013	0.968	−1.00	0.000	0.000	0.000	0.012	0.958	0.883	0.780	0.000	0.002	0.000
M_2																		
	$\beta_{2,0}$	0.80	0.000	0.000	0.000	0.003	0.962	—	—	—	—	—	—	—	—	—	—	—
	$\beta_{2,t}$	−0.05	0.000	0.000	0.000	0.003	0.950	—	—	—	—	—	—	—	—	—	—	—
R																		
	γ^{R}	0.25	0.000	0.002	0.003	0.065	0.994	—	—	—	—	—	—	—	—	—	—	—
	α_1^{R}	−2.00	−0.004	0.007	0.009	0.083	0.954	—	—	—	—	—	—	—	—	—	—	—
	α_2^{R}	−1.00	0.003	0.003	0.004	0.053	0.944	—	—	—	—	—	—	—	—	—	—	—
T_1																		
	γ_1^{T}	0.25	0.006	0.012	0.018	0.114	0.948	0.25	−0.001	0.007	0.009	0.108	0.988	−0.033	0.008	0.011	0.114	0.990
	$\alpha_{1,1}^{T}$	−2.00	−0.096	0.303	0.415	0.563	0.966	−2.00	−0.010	0.103	0.152	0.319	0.938	−5.473	30.216	0.397	0.378	0.000
	$\alpha_{1,2}^{T}$	−1.00	−0.005	0.015	0.021	0.118	0.932	—	—	—				—	—	—	—	—
	α_1^{v}	1.00	0.025	0.040	0.062	0.198	0.938	—	—	—	—	—	—	—	—	—	—	—
T_2																		
	γ_2^{T}	0.25	0.001	0.011	0.014	0.105	0.950	—	—	—	—	—	—	—	—	—	—	—
	$\alpha_{2,1}^{T}$	−2.00	−0.077	0.247	0.330	0.548	0.974	—	—	—	—	—	—	—	—	—	—	—
	$\alpha_{2,2}^{T}$	−1.00	−0.012	0.014	0.020	0.123	0.958	—	—	—	—	—	—	—	—	—	—	—
	α_2^{v}	1.00	0.006	0.044	0.065	0.203	0.932	—	—	—	—	—	—	—	—	—	—	—

Note: 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: CP, coverage probability; ESE, empirical standard error; M_1, 1st longitudinal marker; M_2, 2nd longitudinal marker; MSDev, mean estimated posterior standard deviation; MSE, mean squared error; Param., parameter; R, Recurrent event; T_1, 1st terminal event; T_2, 2nd terminal event.

4 | Application

4.1 | The CFFPR Dataset

The CFFPR dataset 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 [2]. 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. [32]), percentage of green space[1] , and moving-truck density. Previous research demonstrated that environmental and community characteristics, alongside clinical and demographic factors, are critical to comprehensively understand CF progression [34, 35].

BMI and $ppFEV_1$ 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. FEV_1 measures the maximum volume of air that a person can forcefully exhale in the first second of expiration after taking a deep breath. $ppFEV_1$ compares a patient's measured FEV_1 to the expected value for a person of the same age, sex, and height with normal lung function [36]. We assume that $ppFEV_1$ ranges from 0% to 150%, with a value of 100% meaning that the patient's FEV_1 is equal to the expected value for a healthy individual. While it is uncommon, there are instances in which the $ppFEV_1$ is reported as above 100% owing to early intervention and treatment. Lower BMI and $ppFEV_1$ levels are associated with worse clinical outcomes [37]. The median numbers of $ppFEV_1$ 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 $ppFEV_1$ (left panel) and BMI (center panel) evolution experienced by nine randomly selected individuals over time. The profiles exhibit different follow-up durations and diverse non-linear 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% 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 non-ignorable missing data in the measurements of $ppFEV_1$ and BMI.

A PEx is a sudden worsening of CF respiratory symptoms usually caused by an infection or inflammation in the airways [38]. 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.

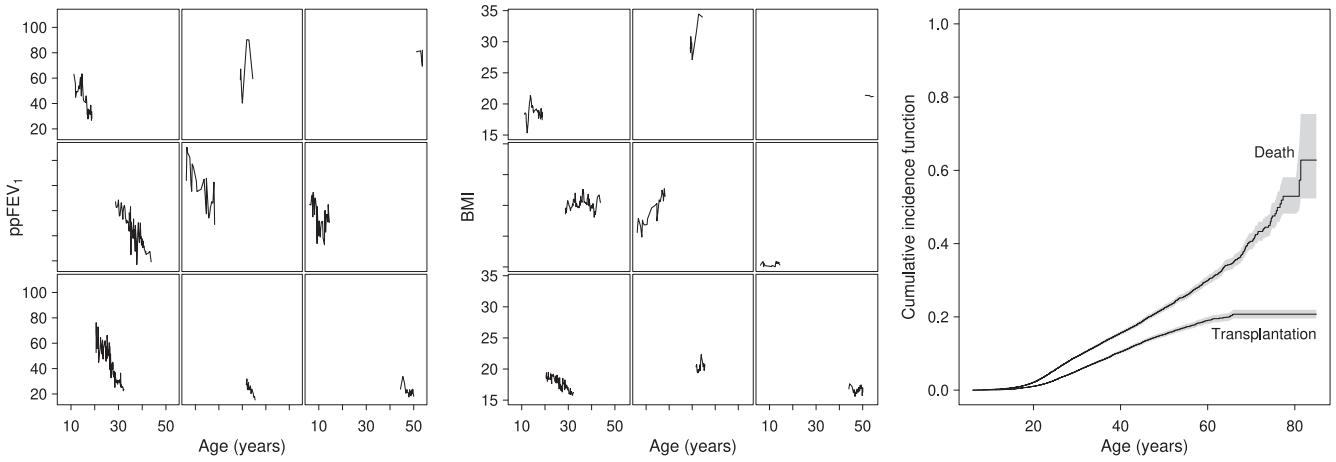

FIGURE 2 | Longitudinal and survival outcomes of interest. Left: $ppFEV_1$ measurements against age for nine randomly selected individuals. Center: BMI measurements against age for the same individuals. Right: Cumulative incidence functions for the competing events of death and lung transplantation, 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 ppFEV_1 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:

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

and

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

for $t > 0$, where $\left(b_{1,0,i}, b_{1,t,i}, b_{2,0,i}, b_{2,1,i}, b_{2,2,i}\right)^\top \sim \mathcal{N}(\mathbf{0}, \mathbf{D})$, and $\varepsilon_i(t) \sim \mathcal{N}(0, \sigma^2_{y_2})$, with the two random variables assumed independent of each other. Here, $\tilde{\text{BMI}}_i(t)$ is the BMI response without error, and $\text{ppFEV}_{1\,i}^{**}(t)$ is the ppFEV_1 response scaled to the interval $(0, 1)$.[2]

For ppFEV_1, we assume a linear average evolution over time, while for BMI, we assume a non-linear 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 ppFEV_1 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), $\text{ethn}_{\text{hisp},i}$; and neighborhood deprivation index, $\text{depr}_i(t)$. Additionally, the average ppFEV_1 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 time specification as that used 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

$$\begin{aligned}h_i^{\text{R}}(t) = h_0^{\text{R}}(t - t_{0_{l,i}})\exp\Big[&\gamma_{\text{PEx}}^{\text{R}}\text{nPEx}_i(t) + \text{ppFEV}_{1\,i}^{**}(t)\alpha_{1,1}^{\text{R}} \\ &+ \frac{1}{t}\int_0^t \tilde{\text{BMI}}_i(s)\,\text{d}s\;\alpha_{2,1}^{\text{R}} + \upsilon_i^{\text{R}}\Big]\end{aligned}$$

$$\begin{aligned}h_{1,i}^{\text{T}}(t) = h_{0_1}^{\text{T}}(t)\exp\Big[&\text{ppFEV}_{1\,i}^{**}(t)\alpha_{1,1,1}^{\text{T}} + \frac{\text{d}\;\text{ppFEV}_{1\,i}^{**}(t)}{\text{d}t}\alpha_{1,1,2}^{\text{T}} \\ &+ \frac{1}{t}\int_0^t \tilde{\text{BMI}}_i(s)\,\text{d}s\;\alpha_{1,2,1}^{\text{T}} + \upsilon_i^{\text{R}}\;\alpha_1^{\upsilon}\Big]\end{aligned}$$

and

$$\begin{aligned}h_{2,i}^{\text{T}}(t) = h_{0_2}^{\text{T}}(t)\exp\Big[&\text{ppFEV}_{1\,i}^{**}(t)\alpha_{2,1,1}^{\text{T}} + \frac{\text{d}\;\text{ppFEV}_{1\,i}^{**}(t)}{\text{d}t}\alpha_{2,1,2}^{\text{T}} \\ &+ \frac{1}{t}\int_0^t \tilde{\text{BMI}}_i(s)\,\text{d}s\;\alpha_{2,2,1}^{\text{T}} + \upsilon_i^{\text{R}}\;\alpha_2^{\upsilon}\Big]\end{aligned}$$

for $t > 0$, where $\upsilon_i^{\text{R}} \sim \mathcal{N}\left(0, \sigma_\upsilon^2\right)$, $\upsilon_i^{\text{R}} \perp \;\perp \left(b_{1,0,i}, b_{1,t,i}, b_{2,0,i}, b_{2,1,i}, b_{2,2,i}\right)$ and $\upsilon_i^{\text{R}} \perp \;\perp \varepsilon_i(t)$. Changes in BMI over time occur relatively slowly, whereas ppFEV_1 can experience sudden declines. Therefore, guided by clinical insights, we include as predictors the ppFEV_1's value, $\text{ppFEV}_1^{**}(t)$, and its rate of change, $\text{d}\;\text{ppFEV}_1^{**}(t)/\text{d}t$, evaluated on its original scale—applying the expit$(\cdot)$ transformation to the linear predictor described in Section 2.1—and the standardized cumulative effect of BMI's underlying value, $\frac{1}{t}\int_0^t \tilde{\text{BMI}}(s)\,\text{d}s$. 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 $\hat{R}$, [31] 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 ppFEV_1 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 ppFEV_1 decline during childhood and adolescence and a more gradual decline thereafter. When modeling ppFEV_1 with a Gaussian distribution and allowing for flexible temporal evolution, the resulting model produces non-feasible negative values (Figure 3, right panel). The observed and predicted longitudinal trajectories for ppFEV_1 and BMI for randomly selected individuals are provided in Supplementary Figure S5.

The model parameter estimates are listed in Table 2. The estimates suggest that lower overall ppFEV_1 values are associated with being female, non-Hispanic, born after 1993, having a CFTR mutation other than F508del, and living in more deprived areas, areas with less green space, or areas with higher moving-truck

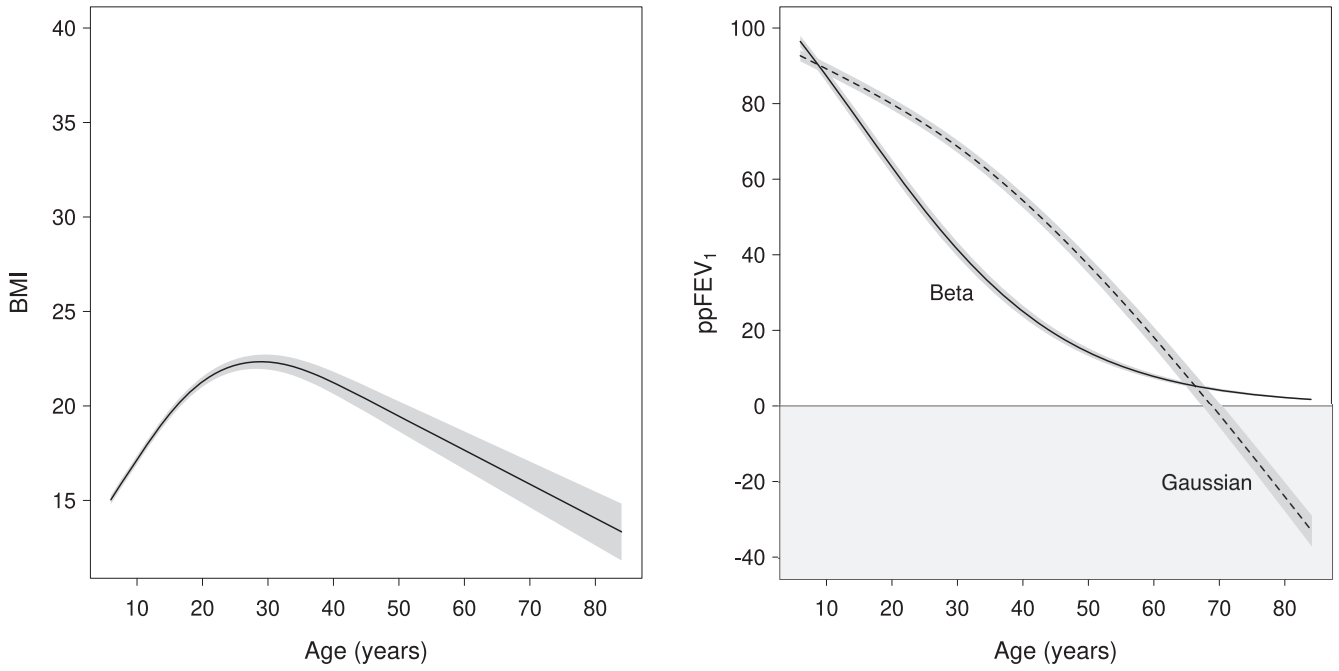

FIGURE 3 | Left: Estimated BMI evolution with age, with associated 95% credible interval, for Hispanic females with CF, F508del homozygotes, who were born before 1993, did not take pancreatic enzymes, and lived in a community with a deprivation index of 0.5. Right: Estimated $ppFEV_1$ evolution with 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 μtruck-meters/m^2. For a Gaussian distribution, the model generates non-feasible negative values despite incorporating flexible temporal evolution via natural cubic splines.

density. Similarly, lower overall BMI values are associated with being female, Hispanic, born before 1993, being F508del homozygous, living in more deprived areas, and not taking enzymes. The risk of a PEx increases with the number of previous episodes. The results suggest that both $ppFEV_1$ and BMI are associated with the risks of experiencing PEx, transplantation, and death. For example, a one-unit decrease in value and one-unit increase in the rate of $ppFEV_1$ 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 estimates of association parameters in Table 2. The estimates for the association between $ppFEV_1$ and the risk of transplantation are different from that between $ppFEV_1$ and death, illustrating the value of modeling both events individually, rather than as a composite endpoint.

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 expressed in percentiles or z-scores. The model is now available in the R package `JMbayes2` [22] 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. [20].

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 $ppFEV_1$, 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. Those features can be evaluated

TABLE 2 | Posterior means, posterior standard deviations, and 95% credible intervals for some of the joint model parameters fitted to the CFFPR dataset. Estimates for the longitudinal submodels are presented on the linear predictor scale.

Model	Parameter/HR	Mean	Std. Dev.	95% CI	
$ppFEV_1$					
	$\beta_{1,0}$	0.591	0.019	(0.554,	0.629)
	$\beta_{1,t}$	−0.065	< 0.001	(−0.065,	−0.064)
	$\beta_{1,\text{male}}$	0.001	0.009	(−0.017,	0.018)
	$\beta_{1,[93,98)}$	−0.157	0.012	(−0.180,	−0.133)
	$\beta_{1,\geq 98}$	−0.125	0.011	(−0.147,	−0.103)
	$\beta_{1,\text{htz}}$	0.019	0.010	(0.001,	0.038)
	$\beta_{1,\text{oth}}$	−0.024	0.013	(−0.050,	0.002)
	$\beta_{1,\text{ethn}}$	0.223	0.017	(0.191,	0.256)
	$\beta_{1,\text{depr}}$	−0.003	0.004	(−0.010,	0.005)
	$\beta_{1,\text{truck}}$	$-4.440 \cdot 10^{-5}$	< 0.001	$(-3.160 \cdot 10^{-4},$	$3.340 \cdot 10^{-4})$
	$\beta_{1,\text{pgrn}}$	−0.266	0.001	(−0.519,	0.029)
BMI					
	$\beta_{2,0}$	15.053	0.098	(14.858,	15.244)
	β_{2,ns_1}	12.867	0.143	(12.585,	13.143)
	β_{2,ns_2}	1.881	0.230	(1.424,	2.330)
	$\beta_{2,\text{male}}$	−0.465	0.043	(−0.548,	−0.378)
	$\beta_{2,[93,98)}$	0.242	0.058	(0.127,	0.356)
	$\beta_{2,\geq 98}$	0.633	0.056	(0.523,	0.743)
	$\beta_{2,\text{htz}}$	0.170	0.046	(0.080,	0.259)
	$\beta_{2,\text{oth}}$	0.269	0.066	(0.140,	0.398)
	$\beta_{2,\text{ethn}}$	−0.191	0.081	(−0.348,	−0.032)
	$\beta_{2,\text{depr}}$	−0.038	0.032	(−0.101,	−0.021)
	$\beta_{2,\text{enzy}}$	0.021	0.003	(0.016,	0.026)
Recurrent PEx					
	$\exp(\gamma^{\text{R}}_{\text{PEx}})$	1.010	0.001	(1.009,	1.011)
	σ_v	0.835	0.007	(0.822,	0.849)
	$\exp(\alpha^{\text{R}}_{1,1}/150)$	0.962	< 0.001	(0.961,	0.962)
	$\exp(\alpha^{\text{R}}_{2,1})$	$1.000^{(40)}$	< 0.001	$(1.000^{(37)},$	$1.000^{(42)})$
Transplantation					
	$\exp(\alpha^{\text{T}}_{1,1,1}/150)$	0.830	0.002	(0.825,	0.835)
	$\exp(\alpha^{\text{T}}_{1,1,2}/150)$	0.863	0.013	(0.839,	0.891)
	$\exp(\alpha^{\text{T}}_{1,2,1})$	1.060	0.008	(1.044,	1.076)
	$\exp(\alpha^{v}_{1})$	1.203	0.042	(1.122,	1.287)
Death					
	$\exp(\alpha^{\text{T}}_{2,1,1}/150)$	0.884	0.001	(0.882,	0.887)
	$\exp(\alpha^{\text{T}}_{2,1,2}/150)$	0.909	0.009	(0.892,	0.925)
	$\exp(\alpha^{\text{T}}_{2,2,1})$	1.071	0.008	(1.054,	1.087)
	$\exp(\alpha^{v}_{2})$	1.326	0.032	(1.266,	1.389)

Note: Superscripts in parentheses indicate additional decimal places.
Abbreviations: BMI, body mass index; CI, credible interval; HR, hazard ratio; PEx, pulmonary exacerbation; $ppFEV_1$, percent predicted forced expiratory volume in one second; Std. Dev., standard deviation.

by visually inspecting the observed values. Nonetheless, when using a Gaussian distribution, it is important to assess the distribution of predicted values to ensure the model does not generate values outside the feasible range.

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 $ppFEV_1$ observations to the interval $[0, 1]$ and subsequently

to the open interval $(0, 1)$ using the transformation proposed by Smithson and Verkuilen [39]. 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 [40] represents an important area of application of the proposed model. To support this, the development of appropriate accuracy assessment tools is imperative for evaluating the model's predictive performance and enabling its translation into clinical practice.

Our comprehensive modeling approach offers a new perspective on studying the progression of CF, and we hope it 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.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available from the 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.

Endnotes

[1] Percentage of greenspace, impervious, and tree canopy areas within the Zone Improvement Plan Code Tabulation Area (ZCTA) derived from the National Land Cover Database [33].

[2] 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 [39].

References

1. P. M. Farrell, B. J. Rosenstein, T. B. White, et al., "Guidelines for Diagnosis of Cystic Fibrosis in Newborns Through Older Adults: Cystic Fibrosis Foundation Consensus Report," *Journal of Pediatrics* 153, no. 2 (2008): S4–S14.

2. E. A. Knapp, A. K. Fink, C. H. Goss, et al., "The Cystic Fibrosis Foundation Patient Registry. Design and Methods of a National Observational Disease Registry," *Annals of the American Thoracic Society* 13, no. 7 (2016): 1173–1179.

3. R. Henderson, P. Diggle, and A. Dobson, "Joint Modelling of Longitudinal Measurements and Event Time Data," *Biostatistics* 1, no. 4 (2000): 465–480.

4. A. A. Tsiatis and M. Davidian, "Joint Modeling of Longitudinal and Time-to-Event Data an Overview," *Statistica Sinica* 14, no. 3 (2004): 809–834.

5. D. Rizopoulos, *Joint Models for Longitudinal and Time-To-Event Data: With Applications in R* (CRC Press, 2012).

6. C. L. Faucett and D. C. Thomas, "Simultaneously Modelling Censored Survival Data and Repeatedly Measured Covariates: A Gibbs Sampling Approach," *Statistics in Medicine* 15, no. 15 (1996): 1663–1685.

7. M. S. Wulfsohn and A. A. Tsiatis, "A Joint Model for Survival and Longitudinal Data Measured With Error," *Biometrics* 53, no. 1 (1997): 330–339.

8. G. L. Hickey, P. Philipson, A. Jorgensen, and R. Kolamunnage-Dona, "Joint Modelling of Time-To-Event and Multivariate Longitudinal Outcomes: Recent Developments and Issues," *BMC Medical Research Methodology* 16 (2016): 1–15.

9. G. L. Hickey, P. Philipson, A. Jorgensen, and R. Kolamunnage-Dona, "Joint Models of Longitudinal and Time-To-Event Data With More Than One Event Time Outcome: A Review," *International Journal of Biostatistics* 14, no. 1 (2018): 20170047, https://doi.org/10.1515/ijb-2017-0047.

10. G. Papageorgiou, K. Mauff, A. Tomer, and D. Rizopoulos, "An Overview of Joint Modeling of Time-To-Event and Longitudinal Outcomes," *Annual Review of Statistics and Its Application* 6 (2019): 223–240.

11. M. Alsefri, M. Sudell, M. García-Fiñana, and R. Kolamunnage-Dona, "Bayesian Joint Modelling of Longitudinal and Time to Event Data: A Methodological Review," *BMC Medical Research Methodology* 20 (2020): 1–17.

12. L. Liu, X. Huang, and J. O'Quigley, "Analysis of Longitudinal Data in the Presence of Informative Observational Times and a Dependent Terminal Event, With Application to Medical Cost Data," *Biometrics* 64, no. 3 (2008): 950–958.

13. L. Liu and X. Huang, "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* 58, no. 1 (2009): 65–81.

14. S. Kim, D. Zeng, L. Chambless, and Y. Li, "Joint Models of Longitudinal Data and Recurrent Events With Informative Terminal Event," *Statistics in Biosciences* 4 (2012): 262–281.

15. A. Król, L. Ferrer, J. P. Pignon, et al., "Joint Model for Left-Censored Longitudinal Data, Recurrent Events and Terminal Event: Predictive Abilities of Tumor Burden for Cancer Evolution With Application to the FFCD 2000–05 Trial," *Biometrics* 72, no. 3 (2016): 907–916.

16. R. M. Elashoff, G. Li, and N. Li, "A Joint Model for Longitudinal Measurements and Survival Data in the Presence of Multiple Failure Types," *Biometrics* 64, no. 3 (2008): 762–771.

17. P. R. Williamson, R. Kolamunnage-Dona, P. Philipson, and A. G. Marson, "Joint Modelling of Longitudinal and Competing Risks Data," *Statistics in Medicine* 27, no. 30 (2008): 6426–6438.

18. E. R. Andrinopoulou, D. Rizopoulos, J. J. Takkenberg, and E. Lesaffre, "Joint Modeling of Two Longitudinal Outcomes and Competing Risk Data," *Statistics in Medicine* 33, no. 18 (2014): 3167–3178.

19. E. R. Andrinopoulou, J. P. Clancy, and R. Szczesniak, "Multivariate Joint Modeling to Identify Markers of Growth and Lung Function Decline That Predict Cystic Fibrosis Pulmonary Exacerbation Onset," *BMC Pulmonary Medicine* 20 (2020): 1–11.

20. P. Miranda Afonso, D. Rizopoulos, A. K. Palipana, et al., "Efficiently Analyzing Large Patient Registries With Bayesian Joint Models for Longitudinal and Time-To-Event Data," *arXiv Preprint arXiv* (2023): 2310.03351, https://arxiv.org/abs/2310.03351.

21. R. Szczesniak, E. R. Andrinopoulou, W. Su, et al., "Lung Function Decline in Cystic Fibrosis: Impact of Data Availability and Modeling Strategies on Clinical Interpretations," *Annals of the American Thoracic Society* 20, no. 9 (2023): 958–968.

22. D. Rizopoulos, G. Papageorgiou, and P. Miranda Afonso, *JMbayes2: Extended Joint Models for Longitudinal and Time-To-Event Data* (CRAN, 2023), http://CRAN.R-project.org/package=JMbayes2, R package Version 0.4-5.

23. I. Sousa, "A Review on Joint Modelling of Longitudinal Measurements and Time-To-Event," *Revstat Statistical Journal* 9 (2011): 57–81.

24. S. Ferrari and F. Cribari-Neto, "Beta Regression for Modelling Rates and Proportions," *Journal of Applied Statistics* 31, no. 7 (2004): 799–815.

25. A. K. Gupta and S. Nadarajah, *Handbook of Beta Distribution and Its Applications* (CRC Press, 2004).

26. D. Rizopoulos, L. A. Hatfield, B. P. Carlin, and J. J. Takkenberg, "Combining Dynamic Predictions From Joint Models for Longitudinal and Time-To-Event Data Using Bayesian Model Averaging," *Journal of the American Statistical Association* 109, no. 508 (2014): 1385–1397.

27. P. H. Eilers and B. D. Marx, "Flexible Smoothing With B-Splines and Penalties," *Statistical Science* 11, no. 2 (1996): 89–121.

28. L. Duchateau, P. Janssen, I. Kezic, and C. Fortpied, "Evolution of Recurrent Asthma Event Rate Over Time in Frailty Models," *Journal of the Royal Statistical Society, Series C* 52, no. 3 (2003): 355–363.

29. K. Mauff, E. W. Steyerberg, G. Nijpels, v. d. A. A. Heijden, and D. Rizopoulos, "Extension of the Association Structure in Joint Models to Include Weighted Cumulative Effects," *Statistics in Medicine* 36, no. 23 (2017): 3746–3759.

30. P. H. Garthwaite, Y. Fan, and S. A. Sisson, "Adaptive Optimal Scaling of Metropolis–Hastings Algorithms Using the Robbins–Monro Process," *Communications in Statistics - Theory and Methods* 45, no. 17 (2016): 5098–5111.

31. A. Gelman and D. B. Rubin, "Inference From Iterative Simulation Using Multiple Sequences," *Statistical Science* 7, no. 4 (1992): 457–472.

32. C. Brokamp, A. F. Beck, N. K. Goyal, P. Ryan, J. M. Greenberg, and E. S. Hall, "Material Community Deprivation and Hospital Utilization During the First Year of Life: An Urban Population–Based Cohort Study," *Annals of Epidemiology* 30 (2019): 37–43.

33. S. Jin, C. Homer, L. Yang, et al., "Overall Methodology Design for the United States National Land Cover Database 2016 Products," *Remote Sensing* 11, no. 24 (2019): 2971.

34. E. Gecili, C. Brokamp, E. Rasnick, et al., "Built Environment Factors Predictive of Early Rapid Lung Function Decline in Cystic Fibrosis," *Pediatric Pulmonology* 58, no. 5 (2023): 1501–1513.

35. A. K. Palipana, A. Vancil, E. Gecili, et al., "Social-Environmental Phenotypes of Rapid Cystic Fibrosis Lung Disease Progression in Adolescents and Young Adults Living in the United States," *Environmental Advances* 14 (2023): 100449.

36. S. Stanojevic, D. Bilton, A. McDonald, et al., "Global Lung Function Initiative Equations Improve Interpretation of FEV1 Decline Among Patients With Cystic Fibrosis," *European Respiratory Journal* 46, no. 1 (2015): 262–264.

37. T. G. Liou, F. R. Adler, S. C. FitzSimmons, B. C. Cahill, J. R. Hibbs, and B. C. Marshall, "Predictive 5-Year Survivorship Model of Cystic Fibrosis," *American Journal of Epidemiology* 153, no. 4 (2001): 345–352.

38. P. A. Flume, P. J. Mogayzel, Jr., K. A. Robinson, et al., "Cystic Fibrosis Pulmonary Guidelines: Treatment of Pulmonary Exacerbations," *American Journal of Respiratory and Critical Care Medicine* 180, no. 9 (2009): 802–808.

39. M. Smithson and J. Verkuilen, "A Better Lemon Squeezer? Maximum-Likelihood Regression With Beta-Distributed Dependent Variables," *Psychological Methods* 11, no. 1 (2006): 54.

40. E. R. Andrinopoulou, M. O. Harhay, S. J. Ratcliffe, and D. Rizopoulos, "Reflection on Modern Methods: Dynamic Prediction Using Joint Models of Longitudinal and Time-To-Event Data," *International Journal of Epidemiology* 50, no. 5 (2021): 1731–1743.

Supporting Information

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